UNIVERSITY OF MANNHEIM
DEPARTMENT OF ECONOMICS

E600 MATHEMATICS

COMPANION SCRIPT

This script collects all contents featured in the classroom discussions of E600 Mathematics and provides some additional detail and reference to further resources. Relative to the remaining course material, the script gives the formal proofs that justify the propositions relevant to this class, and discusses additional, more involved concepts. Specifically, contents exclusively featured in the script are a thorough formal discussion of the vector space concept in Chapter 1 and an investigation of systems of equations where the number of equations does not coincide with the number of unknowns in Chapter 2.

Beyond facilitating preparation of and follow-up study for this course, the script's main purpose is to serve as a primary resource to consult during the courses E601-E603 as well as other Master's classes, especially for issues related to general mathematics. It gives a broad introduction to mathematical notation and objects, from simpler things like quantifiers, sets and functions to more rigorous discussion of spaces, i.e. structured collections of objects such as numbers, vectors and matrices. Next, due to their frequent use in the economic profession, we turn to matrices and their relation to linear equation systems. Lastly, the script lays the foundation for economists' favorite exercise – optimization – by giving a rigorous introduction to and discussion of general techniques of multivariate differential calculus, and also briefly mentions integral calculus. The fourth chapter is dedicated to applying these techniques to optimization problems, and finding necessary and sufficient conditions for existence of solutions. The final chapter gives an introduction to probability theory and econometric methods, key items in the toolbox of the empirical economist.

Script written by:
Martin Reinhard

Course taught by:
Julian Klix*
E-Mail: julian.klix[at]uni-mannheim.de

This version: Fall Semester 2026

*Ph.D. Student at the Center for Doctoral Studies in Economics of the Graduate School of Economics and Social Sciences. This script was developed by my predecessors Johannes Gessner and Martin Reinhard, drawing on the work of his predecessors Simona Helmsmüller and Justin Leduc. I am grateful for their great work.

Contents

-1 Preface

First of all, welcome to the brief journey through the realm of mathematics that is E600! You may ask yourself what precisely E600 Mathematics is and why you, a prospective student of the University of Mannheim's MSc Program in Economics – or generally any student interested in a thorough review of the mathematical foundations of advanced economic studies – should put up with it.

Broadly, the course's purpose is to equip you with the mathematical knowledge and resources necessary to successfully complete the Master's program. The MSc level courses at the University of Mannheim's economics department typically rely on the concepts this course discusses without giving them an explicit, independent treatment. However, this **does not mean** that you need to fully master everything discussed in E600 – rather, this course, and especially this script, should serve you as a resource to consult when you come across issues related to general mathematics, and it should help our students, coming from many different undergraduate schools and fields, catch up to a common ground when it comes to their math skills.

Every chapter of this script concludes with a summary of its key contents and take-aways. You can consult this summary both to compare the contents to your current knowledge and to plan how much time and effort you will need to absorb its contents, and to check how well you understood the contents after reading them. That being said, these summaries are deliberately placed at the *end* of the chapters; it is perfectly fine if you have no or little knowledge of the concepts before reading the chapters. Further, you can find a number of review questions at the end of every topic. In contrast to the problem sets, they are mostly not meant to be that difficult and deepen your understanding, but rather serve as a more basic check of whether you have understood the chapter's key points. The solutions can be found at the end of the script.

The main resources of this course are

(SB) Carl P. Simon / Lawrence Blume (1994): Mathematics for Economists, 1st Edition. W.W. Norton & Company

(dlF) Angel de la Fuente (2000): Mathematical Methods and Models for Economists, Cambridge University Press

and the labels in parentheses will be used to refer to them throughout this script. They may also serve you as an independent resource to consult when coming across issues related to mathematics. The advantage this script offers to you is that all concepts and proofs are written in a way that ensures consistency with our notation and uses only definitions and facts that have been introduced previously. You will find that this is not always the case with other resources, especially when referring to many of them at once.

Finally, as any document that did not go through a formal publication process, this script – as well as the contents of the entire course – may contain (hopefully small) errors or inconsistencies. If you find some, please contact us.

0 Fundamentals of Mathematics

Please try to find the time to read this section or the material for Chapter 0 before coming to the first class!

This introductory section tries to illustrate the value mathematics holds for the economist profession and deals with fundamental, overarching concepts that may be viewed as prerequisites for any mathematical application and study, regardless of its purpose or context.

0.1 Why Mathematics?

Usually, it is easier to find motivation for studying abstract and complex matters if you know why you're doing it. You may ask yourself, "Hey, I signed up for an Economics program! Why should I bother with a Math course, rather than reviews in Micro- or Macroeconomics, or perhaps Econometrics?" The main reason is that you will need mathematical concepts an methods as an economist: to write down, understand and potentially estimate models. If you intend to work in a job related to economics, you will most likely encounter problems that can only be described and solved using mathematics. You will need even more maths if you intend to do research in economics!

0.2 Notation and Logic

Mathematics can be thought of as a language. In this view, the **vocabulary** of mathematics consists of a large set of commonly agreed-upon mathematical notation (numbers, variables, functions, operators for e.g. addition (+) and multiplication (×), etc.) and symbols (e.g. $\Rightarrow$, $\subset$, $\in$, etc.). Any combination of items in the vocabulary constitutes a **statement**, which can be judged whether it is meaningful and if so, also true. These statements can then be combined to form logical **arguments** that convey information about mathematical facts and relationships.

0.2.1 Mathematical Notation (or: Vocabulary)

Mathematical statements can be more efficiently written using symbols and other mathematical notation. Having this common mathematical notation speeds up things substantially! (Imagine writing 100 times "for all natural numbers i" instead of "$i \in \mathbf{N}$".) If you ever come across a symbol that you have never seen, you can check https://www.rapidtables.com/math/symbols/Basic_Math_Symbols.html, there is a good chance that you will find out there what it means. Table 1 gives a brief overview of those symbols most important to economists.[1] Some of the symbols are discussed to more detail below. Don't worry if you can't remember everything just yet, you will see and use these often enough.

In the first column, the word *quantifier*, which refers to the first four symbols, may be new to you. They are important since we frequently make use of quantifying statements, that is, expressions of the form

Quantifier + Considered Elements + Property

[1] Moreover, the sets of numbers $\mathbb{R}, \mathbb{N}, \mathbb{C}, \mathbb{Q}$ and $\mathbb{Z}$, i.e. real, natural, complex, rational numbers and integers, are very important and should be familiar to you.

Basics and Quantifiers		*Logical Statements*	
Symbol	Meaning	Symbol	Meaning
$\exists$	there exists	$\Rightarrow$	implies
$\exists!$	there exists exactly one	$\Leftrightarrow$	is equivalent to
$\nexists$	there does not exist (any)	$\Leftarrow$	is implied by
$\forall$	for all	$\wedge$	logical "and"
:	which/for which/such that	$\vee$	logical "or"
	(alternatively: "it holds that")	$\neg$	logical "not"
$\in$	element of	(…)	delimiters of statement
$\notin$	not an element of		

Table 1: Mathematical symbols and their meaning.

which indicate whether a certain relationship holds for all ($\forall$), some ($\exists$), exactly one ($\exists!$) or none ($\nexists$) of the elements considered. To see this in a simple example, consider the mathematical way of saying that all natural numbers are non-negative:

$$\forall n \in \mathbb{N} : n \geq 0 \quad \text{or equivalently} \quad \nexists n \in \mathbb{N} : n < 0.$$

Further simple examples of quantifying statements (or equivalently, statements in quantifier notation) that you can use to practice are $(\forall n \in \mathbb{N} : n \in \mathbb{R})$, $(\exists x \in \mathbb{R} : x \notin \mathbb{N})$ or $(\nexists n \in \mathbb{N} : n \cdot \pi \in \mathbb{N})$, where the last one states that there does not exist any natural number for which the product with pi is a natural number.[2]

While most symbols in Table 1 are straightforward to understand, the logical operators $\wedge$, $\vee$ and $\neg$ may require some explanation. The logical "or" has a meaning slightly different from standard English, and more precisely translates to "and/or", i.e. it does not preclude that both statements are true, but requires at least one to be true. As an example, consider the statement $(1 \in \mathbb{R} \vee 1 \in \mathbb{N})$, which asserts that the number 1 is an element of the real numbers "or" an element of the natural numbers.[3] The logical "not" reads as "it is not the case that" and inverts the meaning of a statement, asserting the exact opposite. We can frequently re-write these statements as more natural expressions, e.g. $\neg(1.5 \in \mathbb{N})$ as $1.5 \notin \mathbb{N}$, or $\neg(\exists n \in \mathbb{N} : n \cdot \pi \in \mathbb{N})$ as $\nexists n \in \mathbb{N} : n \cdot \pi \in \mathbb{N}$. This simplification abides exactly by the same logic as the one we also use in "normal" English where, for instance, "there are some people who own cats" is a more direct way of saying "not all people do not own cats".

0.2.2 Statements

In our discussion of Table 1 above, we have already considered statements without explicitly defining them. Formally, a statement is any combination of mathematical vocabulary. Statements may or may not be meaningful, i.e. "make sense", whenever reading them out verbally they give you a grammatically correct English sentence.

Furthermore, statements may or may not be true. Even meaningful mathematical statements need not be true; consider e.g. $(5 > 10)$ or $(\forall n \in \mathbb{N} : n < 5)$. A statement that is not meaningful can never be true: if there is no meaning, there is nothing to be contrasted against

[2] For further reading on quantifiers, you may consult https://en.wikipedia.org/wiki/Quantifier_(logic).
[3] The statement $(1 \in \mathbb{R} \wedge 1 \in \mathbb{N})$ is true as well, however, here we require that both partial statements are true.

English	Mathematical	meaningful (M) and true (T)?
Alaska European capital	$a\ EC$	not M
Alaska is a European capital	$a \in EC$	M but not T
Berlin is a European capital	$b \in EC$	M and T

Table 2: Meaningful and true statements in different languages.

the universe of "true" circumstances. To (re-)familiarize yourself with these notions, consider Table 2, where a denotes "Alaska", b "Berlin", and EC the set of European capitals.

Table 3 gives some examples of how basic and logical symbols can be combined to statements. First cover columns 2 and 3 and see whether you can identify the verbal meaning of the mathematical statement and assess whether it is true!

Math. statement	verbal meaning	statement true?
$\exists x \in \mathbb{R} : x \in \mathbb{Q}$	There exists a real number x which is a rational number.	+ (e.g. 1)
$\forall x \in \mathbb{R} : x \in \mathbb{Q}$	For all real numbers x, it holds that x is a rational number.	– (e.g. π)
$(x \geq 5 \wedge y \leq 4) \Rightarrow x \geq y$	x being greater than 5 and y being smaller than 4 implies x being greater than y.	+
$(x \geq 5 \vee y \leq 4) \Rightarrow x > y$	x being greater than 5 or y being smaller than 4 implies x being strictly greater than y.	–
$\forall x \in \mathbb{R} : (\exists! y \in \mathbb{R} : x = y)$	For all real numbers x it holds that there exists exactly one real number y for which x is equal to y.	+

Table 3: Mathematical statements using symbols.

Note that the delimiting brackets are of crucial importance in the third and fourth statement, because only with them, it is clear that the implication ($\Rightarrow$) refers to the whole statement in parentheses rather than just to "$y \leq 4$". In the last example, on the other hand, they are just there to increase clarity and would probably be left out by many. Second, the use of colons between multiple quantifiers is rather uncommon (despite, strictly speaking, more correct). Thus, you would commonly expect to read the last statement as $\forall x \in \mathbb{R} \exists! y \in \mathbb{R} : x = y$. Third, when we say "greater/smaller than", we typically include equality, and more precisely mean "weakly greater/smaller than" or respectively, "greater/smaller than or equal to". When we do not include equality (e.g. $x > 5$), we always say "strictly greater/smaller than". So, whenever you don't read the word "strictly", assume that equality is included! Finally, while it may take some time to get used to the notation, you should be able to clearly see the notation's value added by simply comparing the space needed for the mathematical and verbal statements.

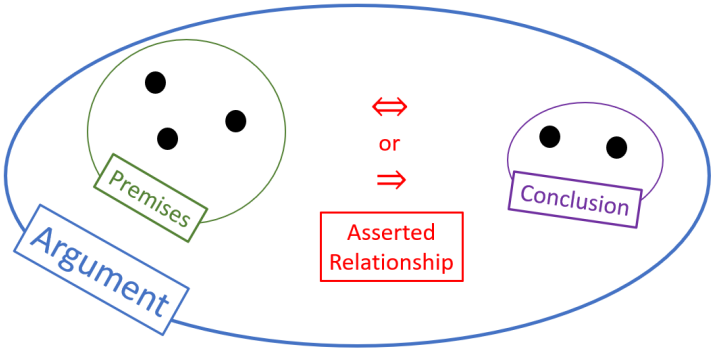


Figure 1: General structure of a mathematical argument. Black dots indicate individual statements.

0.2.3 From Statements to Arguments and Reasoning

Most of the time, it is obvious to assess whether individual *statements*, as we have seen in Table 2 and statements 1, 2, and 5 in Table 3, are true. The more essential part of mathematical analysis is how certain statements relate to each other. We call an *argument* an assertion of a relationship between a set of *premises* and a *conclusion* (see Figure 1). Typically, the assertion is either that the premises imply ("$\Rightarrow$") the conclusion or that they are equivalent to it ("$\Leftrightarrow$"). Indeed, you have already seen two arguments in Table 3, namely statements 3 and 4. Here, you have also seen that the individual premises "$x \geq 5$" and "$y \leq 4$" could be combined to a single statement, namely "$(x \geq 5 \wedge y \leq 4)$". Indeed, this is what we always do to collect our premises when expressing the argument as a mathematical statement. From the examples above, you can also see that the argument is nothing but a special type of statement, namely one that asserts a relationship between individual sub-statements!

We can use *basic logic* to internally investigate whether the argument "makes sense", i.e. whether the asserted relationship between premises and conclusion holds, while remaining agnostic about the plausibility of the premises. If it does, we call the argument **valid**. The fundamental definitions of mathematics always unambiguously determine whether an argument is valid or not. For instance, the argument that the premises "ants are taller than humans" and "humans are taller than elefants" imply the conclusion that "ants are taller than elefants" is indeed valid, since the conclusion logically follows from the premises.

If a valid argument *additionally* has true premises, it is called **sound**. It is worthwhile to stress that validity is always needed for soundness – an invalid argument can never be sound! Unlike with validity, the assessment of soundness, i.e. whether or not some premises are true, may be context-specific. For instance, the argument *if (Premise 1) "f is a differentiable function" and if (Premise 2) "any differentiable function is continuous", then (Conclusion) "f is a continuous function"* is valid. Whether or not it is sound depends on the premises - the latter premise is, as we will see in Chapter 3, a general statement that is always true, whereas the former depends on how the concrete function f is defined in the context we are concerned with.

So what precisely is basic logic? Basic logic can be thought of as the fundamental rules that determine whether a certain mathematical argument is true, or in technical terms, valid. It is hard-wired in the mathematical universe and, roughly speaking, refers to the rules implied by the general definitions of mathematical objects, for instance logical operators, set operators, the definition of sets of numbers and operations on its elements, etc.

To see this abstract elaboration in action, let us consider how basic logic helps us in the example of ants, humans and elephants, we can in fact proceed intuitively. Remaining with the example of standard English, we know *logically* that if one thing is larger than the other, and this other thing is again larger than a third thing, then the first thing must also be larger than the third - this is just *common sense*. Mathematically, all we do in this example is to compare positive real numbers with each other: the elements in the sets of ants' (A), humans' (H) and elephants' (E) numerical heights. We consider the argument

$$(\forall a \in A : (\forall h \in H : a > h) \wedge \forall h \in G : (\forall e \in E : h > E)) \Rightarrow (\forall a \in A : (\forall e \in E : a > e))$$

The basic mathematical reason that this relationship is true (i.e. the "mathematical common sense" that justifies the argument) is transitivity of the "strictly-greater-than" relation on the real numbers, namely that if for $x, y, z \in \mathbb{R}$, $x > y$ and $y > z$, then also $x > z$.

In this rather simple example, transitivity of the ">"-relation is the entire fundamental mathematical reason why this argument is valid. Still, it is already non-obvious how exactly this circumstance justifies validity of the argument. This is especially true for more complex arguments that depend on a multitude of fundamental mathematical facts. This is typically where *mathematical proofs* come in: they provide a step-wise decomposition of how fundamental mathematical circumstances make certain arguments valid or invalid.[4] Proofs are not only helpful to establish validity of an argument, but also invalidity. In fact, for quantifying statements using $\forall$, these proofs are usually much easier to come up with as it suffices to give a specific example where the asserted relationship does not hold.[5]

Excursion: Inductive and Deductive Reasoning and Proof by Induction.

Most proofs we typically consider are *inductive*, which means that we establish an argument as a whole by showing that the asserted relationship holds for any specific case. For instance, if the argument starts with "$\forall x \in \mathbb{R} : \ldots$", then our proof would start out as "Let $x \in \mathbb{R}$", where investigate whether the relationship holds for this fixed but arbitrary number. Moreover, when concerned with the natural numbers, a common approach is the so-called proof by (complete) induction, where you show that a statement "$\forall n \in \mathbb{N} : \ldots$" or "$\forall n \in \mathbb{N}\backslash\{0\} : \ldots$"[a] holds by (i) establishing the relationship for the smallest element considered in the argument, i.e. either zero or one, and (ii) showing that if the relationship holds for any fixed $n \in \mathbb{N}$, then it also holds for $n' = n + 1$. Like this, if starting at 1, these two steps prove that the relationship holds for $n = 1$ and thus for $n' = 1 + 1 = 2$, and thus for $n' = 2 + 1 = 3$, and so forth, which establishes the statement as a whole for all $n \in \mathbb{N}$. The opposite of the inductive proof is the *deductive* proof, which builds on the realization that the argument is a special case of an overarching truth that is readily established, however, you will come across this approach much less frequently.

Nonetheless and interestingly, most theoretical economic reasoning is more deductive than inductive, because we write down a general model for a class of interactions to conclude on individual real economic processes that may be, with some level of abstraction, viewed as a special case of the universe of modeled interactions![b]

Still, mathematical proofs and solid (economic) reasoning have much in common: both address a circumstance that is non-obvious (otherwise, no need to prove/discuss), suggest a clear conclusion and establish it via a sequence of small steps that are readily comprehensible for anyone familiar with the key underlying concepts and terminology.

[a] We adopt the convention that zero is included in the natural numbers.

[4] In case you wonder how we would proceed for the example of ants, humans and elephants, here is the formal proof: Pick an arbitrary $a \in A$, and then an arbitrary $e \in E$. Also, pick an arbitrary $h \in H$. By premise 1, $a > h$ and by premise 2, $h > e$. By transitivity of the strictly-greater-than relation, $a > e$. This establishes that for any $a \in A$ and any $e \in E$, $a > e$, and therefore the conclusion.

[5] Suppose you are told to prove or disprove that the premises $x > 5$ and $y < 4$ imply that $x + y < 9$. Here, you can choose $x = 10$ and $y = 0$ as a contradiction to the conclusion $x + y < 9$ that is perfectly consistent with the premises. Thus, the argument "$(x > 5 \wedge y < 4) \Rightarrow x + y < 9$" is invalid.

[b] Of course, any model is a facilitated representation of the real world that does not capture all its nuances, and strictly speaking, no real-world interaction is a special case of an economic model. But indeed, this is the whole point of a model: to reduce a complex phenomenon to a number of key aspects and study a tractable scenario. A frequent example of why this may be very useful are maps: of course, putting the world into a 2-dimensional picture and omitting e.g. houses and trees is simplistic and loses information, as does depicting everything on, say, a 1:10,000 scale. But it is exactly this simplified and more compact perspective that allows the reader of a map to navigate!

Nr.	Premise 1	Premise 2	ass. rel.	Conclusion	valid	sound
1	Berlin is the German capital	Germany is part of Europe	$\Rightarrow$	Berlin is a European capital	+	+
2	Berlin is the Chinese capital	China is part of Europe	$\Rightarrow$	Berlin is a European capital	+	-
3	Berlin is the French capital	Italy is part of Asia	$\Rightarrow$	Berlin is a European capital	-	
4	Berlin is the German capital	Germany is part of Europe	$\Leftrightarrow$	Berlin is a European capital	-	
5	I was born blind	My blindness has never been healed	$\Leftrightarrow$	I have always been blind	+	-

Table 4: Arguments with 2 premises and one conclusion. Note that invalid arguments can not be investigated with respect to their soundness.

Table 4 gives some examples of arguments and their validity and soundness.[6] A few comments deem worthwhile. First, for Nr. 3, the premises do not preclude the conclusion. Therefore, provided that the premises are true, the conclusion may still be true as well. However, the argument states that the premises *necessarily imply* the conclusion, which makes it invalid, as for this, not enough information is given in the premises. The invalidity of Nr. 4 is for a similar reason, try to find out why exactly. Finally, Nr. 5 serves as an example for a valid statement of equivalence and a case where one statement (here: the conclusion) has multiple implications (here: the premises). If you are very bored, try to define mathematical notation for examples of Table 4 (e.g. start with b and EC as above) and see if you can fit all arguments in one line.

Typically, mathematical theory is more concerned with argument validity rather than soundness. Theory provides us with theorems and propositions that tell us that "if this and that is true, then also some other property will be true". You can find an abundance of examples in the remainder of the course, but to make the point very clear, let's consider the so-called Weierstrass Extreme Value Theorem (its content is not important at this point, do not worry if this does not make sense yet), which states that "If (premise 1) f is a continuous function and (premise 2) f has a compact domain then (conclusion) f must assume a global maximum and minimum." Here, f is an unspecified, hypothetical function. For concrete functions, the premises may or may not be true, but this is not essential for the usefulness of this theorem and the validity of the statement.

On the other hand, if you are working on some exercise problems or writing an exam, you will frequently be given concrete contexts (in the example of the Weierstrass theorem: concrete functions) to work with. Then, you will likely refer to all the valid arguments that you know from your textbooks and try to make sound arguments with them. Say, for instance, you are given some utility function and are asked whether it has a global maximum. Then, if your argument is that by the Weierstrass Extreme Value Theorem, this function must have a global maximum, it depends on the precise function that you are given whether your argument is sound or not.

[6] Note that strictly speaking, validity in these examples relies on many implicit truths, e.g. that European capitals are capitals of countries that are part of Europe, or that being "the German capital" is equivalent to being "the capital of Germany". When our context is the real world, this is of course the case, but at higher levels of mathematical abstraction, we need to apply some more care when using expressions that are not identical.

0.3 Implications, Equivalence, and Necessary and Sufficient Conditions

Throughout their career, any economist will hear the words "necessary" and "sufficient" quite a lot, and as they are closely related to the mathematical context, a brief introduction is given here. If you have been thinking thoroughly about the three arrows in Table 1, you will not have a hard time to understand what follows.

Suppose we are interested in some statement S. A necessary condition for S must hold for S to be true. It need not guarantee truth of S. Thus, *S is true **only if** the necessary condition is satisfied*. A sufficient condition for S guarantees that S is true. However, it need not hold for S to be true. Thus, *S is true **if** the sufficient condition is satisfied*. Finally, a necessary and sufficient (or equivalent) condition for S (i) must hold for S to be true and (ii) guarantees that S is true. Thus, *S is true **if and only if** the equivalent condition is satisfied*.

In terms of our logical arrows, let C be the condition. If C is necessary for S, then C is implied by S: $C \Leftarrow S$. If instead, C is sufficient for S, then C implies S: $C \Rightarrow S$. And if C is an equivalent condition for S – you guessed it – then C is equivalent to S: $C \Leftrightarrow S$. Note that sufficient and equivalent conditions typically make us happy if we want to establish S: their truth is enough to know that S is true. With a necessary condition C, on the other hand, we only know that S cannot be true unless C is also true. This may help disprove S: if C is not true, or respectively, the opposite of C, $\neg C$, is true, then S is not true (and $\neg S$ is true): $\neg C \Rightarrow \neg S$. Thus, violation of the necessary condition implies violation of the statement of interest.[7]

To give a practical example, consider the United Kingdom (UK)'s definition of an economic recession, which states that an economy is in the state of recession whenever GDP growth has been negative for at least two quarters. Consider the following conditions:

1. German GDP growth was negative in the last quarter.
2. German GDP growth was at -1% constantly throughout the last year.
3. The average of German GDP growth during the last two quarters was -0.25%.
4. The average of German GDP growth during the last two quarters was below zero.
5. German GDP growth was below zero in the last two quarters.

Try to figure out which of them are necessary, sufficient, equivalent or nothing at all for the German economy currently being in a recession. The answers are in this footnote.[8]

In the mathematical context relevant to economists, you come across necessary and sufficient conditions mostly in optimization, where we frequently deal with them (mostly related to second derivatives) when investigating whether a solution constitutes a maximum, a minimum, or neither. Therefore, they are at the heart of, amongst others, utility or profit maximization, cost minimization, and also error/deviation minimization of statistical estimators. As a final remark, a basic principle of proper communication is to always pick the word most adequate

[7] Note that this means that if you negate the statements (using $\neg$), you can always "flip" the implication arrow!
[8] 1: necessary, 2: sufficient, 3: nothing, 4: necessary, 5: equivalent.

to the information you wish to transmit. Conversely, do not read into a sentence more than is written. A necessary condition may also be sufficient, but if one refers to it as a necessary condition, then you should not read more in the sentence. The only relevant information is that which is clearly transmitted.

If you feel that you need further reading on what has been discussed so far, you can visit https://gowers.wordpress.com/2011/10/09/basic-logic-summary/#more-3332 for an extensive overview of mathematical notation and basic logic. You can also browse through the individual posts on this blog for even more extensive reading. At the bottom of the page, you can click a link that directs you to several quizzes, where you can test your ability to draw logical conclusions.

Excursion: Theorems, Propositions, Lemmas and Corollaries

When reading mathematical texts, you come across a range of "facts" with different names. If you are interested, you can find a brief overview of what sets apart Theorems, Propositions, Lemmas and Corollaries, which make up for almost all of these facts, below. This excursion is purely purposed to satisfy your curiosity, however. If you are short in time or simply don't find this distinction interesting, feel free to skip to the next section. Also, if you don't find it too accessible, you can wait until you have seen more Theorems and Propositions in the upcoming chapters and come back here after having become more familiar with the concepts.

All examples of the concepts introduced here usually contain one or multiple mathematical statements and/or arguments. An example that we have already seen before is the Weierstrass Extreme Value Theorem, which claimed that two properties of a function (continuity and compact support) always imply a third property, namely the existence of global extreme values. The claims made can be far more complex or simple, lengthy or short. However, this is not essential for how we call it, but rather, the name depends upon its purpose and the importance.

The most common "fact" is the proposition. It is a statement that is "interesting" by itself, and usually contains at least an important part or "setup"-result for the purpose of the text. Since by the nature of the word, some fact is proposed, propositions are always expected to come with a proof (but of course not references to propositions in other texts, as in "see Proposition 5 of Textbook XY"). Accordingly, all results labeled as "proposition" in this script feature a proof allowing you to understand step by step why they are true.

A theorem is similar but distinct, as theorems are typically of greater importance than the proposition, either to the text itself or in the relevant mathematical context. For instance, a mathematical paper would probably call its two to three main results theorems and other related, more technical insights propositions. Moreover, any fact of central importance to a mathematical (sub-) field is likely to be called a theorem, take again the example of the Weierstrass Extreme Value Theorem.

Next, a lemma typically has no immediate value for the insights to be taken away from a text, but rather, it provides a "helper fact" that facilitates proving a proposition. As such, lemmas most frequently occur directly before propositions requiring rather complex,

multi-step proofs, and their predominant value lies in organizing the structure of the line of reasoning presented as proof.

Finally, a corollary is something that follows rather immediately – without any or at most with one to two lines of proof – from one or more other facts. But just because corollaries are easy to establish given previous considerations does not mean they are not important, and some very important theorems are indeed corollaries!

What is true for all of the concepts mentioned here is that they give you a mathematical fact. These facts can be complex and/or unintuitive and it may be hard to immediately see why they are true. Naturally, you may therefore ask, when would we expect to see a proof? Well, for any of the concepts, when a text states them for the first time, they are expected to come with a proof immediately below it to allow the reader to judge upon their validity. Further, if the text's main purpose is educational, proofs are also given for existing results so that they don't fall from the sky for the reader. If the proof is not too essential for what the text wants to convey, you will frequently see reference to a resource giving the proof. Only if results are sufficiently well-established in the relevant mathematical context (e.g. the Weierstrass Extreme Value Theorem in the context of optimization), you will find that no proof is given at all.

0.4 Set Theory

As any good mathematical text should, let us begin our discussion of sets by defining what precisely we will be studying. Since this will be our first definition, the discussion below also outlines some general key insights into reading mathematical definitions. You should have in mind the following idea: Every word counts, missing words too!

Definition 1. ***(Element, Set)*** *A set, S, is a collection of distinct objects, considered as a whole. An object s in a set S is called an element or member of S, denoted $s \in S$. For an object s' that is no element of S, we write $s' \notin S$.*

Given our view of mathematics as a distinct language, some of its words can take a specific meaning and need not exactly coincide with the one we know from regular English, and it is perfectly possible and common for mathematicians to define distinct meanings for words that already exist in the English language, like the word "set" in this case. It is important to make sure you know **the** meaning of every word in a definition. The emphasis here is on "the", for mathematical expressions rarely have several meanings, as that could generate misunderstandings.[9] The knowledge of these meanings is mostly gained by regular interaction with the words. In the above definition, for instance, the word "object" should be understood as "any entity that is potentially of interest to the modeler." Therefore, depending on the context, objects can be real numbers, but also functions, matrices, geometrical figures, or even sets themselves! The word "collection" is to be understood as one of its standard English meanings, i.e., as group of things.

Moreover, in good mathematical definitions, no word is redundant, and the meaning does not go beyond what is written. In our example, the word "distinct" suggests that sets do not

[9] The converse is not true, as one can see from our definition where "element" and "member" are synonyms.

contain duplicates: thus, the collection $\{1, 2, \pi\}$ may represent a set, while the one $\{1, 2, \pi, \pi\}$ may not. Moreover, "considered as a whole" suggests that the set itself should be seen as a distinct object. Conversely, the definition says nothing about the order of elements in a set, so that we may infer that the sets $\{1, 2, \pi\}$ and $\{2, \pi, 1\}$ are identical.

In terms of notation, familiar with the way the sets above are written: two curly braces, and within them the characterization of its elements. The word "characterization" is used deliberately rather than "list", because typically, the sets we deal with are too big to list all its elements or even contain infinitely many of them, consider e.g. the set of natural numbers $\mathbb{N}$. More generally, we define sets by a mathematical *statement* as introduced above that characterizes the elements. How this works exactly can readily be seen from the definition of intervals given below. Note that it refers to the extended set of real numbers $\bar{\mathbb{R}}$ that encompasses all real numbers, $\mathbb{R}$, as well as $\{-\infty, +\infty\}$.

Definition 2. ***(Real-valued Interval)*** *A real-valued interval, I, is a set that contains all $x \in \mathbb{R}$ in between two thresholds. With $a, b \in \bar{\mathbb{R}}$, $a \leq b$, we denote*

$$[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}, \quad [a, b) := \{x \in \mathbb{R} : a \leq x < b\},$$
$$(a, b] := \{x \in \mathbb{R} : a < x \leq b\}, \quad (a, b) := \{x \in \mathbb{R} : a < x < b\}.$$

If $I = (a, b)$, I is called open, and if $I = [a, b]$, we call I closed. Else, we call I semi-open. If $a = -\infty$, then the lower bound must be open. Conversely, if $b = \infty$, the upper bound must be open.

When reading this definition, note that "in between" says nothing about whether the thresholds are included in the interval or not – indeed, this is inessential for our understanding of what constitutes an interval. As can be seen, in terms of notation, a round bracket indicates that the threshold value is *not* included in the interval, whereas a square bracket indicates its inclusion.

In set theory, a key concept is the **subset**. For the sets A and X, we say that A is a subset of X, denoted by $A \subseteq X$, whenever all elements of A are contained in X, formally $(\forall x \in A : x \in X)$. A is a *proper subset* of X, $A \subset X$, if all elements of A are contained in X but there is at least one element in X that is not an element of A, i.e. $(A \subseteq X \wedge \exists x \in X : x \notin A)$.[10] The approach we adopt toward set theory is the so called "naive" approach. It is naive in the sense that it is not axiomatic. For an economist, there are no costs but many benefits to follow this simpler approach. In order to avoid paradoxes,[11] however, one needs to assume that every set we consider is itself a subset of a fixed, all encompassing set called the *universal (super)set*, which we denote by X. In addition, one defines an "encompassed by all" subset, called the *empty set*, and conventionally denoted $\emptyset$. For every set A, we thus have $\emptyset \subseteq A \subseteq X$. The empty set is always the same and contains no elements, while the universal set varies across applications, so that we may have $X = \mathbb{R}$ when considering sets of real numbers, and $X = \mathbb{R}^n$ for sets of real-valued vectors of length $n \in \mathbb{N}$.

0.4.1 Basic Concepts

[10] Alternatively, some use $\subset$ as the subset, and indicate a proper subset using $\subsetneq$. Be aware of this, and critically think about the meaning of "$\subset$" when you see the symbol.

[11] The interested reader can have a look at Russell's paradox.

Now it is time to consider some key concepts related to sets. To define them, let $A, B \subseteq X$, where X is the universal superset.

- *Set equality.* A and B are said to be equal whenever they contain the same elements, i.e. set equality "$A = B$" is equivalent to $\forall x \in X : (x \in A \Leftrightarrow x \in B)$.
- *Disjoint sets.* A and B are said to be disjoint whenever they have no elements in common, i.e. $\forall x \in X : ((x \in A \Rightarrow x \notin B) \wedge (x \in B \Rightarrow x \notin A))$ (recall that "$\wedge$" is the logical "and").
- *Superset.* B is a superset of A whenever A is a subset of B: $B \supseteq A \Leftrightarrow A \subseteq B$.
- *Complement.* B is the complement of A (with respect to X) whenever it contains all those elements of X that are not contained in A: $B = \{x \in X : x \notin A\}$. We usually denote the complement of A as A^c.

As with real numbers, we can perform *operations* on sets:

- *Union.* $A \cup B := \{x \in X : (x \in A \vee x \in B)\}$ (with $\vee$ as the logical "or"). The union contains all elements that are contained in A, B, or both.
- *Intersection.* $A \cap B := \{x \in X : (x \in A \wedge x \in B)\}$. The intersection contains all elements that are contained in both A and B.
- *Difference.* $A \backslash B := \{x \in X : (x \in A \wedge x \notin B)\}$. The difference of A and B contains all elements of A that are not contained in B.

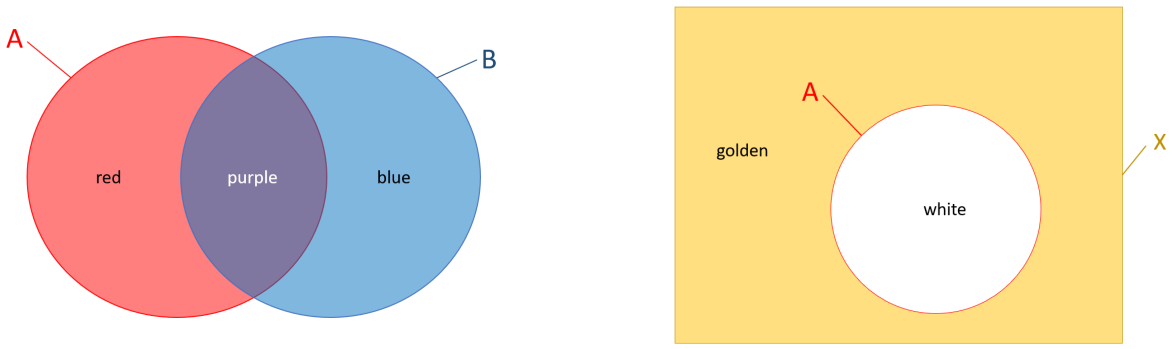


Figure 2: Left: red area = set difference $A \backslash B$, blue = set difference $B \backslash A$, purple = intersection $A \cap B$, and the union $A \cup B$ contains all three of the colored areas. Right: golden = complement A^c of A with respect to the universal superset X.

An illustration of the characterizations and operations is given in Figure 2. These operations facilitate our lives greatly in many dimensions: e.g. the somewhat awkward definition of disjoint sets above, where we required that $\forall x \in X : ((x \in A \Rightarrow x \notin B) \wedge (x \in B \Rightarrow x \notin A))$, can be simply re-written as $A \cap B = \emptyset$. The symbol ":=" indicates a *defining equality*, and is used whenever we introduce a new object of interest.[12] Alternatively, you will sometimes see "$\equiv$".

An overview of basic set notation is given in Table 5.

[12] In accordance with the introduction of ":" in Table 1, you can read "Let $S := \{\ldots\}$" as "let S (be) *such that* it is equal to the set …". In this sense, ":=" is not a new symbol, but rather a combination of two familiar ones!

$\varnothing$	the empty set	$\in$	element of
$\subseteq$	is a subset of	$\notin$	not an element of
$\subset$	is a proper subset of	$\setminus$	set difference
$\supseteq$	is a superset of	$\cup$	union
$\supset$	is a proper superset of	$\cap$	intersection

Table 5: Set notation: overview.

0.4.2 Sets of Sets and Index Sets

So far, we have not yet explicitly addressed that elements of sets may be anything else than usual, real numbers. To address this aspect, consider the power set: When A ($A \subseteq X$) denotes a set, then the power set of A is $\mathcal{P}(A) := \{S \subseteq X : S \subseteq A\}$, i.e. the set of subsets of A. Note that $\varnothing \in \mathcal{P}(A)$ for any A, as the empty set is the "encompassed-by-all" set, as introduced above. To give an example, $\mathcal{P}(\{1,2\}) = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$. Note that this class of sets is subject to a different universal set than $\{1,2\}$. However, it is easily verified that $\mathcal{P}(X)$ is a suitable universal set for the power sets of A, $A \subseteq X$, as

$$\mathcal{P}(A) = \{S \subseteq X : S \subseteq A\} \subseteq \{S \subseteq X\} = \mathcal{P}(X) \qquad \forall A \subseteq X.$$

At times, it may be convenient to give the individual objects in the set an index, so that we may write $A = \{S_1, S_2, \ldots, S_n\}$, $n \in \mathbb{N}$, or equivalently $A = \{S_i : i \in \{1, \ldots, n\}\} = \{S_i\}_{i \in \{1,\ldots,n\}}$.[13] Of course, we can use a more general index set, denote it by I, that need not be equal to $\{1, \ldots, n\}$ for an $n \in \mathbb{N}$.[14] When the elements of A are indeed sets, we can elegantly use the index set for short notations for multiple intersections or unions:

$$\cup A := \bigcup_{i \in I} S_i := \{x \in X : (\exists i \in I : x \in S_i)\}, \quad \cap A := \bigcap_{i \in I} S_i := \{x \in X : (\forall i \in I : x \in S_i)\}.$$

Finally, we say that the collection $A = \{S_i : i \in I\}$ of sets S_i is *pairwise disjoint* whenever any two elements of A are disjoint, i.e. $\forall i, j \in I : (i \neq j \Rightarrow S_i \cap S_j = \varnothing)$.

As with the operations on real numbers, it is possible to establish a range of properties that set operations satisfy. Let us have a look at the ones most frequently used in economics:

Theorem 1. ***(Properties of Set Operations)***
Let $A, B, C \subseteq X$ for a universal set X, and $\mathcal{S} = \{S_i : i \in I\}$ for an index set I, where $S_i \subseteq X$ $\forall i \in I$. The following properties hold:
(i) Commutativity: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
(ii) Associativity: $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.
(iii) Distributivity: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
(iv) Simple De Morgan Laws: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.
(v) General De Morgan Laws: $(\bigcup_{i \in I} S_i)^c = \bigcap_{i \in I} S_i^c$ and $(\bigcap_{i \in I} S_i)^c = \bigcup_{i \in I} S_i^c$.

Of course, you don't need to memorize these facts by heart. But when facing a problem

[13] Depending on your econometrics background, you may have seen that one writes the random variables associated with a sample of size n in similar fashion, namely, $\{(Y_1, X_1), \ldots, (Y_n, X_n)\} = \{(Y_i, X_i)\}_{i \in \{1,\ldots,n\}}$.

[14] We distinguish finite, countable and uncountable index sets. The set is finite if (and only if) it contains only finitely many elements. The distinction between "countable" and "uncountable" is not too important here.

related to sets, you may want to look these up, because they oftentimes make your life much easier. Also, these rules are a good opportunity to re-familiarize yourself with the expressions Commutativity, Associativity and Distributivity, and they may also be helpful in developing a better intuition for sets using the circle-approach introduced in Figure 2 - take a piece of paper and see whether you can visually "prove" the simple De Morgan laws!

0.5 Functions, Relations and Limits

The last introductory section is concerned with functions and limits. This script gives an introduction to functions using the concept of relations, partly for formal precision, but also to remind you that relations, as you may come across in your micro-oriented classes when studying (consumer) preference, are nothing fancy, just a generalization of the concept of functions.

0.5.1 Functions and Relations

To understand the concept of relations,[15] consider the *Cartesian product* $X \times Y$ of two sets X and Y, defined as

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

Then, a binary relation R from X to Y is nothing but a **subset** of $X \times Y$: $R \subseteq X \times Y$, and if $(x, y) \in R$, we say that y is an *image* of x under the relation R,[16] and write xRy or $y \in R(x)$, where

$$R(x) = \{y \in Y : xRy\} = \{y \in Y : (x, y) \in R\} \qquad \text{for any } x \in X. \tag{1}$$

Note that the sets $R(x)$, $x \in X$, are a complete characterization of the relation R, this will be important in a second. Moreover, for any fixed $x \in X$, $R(x)$ can be empty or contain multiple arguments. As an example, consider $X = Y = [0, 1]$, where the relation R_1 is defined as **the set**

$$R_1 = \{(x, y) \in [0, 1]^2 : x > y\}$$

where we use the common notation $[0, 1]^2 := [0, 1] \times [0, 1]$. Then, $R_1(x) = \{y \in [0, 1] : y < x\}$, so that $R_1(0) = \varnothing$ and $R_1(x) = [0, x)$ for any $x > 0$. Another example that is frequently discussed in undergraduate economics courses (with varying degree of formality) are *preference relations*, where $X = Y$ contains vectors of goods quantities, and for a consumer i, the relation is given by $\{(x_1, x_2) \in X \times X : x_1 \succsim_i x_2\}$.[17]

Intuitively, it should be clear that we can view a "function" as introduced in high-school courses as a relation, since the values for x are related to $y = f(x)$ through the function. Indeed, this is what we call a function also more formally: any relation that assigns **exactly one value** y to every argument x. So, if we call f a function, that means that for any x, $f(x)$ must be a single object (e.g. real number, but also vectors, matrices, etc., as we will see later), and not a set!

Let us go over the line of reasoning defining a function as a relation step by step. Once you

[15] For more detail, see dlF, p. 15.

[16] The relation is binary because any (x, y) is either an element of R or not, and there is no (continuous) "degree of relatedness".

[17] The interested reader can review under what conditions a relation is reflexive, symmetric, antisymmetric or transitive; these concepts are quite central to the study of preference relations in microeconomics.

have understood this, you will be familiar with the names and nature of all the fundamental concepts relevant for a function, including the *domain*, *codomain*, *image* and *graph*, which are very important for everything to follow!

The classical form of a function f that you are likely well-familiar with is the one of a rule which associates every element $x \in X$ in the *domain* X of f with a single element $y \in Y$ in the *codomain* Y of f. We write

$$f : X \mapsto Y, x \mapsto y =: f(x).$$

This statement is a concise summary of all relevant aspects of f: the domain X, the codomain Y, and the rule $f(x)$ that maps x's into y's. Note that two functions are identical if and only if the mapping $x \mapsto y$ and the domain X coincide; the codomain may well be different (consider e.g. $f_1 : \mathbb{R} \mapsto \mathbb{R}, x \mapsto x^2$ and $f_2 : \mathbb{R} \mapsto \mathbb{R}_+, x \mapsto x^2$ where $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ is the set of non-negative reals. Then, f_1 and f_2 are clearly identical). To see the connection to relations, consider the *graph* $G(f)$ of f,

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

Clearly, $G(f)$ is a subset of $X \times Y$, since it contains only elements in $X \times Y$ and adds the restriction $y = f(x)$, which may exclude some elements. Like this, we can view the graph $G(f)$ as the relation from the domain X to the codomain Y, since the set of y's related to any fixed $x \in X$ under $G(f)$, denoted $R(x)$ above (cf. equation (1)), is simply

$$\{y \in Y : xG(f)y\} = \{y \in Y : (x, y) \in G(f)\} = \{y \in Y : y = f(x)\} = f(x)$$

where the last equality is because Y is a set and sets do not contain duplicate elements. This highlights the function property that it assigns only one image $f(x) = y$ to any one $x \in X$. Conversely, this highlights that when viewing relations as a generalization of functions, the set $R(x)$ can be interpreted as a generalized image of x under R, in a fashion very similar to standard functions. As take-away, one may summarize

1. The graph $G(f)$ of the function f is **a set** and defines f as a relation. (The graph contains the combinations $(x, f(x))$ and is **not** just a collection/picture of the $f(x)$'s!)

2. The relation is fully characterized by *(i)* the rule $x \mapsto y = f(x)$ that maps domain objects $x \in X$ onto codomain objects $y \in Y$, and *(ii)* the domain X.

Before moving on, a conceptual note. You may be used to calling "$f(x)$" a function, e.g. from high school. If so, you should **stop doing this now**. Indeed, people sometimes do this, especially at lower levels of mathematics, but this is arguably imprecise/wrong. $f(x)$ may refer to a specific element in the codomain of f, the value of f when evaluated at a concrete $x \in X$, or, when considering x as a variable, as the mapping rule $x \mapsto y = f(x)$ of the function f.[18] However, neither case provides sufficient information to fully characterize f (in the latter, it is still unclear what domain and codomain are), and you run into troubles related to notation

[18] You may be familiar with this case from specific representations like "$f(x) = x^2 + \sin(x)$", which unambiguously summarizes the mapping rule.

when it comes to differentiation (see also the discussion in Chapter 3). To be formally precise, in everything to follow, we will call f the function, $x \in X$ an argument and the object $f(x)$ in the codomain of f the value of f at x, and $x \mapsto y = f(x)$ the mapping rule of f.

0.5.2 Key Concepts related to Functions

Finally, let us consider some further important concepts that you will come across frequently in the function context. Again, you don't need to memorize this by heart by now - just try to become familiar with the expressions and keep in mind that you can look them up here if you come across them and are unsure as to what they mean precisely.

For what follows, let $f : X \mapsto Y$ be a function as defined above, and in addition, let $g : Y \mapsto Z$ be another function. Then,

- For a set $A \subseteq X$, the *image of A under f* is $f[A] := \{y \in Y : (\exists x \in A : f(x) = y)\}$, i.e. the set of $y \in Y$ to which f maps.
- For a set $B \subseteq Y$, the *preimage of B under f* is $f^{-1}[B] := \{x \in X : (\exists y \in B : f(x) = y)\}$, i.e. the set of $x \in X$ that are mapped onto elements in B by f.
- If for any $y \in Y$, there exists exactly one $x \in X$ so that $f(x) = y$, (in quantifier notation: $\forall y \in Y : (\exists! x \in X : f(x) = y)$), then we can define the *inverse function* $f^{-1} : Y \mapsto X, y \mapsto x = f^{-1}(y)$ where $f^{-1}(y) \in X$ is such that $f(f^{-1}(y)) = y$.
- The *composition* $h = g \circ f$ of g and f is defined as $h : X \mapsto Z, x \mapsto g(f(x))$.
- f is called *monotonically increasing (decreasing)*, if for any $x_1, x_2 \in X$, $x_1 \geq x_2$ implies $f(x_1) \geq f(x_2)$ ($f(x_1) \leq f(x_2)$), and *strictly monotonically increasing (decreasing)* if for any $x_1, x_2 \in X$, $x_1 > x_2$ implies $f(x_1) > f(x_2)$ ($f(x_1) < f(x_2)$).

That the word "range" is frequently used synonymously for the image of X under f (also denoted as $\text{im}(f)$) and rarely also for the codomain, where the latter is however technically imprecise. Thus, make sure to thoroughly question its meaning when you come across this word! Next, an alternative name for the preimage is "inverse image", which may be somewhat misleading and easily confused with the image of the inverse function. Thus, let us not use this label, but be aware that some other texts and courses may do so. The inverse function will be investigated more thoroughly later, but you can already note that (i) its existence depends crucially on the definition of the codomain Y as well as the mapping $x \mapsto y$, and (ii) that despite looking quite similar, **the expressions $f^{-1}(y)$ and $f^{-1}[\{y\}]$ refer to fundamentally different concepts!** To tell them apart more easily, one sometimes uses square brackets for (pre-)images of sets and round ones for (inverse) images of single elements, as we have done above. Make sure that you understand this difference!

As a last note on functions, Table 6 gives common rules for derivatives of functions where both domain and codomain are $\mathbb{R}$. Since here, f only has a single (uni-dimensional) argument and it is clear with respect to which variable the derivative is taken, we may simply write $f'(x)$.[19] Note that any letters other than x refer to constants in $\mathbb{R}$ and are *not* arguments of

[19] As we will see later, when $f(x) = f(x_1, \ldots, x_k)$, we need to specify the variable(s) with respect to which we take the derivative.

Derivatives of specific functions

Function $f(x)$	Derivative $f'(x)$	Function $f(x)$	Derivative $f'(x)$
c	0	$c \cdot x$	c
$\ln(x)$	$\frac{1}{x}$	$\exp(x)$	$\exp(x)$
$\sin(x)$	$\cos(x)$	$\cos(x)$	$-\sin(x)$
c^x	$\ln(c) \cdot c^x$	x^c $(c \neq 0)$	$c \cdot x^{c-1}$

Rules for Derivatives

Name	Function	Derivative
Sums Rule	$f(x) + g(x)$	$f'(x) + g'(x)$
Product Rule	$f(x) \cdot g(x)$	$f'(x) \cdot g(x) + f(x) \cdot g'(x)$
Quotient Rule	$f(x)/g(x)$	$(f'(x) \cdot g(x) - f(x) \cdot g'(x))/(g(x))^2$
Chain Rule	$(g \circ f)(x) = g(f(x))$	$f'(x) \cdot g'(f(x))$

Table 6: Important derivatives.

the function. You should have a good command of these rules, as being able to apply them correctly will make your life a lot easier in virtually any economics class.

0.6 Limits and Continuity in $\mathbb{R}$

To conclude our investigations into the fundamental background concepts of mathematics that are relevant to the context of the economist, we consider the limit concept in relation to the real line, both for sequences of numbers and univariate, real-valued functions.

0.6.1 Limits of Sequences

Let $\{x_n\}_{n\in\mathbb{N}}$ be a *sequence* of real numbers, i.e. $\forall n \in \mathbb{N} : x_n \in \mathbb{R}$. Then, we call $x \in \mathbb{R}$ the limit of this sequence if[20]

$$\forall \varepsilon > 0 \exists N \in \mathbb{N} : (\forall n \in \mathbb{N} : (n \geq N \Rightarrow |x_n - x| < \varepsilon)).$$

Verbally, for any, and thus especially any arbitrarily small number ε, there exists a threshold N after which the sequence elements only deviate from x by less than ε, such that eventually, as $n \to \infty$, the sequence elements will lie arbitrarily close to x. If the limit x of the sequence $\{x_n\}_{n\in\mathbb{N}}$ exists, we write $x = \lim_{n\to\infty} x_n$. Crucially, we also write that $\lim_{n\to\infty} x_n = \infty$ if

$$\forall x \in \mathbb{R} \exists N \in \mathbb{N} : (\forall n \in \mathbb{N} : (n \geq N \Rightarrow x_n > x)),$$

i.e. if the sequence elements eventually exceed any arbitrarily large but fixed number x. A similar characterization can be written down for $\lim_{n\to\infty} x_n = -\infty$ (see footnote, but try writing it down on your own first!).[21]

Because function limits usually receive relatively less attention in undergraduate economics programs, let us now study this issue, which is, as you will shortly see, highly similar to the sequence limit concept.

[20] It is common to omit the expression "if and only if" in definitions because it is clear that the statement is *defining* of the property to be defined and thus equivalent to it. Thus, don't be surprised to only read "if" in mathematical definitions although the concepts are equivalent.

[21] $\forall x \in \mathbb{R} \exists N \in \mathbb{N} : (\forall n \in \mathbb{N} : (n \geq N \Rightarrow x_n < x))$.

0.6.2 Limits of Functions

When $X, Y \subseteq \mathbb{R}$, we call $f_a \in \mathbb{R}$ the limit of the function $f : X \mapsto Y$ at $a \in \mathbb{R}$, if

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in X : (|x - a| \in (0, \delta) \Rightarrow |f(x) - f_a| < \varepsilon)).$$

The concept is similar to the standard limit of a sequence: for any arbitrarily small $\varepsilon > 0$, there must exist a *neighborhood* $N = (a - \delta, a + \delta)$, $\delta > 0$ such that f deviates from f_a by less than ε on N. In other words, by choosing x sufficiently close to a, one may ensure that f deviates from f_a no more than ε. We write $f_a = \lim_{x \to a} f(x)$. Note that we need not have $a \in X$, so that a can either be a boundary point (e.g. $a = 0$ when f is defined on $(0, \infty)$) or a point where f is not defined (e.g. $a = 2$ when $f(x) = 1/(x - 2)$). Further, we adopt the convention that if for any sequence $\{x_n\}_{n \in \mathbb{N}}$, where $x_n \in X \ \forall n \in \mathbb{N}$, so that $\lim_{n \to \infty} x_n = a$, it holds that $\lim_{n \to \infty} f(x_n) = \infty$ ($\lim_{n \to \infty} f(x_n) = -\infty$), then we write $\lim_{x \to a} f(x) = \infty$ ($\lim_{x \to a} f(x) = -\infty$).

To characterize the asymptotic behavior of a function f with domain $\mathbb{R}$ or intervals unbounded to one side (e.g. $(-\infty, a], (b, \infty)$, etc.), one frequently considers the limits $\lim_{x \to \infty} f(x)$ and $\lim_{x \to -\infty} f(x)$. Here, it is important to know how these quantities are defined. We write $\lim_{x \to \infty} f(x) = c$ for a $c \in \mathbb{R}$ if

$$\forall \varepsilon > 0 \exists x_\varepsilon \in \mathbb{R} : (\forall x > x_\varepsilon : |f(x) - c| < \varepsilon)$$

Try to write down the analogous formal statement that defines $b \in \mathbb{R}$ as the left asymptote of f, $\lim_{x \to -\infty} f(x)$.

ANSWER (toggle show/hide): $\forall \varepsilon > 0 \exists x_\varepsilon \in \mathbb{R} : (\forall x < x_\varepsilon : |f(x) - b| < \varepsilon)$ TOGGLE ANSWER END

As with the limit at a point $a \in \mathbb{R}$, we write $\lim_{x \to \infty} f(x) = \infty$ ($\lim_{x \to \infty} f(x) = -\infty$) if for any sequence $\{x_n\}_{n \in \mathbb{N}}$, where $x_n \in X \ \forall n \in \mathbb{N}$, so that $\lim_{n \to \infty} x_n = \infty$, it holds that $\lim_{n \to \infty} f(x_n) = \infty$ ($\lim_{n \to \infty} f(x_n) = -\infty$), and analogously for $\lim_{x \to -\infty} f(x) = \infty$ ($\lim_{x \to -\infty} f(x) = -\infty$).

An important point is that $\lim_{x \to a} f(x) = f(a)$ need not necessarily hold. Consider, for instance, $a = 0$ and $f(x) = 1/x$, where $f(0)$ is not even defined ($a \notin X$). Next, consider the indicator function $f(x) = \mathbb{1}[x > 0]$ on $\mathbb{R}$ that is equal to 1 if $x > 0$ and zero else. It is defined at $x = 0$, i.e. $a = 0 \in X$, but for any $f_a \in \mathbb{R}$ and any $\varepsilon < 1$, there exists no $\delta > 0$ such that $|f(x) - f_a| < \varepsilon$ for all $x \in (-\delta, \delta)$ because $f(x) = 0$ for $x \in (-\delta, 0]$ and $f(x) = 1$ for $x \in (0, \delta)$. Thus, $\lim_{x \to 0} f(x)$ does not exist, and especially, $\lim_{x \to a} f(x) = f(a)$ does not hold. Finally, even if the limit exists, the equation need not hold. Look at the function f with $f(x) = \mathbb{1}[x = 0]$ that is equal to 1 at $x = 0$ and zero else. Then $\lim_{x \to 0} f(x) = 0 \neq f(0)$.

Indeed, if $\lim_{x \to a} f(x) = f(a)$, then f features a desirable property called *continuity*. We will have a more rigorous discussion of it later, but keep this important characterization in mind!

Definition 3. ***(Continuity of Real Functions)*** *Consider a function $f : X \mapsto Y$, $X, Y \subseteq \mathbb{R}$. Then,*

(i) f is called continuous at $a \in X$ if $\lim_{x \to a} f(x) = f(a)$.

(ii) f is called continuous on the interval $I \subseteq X$ if $\forall a \in I$: $\lim_{x \to a} f(x) = f(a)$.

Function $f(x)$	Limit $\lim_{x\to a} f(x)$	Function $h(x)$	Limit $\lim_{x\to a} h(x)$
c	c	$f(x)+g(x)$	$\lim_{x\to a} f(x) + \lim_{x\to a} g(x)$
$c\cdot x$	$c\cdot a$	$f(x)\cdot g(x)$	$(\lim_{x\to a} f(x))\cdot(\lim_{x\to a} g(x))$

Table 7: Rules for limits. The right column assumes that the respective limits exist.

An further concept that you may come across frequently is the one of left and right limits. The left (right) limit of f at a is the value f takes "when moving towards a from the left (right)". This is useful for two reasons: (i) we can characterize the behavior of functions like $f(x) = \mathbb{1}[x > 0]$ at points a, here $a = 0$, where the limit $x \to a$ is undefined, and (ii) the concept provides a rather straightforward method to disprove existence of the limit of f at a. Formally, we say that f_a^+ is the right limit of f at a if

$$\forall \varepsilon > 0 \exists \delta^+ > 0 : (\forall x \in X : (x - a \in (0, \delta) \Rightarrow |f(x) - f_a^+| < \varepsilon)),$$

and f_a^- is the left limit of f at a if

$$\forall \varepsilon > 0 \exists \delta^- > 0 : (\forall x \in X : (x - a \in (-\delta, 0) \Rightarrow |f(x) - f_a^-| < \varepsilon)).$$

We write $\lim_{x\to a^+} f(x) = f_a^+$ and $\lim_{x\to a^-} f(x) = f_a^-$. Then, it is easily verified (for $\varepsilon > 0$, choose $\delta = \min\{\delta^+, \delta^-\}$, or respectively $\delta^+ = \delta^- = \delta$) that the limit of f at a exists and is equal to f_a if and only if the right and left limits exist and $f_a^+ = f_a^- = f_a$. Conversely, this implies that whenever $f_a^+ \neq f_a^-$ or either limit does not exist, then f_a does not exist as well. Try to use this method to show non-existence of the limit of f at a for the specific example of $f(x) = \mathbb{1}[x > 0]$ and $a = 0$. As a final remark, if they exist, proper limits (f_a) as well as left and right limits (f_a^+, f_a^-) are unique.

To conclude this introductory chapter, let us consider some rules for limits. The simple ones can be found in Table 7. Further, if f is continuous, then $\lim_{x\to a} f(g(x)) = f(\lim_{x\to a} g(x))$. Thus, if also g is continuous, then $\lim_{x\to a} f(g(x)) = f(g(a))$. A further important rule is *L'Hôpital's rule* for the limit of ratios:

Theorem 2. ***(L'Hôpital's Rule)*** *Let f and g be two real valued differentiable functions on an open interval I and $a \in I$, or $a \in \{\pm\infty\}$. Let $g'(x) \neq 0$ for all $x \in I$, $x \neq a$. Suppose that $\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = 0$ or $\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = \pm\infty$. Then, if $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ exists, it holds that*

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

Thus, we can use derivatives and L'Hôpital's rule if the product rule does not apply because at least one limit does not exist. Note that when the functions are sufficiently differentiable, you can apply this rule multiple times (i.e., higher order derivatives). An example is given in the recap questions.

A final, important rule with a quite memorable name is the following:

Theorem 3. ***(Sandwich Theorem (Sequences))*** *Consider three real-valued sequences $\{y_n\}_{n\in\mathbb{N}}$, $\{z_n\}_{n\in\mathbb{N}}$ and $\{x_n\}_{n\in\mathbb{N}}$ such that $\{y_n\}_{n\in\mathbb{N}}$ and $\{z_n\}_{n\in\mathbb{N}}$ are convergent and $\lim_{n\to\infty} y_n = \lim_{n\to\infty} z_n = \bar{x} \in \mathbb{R}$.*

Further, suppose that there exists $N \in \mathbb{N}$ *such that* $\forall n \in \mathbb{N} : (n \geq N \Rightarrow y_n \leq x_n \leq z_n)$*. Then,* $\{x_n\}_{n\in\mathbb{N}}$ *is convergent with* $\lim_{n\to\infty} x_n = \bar{x}$*.*

This theorem is frequently used to avoid involved mathematical considerations using the ε/δ approach from the definition of the limit. Note that y_n or z_n need not necessarily depend on n, for instance, if we have $0 \leq x_n \leq z_n$ with $\lim_{n\to\infty} z_n = 0$ for all $n \geq N \in \mathbb{N}$, then we can also establish $\lim_{n\to\infty} x_n = 0$ from the sandwich theorem. Finally, the "$N \in \mathbb{N}$" part just tells us that it doesn't matter for the limit if the inequality does not hold for some "early" elements of the sequences, in most applications, you might be lucky enough to choose $N = 1$, i.e. the inequality holds for all $n \in \mathbb{N}$. As an example, consider the sequence $x_n = -\frac{1}{n^2+4n+25}$ for $n \in \mathbb{N}$. We can bound

$$-\frac{1}{n} \leq -\frac{1}{n^2+4n+25} \leq 0 \qquad \forall n \in \mathbb{N}$$

and since $\lim_{n\to\infty} -1/n = 0$, the sandwich theorem allows us to conclude that $\lim_{n\to\infty} -\frac{1}{n^2+4n+25} = 0$.

This theorem holds for limits of functions in an analogous way:

Theorem 4. ***(Sandwich Theorem (Functions))*** *Consider three real-valued functions,* g*,* h *and* f *such that for a value* x_0 *in their domain,* $\lim_{x\to x_0} g(x)$ *and* $\lim_{x\to x_0} h(x)$ *exist with* $\lim_{x\to x_0} g(x) = \lim_{x\to x_0} h(x) = f_0$*. Further, suppose that for any* x *in proximity to* x_0*, it holds that* $g(x) \leq f(x) \leq h(x)$*. Then,* $\lim_{x\to x_0} f(x)$ *exists, and* $\lim_{x\to x_0} f(x) = f_0$*.*

Here, it may not be too clear what "in proximity to x_0" means precisely, at least not formally. To express this fact more formally, we need the distance concept we are to touch upon in the next chapter. As this has not been introduced this point, the vague statement given above shall suffice for now.

0.7 Contents and Take-Aways

Chapter 0: Fundamentals of Mathematics discusses

- the formal basics of mathematical texts, that is, notation as well as statements and arguments
- mathematical implication and its relation to necessary, sufficient and equivalent conditions
- notation and central concepts of set theory
- the formal foundation of functions and key concepts and properties
- the univariate limit concept, both in the context of sequences and functions

Someone with profound knowledge of the contents of this chapter should

- be familiar with the central mathematical symbols
- know what quantifiers are and why they are useful
- be able to connect the concepts of necessity, sufficiency and equivalence to mathematical implication
- be familiar with the mathematical statement concept, including what validity and soundness refer to in this context
- know what a mathematical set is in the formal sense, and be able to handle central concepts of set theory, e.g. intervals and index sets, empty set and universal superset, etc.
- be familiar with relations between and operations on sets, e.g. subset, complement, union and intersection, set difference and power set, and be able to deal with these concepts in concrete applications
- be able to explain how the mathematical concept of relations is useful in thinking about functions formally
- be familiar with the building blocs of the function concept: domain, codomain, mapping rule and graph
- know about inverse functions and preimages, and the difference between these concepts
- know basic rules for differentiating common functions (e.g. chain rule, product rule)
- know the formal definition of the limit concept for both functions and sequences, and how continuity of a function relates to it
- be familiar with simple rules for the limit, and further with L'Hôpital's Rule and the Sandwich Theorem

and be able to answer a number of related questions, including

- What is the value of mathematics for the economist profession?
- Can a logical statement be true if it is not meaningful?
- If N is a necessary condition for S, can S be true when N is violated?
- How can quantifying statements be negated (using for all/exists)?
- What is a pairwise disjoint sequence of sets?
- Can an object be contained in a set more than once?
- What is the set difference $A \backslash B$ when $A := \{1, 2, 3, \ldots, 10\}$ and $B = \{n \in \mathbb{N} : n < 7\}$?
- What is the difference between a subset and a proper subset?
- What is the graph of a function in the formal sense? How does it relate to the cartesian product of the function's domain and codomain?
- When $f : [0, \infty) \mapsto \mathbb{R}, x \mapsto \sqrt{x} + 1$, what are the domain and codomain of f? What is the range? Can you draw the graph?
- If at $a = 2$, the right limit f_a^+ is not equal to the left limit f_a^- for some function f, can $\lim_{x \to a} f(x)$ exist?

0.8 Recap Questions

1. Define what we mean when talking about a mathematical argument.
2. What is the difference between a sound argument and a valid argument?
3. Are the following statements true?
   (a) $4 \in (3,5)$.
   (b) $5 \in (3,5)$.
   (c) $4 \in \{x \in \mathbb{R} : (\exists n \in \mathbb{N} : x = 2n+1)\}$. (How can this set be described verbally?)
   (d) $10 > \max(\{x \in \mathbb{R} : (\exists n \in \mathbb{N} : x = 2n+1)\} \cap [0,10])$
4. Negate the following statements ($A, B \subseteq X$ are sets):
   (a) $\forall x \in A : x \in B$.
   (b) $A \subseteq B$.
   (c) $\forall x \in X : (\exists y \in X : y > x)$.
   (d) $\exists x \in X : (x \in A \wedge x \in B^c)$.
5. What is $\mathcal{P}(\mathcal{P}(A))$ when $A = \{1, \pi\}$? What is the universal superset for $\mathcal{P}(\mathcal{P}(A))$ when $\mathbb{R}$ is the universal superset of A?
6. Let $f : \mathbb{R} \mapsto \mathbb{R}_+, x \mapsto \exp(|x|)$, where $|\cdot|$ is the absolute value, i.e. $|x| = \begin{cases} x & x \geq 0, \\ -x & x < 0. \end{cases}$
   (a) Determine $f(\ln(10))$ (hint: $\ln(10) > 0$; indeed, $\ln(x) > 0 \ \forall x > 1$).
   (b) Determine the range of f, i.e. the image of X under f. (You may want to have read the section on continuity and limits before solving this one.)
   (c) Determine $f^{-1}[\{2\}]$.
   (d) Does $f^{-1}(2)$ exist?
7. Take the derivatives of the following functions.
   (a) $f(x) = 3\exp(x)$.
   (b) $f(x) = \sin(x)/x^4$.
   (c) $f(x) = \ln(3^x)$.
   (d) $f(x) = \sin(\cos(x^2))$.
8. Determine the following limits. You may want to use L'Hôpital's rule for some subquestions.
   (a) $\lim_{x \to 0} \frac{\sin(x)}{x}$.
   (b) $\lim_{x \to \pi} \frac{\cos(x)}{1+x}$.
   (c) $\lim_{x \to 0} \frac{\exp(x)-1-x}{x^2}$.

1 Introduction to Vector Spaces

Much of what is usually done in undergraduate economics is restricted to the plane $\mathbb{R}^2$, which has the advantage of being easily displayed graphically. For example, you will likely remember how to look for the utility-maximizing consumption bundle of two goods when given the budget restriction and the indifference curve. However, when considering more complex problems (e.g. more than two goods/inputs or uncertainty through stochastic components), a more general concept is needed. The main objective of the theory of vector spaces is sometimes described as follows: *Geometrical insights at hand with 2- or 3-dimensional real vectors are really helpful. Can we, in some way, generalize these insights to other mathematical objects, for which a geometric picture is not available?*

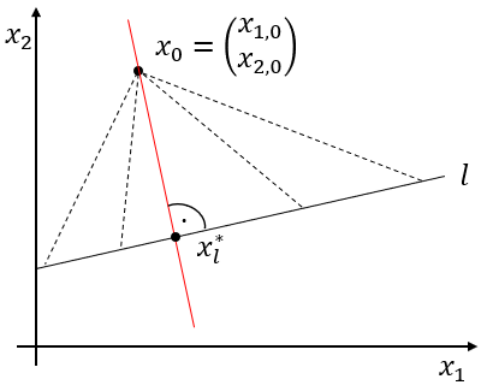


Figure 3: Distance minimization in the $\mathbb{R}^2$.

In the $\mathbb{R}^2$, geometrical representations give us simple and fundamental principles that help us solving various problems. For instance, it is easily seen that when minimizing the (Euclidean) distance of a point $x_0 = (x_{1,0}, x_{2,0})$ to a line l defined by $x_2 = cx_1 + d$, i.e. when choosing a point x^* on l so that the distance $d(x^*, x_0)$ is smaller than $d(x_l, x_0)$ for any point x_l on l, then the line through x^* and x_0 is orthogonal to l. Without this intuitive geometrical representation, however, the result would probably have been very hard to find out. Indeed, the result's beauty lies within the insight that even when considering higher-dimensional spaces (e.g. of vectors $x = (x_1, x_2, \ldots, x_n)$, $n \in \mathbb{N}$, that is, the $\mathbb{R}^n$), the *least-squares* solution that minimizes the Euclidean distance $\|x_0 - x_l\|_2 = \left(\sum_{i=1}^{n} (x_{0,i} - x_{l,i})^2\right)^{1/2}$ continues to satisfy this orthogonality property![1]

First things first - to begin, let us define what we mean precisely by a vector.

Definition 4. ***(Vector)*** *A row vector x of length $n \in \mathbb{N}$ is an ordered tuple of elements x_i. We write $x = (x_1, x_2, \ldots, x_n)$. A column vector x stacks the elements in a column, i.e.* $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = (x_1, x_2, \ldots, x_n)'$

where $(\cdot)'$ indicates vector transposition. Hence, a row vector x is such that x' is a column vector. A "vector" typically refers to a column vector.

In distinction to the set, the order of elements in a vector matters, such that $x = (1, 2)$ and $y = (2, 1)$ are distinct! Also, the vector can contain an element multiple times, consider e.g. the origin of the $\mathbb{R}^2$, $x^o = (0, 0)'$. As with sets, however, the definition does not restrict elements to be real numbers. **Thus, be aware that even though we predominantly deal with vectors of real numbers, the concept is much broader** and may also refer to collections of functions, matrices, sets, vectors, etc.

[1] We later discuss how to define orthogonality formally, for the $\mathbb{R}^2$ as well as for the $\mathbb{R}^n$ with arbitrary $n \in \mathbb{N}$.

While this is good news because it means that once you understand how to deal with vectors of real numbers, you can apply the same concepts to much more general problems and mathematical investigations, due to the broad variety of object that constitute a "vector" according to Definition 4, we require a unified notion of how to handle vectors generally, and how sets of vectors must be structured so that we can make use of them – enter vector *spaces*. Due to the focus of economic applications on vectors of real numbers (or "real vectors", vectors in $\mathbb{R}^n$), while introducing the concept of vector spaces generally, the narrative also extensively discusses the specific context of real vectors.

1.1 The Algebraic Structure of Vector Spaces

1.1.1 Definitions

In the following, consider a set of vectors X, for instance

$$\mathbb{R}^n := \{(x_1, \dots, x_n)' : (\forall i \in \{1, \dots, n\} : x_i \in \mathbb{R})\},\ n \in \mathbb{N}.$$

Note that as a convention, we view the $\mathbb{R}^n$ as the collection of *column* vectors of length n. Without any further structure, this is just a set, and there is little we can do with it – indeed, we are yet a long way from computing the distance of two elements of $\mathbb{R}^n$, let alone discuss orthogonality, as the simple example above has done.

As you may already have done in school, it is useful to think of real vectors as an entity with *direction* and *magnitude*. To do so, one writes a vector $x = (x_1, x_2)' \in \mathbb{R}^2$ as the product of a direction vector (e.g. of unit length)[2] and an augmenting magnitude coefficient: e.g. one writes $x = (0,4)'$ as $x = 4 \cdot (0,1)'$, and $y = (2,-4)'$ as $y = 6 \cdot (1/3, -2/3)'$. Then, x and y have magnitude coefficients 4 and 6, and directionality $(0,1)'$ and $(1/3,-2/3)'$. Indeed, this concept is the first fundamental building block of the structure we assign to sets X of vectors to do algebra with them: *scalar multiplication*.[3] As you likely know already, for $X = \mathbb{R}^n$, multiplication of $x \in X$ with a scalar $\lambda \in \mathbb{R}$ is defined as $\lambda \cdot x = (\lambda \cdot x_1, \dots, \lambda \cdot x_n)'$. The second is the one of vector addition, which for real vectors $x, y \in \mathbb{R}^n$ is defined as $x + y = (x_1 + y_1, \dots, x_n + y_n)'$. With these two concepts, we just define a *vector space of real numbers* as the collection $(\mathbb{R}^n, +, \cdot)$, i.e. the set of real vectors of length n, endowed with the operations vector addition "+" and scalar multiplication "$\cdot$". Writing the space as this collection just ensures that it is clear what we mean when talking about addition and scalar multiplication.

For our example of the $\mathbb{R}^2$, note that there are two fundamental directions: the horizontal and vertical axes. In the notion of the above, these directions are expressed by the vectors $e_1 = (1,0)'$ and $e_2 = (0,1)'$. Then, the directionality vector v_x of any $x \in \mathbb{R}^2$ is a linear combination $v_x = \mu_{x,1} e_1 + \mu_{x,2} e_2$, $\mu_{x,1}, \mu_{x,2} \in [-1,1]$ and $|\mu_{x,1}| + |\mu_{x,2}| = 1$,[4] and we can use the scalar product of v_x with the magnitude coefficient, denote it by $\lambda_x \geq 0$, to write x as $x = \lambda_x \cdot v_x$. For the $\mathbb{R}^n$, the idea is identical, we simply need more basis vectors!

[2] The concept of "length" is discussed more precisely in the next section. In the notion introduced here, the elements of v_x need to sum to one in *absolute* value, so that e.g. $(1/3, -2/3)$ has unit length, too!

[3] A scalar is defined as an element of the *field* underlying the vector space. Focused on the context most relevant to economics, you can think of scalars simply as real numbers in the following.

[4] For our examples $x = (0,4)'$ and $y = (2,-4)'$, we have $\mu_{x,1} = 0$ and $\mu_{x,2} = 1$, and $\mu_{y,1} = 1/3$ and $\mu_{y,2} = -2/3$.

The beauty of the representation discussed above lies in the fact that we have now tamed the huge set $\mathbb{R}^n$ to tractable expressions that characterize its elements using only a finite combination of basis vectors and the real line $\mathbb{R}$. This reduction is especially powerful when considering vectors of matrices, functions, etc. Consequently, to represent and manipulate general vectors, we endow them with similar operations, i.e. the vector addition and multiplication with a scalar. However, without restricting the class of elements in vectors, we can not ensure that addition and scalar multiplication work in the same way as with real vectors. Thus, we define these operations by *axioms* that leave the precise nature of the operation unspecified, but require it to satisfy a number of conditions that ensure transferability of the structure and properties of vector spaces of real numbers.

Before proceeding, it shall be stressed that everything that follows in this section 1.1 is just a generalization and formalization of the concepts discussed in the above three paragraphs. Thus, make sure you thoroughly understand this intuition, then, despite admittedly being formally quite intensive, the following concepts should be well comprehensible to you. That being said, here is the general definition of a real vector space:

Definition 5. ***(Real Vector Space)*** *Let X be a set of vectors and $\mathbb{X} := (X, +, \cdot)$ be the collection of this set together with two operations, called vector addition and scalar multiplication, which associates to any scalar $\lambda \in \mathbb{R}$ and any $x \in X$ the vector $\lambda \cdot x$. Then, $\mathbb{X}$ is called a* ***vector space*** *if the following properties hold:*

(i) X *is* ***closed*** *with respect to the operations:* $\forall x, y \in X : x + y \in X$, *and* $\forall x \in X \forall \lambda \in \mathbb{R} : \lambda \cdot x \in X$.

(ii) Vector addition is ***commutative****:* $\forall x, y \in X : x + y = y + x$

(iii) Vector addition is ***associative****:* $\forall x, y, z \in X : x + (y + z) = (x + y) + z$.

(iv) There exists a neutral element $\mathbf{0} \in X$ *("additive identity") such that* $\forall x \in X : x + \mathbf{0} = x$.

(v) Scalar multiplication is ***associative****:* $\forall \lambda, \mu \in \mathbb{R}\ x \in X : \lambda \cdot (\mu \cdot x) = (\lambda\mu) \cdot x$

(vi) Scalar multiplication is ***distributive*** *over vector and scalar addition:*

$$\forall \lambda \in \mathbb{R} \forall x, y \in X : \lambda(x + y) = \lambda x + \lambda y$$

$$\forall \lambda, \mu \in \mathbb{R} \forall x \in X : (\lambda + \mu)x = \lambda x + \mu x$$

(vii) If 1 denotes the scalar multiplicative identity and 0 the scalar zero, then:

$$\forall x \in X : (1 \cdot x = x \wedge 0 \cdot x = \mathbf{0}).$$

A few comments on this more general definitions of vector spaces deem worthwhile. First, the definition defines a *real* vector space not because X necessarily contains vectors of real numbers, but rather because the scalars used here necessarily lie in $\mathbb{R}$. Then, $(\mathbb{R}^n, +, \cdot)$ with vector addition and scalar multiplication as previously discussed is clearly a real vector space, because the properties of these precisely operations were used to construct the axioms in Definition 5. Moreover, closedness as imposed by *(i)* is needed to ensure that the operations are

well-defined, i.e. that the expressions $x+y$ or $\lambda \cdot x$ are meaningful – as a counterexample, think of the set $X = \{(r,e,d,0,0)', (g,r,e,e,n)', (b,l,u,e,0)'\} =: \{c_1, c_2, c_3\}$ of vectors of letters that indicate primary colors, where it is unclear what e.g. $c_1 + c_2$ is supposed to be. Next, note that in *(v)*, while $\mu \cdot x$ and $\lambda \cdot (\mu \cdot x)$ refers to scalar multiplication, $\lambda\mu$ does **not**, because neither λ nor μ are vectors. Similarly, $\lambda + \mu$ refers to standard addition of real numbers in *(vi)*. So pay attention to the context in which addition and multiplication operators are used, and which concept they refer to! For *(vii)*, since we only consider real numbers as scalars in this class, 1 and 0 have their normal interpretation, there is nothing fancy here and this is just a more general wording.

Having defined a real vector space, we have moved from a set X of vectors that we could not say much about to a space in which all objects have a clearly defined *position*, plus are subject to some very basic algebraic operations ensuring that you can take *linear combinations* of elements. In simpler words, we now not only know the elements in the set X, but also how we can **add them to each other** and **multiply them with a constant**. Indeed, this is *all* that has happened so far! We still have to address (and will do so in the next section) what we precisely mean by a "distance", and even the concept of multiplying vectors with each other has not yet been discussed. Still, defining real vector spaces as we have done allows to more fundamentally understand the properties characteristic of real vectors as we are familiar with, and to comprehend that (and how) this characteristic structure can be extended to a broader class of vectors. Regarding the latter point, it is a good exercise to verify that the following sets, endowed with proper operations, can also be considered as vector spaces:

- $X = \{x = (x_1, \dots, x_5)' \in \mathbb{R}^5 : x_3 = 0\}$,
- $F_{[a,b]} = \{f : [a,b] \mapsto \mathbb{R}\}$ with $-\infty < a < b < \infty$,
- $S = \{x = \{x_k\}_{k=1}^{\infty}\}$ the set of infinite sequences of real numbers,
- $\mathcal{M}_{\mathbf{m}\times\mathbf{n}}$, the set of $m \times n$ matrices where $m, n \in \mathbb{N}$,
- ... And many others!

The idea here is to come up with operations "$+$" and "$\cdot$" for addition and scalar multiplication (don't be scared, this is quite intuitive!) and verify the properties *(i)* through *(vii)*.

You may wonder why we included exactly these properties in the definition and not, say, the unique existence of the inverse element, i.e. that $\forall x \in X \exists! (-x) \in X : x + (-x) = \mathbf{0}$. Here, this additional statement is not needed in the definition, because it follows from the axioms as property of vector spaces: by *(i)*, $\forall x \in X : (-1)\cdot x \in X$, and by *(vi)*, $x + (-1)\cdot x = (1 + (-1))\cdot x = 0 \cdot x = \mathbf{0}$, where the last equality uses *(vii)*.[5] Indeed, the axioms included in the definition are such that they imply the important characterizations of vector spaces of real numbers.

Elementary but important properties of vector spaces that facilitate handling *arbitrary vectors* in a fashion similar to real vectors or even real numbers are the cancellation laws:

Theorem 5. ***(Cancellation Laws)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, $x, y, z \in X$, and $\lambda, \mu \in \mathbb{R}$. Then,*

[5] Uniqueness is verified as $x + y = 0 \Leftrightarrow x + y + (-1)\cdot x = (-1)\cdot x \Leftrightarrow y = (-1)\cdot x$.

(i) *If $x + y = x + z$, then $y = z$.*

(ii) *If $\lambda x = \lambda y$ and $\lambda \neq 0$, then $x = y$.*

(iii) *If $\lambda x = \mu x$ and $x \neq \mathbf{0}$, then $\lambda = \mu$.*

Another very important concept – that we have already seen in Chapter 0 – is the Cartesian product:

Definition 6. ***(Cartesian Product)*** *Let $\mathbb{X} := (X, +_X, \cdot_X)$ and $\mathbb{Y} := (Y, +_Y, \cdot_Y)$ be two real vector spaces. Then, the Cartesian product of $\mathbb{X}$ and $\mathbb{Y}$, denoted $\mathbb{X} \times \mathbb{Y}$, is the collection of ordered pairs (x, y) with elements $x \in X$ and $y \in Y$ together with addition and scalar multiplication, respectively defined as $(x_1, y_1) + (x_2, y_2) = (x_1 +_X x_2, y_1 +_Y y_2)$ and $\lambda \cdot (x, y) = (\lambda \cdot_X x, \lambda \cdot_Y y)$.*

Indeed, the Cartesian product of two real vector spaces is itself a real vector space. It may be a good exercise to verify this! As a very simple example, note that we can write $\mathbb{R}^5 = \mathbb{R}^3 \times \mathbb{R}^2$.

To conclude the section on definitions, let us define a very special vector operation that is extensively used in all economic disciplines: the *scalar product* (alternative names are dot product, inner product or vector product):

Definition 7. ***(Scalar product)*** *Let $x = (x_1, ..., x_n)', y = (y_1, ..., y_n)' \in \mathbb{R}^n$. Then, the scalar product $\cdot$ is defined as*

$$x \cdot y = \sum_{i=1}^{n} (x_i \cdot y_i) = x_1 \cdot y_1 + ... + x_n \cdot y_n.$$

Note that the scalar product as stated here is defined only for the $\mathbb{R}^n$[6] and not more general vector spaces where vectors potentially contain elements other than real numbers. This is because multiplying elements within vector spaces is too context-specific to be generalized into a broad concept, as we have done with addition and scalar multiplication. An alternative notation for the scalar product of x and y that you may see more frequently in math textbooks is $\langle x, y \rangle$. As you will see in the matrix chapter, a further convenient notation is $x'y$. It may have the advantage to reduce confusion with what is meant with the symbol "$\cdot$" that already refers to scalar multiplication and the product of real numbers, but in economics, we nevertheless mostly use "$\cdot$" for the scalar product as well. On the other hand, it is still clear what we mean by "$\cdot$" given the objects that are to be multiplied with each other, and sticking to the symbols "$+$" and "$\cdot$" highlights that we're still doing nothing more than adding and multiplying, these operations are just no longer confined to standard real numbers! Note that the scalar product is commutative, distributive over vector addition and associative w.r.t. scalar multiplication, but not generally associative (why not?). Again, be sure to not confuse the scalar product with scalar multiplication, those are fundamentally different things!

The Scalar product is also used to define *orthogonality*: In the $\mathbb{R}^2$, a line l with slope c and intercept d in $\mathbb{R}^2$ may be written as the set $l = \left\{ \begin{pmatrix} x \\ cx + d \end{pmatrix} = \begin{pmatrix} 1 \\ c \end{pmatrix} x + \begin{pmatrix} 0 \\ d \end{pmatrix} : x \in \mathbb{R} \right\}$. Two lines l_1 and l_2 with slopes c_1 and c_2 are orthogonal if $(1, c_1)' \cdot (1, c_2)' = 0$. More generally, we say that $x, y \in \mathbb{R}^n$ are orthogonal if $x \cdot y = 0$.

[6] More precisely, it is defined for the space $(R^n, +, \cdot)$ with the common operations "$+$" and "$\cdot$" – if not explicitly stated otherwise, when just speaking of "the $\mathbb{R}^n$", we always assume that it is endowed with these two operations.

Before moving on, a final comment on adding and multiplying: you may wonder why we have not yet talked about subtraction and division at all. The simple reason is that in closed vector spaces, subtraction is the same as addition of the additive inverse, i.e. for $x, y \in X$, $x - y = x + (-1) \cdot y$, where $(-1) \cdot y \in X$, and division by the scalar $\lambda \in \mathbb{R}$ is equivalent to scalar multiplication with $1/\lambda$. Thus, it is common to reduce attention to addition and multiplication in more advanced studies of mathematics.

1.1.2 Subspaces

Considering sub*sets* of a universal superset often proves useful in mathematics. For instance, we may like to consider only the integers and not the whole real line. It is possible to generalize the concept of a subset to the context of *spaces* (i.e. algebraically structured sets). If we are to do so, however, we do not want to loose the very structure we looked for when moving from the notion of a set to that of a space.[7] Like with the universal superset discussed in Chapter 0, we will now start from an "all-encompassing" space $\mathbb{X} := (X, +, \cdot)$ and consider whether we can generalize subsets $Y \subseteq X$ to vector spaces. We will always focus on subspaces $\mathbb{Y} := (Y, +, \cdot)$ that use the same operations "$+$" and "$\cdot$" as $\mathbb{X}$. Luckily, when starting from such a space $\mathbb{X}$, we only need to check a single condition to verify that the vector space structure is maintained in for the subset Y when endowed with the operations of $\mathbb{X}$.

Definition 8. ***(Closure of Subsets under Operations)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space. We say that $Y \subseteq X$ is closed under vector addition if $\forall y_1, y_2 \in Y : y_1 + y_2 \in Y$. Similarly, Y is closed under scalar multiplication if $\forall y \in Y \forall \lambda \in \mathbb{R} : \lambda \cdot y_1 \in Y$.*

Indeed, this concept is not new, rather, we have already used it as an axiom in defining the real vector space. However, an important subtlety is that here, it is clear how "$+$" and "$\cdot$" are defined from the definition of $\mathbb{X}$, and thus, when saying that Y is closed under scalar multiplication, we implicitly refer to the respective operation $\mathbb{X}$ is endowed with!

Definition 9. ***((Real) Vector Subspace)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space and Y a non empty subset of X, i.e. $\emptyset \neq Y \subseteq X$. We say that $\mathbb{Y} := (Y, +, \cdot)$ is a subspace of $\mathbb{X}$ if Y is closed under vector addition and scalar multiplication.*

As discussed above, note that $\mathbb{X}$ and $\mathbb{Y}$ necessarily use the same operations "$+$" and "$\cdot$". Instead of imposing closure under both operations separately, one may also refer to a more elegant and perhaps intuitive characterization:

Definition 10. ***(Linear Combination)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, $x, y \in X$ and $\lambda_x, \lambda_y \in \mathbb{R}$. Then, the linear combination z of x and y with coefficients λ_x and λ_y is $z = \lambda_x \cdot x + \lambda_y \cdot y$. More generally, for $k \in \mathbb{N}$, the linear combination of $x_1, \ldots, x_k \in X$ with coefficients $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ is $\sum_{j=1}^{k} \lambda_j \cdot x_j$.*
We say that $\mathbb{X}$ is closed under linear combination if $\forall x, y \in X \forall \lambda_x, \lambda_y \in \mathbb{R} : \lambda_x \cdot x + \lambda_y \cdot y \in X$.

Note that closure under linear combination implies that any linear combination of the more general form $\sum_{j=1}^{k} \lambda_j \cdot x_j$ is an element of X.[8] It is straightforward to verify (try it! – or see

[7] Remember, the structure will guarantee the valid extension of our geometrical insights!

[8] To see this, note that we can just apply the linear combination with two vectors iteratively $k-1$ times until we arrive at one with k components for any $k \in \mathbb{N}$.

footnote[9]) that closure under linear combination is equivalent to closure under both vector addition and scalar multiplication. Thus, an equivalent definition of a vector subspace is

Definition 11. ***((Real) Vector Subspace – linear combination definition)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space and Y a non empty subset of X, i.e. $\emptyset \neq Y \subseteq X$. We say that $\mathbb{Y} := (Y, +, \cdot)$ is a subspace of $\mathbb{X}$ if Y is closed under linear combination.*

This may be more intuitive to remember: given a real vector space $\mathbb{X} := (X, +, \cdot)$, any subset Y gives rise to a subspace given the operations of $\mathbb{X}$, $\mathbb{Y} = (Y, +, \cdot)$, so long as for any two elements $y_1, y_2 \in Y$, any of their linear combinations lies in Y![10]

As a further remark, any subspace of a vector space is itself a vector space, this follows directly from Definition 5 as "$+$" and "$\cdot$" satisfy *(ii)* through *(vii)*. Further, note that the entire space $\mathbb{X}$ is a subspace of $\mathbb{X}$ as $\mathbb{X}$ is by definition a subset of itself and is closed under scalar multiplication and addition. A subspace not equal to the entire space is called a *proper subspace*.

An example of a proper subspace is the following:

Proposition 1. *The space of convergent real sequences L, i.e.*

$$L := \{x = \{x_n\}_{n\in\mathbb{N}} : x_n \in \mathbb{R} \forall n \in \mathbb{N} \wedge (\exists x_0 \in \mathbb{R} : \lim_{n\to\infty} x_n = x)\},$$

constitutes a proper subspace of $(S, +, \cdot)$, the space of real sequences S endowed with addition $x + y = \{x_n + y_n\}_{n\in\mathbb{N}}$ and scalar multiplication $\lambda \cdot x = \{\lambda \cdot x_n\}_{n\in\mathbb{N}}$ for $x, y \in S, \lambda \in \mathbb{R}$.

Proof. For simplicity, we will take for granted that $(S, +, \cdot)$ constitutes a real vector space; then, it suffices to establish the subspace property. Let $a_n, b_n \in L$, $\lambda, \mu \in \mathbb{R}$. Then, $\exists a, b \in \mathbb{R} : (a = \lim_{n\to\infty} a_n \wedge b = \lim_{n\to\infty} b_n)$. From the laws on limits of sequences, it follows that $\lim_{n\to\infty} \lambda a_n = \lambda a \in \mathbb{R}$ and $\lim_{n\to\infty} \mu b_n = \mu b \in \mathbb{R}$, and lastly $\lim_{n\to\infty}(\lambda a_n + \mu b_n) = \lambda a + \mu b \in \mathbb{R}$. Hence, any linear combination $\lambda x + \mu y$ of convergent real sequences x and y is also a convergent real sequence and hence an element of L: $\lambda x + \mu y \in L$. Lastly, the existence of divergent sequences establishes $L \neq S$, so that $(L, +, \cdot)$ is a proper subspace of $(S, +, \cdot)$. □

The next result is just given for completeness. It is not essential for our purposes, but given nonetheless in case you may ever wonder whether it holds in future coursework.

Theorem 6. ***(Intersection and Addition of Subspaces)*** *Let $M, N \subseteq X$ be such that they give rise to subspaces of a real vector space $\mathbb{X} = (X, +, \cdot)$. Then,*

(i) their intersection, $M \cap N$, gives rise to a subspace of $\mathbb{X}$.

(ii) their sum, $M + N := \{m + n : m \in M, n \in N\}$, gives rise to a subspace of $\mathbb{X}$.

[9] "$\Rightarrow$" Suppose that $\mathbb{X}$ is closed under vector addition and scalar multiplication. Let $x, y \in \mathbb{X}$, $\lambda_x, \lambda_y \in \mathbb{R}$. By closure under scalar multiplication, $\lambda_x \cdot x \in X$ and $\lambda_y \cdot y \in X$. By closure under vector addition, $\lambda_x \cdot x + \lambda_y \cdot y \in X$. Thus, $\mathbb{X}$ is closed under linear combination.
"$\Leftarrow$" Suppose that $\mathbb{X}$ is closed under linear combination. Let $x, y \in \mathbb{X}$. With $\lambda_x = \lambda_y = 1$, it follows that $x + y \in X$. Thus, $\mathbb{X}$ is closed under vector addition. Let $\lambda_x \in \mathbb{R}$. With $\lambda_y = 0$, $\lambda_x \cdot x \in \mathbb{R}$. Thus, $\mathbb{X}$ is closed under scalar multiplication. □

[10] This requirement can be quite restrictive, however, and e.g. applies to none of $\mathbb{R}^n$'s subsets $\mathbb{N}^n$, $\mathbb{Z}^n$ and even $\mathbb{Q}^n$: when $n = 1$, 1 and 0 are elements of the subsets, but with $\lambda_x = \pi$ and an arbitrary λ_y, $\pi \cdot 1 + \lambda_y \cdot 0 = \pi$ which is neither contained in $\mathbb{N}, \mathbb{Z}$ or $\mathbb{Q}$. To work with such spaces, one would define vector spaces from different scalars.

Note that nothing is said about the *union* $M \cup N$! (For a counterexample, one may think about $m \in M \backslash N$ and $n \in N \backslash M$, where nothing ensures that the linear combination lies in either M or N, even though both give rise to subspaces.)

As a take-away, until now, we know vector spaces as abstract entities that preserve our neat graphical intuitions from the $\mathbb{R}^2$ by generalizing addition and scalar multiplication in a suitable way, and have shown that closure of subsets under linear combination implies that they give rise to subspaces. In what follows, we pursue a converse approach: starting from a real vector space $\mathbb{X} = (X, +, \cdot)$ and an arbitrary subset $Y \subseteq X$, can we define a (sub)space on basis of Y via linear combinations? The motivation is transferring the intuition of magnitude and direction, which as you may recall, allowed us to arrive at simple representations using only a finite basis set – the subset $Y \subseteq X$ – and the real line $\mathbb{R}$!

1.1.3 Span, Bases and Linear Dependence

A key concept for the following is the *span operator*:

Definition 12. ***(Span Operator)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, and let $Y \subseteq X$. Then, the we define*

$$Span(Y) = \{z \in X : (\exists k \in \mathbb{N} : \exists y_1, \dots, y_k \in Y, \lambda_1, \dots, \lambda_k \in \mathbb{R} : z = \sum_{j=1}^{k} \lambda_j \cdot y_j,)\}$$

where "$+$" and "$\cdot$" are the operations of $\mathbb{X}$.

In words, $Span(Y)$ is the *set of linear combinations* that we can take using the elements in Y.

Theorem 7. ***(Span as Generated Subspace)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, and let $Y \subseteq X$. Then, $Span(Y)$ endowed with the operations of $\mathbb{X}$ (so $(Span(Y), +, \cdot)$) is a subspace of $\mathbb{X}$. It is called the subspace generated by Y or the **span** of Y and is the smallest subspace which contains Y.*

Note the following important distinction: In Definition 12, we defined $Span(Y)$ as a *set*, whereas in Theorem 7, the word "span" refers to a *vector space* (based on $Span(Y)$). Recall that early in Subsection 1.1.2, we noted that the "universal" space $\mathbb{X} = (X, +, \cdot)$ is always in the background of subspace considerations and that thus, it is implicitly clear which operations "$+$" and "$\cdot$" we refer to when looking at subspace candidates. Therefore, in the subspace context, there exists a one-to-one relationship between subsets closed under linear combination and subspaces, so that the concepts are almost equivalent! Nonetheless, you should take away that (i) "the span" is a vector space, but that (ii) the formal expression "$Span(Y)$" refers to a set, and generally (iii) what distinguishes a sub*space* from a sub*set* closed under some operations.

To prove Theorem 7, one must establish two parts, (i) that $Span(Y)$, with the operations of $\mathbb{X}$, gives rise to a subspace, and (ii) that if $Z \subset Span(Y)$, i.e. Z is a proper subset of $Span(Y)$, then Z does not give rise to a subspace. However, both parts more or less immediately follow from Definition 11, the linear combination definition of a vector space, and construction of the $Span(\cdot)$-operator in Definition 12.[11] Should you feel the need to improve your proving skills

[11] For (i), a linear combination of linear combinations is again a linear combination. For (ii), if a linear combination of elements of Y is excluded from $Span(Y)$ in a set Z, since all elements of Y are elements of $Span(Y)$, there is a linear combination of elements of Z that are no longer contained in Z.

you may try to verify this, however, the span concept given in Theorem 7 is far more relevant to our purposes than the proof.

You may wonder why we need this concept. Recall that earlier, we discussed $e_1 = (1,0)'$ and $e_2 = (0,1)'$ as "basis vectors" for the directionality of any vector in $\mathbb{R}^2$. Indeed, when $x \in \mathbb{R}^2$ has directionality $v_x = \mu_x \cdot e_1 + (1-\mu_x) \cdot e_2$, $\mu_x \in [0,1]$ and magnitude $\lambda_x \in \mathbb{R}$, then

$$x = \lambda_x v_x = \underbrace{\lambda_x \mu_x}_{=:\alpha_1} \cdot e_1 + \underbrace{\lambda_x (1-\mu_x)}_{=:\alpha_2} \cdot e_2 = \alpha_1 e_1 + \alpha_2 e_2.$$

Thus, we can write any $x \in \mathbb{R}^2$ as a linear combination of these two basis vectors, which gives $\mathbb{R}^2 = Span(\{e_1, e_2\})$, and the space $(\mathbb{R}^2, +, \cdot)$ is nothing but the span of $\{e_1, e_2\}$! Hence, we say that e_1 and e_2 *span* the space of $\mathbb{R}^2$. However, note that we also have $Span(\{(2,0)', (0,2)'\}) = Span(\{(1,0)', (2,8)', (0,1/4)'\}) = Span(\mathbb{R}^2) = \mathbb{R}^2$. In the following, we aim to find the *smallest* set(s) which spans a vector space, which we will call the *basis* of the space, a concept that we already used informally. Recall that the value of basis vectors lies in capturing the independent directions into which elements in the space can extend and that they greatly facilitate representing complex spaces by offering a comparably small set of elementary objects that any arbitrarily complex object in the space is just a mere (linear) combination of.[12]

To find a "smallest" set, i.e. a set with as few elements as possible, that spans a space $\mathbb{X}$, the key concept needed is linear independence.

Definition 13. ***(Linear Dependence, Linear Independence)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, and let $S \subseteq X$, $x \in X$. x is said to be linearly dependent upon the set S if it can be expressed as a linear combination of its elements, i.e. $\exists k \in \mathbb{N} \exists s_1, \dots, s_k \in S, \lambda_1, \dots, \lambda_k \in \mathbb{R} : x = \sum_{j=1}^{k} \lambda_j \cdot s_j$. Equivalently, x is linearly dependent upon S if and only if $x \in Span(S)$.[13] Otherwise, the vector x is said to be linearly independent of S. Finally, a set $B \subseteq X$ is said to be linearly independent if each vector in the set is linearly independent of the remainder of the set, i.e. if $\forall b \in B : (b$ is lin. indep. of $B \backslash \{b\})$.*

In words, x is linearly dependent of S if the elements in S can be (linearly) combined to obtain x. Then, x does not add a new, *independent* direction, which is why we call it dependent of (the directions in) S. In terms of the basis, such vectors are redundant: If we can already account for the directions of x in a set S that is supposed to be a basis of the space $\mathbb{X}$, x does not add any value to the elements already contained in S. Accordingly, a basis set B should not contain any linearly dependent, or respectively, only linearly independent vectors! It is thus crucial how we may establish linear independence of a basis candidate set B.

Theorem 8. ***(Testing Linear Independence)*** *An equivalent condition for linear independence of the set of vectors $B = \{b_1, b_2, ..., b_k\}$ is that*

$$\sum_{j=1}^{k} \lambda_j b_j = \mathbf{0} \Rightarrow (\forall j \in \{1, ..., k\} : \lambda_j = 0). \tag{2}$$

[12] While this may seem a bit artificial for the $\mathbb{R}^2$ or the $\mathbb{R}^n$ which are rather comprehensible spaces to begin with, when considering more complex or abstract spaces, the basis concept greatly aids clarification and simplification, and beyond ensures transfer of the graphical intuition of direction and magnitude!

[13] This is due to the definition of $Span(\cdot)$.

This result is really important and in fact a key take-away of this chapter! Thus, feel encouraged to go over the proof given below, as doing so may give you a more thorough understanding of this central result.

Proof of Theorem 8. "$\Rightarrow$" (necessary/"implied by" part). Suppose that $\sum_{j=1}^{k} \lambda_j b_j = \mathbf{0}$ and $\exists l \in \{1,\dots,k\} : \lambda_l \neq 0$. Then, $b_l = (-1/\lambda_l) \cdot \sum_{j=1, j\neq l}^{k} \lambda_j b_j$, and b_l is linearly dependent of $B\backslash\{b_l\}$. Thus, B can not be linearly independent unless equation (2) holds, i.e. (2) $\Leftarrow$ (B is lin. indep.).[14] "$\Leftarrow$" (sufficient/"implies" part). Again using the same method as above ("contrapositive"): Suppose that B is a linearly dependent set. Then, there exists $l \in \{1,\dots,k\}$ so that $b_l = \sum_{j=1, j\neq l}^{k} \lambda_j b_j$ or respectively, $\sum_{j=1, j\neq l}^{k} \lambda_j b_j + (-1) b_l = 0$. Because the l-th coefficient of this linear combination is non-zero, equation (2) does not hold. Thus, if equation (2) holds, B must be a linearly independent set. □

Finally, we are ready to express our intuition of a "basis" formally.

Definition 14. ***(Basis and Space Dimension)*** *A finite set B of linearly independent vectors is said to be a basis for the space $\mathbb{X} = (X, +, \cdot)$ if B generates $\mathbb{X}$ i.e. if $X = Span(B)$. The cardinality of a vector space $\mathbb{X}$'s basis, i.e. the number of its elements, is called the dimension of $\mathbb{X}$, denoted* $\dim(\mathbb{X})$. *If $\dim(\mathbb{X}) < \infty$, $\mathbb{X}$ is said to be finite dimensional. Otherwise, it is said to be infinite dimensional.*

An important technical detail is that $\mathbf{0}$ can not be an element of a basis B, because $\mathbf{0}$ is linearly dependent of *any* set of vectors $\{x_1,\dots,x_m\}$: by simply setting $\lambda_1 = \dots = \lambda_m = 0$, we have $\mathbf{0} = \sum_{k=1^m} \lambda_k x_k$. Note that because there is not only one basis (indeed, typically there are infinitely many, you can e.g. check that $\{(1,0)', (0,1)'\}$, $\{(5,0)', (0,3)'\}$ or $\{(1,0)', (1,1)'\}$ are all bases of the $\mathbb{R}^2$) it is not ex-ante guaranteed that the dimension of the space is uniquely defined. However, this is indeed the case, as can be shown:

Theorem 9. ***(Uniqueness of the Dimension)*** *Any two bases for a finite dimensional vector space $\mathbb{X} = (X, +, \cdot)$ contain the same number of elements.*

The result is quite intuitive: the idea is that if you can span the whole space using n elements, then $m > n$ elements will necessarily include some redundant ones one may leave out. However, the proof is quite tricky and beyond the scope of this course, the interested reader may consult https://en.wikipedia.org/wiki/Dimension_theorem_for_vector_spaces.

For the $\mathbb{R}^n$, the dimension is more or less obviously equal to n (recall e.g. $\{e_1, e_2\}$ as the basis of $\mathbb{R}^2$), however, with the general definition in Definition 14, we may consistently extrapolate the concept to general and more complex spaces where the dimension is less obvious. Finally, in the context of the $\mathbb{R}^n$, we define as the *canonical* basis the basis containing the fundamental directions: $B^{canon} = \{e_1,\dots,e_n\}$ where $e_{ji} = \begin{cases} 1 & j = i \\ 0 & \text{else} \end{cases} = (0,\dots,0,\underbrace{1}_{\text{Pos. } j},0,\dots,0)'$. Indeed, the basis $\{e_1, e_2\}$ discussed so far for the $\mathbb{R}^2$ is its canonical basis! The feature of the canonical basis, beyond intuitively relating to the "isolated" directions into which vectors in $\mathbb{R}^n$ can extend, is that, in contrast to an arbitrary basis, it is unique. However, the concept is somewhat difficult to extrapolate to more general spaces beyond the context of $\mathbb{R}^n$.

[14] A technical comment for the mathematically curious: This is a so-called "contrapositive proof". Recall that we said that $(\neg Q \Rightarrow \neg P)$ is equivalent to $(P \Rightarrow Q)$, i.e. that negation "flips" the implication arrow. So, to establish that Q is implied by P, we can equivalently prove that the negation of Q implies the negation of P.

A last property of the span that is frequently useful is the following:

Proposition 2. ***(Span of a Union)*** *Let $\mathbb{X} := (X, +, \cdot)$ be a real vector space, and let $Y_1, Y_2 \subseteq X$. Then, $Span(Y_1) \cup Span(Y_2) \subseteq Span(Y_1 \cup Y_2)$.*

Proof. Let $x \in Span(Y_1) \cup Span(Y_2)$. Then, there exist $\lambda_1^1, \dots, \lambda_n^1 \in \mathbb{R}$ and $y_1^1, \dots, y_n^1 \in Y_1$ or $\lambda_1^2, \dots, \lambda_n^2 \in \mathbb{R}$ and $y_1^2, \dots, y_n^2 \in Y_2$ so that $\sum_{i=1}^n \lambda_i^j y_i^j = x$, $j \in \{1, 2\}$. Because either way, $y_1^j, \dots, y_n^j \in Y_1 \cup Y_2$, x is a linear combination of elements of $Y_1 \cup Y_2$ and thus $x \in Span(Y_1 \cup Y_2)$. It results that $Span(Y_1) \cup Span(Y_2) \subseteq Span(Y_1 \cup Y_2)$. □

Note that the converse need not be true: let $x \in Span(Y_1 \cup Y_2)$. Then, there exist $\lambda_1^1, \dots, \lambda_n^1$, $\lambda_1^2, \dots, \lambda_m^2 \in \mathbb{R}$ and $y_1^1, \dots, y_n^1 \in Y_1, y_1^2, \dots, y_m^2 \in Y_2$ so that

$$x = \underbrace{\sum_{i=1}^{n} \lambda_i^1 y_i^1}_{=:x_1} + \underbrace{\sum_{i=1}^{m} \lambda_i^2 y_i^2}_{=:x_2} = x_1 + x_2$$

with $x_1 \in Span(Y_1)$ and $x_2 \in Span(Y_2)$. However, nothing guarantees that the sum $x_1 + x_2$ lies in the union of the spans![15]

To summarize, if for a space $\mathbb{X} = (X, +, \cdot)$, we find a set Y such that $X = Span(Y)$, then we found a set of directions that *span* the space $\mathbb{X}$. If Y is a basis, we may interpret its elements as *independent* directions into which the space extends. Finally, if $X = \mathbb{R}^n$ and Y is the canonical basis, its elements are the *fundamental* independent directions of vectors in the space. As a wrap-up, in addition to knowing the ("position of") elements in a set of vectors and endowing them with useful concepts of addition and multiplication, we now know how we may restrict attention to subsets while ensuring that these concepts are still applicable in the desired way. Moreover, we have learned how to define a basis of any vector space $\mathbb{X}$ that allows for an intuitive and concise representation of its elements. Since subspaces are also vector spaces, we also know how to obtain a basis representation for them.

1.2 Normed Vector Spaces and Continuity

The previous section has demonstrated how for very general and abstract sets of vectors, we can define a space, where we can find a helpful *spatial representation* – as with the simple two-dimensional plane $\mathbb{R}^2$ – by characterizing the "position of" elements in a set of vectors using nothing but addition and multiplication in a fashion similar to real numbers, and a manageable representation of basis elements. Moreover, we have seen how one may restrict attention to subsets while ensuring that these concepts are still applicable in the desired way. As we have seen in Figure 3, however, there is much more we can graphically do in the $\mathbb{R}^2$, and thus much more to generalize! Hence, building on the previous insights, this section addresses how we formally define the "length" (or: magnitude) of a vector and, even more important, how we

[15] As an example, consider $Y_1 = \{\begin{pmatrix}1\\0\end{pmatrix}\}$ and $Y_2 = \{\begin{pmatrix}0\\1\end{pmatrix}\}$. Then,

$$Span(Y_1) \cup Span(Y_2) = \left\{\begin{pmatrix}x_1\\x_2\end{pmatrix} \in \mathbb{R}^2 : (x_1 = 0 \vee x_2 = 0)\right\} \subset \mathbb{R}^2 = Span(Y_1 \cup Y_2).$$

assess the distance of two points in general vector spaces. Furthermore, we will learn how to use this distance concept transfer the standard definition of continuity of simple functions mapping real numbers on real numbers to much more general functions.

Before digging into any formal details, let us consider an easy, intuitive example of what will be going on formally and more abstractly below. Hopefully, this will convince you that distances are indeed very intuitive and straightforward concepts! As you may know, the city core of Mannheim, similar to Manhattan, is organized in squares. Roughly, if you move north, the street names are increasing in letters (e.g. L1, M1, N1, etc.) whereas when moving east, they increase in numbers (L1, L2, L3, ...). Accordingly, a map of Mannheim can be thought of as the $\mathbb{R}^2$ with fundamental directions "north" and "east" (south is "negative north" and west "negative east"). If you have trouble imagining this, try to draw the map on a piece of paper; any online map service will also help. So, suppose you are in the econ building in L7 and tired of studying, so you wish to go see a movie in the Cineplex in P4. Then, regardless of how you walk precisely, you will have to go four blocks north and three blocks west, so a total number of seven blocks. This simple calculation (going only "zig-zag") is called the "Manhattan metric", a commonly used mathematical distance measure! Conversely, if you were a bird and could fly there, you would probably go the direct way (so the minimum distance necessary). Recalling the Pythagorean theorem, this distance is $\sqrt{3^2+4^2} = \sqrt{25} = 5$. This is what we call the *Euclidean* distance, that we already have heard of above! The following shows you how we can generalize these intuitive concepts and introduce them to our more abstract framework.

1.2.1 Metric and Norm in a Vector Space

Many basic mathematical concepts are very intuitive; this is especially true for the concept of a *metric* or distance function. Consider two objects that stand nearby you, and ask yourself what properties you would like the "distance" between these two objects to have. Clearly, the distance should not below zero or respectively *non-negative*, and zero if and only if the objects are in fact in the exact same location (e.g. same building but different level in the maps example). Second, it seems natural that the distance should be the same from object 1 to object 2 as for the other way around, i.e. that the distance measure is *symmetric*. Finally, a third natural requirement is the following: when asked to measure the distance traveled from object 1 and 2 (i) directly and (ii) while passing by some arbitrarily located object 3, one should hope the outcome from (i) to be, in some sense, "smaller" than the outcome from (ii). As the following formal definition will show, these three properties are exactly what defines, in the eyes of mathematicians, a distance function.

In the following, we will assume that we consider vectors in some vector space $\mathbb{X} = (X, +, \cdot)$. Thus, you can assume that vector addition and scalar multiplication are well-defined even when they are not explicitly introduced in definitions. As a new (and slightly informal) notation, we will also use $x \in \mathbb{X}$ interchangeably with $x \in X$.

Definition 15. ***(Metric and Metric Space)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space. Then, a function $d : X \times X \mapsto \mathbb{R}$ defines a **metric** on X if it satisfies the following three properties:*

	Condition	*Name*
(i)	$\forall x, y \in X : d(x,y) \geq 0$, *and* $d(x,y) = 0 \Leftrightarrow x = y$	*non-negativity*
(ii)	$\forall x, y \in X : d(x,y) = d(y,x)$	*symmetry*
(iii)	$\forall x, y, z \in X : d(x,y) \leq d(x,z) + d(z,y)$	*triangle inequality*

If d defines a metric on X, we call $(\mathbb{X}, d)$ a ***metric space****.*

The metric is defined on the Cartesian product of X with itself, because the metric takes two elements of X and assesses their distance (i.e. the first element must be in X and also the second, and thus the complete input must lie in the Cartesian product)![16]

The metric is the most crude distance concept that we usually consider. It is crude in the sense that it is based on satisfaction of only on a set of minimum requirements that already rules out a number of erratically behaving functions as distance measures, but, as you will also see below, still leaves a high degree of freedom of how a metric may be defined, and in consequence also some room for properties that may frequently be viewed as inconvenient in applications.

First, a further characteristic of the intuitive distance concept is that, when starting from two objects, call their positions x and y, then moving them in the exact same fashion, e.g. $\tilde{x} = x + z$, $\tilde{y} = y + z$, should leave the measured distance unaffected. In terms of a measure d, this means that $d(x,y) = d(x+z, y+z)$. This property is called *translation invariance*. It is not part of our metric definition, and indeed, it is not ensured to hold for any function that we may call a metric according to Definition 15.

Another (related but distinct) issue of the metric concept is the one of scaling or "distance from the origin". Suppose that we are considering some "origin point"; for the $\mathbb{R}^n$, this would usually be the zero vector $\mathbf{0} = (0, 0, \ldots, 0)'$. The origin point is special because it has no magnitude, that is, it does not extend into any of the vector space's directions. Typically, we find it practical to think of the length of a vector x as its distance from the origin, i.e. $d(x, \mathbf{0})$. Then, intuitively, when doubling the magnitude (e.g. "zooming in", if we imagine the $\mathbb{R}^2$ as a map) of x, we should double its length, so that $d(2 \cdot x, \mathbf{0}) = 2d(x, \mathbf{0})$. Like translation invariance, this is neither part of the definition of a metric nor ensured by it.

These two points motivate the following concept:

Definition 16. ***(Norm and Normed Vector Space)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space. Then, a function $\|\cdot\| : X \mapsto \mathbb{R}$ defines a* ***norm*** *on X if it satisfies the following three properties:*

	Condition	*Name*
(i)	$\forall x \in X : \|x\| \geq 0$, *and* $\|x\| = 0 \Leftrightarrow x = \mathbf{0}$	*non-negativity*
(ii)	$\forall x, y \in X : \|x + y\| \leq \|x\| + \|y\|$	*triangle inequality*
(iii)	$\forall x \in X, \lambda \in \mathbb{R} : \|\lambda \cdot x\| = \lvert\lambda\rvert \cdot \|x\|$	*absolute homogeneity*

If $\|\cdot\|$ defines a norm on $\mathbb{X}$, we call $(\mathbb{X}, \|\cdot\|)$ a ***normed vector space****.*

In contrast to Definition 15, here, we explicitly need the operations from $\mathbb{X}$, vector addition in *(ii)* and scalar multiplication in *(iii)*. Thus, the norm concept is more closely linked to the

[16] Note that the definition of a metric only refers to the *set* X and does *not* need the operations "+" and "·" of $\mathbb{X}$. The space is just there in the definition to simultaneously define a metric space.

vector space as the metric! Don't worry if you don't yet have an intuition for the name of *(iii)*, you'll understand the label by chapter 3. Keep the name in mind! The following definition shows how this concept helps us with the "location" problem.

Definition 17. ***(Norm-induced Metric)*** *Let $(\mathbb{X}, \|\cdot\|)$ be a normed vector space. Then, the metric induced by $\|\cdot\|$ is $d_N : X \times X \mapsto \mathbb{R}, (x, y) \mapsto \|x - y\|$.*

To deepen your understanding of norm and metric, it may be a useful exercise for you to verify that the norm-induced metric is, indeed, a metric. This task is also given as a review question, so that you can check how to proceed using the solutions.

Proposition 3. ***(Inverse Triangle Inequality)*** *Let $(\mathbb{X}, \|\cdot\|)$ be a normed vector space. Then, $\forall x, y \in \mathbb{X} : \|x - y\| \geq |\|x\| - \|y\||$.*

If you have time, cover the proof and try it on your own to practice dealing with norms.

Proof. Let $x, y \in \mathbb{X}$. Then,

$$\|x\| = \|x - y + y\| \overset{\triangle - ineq.}{\leq} \|x - y\| + \|y\| \quad \Leftrightarrow \quad \|x\| - \|y\| \leq \|x - y\|.$$

By the way, "false zeros" (as here, "$-y + y$" at the first equality) are always worth a try if you're attempting to prove an (in)equality! In the exact same fashion (simply inverting the roles of x and y, you can show that

$$\|y\| - \|x\| \leq \|y - x\| = \|(-1) \cdot (x - y)\| \overset{(\star)}{=} |-1| \cdot \|x - y\| = \|x - y\|.$$

where we used absolute homogeneity at $(\star)$. Thus,

$$\|x - y\| \geq \max\{\|y\| - \|x\|, \|x\| - \|y\|\} = |\|x\| - \|y\||. \qquad \square$$

Above, we had motivated the norm concept with the metric's crudeness, and especially its inability to guarantee translation invariance and robustness to scaling. We are now set up to establish that the norm-induced metric indeed always guarantees these desirable properties.

Theorem 10. ***(Norm vs. Metric)*** *Let $(\mathbb{X}, \|.\|)$ be a normed vector space, $\mathbb{X} = (X, +, \cdot)$, and d_N the metric induced by $\|\cdot\|$. Then, d_N defines a metric on $\mathbb{X}$. Further, d_N exhibits the following extra properties:*

	Property	*Name*
(i)	$\forall x, y \in X \; \forall \lambda \in \mathbb{R} \; d(\lambda x, \lambda y) = \lvert\lambda\rvert d(x, y)$	*absolute homogeneity*
(ii)	$\forall x, y, z \in X \; d(x + z, y + z) = d(x, y)$	*translation invariance*

Conversely, from a metric d that satisfies absolute homogeneity and translation invariance, the mapping $n : X \mapsto \mathbb{R}, x \mapsto d(x, \mathbf{0})$, i.e. the distance to the origin $\mathbf{0}$, defines a norm.

Proof. The first part is already clear from the above; absolute homogeneity is immediate from absolute homogeneity of the norm and construction of $d_N(x, y) = \|x - y\|$. To show that a absolutely homogeneous and translation invariant metric d induces the norm $n(x) = d(x, \mathbf{0})$ (try it yourself if you need practice, it's not too hard!), note that by definition of d as a metric

(i) For all $x \in X$, $n(x) \geq 0$, and $n(x) = 0 \Leftrightarrow d(x, \mathbf{0}) = 0 \Leftrightarrow x = \mathbf{0}$,

(ii) For all $x, y \in X$, $n(x+y) = d(x+y, \mathbf{0})$. By translation invariance, $n(x+y) = d(x, -y)$. By triangular inequality, $n(x+y) \leq d(x,0) + d(0,-y) = d(x, 0 + d(-y, 0)$. By absolute homogeneity d, $d(-y,0) = |-1| \cdot d(y,0) = d(y,0)$, and thus $n(x+y) \leq d(x,0) + d(y,0) = n(x) + n(y)$.

(iii) For all $x \in X$, $\lambda \in \mathbb{R}$: $n(\lambda \cdot x) = d(\lambda \cdot x, 0) = |\lambda| \cdot d(x,0) = |\lambda| \cdot n(x)$ by absolute homogeneity of d. □

From this result, it becomes clear that norms are extremely helpful in defining distance functions with a broader range of appealing properties. Indeed, in almost all applications relevant to economists, **norm-induced metrics are our go-to way of defining a distance** in the mathematical sense!

Definition 18. ***(p-Norm and Euclidean space)*** *Consider the real vector space $(\mathbb{R}^n, +, \cdot)$. Then, the p-norm over $\mathbb{R}^n$ with $p \in \mathbb{N}$ is the norm $\|\cdot\|_p : \mathbb{R}^n \mapsto \mathbb{R}, x \mapsto \left(\sum_{k=1}^n |x_k|^p\right)^{1/p}$. Moreover, we define $\|\cdot\|_\infty : \mathbb{R}^n \mapsto \mathbb{R}, x \mapsto \max_{1 \leq k \leq n} |x_k|$ as the **maximum norm**. When d_N^2 is the metric induced by the 2-norm ("Euclidean norm"), we call $((\mathbb{R}^n, +, \cdot), d_N^2)$ the **Euclidean space** of dimension n.*

Note that when $n = 1$, i.e. when considering $\mathbb{R}$ rather than an actual vector space, all p-norms are simply equal to the absolute value. Indeed, the resulting metric, $d(x,y) = |x - y|$ for $x, y \in \mathbb{R}$, is the so-called *natural* metric of the $\mathbb{R}$, and is the common metric used to measure distances between points in $\mathbb{R}$. The interested reader may want to verify that the p-norm indeed constitutes a norm. You can use that the mapping $x \mapsto x^{1/p}$ is concave for $p \geq 1$, and that thus, $(x+y)^{1/p} \leq x^{1/p} + y^{1/p}$, then this should be a simple exercise. The classical spaces considered in economics are metric spaces $(\mathbb{R}^n, +, \cdot)$ endowed with norm-induced metrics to have all the intuitive properties we are interested in. For instance, the "zig-zag" Manhattan-metric discussed earlier corresponds to the metric induced by the 1-norm, and the "direct way" Euclidean metric to one induced by the 2-norm. Mostly, we are interested in the "direct" or "shortest" distance, so that we consider the Euclidean space.[17] A nice relationship that you may want to be aware of is the following:

Proposition 4. ***(p-Norm and Maximum-Norm)*** *Consider the vector space $(\mathbb{R}^n, +, \cdot)$, $n \in \mathbb{N}$, and let $p < \infty$. Then, for any $x \in \mathbb{R}^n$, $\|x\|_\infty \leq \|x\|_p \leq n^{1/p} \cdot \|x\|_\infty$.*

The proof is rather simple, try it first on your own if you have time!

Proof. Let $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$, and $m \in \arg\max_{j \in \{1,\ldots,n\}} |x_j|$, i.e. $m \in \{1, \ldots, n\}$ so that $|x_m| = \max_{j \in \{1,\ldots,n\}} |x_j|$. Then,

$$\|x\|_\infty = \left(\|x\|_\infty^p\right)^{1/p} = (|x_m|^p)^{1/p} \leq \left(\sum_{k=1}^n |x_k|^p\right)^{1/p} = \|x\|_p,$$

[17] For a more thorough introduction to the Euclidean space ($\mathbb{R}^n$ + Euclidean norm), see SB, Chapter 10. Especially the exercises after sections 10.3 and 10.4 may be worth exploring. Solutions can be found online.

and

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p} \leq \left(\sum_{k=1}^{n} \max\{|x_j|^p : j \in \{1,\dots,n\}\}\right)^{1/p}$$
$$= \left(n \cdot \max\{|x_j| : j \in \{1,\dots,n\}\}^p\right)^{1/p} = n^{1/p}\left(\|x\|_\infty^p\right)^{1/p} = n^{1/p}\|x\|_\infty$$

which establishes the proposition. □

As a take-away, if you define a distance measure (a "metric") from a norm, you are guaranteed a broad set appealing, intuitive properties. For the $\mathbb{R}^n$, norms are rather easy to come by, and can e.g. be constructed as p-norms. We usually deal with the Euclidean norm, a special p-norm with p=2, because it has an intuitive "direct distance" interpretation in the $\mathbb{R}^2$.

Indeed, if you do not rely on norm-induced metrics, these properties are not guaranteed. An example for violation of translation invariance is the so-called French Railway metric: Consider the $\mathbb{R}^2$

$$d_{FR}(x,y) = \begin{cases} \|x-y\|_2 & \text{if } x = \lambda y \text{ for a } \lambda \in \mathbb{R}, \\ \|x\|_2 + \|y\|_2 & \text{else.} \end{cases}$$

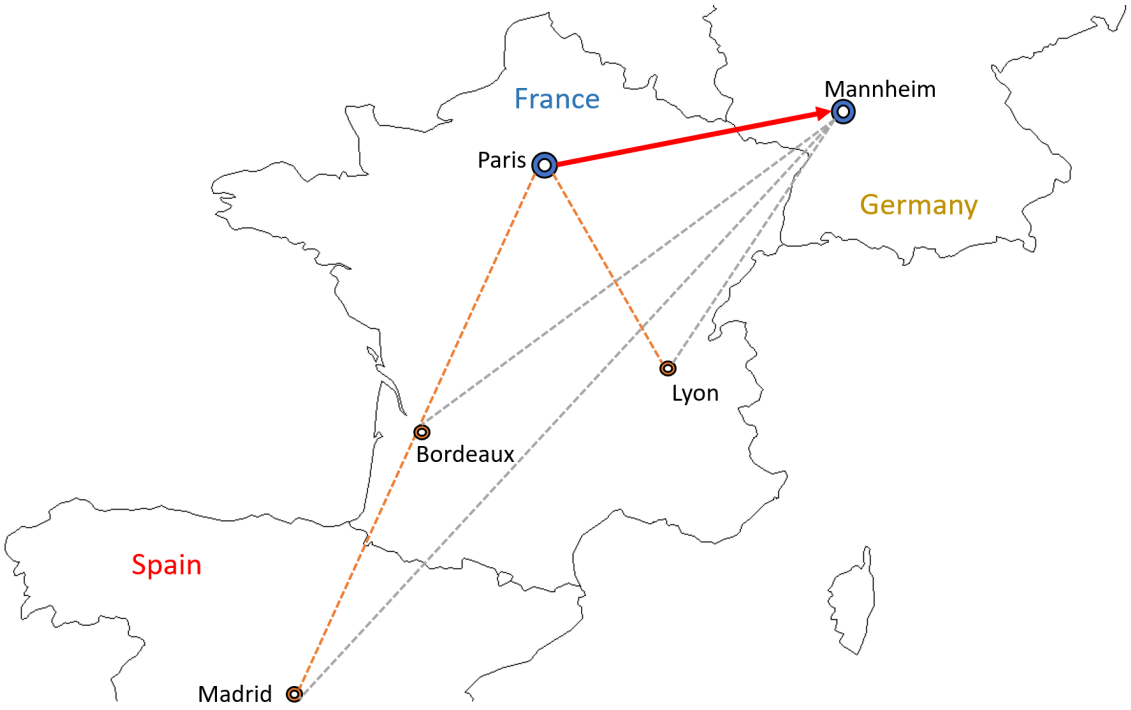


Figure 4: Illustration of the French Railway metric.

The intuition of this metric is sketched in Figure 4. d_{FR} imposes to to pass by the origin (Paris) when measuring the distance between two points that are not contained in a single ray from the origin (orange lines). It is called the French Railway Metric because it used to be almost true that, in France, if you were to travel between two cities that did not lie on a single ray from Paris, then you had to travel through Paris. For instance, in the figure you see that for going from Bordeaux to Madrid you can proceed without going through Paris, while to go from Bordeaux to Lyon, you need to go through Paris. Accordingly, d_{FR} imposes to to pass by the origin (Paris) when measuring the distance between two points that are not contained in a

single ray from the origin.

Then, if one translates the origin to, say, Mannheim, i.e. one imposes that all travels must go through Mannheim unless they are contained on the same line to Mannheim (rather than Paris; gray lines), the distance between Bordeaux and Madrid, or between Toulouse and Bordeaux, as measured by our railway metric, will change! This intuition is easily verified mathematically: Let $x = (1,0)', y = (2,0)'$ and $z = (1,1)'$. Then, $y = 2 \cdot x$, so that

$$d_{FR}(x,y) = \|x-y\|_2 = \|x-2x\|_2 = |-1|\|x\|_2 = (1^2+0^2)^{1/2} = 1.$$

However, $x+z = (2,1)'$ and $y+z = (3,1)'$ are linearly independent (recall our theorem for testing linear independence - this can be verified by solving $\lambda_1(x+z)+\lambda_2(y+z) = 0$ for λ_1, λ_2), so that

$$\begin{aligned} d_{FR}(x+z,y+z) &= \|x+z\|_2 + \|y+z\|_2 = \|(2,1)'\|_2 + \|(3,1)'\|_2 \\ &= \sqrt{2^2+1^2} + \sqrt{3^2+1^2} = \sqrt{5}+\sqrt{10} \\ &> 1 = d_{FR}(x,y). \end{aligned}$$

Therefore, it can not be the case that for any $x, y \in \mathbb{R}^2$ and for any $z \in \mathbb{R}^2$: $d_{FR}(x+z,y+z) = d_{FR}(x,y)$, because we have found a specific counterexample! If you wonder how precisely to come up with such a counterexample, think again about the intuition. If two points were on the same line before moving the origin, the travel distance will be longer if we move the origin such that they are no more. Thus, start from to linearly dependent vectors and move them in a way that they are no longer linearly dependent.

1.2.2 Open, Closed and Compact Sets

Now that we are formally familiar with distances in vector spaces, i.e. with metric spaces (hopefully at least to some degree :-)!), we move on to some definitions and characterizations of sets in vector spaces that can be obtained from the distance measure. These concepts are quite important for economists and are fundamentals of mathematical analysis, so even though they might not be 100% intuitive right now, it is worthwhile developing a firm understanding and good intuition for their meaning. For later courses (and even for later in this one), there is a lot to gain from doing so now.

Definition 19. *(ε-**Ball, Neighborhood**) Let $\mathbb{X} = (X,+,\cdot)$ be a real vector space and $(\mathbb{X},d)$ be a metric space. Further, let $x_0 \in X$, and $\varepsilon > 0$. The ε-**open ball** or **neighborhood** $B_\varepsilon(x_0)$ centered at x_0 is the set of points whose distance from x_0 is strictly smaller than ε, that is:*

$$B_\varepsilon(x_0) = \{x : x \in \mathbb{X}, d(x,x_0) < \varepsilon\}.$$

*Conversely, the ε-**closed ball** $\bar{B}_\varepsilon(x_0)$ centered at x_0 is the set of points whose distance from x_0 is not larger than ε:*

$$\bar{B}_\varepsilon(x_0) = \{x : x \in \mathbb{X}, d(x,x_0) \leq \varepsilon\}.$$

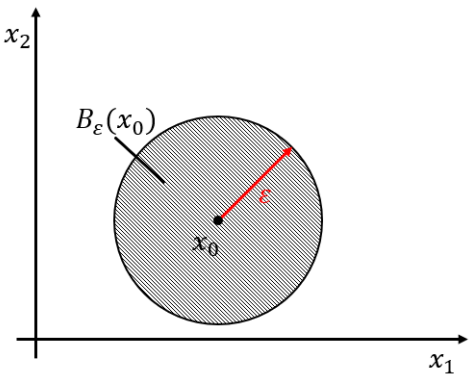


Figure 5: ε-ball in the $\mathbb{R}^2$.

Note that we only call open balls "neighborhoods". The label "ball" again comes from the $\mathbb{R}^2$, especially the Euclidean space (where we, recall, use the metric induced by the Euclidean norm). Here, you can check that an ε-ball around x_0 is merely a circle with radius ε; you may find it easiest to do so with $\varepsilon = 1$ (looking up the definition of the unit circle may help). Then, whether the ball is closed or open is just a matter of whether the *boundary* (defined below) is included in the set, or not. Recall also that we said that in $\mathbb{R}$, all p-norms reduce to the absolute value. Thus, in $\mathbb{R}$ with the metric induced by the absolute value, $d(x,y) = |x-y|$, the balls are just intervals around their middle point: $B_\varepsilon(x_0) = (x_0 - \varepsilon, x_0 + \varepsilon)$ and $\bar{B}_\varepsilon(x_0) = [x_0 - \varepsilon, x_0 + \varepsilon]$. In Figure 5, you can immediately imagine what we mean by the "interior" and the "boundary" of a ball. As we did with addition and scalar multiplication previously, we now extend these concepts to general metric spaces, which allows to generalize this graphical intuition to more abstract scenarios that we can not sketch.

Definition 20. ***(Interior Point, Interior)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, $a \in A$ is said to be an interior point of A if there exists $\varepsilon > 0$ such that the ε-open ball centered at a lies entirely inside of A, i.e. $\exists \varepsilon > 0 : B_\varepsilon(a) \subseteq A$. The set of interior points of A is called the interior of A, denoted $int(A)$ or $\mathring{A}$, i.e. $int(A) = \{a \in A : (\exists \varepsilon > 0 : B_\varepsilon(a) \subseteq A)\}$.*

Definition 21. ***(Open Set)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, A is said to be an open set if $A = int(A)$.*

Intuitively, the open set has only interior points, which is easily visualized with the help of Figure 5. Note that trivially, $int(A) \subseteq A$. Hence, any set A contains its interior, but the converse is true if and only if A is open. Indeed, this is the key take-away also for proofs: to check for openness of a set A, it suffices to check that any point $a \in A$ is also contained in $int(A)$! An example of such proof is given below, where this may become more clear again.

Definition 22. ***(Closure Point, Closure)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, $x \in X$ is said to be a closure point of A if , for every $\varepsilon > 0$, the ε-open ball centered at x contains at least one point a that belongs to A, i.e. $\forall \varepsilon > 0 \exists a \in B_\varepsilon(x) : a \in A$. The set of closure points of A is called the closure of A, denoted $\bar{A}$, i.e. $\bar{A} = \{x \in X : (\forall \varepsilon > 0 \exists a \in B_\varepsilon(x) : a \in A)\}$.*

Intuitively, closure points are either elements of A, or they lie outside of A but "touch" A in the sense that no matter how small a ball we choose around them, they still contain elements of A. Graphically, the latter type of points corresponds to the boundary of the ball illustrated in Figure 5. Like with the closed ball, more generally, a closed set needs to equal the set of all such closure points:

Definition 23. ***(Closed Set)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, A is said to be a closed set if $A = \bar{A}$.*

Hence, any set A is included in its closure, but the converse is true if and only if A is closed. Transferring the intuition of Figure 5 more directly, we now can characterize the boundary as the a set of elements such that, if they all belong to A, then A is closed, and, if none of them belong to A, then A is open.

Definition 24. ***(Boundary Point, Boundary)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, $x \in X$ is said to be a boundary point of A if , for every $\varepsilon > 0$, the ε-open ball centered on x contains both points that belong to A and ones that do not, i.e. $\forall \varepsilon > 0 : (B_\varepsilon(a) \cap A \neq \emptyset \wedge B_\varepsilon(a) \cap (X \backslash A) \neq \emptyset)$. The set of boundary points of A is called the boundary of A and denoted ∂A, i.e. $\partial A = \{a \in A : (\forall \varepsilon > 0 : (B_\varepsilon(a) \cap A \neq \emptyset \wedge B_\varepsilon(a) \cap (X \backslash A) \neq \emptyset))\}$.*

By the way, don't freak out about the notation, it is just there for you to practice. The definition is also fully comprehensible when ignoring all the long symbol expressions ;-). The intuitive characterization of the boundary also follows from Fig. 5, where it corresponds to the set difference of the open and the closed ball, and draws the line between the interior of A and the points that lie outside A. For our usual metric space and $(\mathbb{X}, d)$ and $A \subseteq X$, we may now rephrase our concepts of open and closed sets as follows:

- A is open if and only if *none* of the boundary points of A lie in A: $A \cap \partial A = \emptyset$.
- A is closed if and only if *all* the boundary points of A lie in A: $A \cap \partial A = \partial A$.

Note that a set may be neither open nor closed, namely, if only a fraction of boundary points lie in the set.

Now we have a (rough) idea of what closed and open sets and balls are, and it will soon become evident that they are very useful shortly when studying functions and characterizing the behavior and properties. However, when given a specific set (e.g. think about the budget set $\{x = (x_1, x_2)' \in \mathbb{R}^2_+ : p_1 x_1 + p_2 x_2 \leq y\}$ with income y and prices p_1, p_2),[18] it is typically not immediately clear to determine whether it is open, closed, or neither, and directly proving the definitions above may be cumbersome. To overcome this issue, the next results provide equivalent conditions for openness and closedness of sets.

Theorem 11. ***(Properties of Open Sets)*** *In a metric space $(\mathbb{X}, d)$,*

(i) $\emptyset$ and X are open in $\mathbb{X}$.

(ii) A set $A \subseteq X$ is open if and only if its complement $A^c = X \backslash A$ is closed.

(iii) The union of an arbitrary (possibly infinite) collection of open sets is open.

(iv) The intersection of a finite collection of open sets is open.

Theorem 12. ***(Properties of Closed Sets)*** *In a metric space $(\mathbb{X}, d)$,*

(i) $\emptyset$ and X are closed in $\mathbb{X}$.

(ii) A set $A \subseteq X$ is closed if and only if its complement $A^c = X \backslash A$ is open.

(iii) The union of a finite collection of closed sets is closed.

(iv) The intersection of an arbitrary (possibly infinite) collection of closed sets is closed.

[18] We write $x \in \mathbb{R}^2_+$ because we rule out negative consumption $x_1 < 0$ or $x_2 < 0$.

Note that $\varnothing$ and X are *both* open and closed, so that the properties are not necessarily exclusive! If you want to apply these rules, the union and intersection laws are very helpful to decompose an unwieldy set into a number of sets where openness and closedness are more obvious, and especially Theorem 12 *(iii)* is immensely useful as well, as openness is also not too difficult to check directly.

Two more theorems might be helpful at times to establish closedness:

Theorem 13. ***(Weak Inequalities and the Limit: Functions)*** *Suppose that $\mathbb{X} = (X, +, \cdot)$ is a real vector space, $f : X \mapsto \mathbb{R}$ and $g : X \mapsto \mathbb{R}$ so that $\forall x \in X$: $f(x) \leq g(x)$ (in function notation, we would write $f \leq g$). Let $x_0 \in X$, and suppose that $\exists f_0, g_0 \in \mathbb{R}$ so that $\lim_{x\to x_0} f(x) = f_0$, $\lim_{x\to x_0} g(x) = g_0$. Then, it holds that $f_0 \leq g_0$.*

Proof. Via the contrapositive method: we show that for some given functions, $\neg(f_0 \leq g_0) \Rightarrow \neg(\forall x \in X\text{: } f(x) \leq g(x))$. So, suppose that $f_0 > g_0$. Thus,

$$0 < f_0 - g_0 = \lim_{x\to x_0} (f(x) - g(x)).$$

By the definition of the limit, this means that

$$\forall \varepsilon > 0 : (\exists \delta > 0 : \forall x \in B_\delta(x_0)\backslash\{x_0\} : |(f(x) - g(x)) - (f_0 - g_0)| < \varepsilon.$$

The statement $|(f(x) - g(x)) - (f_0 - g_0)| < \varepsilon$ is equivalent to

$$-\varepsilon < (f(x) - g(x)) - (f_0 - g_0) < \varepsilon.$$

Focusing on the LHS inequality and adding $(f_0 - g_0)$ on both sides, this implies

$$-\varepsilon + (f_0 - g_0) < f(x) - g(x).$$

Choose $\varepsilon = \frac{f_0 - g_0}{2} > 0$, then

$$f(x) - g(x) > (f_0 - g_0) - \frac{f_0 - g_0}{2} > 0,$$

so that for all $x \in B_\delta(x_0)\backslash\{x_0\}$, where $\delta > 0$ is the δ for $\varepsilon = \frac{f_0 - g_0}{2} > 0$ such that the continuity-defining statement holds, we have $f(x) > g(x)$. Thus, we have found $x \in X$ that violate $f(x) \leq g(x)$,[19] i.e. we have found that $\neg(\forall x \in X\text{: } f(x) \leq g(x))$. This concludes our contrapositive proof.□

You can indeed show the analogous theorem for sequences, using the same logic, simply replace $f(x)$ by x_n, $g(x)$ by y_n, and the limits by x and y, respectively. Actually, it may be a very good exercise to practice your understanding of this proof re-doing it for sequences, although admittedly, the result is far more important than understanding why precisely it's true. Let's state the sequence result for completeness:

Theorem 14. ***(Weak Inequalities and the Limit: Sequences)*** *Suppose that $\mathbb{X} = (X, +, \cdot)$ is a real vector space. Let $\{x_n\}_{n\in\mathbb{N}}$ and $\{y_n\}_{n\in\mathbb{N}}$ be convergent sequences over X, i.e. $\forall n \in \mathbb{N} : x_n, y_n \in B$, with limits $x \in X$ and $y \in X$, respectively. If $\forall n \in \mathbb{N}$, it holds that $x_n \leq y_n$, then, we also have $x \leq y$.*

[19] The δ-open ball around x_0 must necessarily be contained in X, i.e. $B_\delta(x_0) \subseteq X$, because the definition of continuity also requires $f(x)$ and $g(x)$ to be defined at any $x \in B_\delta(x_0)\backslash\{x_0\}$!

This fact is extremely useful for the context of set closedness and openness because:

Theorem 15. ***(Closedness and Sequences)*** *Suppose that $\mathbb{X} = (X, +, \cdot)$ is a real vector space, and let $B \subseteq X$. Then, B is closed if and only if, for any convergent sequence $\{x_n\}_{n\in\mathbb{N}}$ over B, i.e. $\forall n \in \mathbb{N} : x_n \in B$, it holds that $\lim_{n\to\infty} x_n \in B$.*

A proof of this theorem can be found at https://math.stackexchange.com/questions/153355/sequential-characterization-of-closedness-of-the-set. To see the intuition, consider again Figure 5. As $n \to \infty$, convergent sequences $\{x_n\}_{n\in\mathbb{N}}$ are restricted to an ever smaller ball around the limit point $x = \lim_{n\to\infty} x_n$. Therefore, if the sequence is over the set B, it will reduce to an ever smaller ball "in proximity to" points of B, if not in B – the precise definition of the closure. This means that either, the limit point lies in the interior or on the boundary. However, for the point x to be certainly included in the set, next to all interior points, any boundary point must be contained in the set – which precisely describes boundedness!

For most mathematical relationships, there is more than one way of establishing them. If you want to make your life easy, you should always think a second about which way will be the shortest and simplest, before deciding on how you're going to prove something. This applies especially (but of course not only) to proofs of set closedness/openness, which is why now is an excellent opportunity to study an example of how one may establish set closedness, to both get some practice with our novel set concepts, and to convince yourself that choosing wisely the approach of proof can save us much time.

Proposition 5. ***(Budget Set Closedness)*** *Consider the metric space $(\mathbb{R}^2, d)$ where d is the metric induced by a p-norm, $d(x,z) = (|x_1 - z_1|^p + |x_2 - z_2|^p)^{1/p}$ for $x = (x_1, x_2)', z = (z_1, z_2)' \in \mathbb{R}$. Let $B(y|p_1, p_2) := \{x = (x_1, x_2)' \in \mathbb{R}^2 : (p_1 x_1 + p_2 x_2 \leq y \wedge x_1, x_2 \geq 0)\}$ denote the budget set for income $y > 0$ and prices $p_1, p_2 \geq 0$. Then, $B(y|p_1, p_2)$ is closed.*

The proof below applies uses the definition of closedness directly to demonstrate how this "direct" approach of proof, which is applicable very generally. The Theorems 14 and 15 can be used to arrive at a very elegant and far more simplistic proof that is sketched below. If you are short on time, you may skip directly to the more elegant variant of the proof.

Proof. Without loss of generality, assume that $p_1 \geq p_2$ (if not, re-label the goods). If $p_1 = 0$, then $B(y|p_1, p_2) = \mathbb{R}^2$, which is closed by Theorem 12 *(i)*. If instead $p_1 > 0$, note that the complement of $B(y|p_1, p_2)$ is $S = \{x = (x_1, x_2)' \in \mathbb{R}^2 : p_1 x_1 + p_2 x_2 > y\}$. As discussed above, clearly, $int(S) \subseteq S$, and to establish $int(S) = S$, it suffices to show $S \subseteq int(S)$.
Let $x_0 = (x_{0,1}, x_{0,2})' \in S$, $\varepsilon > 0$, and $x = (x_1, x_2)' \in B_\varepsilon(x_0)$, i.e. $(|x_1 - x_{0,1}|^p + |x_2 - x_{0,2}|^p)^{1/p} < \varepsilon$. Finally, let $d_0 := p_1 x_{0,1} + p_2 x_{0,2} - y$. Then,

$$\begin{aligned} p_1 x_1 + p_2 x_2 &= p_1(x_1 - x_{0,1}) + p_2(x_2 - x_{0,2}) + p_1 x_{0,1} + p_2 x_{0,2} \\ &> y \quad \text{if and only if} \quad -[p_1(x_1 - x_{0,1}) + p_2(x_2 - x_{0,2})] < d_0. \end{aligned}$$

Let's find a sufficient condition for this to hold. Note that

$$\begin{aligned} -[p_1(x_1 - x_{0,1}) + p_2(x_2 - x_{0,2})] &\leq p_1|x_1 - x_{0,1}| + p_2|x_2 - x_{0,2}| \leq p_1(|x_1 - x_{0,1}| + |x_2 - x_{0,2}|) \\ &\leq 2p_1(\max_{i\in\{1,2\}} |x_i - x_{0,i}|) = 2p_1\|x - x_0\|_\infty \leq 2p_1\|x - x_0\|_p \end{aligned}$$

Because $p_1 > 0$, the last inequality yields the requirement $\|x - x_0\|_\infty \leq d_0/(2p_1) =: \varepsilon$. Because $d_0 > 0$, it holds that $\varepsilon > 0$, and $\forall x \in B_\varepsilon(x_0) : x \in S$ or respectively, $B_\varepsilon(x_0) \subseteq S$. Hence, we have shown that $\exists \varepsilon > 0 : B_\varepsilon(x_0) \subseteq S$ and thus $x_0 \in int(S)$. Because $x_0 \in S$ was chosen arbitrarily, we have that $\forall x_0 \in S : (\exists \varepsilon > 0 : B_\varepsilon(x_0) \subseteq S)$, which is precisely the definition of an interior point: $x_0 \in int(S)$. Thus, $(x_0 \in S \Rightarrow x_0 \in int(S))$, and we conclude $S \subseteq int(S)$ and thus $S = int(S)$. Hence, S is open, and by Theorem 12 *(iii)*, its complement $B(y|p_1, p_2)$ is closed. □

The proof also shows that p-norms and p-norm-induced metrics are quite appealing concepts, as all metrics in this class consistently attribute the closedness label to the budget set. In general, be aware that the concepts of openness and closedness refer to specific metric spaces, so that assigning them to a set need *not* always be robust to variation in the metric used!

As mentioned above, you can proceed far more simplistically and use the Theorems 14 and 15, if you take for granted that the dot product is continuous. Simply pick a sequence $\{x_n\}_{n\in\mathbb{N}}$ over the budget set, i.e. $\forall n \in \mathbb{N} : x_n \in B(p, y)$, and assume that this sequence is convergent with limit $x \in \mathbb{R}^k$. We know that for all $n \in \mathbb{N}$, $p \cdot x_n \leq y$. By continuity of the dot product, $\{p \cdot x_n\}_{n\in\mathbb{N}}$ is a convergent sequence over $\mathbb{R}$ with limit $p \cdot x$. By Theorem 14, $p \cdot x \leq y$, i.e. $x \in B(p, y)$. By Theorem 15, this tells us that the budget set is closed.

Two more concepts regarding sets are of major importance to the mathematic applications economists are concerned with:

Definition 25. ***(Bounded Set)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, A is said to be a bounded set if it is contained in an open ball of finite radius r, i.e. $\exists x_0 \in X \exists r : 0 \leq r < \infty : A \subseteq B_r(x_0)$.*

Verbally, the set is bounded if the distance between its points can not get arbitrarily large, but rather, it is *bounded* by $r < \infty$. It is easily verified that the definition is equivalent to $\exists r^* : 0 \leq r^* < \infty : (\forall a_1, a_2 \in A : d(a_1, a_2) < r^*)$, which more explicitly highlights this interpretation.[20]

Proposition 6. ***(Checking Boundedness with a Norm-induced Metric)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space such that d is norm-induced, i.e. for $x, y \in X$, $d(x, y) = \|x - y\|_p$. Let $A \subseteq X$. Then, A is bounded if the norm is bounded on A, i.e. $\exists b < \infty : (\forall x \in A : \|x\| < b)$.*

Proof. This proposition is very important and in fact, it is easily established. For any $x, y \in A$, the triangle inequality of the metric gives $d(x, y) \leq d(x, 0) + d(y, 0) = \|x\| + \|y\|$. If the norm is bounded by $b < \infty$ on A, then $d(x, y) < 2b$. Thus, for an arbitrary $x_0 \in A$, there exists $r^* = 2b$ so that $\forall x \in A$: $d(x, x_0) < r^*$, and A is bounded. □

Therefore, in the scenario of a p-norm-induced metric, the case usually of interest to us (!), this proposition provides a rather straightforward way to check for boundedness! Indeed, this is likely the most common method that you will want to use when asked to show that a set is bounded in your courses.

Examples of bounded sets are any interval in $\mathbb{R}$ (note that the p-norm metric was $d(x, y) = |x - y|$ for $x, y \in \mathbb{R}$) where $-\infty < a < b < \infty$ on $\mathbb{R}$, or the budget set in the $\mathbb{R}^2$, provided that prices

[20] For completeness, one would proceed as follows to prove completeness here:
"⇒" Suppose A is bounded in the sense of Def. 25. Let $x_0 \in X$, $0 \leq r < \infty$ so that $A \subseteq B_r(x_0)$, and let $a_1, a_2 \in A$. Then, $d(a_1, a_2) \leq d(a_1, x_0) + d(x_0, a_2) < 2r =: r^*$ by the triangle inequality.
"⇐" Suppose that $\exists r^* : 0 \leq r^* < \infty : (\forall a_1, a_2 \in A : d(a_1, a_2) < r^*)$. Then, let $x_0 \in A$ and $r = r^*$. Clearly, $\forall a \in A : d(a, x_0) < r^* = r$. Thus, A is bounded in the sense of Def. 25. □

are strictly positive: $\min\{p_1, p_2\} > 0$. If you're interested, for the budget set, we would proceed as follows. Since the metric is norm-induced, it is sufficient to establish that $\exists b < \infty : \forall x \in B(y|p_1, p_2) : \|x\| < b$. Again, without loss of generality, assume $p_1 \geq p_2$. Using Proposition 4 at the first inequality, note that for $x \in B(y|p_1, p_2)$,

$$\begin{aligned}\|x\| \leq 2^{1/p}\|x\|_\infty = 2^{1/p}\max\{x_1, x_2\} &= \frac{2^{1/p}}{p_2}\max\{p_2x_1, p_2x_2\} \\ &\leq \frac{2^{1/p}}{p_2}(p_2x_1 + p_2x_2) \leq \frac{2^{1/p}}{p_2}(p_1x_1 + p_2x_2) \\ &\leq \frac{2^{1/p}}{p_2}y.\end{aligned}$$

Thus, with $b = (2^{1/p} \cdot y)/p_2 + 1$, $\forall x \in B(y|p_1, p_2) : \|x\| < b$, and $B(y|p_1, p_2)$ is bounded by Proposition 6.

The last concept to be discussed here is compactness. Don't worry about the definition which is rather abstract, it's just given for completeness, the intuition and how we can prove it, as discussed below, are far more important.

Definition 26. ***(Compact Set)*** *Let $\mathbb{X} = (X, +, \cdot)$ be a real vector space and $(\mathbb{X}, d)$ be a metric space. Let $A \subseteq X$. Then, A is said to be compact if every open covering $\{U_i\}_{i\in I}$ with index set I, i.e. $\{U_i\}_{i\in I}$ such that U_i is open $\forall i \in I$ and $\bigcup_{i\in I} U_i \supseteq A$, has a finite subcovering, i.e. $\exists I^* \subseteq I$ such that I^* contains finitely many elements, and $\bigcup_{i\in I^*} U_i \supseteq A$.*

Intuitively, this merely says that regardless of the class of sets you come up with, as long as the whole class covers X, then also a finite number of those sets will cover X, i.e. their union will contain X. This says that X can not be "too large" (similar to the sense of boundedness), and in fact also that X must be closed[21] Indeed, for the $\mathbb{R}^n$, the following equivalence holds:

Theorem 16. ***(Heine-Borel)*** *Consider the metric space $(\mathbb{R}^n, d)$, where d is induced by a p-norm, and let $A \subseteq \mathbb{R}^n$. Then, A is compact if and only if A is closed and bounded.*

Indeed, about 99% of compactness proofs you will see use Heine-Borel's theorem, so that when asked to show compactness, it is **the** starting point for you. To apply it, one separately shows closedness and boundedness. To see the value of compact sets, consider the $\mathbb{R}$, where intervals $[a, b]$ are a special form of closed and bounded and thus compact sets. Clearly, any continuous function f defined on the whole interval will assume a maximum and minimum on such a set, either in the interior (a, b), or otherwise at a or b![22] As we will see, similar reasoning applies to more general spaces, and compact sets are a powerful concept for functional analysis and optimization!

1.2.3 Continuity and Convergence

If you remember the preliminary chapter, using the limit concept for real-valued functions, we associated continuity with the requirement that as two points become ever *closer*, their images

[21] Else, too many sets exist at the boundary, think e.g. of the interval $(0, 1)$, where $(0, 1) = \bigcup_{n\in\mathbb{N}}(1/n, 1 - 1/n)$, but no finite subset of $\{(1/n, 1 - 1/n) : n \in \mathbb{N}\}$ covers $(0, 1)$.

[22] We will establish this formally in Chapter 4.

should not be *too far* apart one from another. Now, we know how to mathematically handle distances more generally, it is time to formalize and generalize the continuity concept, which is the purpose of this section. Start again from the $\mathbb{R}$ and a function $f : X \mapsto Y$ with $X, Y \subseteq \mathbb{R}$, where we call $f_a \in \mathbb{R}$ the limit of f at $a \in \mathbb{R}$, if

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in X : (|x - a| \in (0, \delta) \Rightarrow |f(x) - f_a| < \varepsilon)).$$

The continuity requirement for f at $x_0 \in X$, $f(x_0) = \lim_{x \to x_0} f(x)$ can be written as

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in X : (|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon)).$$

Now, we have repeatedly said that the common metrics that we will use for the $\mathbb{R}^n$ are p-norm-induced, and that for the $\mathbb{R}^n$, any p-norm is equal to the absolute value. Recall also that we said that for the $\mathbb{R}$, we therefore commonly use the so-called *natural* metric $d(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. Then, the definition of continuity at $x_0 \in X$ is equivalent to

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in X : (d(x, x_0) < \delta \Rightarrow d(f(x), f(x_0)) < \varepsilon)).$$

This step is indeed all that is necessary to generalize the continuity concept to arbitrary metric spaces:

Definition 27. ***(Continuous Function)*** *Let $(\mathbb{X}, d_X)$ and $(\mathbb{Y}, d_Y)$ be metric spaces based on the sets X and Y, respectively. Then, a function $f : X \mapsto Y$ is continuous at $x_0 \in \mathbb{X}$ if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that the image of the δ-open ball around x_0 is contained in the ε-open ball around $f(x_0)$, i.e.*

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in X : (d_X(x, x_0) < \delta \Rightarrow d_Y(f(x), f(x_0)) < \varepsilon)).$$

A function that is continuous at every point of its domain is said to be continuous.

Be sure to understand how the statement in quantifiers relates to the verbal statement referring to the open balls. Note also that a function can not be continuous at x_0 if in any δ-open ball around x_0, there is a point at which f is not defined!

Similarly to continuity, we can also straightforwardly generalize convergence of sequences in metric spaces: recall that in $\mathbb{R}$, x is the limit of a sequence $\{x_n\}_{n \in \mathbb{N}}$ if

$$\forall \varepsilon > 0 \exists N \in \mathbb{N} : (\forall n \in \mathbb{N} : (n \geq N \Rightarrow |x_n - x| < \varepsilon)).$$

Using again the natural metric of $\mathbb{R}$, the condition in brackets can be equivalently written as $\forall n \in \mathbb{N} : (n \geq N \Rightarrow d(x_n, x) < \varepsilon)$. Exploiting this intuition, we can define convergence of sequences more generally:

Definition 28. ***(Convergent Sequence)*** *Let $(\mathbb{X}, d)$ be a metric space based on the sets X, and let $\mathbf{x} := \{x_n\}_{n \in \mathbb{N}}$ be a sequence in X, i.e. $\forall n \in \mathbb{N} : x_n \in X$. Then, $\mathbf{x}$ is said to be convergent if*

$$\exists x \in X : (\forall \varepsilon > 0 \exists N \in \mathbb{N} : (\forall n \in \mathbb{N} : (n \geq N \Rightarrow d(x_n, x) < \varepsilon))).$$

If $\mathbf{x}$ is convergent, the point x satisfying this condition is called the limit of $\mathbf{x}$, denoted $x = \lim_{n \to \infty} x_n$.

The fact that the limit of a converging sequence is unique in a metric space is easily shown using the nonnegativity and triangle inequality properties of a metric. Suppose $x_n \to x$ and $x_n \to y$, then $\forall n \in \mathbb{N}\ 0 \leq d(x,y) \leq d(x,x_n) + d(x_n,y)$ and the convergence of the left and right hand side suffices to establish the result (squeeze theorem a.k.a. sandwich theorem[23]).

A theorem that combines the concepts of limits and continuity shall conclude this section. This theorem is perhaps the most important tool to *disprove* continuity, so you can greatly gain from being familiar to it.

Theorem 17. ***(Sequence Characterization of Continuity)*** *Let $(\mathbb{X}, d_X)$ and $(\mathbb{Y}, d_Y)$ be metric spaces based on the sets X and Y, respectively. Then, the function $f : X \mapsto Y$ is continuous at $x_0 \in \mathbb{X}$ if and only if for every sequence $\mathbf{x} := \{x_n\}_{n\in\mathbb{N}}$ in X, $f(x_0) = \lim_{n\to\infty} f(x_n)$.*

Thus, to establish that f is not continuous at x_0, it suffices to find a sequence $\mathbf{x} := \{x_n\}_{n\in\mathbb{N}}$ so that $\lim_{n\to\infty} x_n = x_0$ and either $\lim_{n\to\infty} f(x_n)$ does not exist, or it does but $\lim_{n\to\infty} f(x_n) \neq f(x_0)$.

1.3 Convex Sets and the Separating Hyperplane Theorem

Let's take a step back: we defined vector spaces and subspaces and covered a range of properties. Then we moved on to the concepts of distance and norms. The Euclidean norm is closely linked to the realm of geometry, and it is here that we now open a bracket and cover some topics which you will encounter frequently in your economic classes and for which it is possible to develop a geometric intuition.

While we have argued that by considering a subspace, all appealing properties of a space are maintained, in economic applications, it may not always be possible to restrict attention to a sub*space*. For instance, when maximizing utility subject to a budget constraint, the set we focus on (i.e., the budget set) does generally not give rise to a subspace.[24] This is easily seen as for $x \in B(y|p_1,p_2)$, for $\lambda \in \mathbb{R}$ large enough, we will eventually have $p_1(\lambda x_1) + p_2(\lambda x_2) > y$ and thus $\lambda x \notin B(y|p_1,p_2)$. More generally, economic considerations are motivated by the real world and not by conserving mathematical structures (or at least, they should be ideally, some counter-examples, even well-published ones unfortunately exist), so that nothing guarantees that we may operate in a subspace. Thus, we require additional characterizations of the subsets of vector spaces, regardless of whether they constitute a subspace or not. Beyond a set being open, closed and/or compact, an important concept is set convexity, which builds on a very similar idea as the subspace but applies to a wider range of sets, which will help us analyze a wider range of sets potentially of interest to economic applications along a "geometrical" dimension.

1.3.1 Convex Sets

Hopefully, you perceived that earlier, we established a strong relation between linear combinations and subspaces. We now consider the concept of *convex combinations*, which relates to

[23] If you have a sandwich inequality $a \leq x \leq b$ and a and b converge to the same limit, then x converges to that same limit too.

[24] Exceptions are the trivial cases $p_1 = p_2 = 0$, i.e. consumption is free, or $y = \infty$, i.e. the consumer is actually not budget-constrained, making consumption effectively free again.

convex sets. Compared to subspaces, in a convex set, we restrict the class of linear combinations that we require to be included to the set. This of course comes as a loss in the algebraic dimension. In economics, many optimization problems come with inequality constraints (e.g. the budget set, where money spent on consumption must be smaller or equal to the monetary income). The concept of convex sets helps us deal with that issue, at least in many relevant cases. It is associated with the concept of convex combination and arise naturally when the constraints of our optimization problem are *convex functions*.[25]

Definition 29. ***(Convex Combination and Convex Set)*** *Let $\mathbb{X}$ be a real vector space based on the set X. A convex combination x^c of the vectors $x_1,\dots,x_n \in X$ is a linear combination $x^c = \sum_{i=1}^n \lambda_i x_i$, for which $\forall i \in \{1,\dots,n\} : \lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$.*
A set $A \subseteq X$ is convex if it contains all convex combinations of any two of its elements, i.e. $\forall a_1, a_2 \in A \forall \lambda \in [0,1] : \lambda a_1 + (1-\lambda)a_2 \in A$.

On one hand, the convex combination has the interpretation of a weighted average of the combined vectors, such that in opposition to the general linear combination, no single element can be arbitrarily amplified. On the other, any convex combination of points x_1 and x_2 lies on the line connecting x_1 and x_2. Hence, to verify whether a set is convex or not, one has to make sure that any *line segment* going through two different points of the set is fully contained in it! You can test your understanding using the examples in Figure 6 (see footnote[26] for the answers).

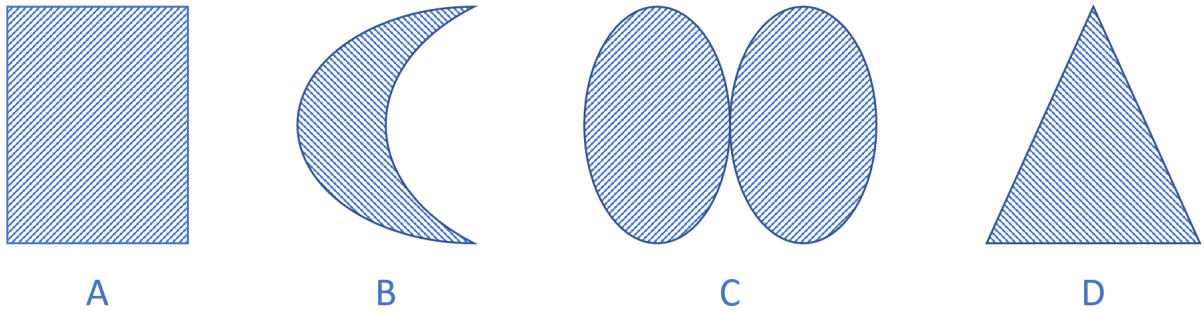


Figure 6: Convex and non-convex subsets of the $\mathbb{R}^2$. Can you tell which are convex and which are not?

Similar to the span that formed the smallest possible subspace from any initial set A, we can also form a convex set. The resulting set is the following:

Definition 30. ***(Convex Hull)*** *Let $\mathbb{X}$ be a vector space based on the set X, and let $A \subseteq X$. The convex hull, denoted $Co(A)$ is the smallest convex set containing A.*

Similar to the span, we construct the convex hull as

$$\mathrm{Co}(A) = \left\{ x \in X : \exists a_1,\dots,a_n \in A \exists \lambda_1,\dots,\lambda_n \geq 0 : \left(\sum_{i=1}^n \lambda_i = 1 \wedge x = \sum_{i=1}^n \lambda_i a_i \right) \right\}.$$

In establishing convexity, the following may be helpful:

[25] We'll define what that means in the next chapter.
[26] A and D are convex, the others are not.

Proposition 7. ***(Convexity-preserving Operations)*** *Let $\mathbb{X}$ be a vector space based on the set X. Then,*

(i) *$\varnothing$ and X are convex.*

(ii) *if $A \subseteq X$ is convex, then so is $\alpha A := \{\alpha \cdot a : a \in A\}$ for any $\alpha \in \mathbb{R}$.*

(iii) *if $A, B \subseteq X$ are convex, then so is $A + B := \{a + b : a \in A, b \in B\}$.*

(iv) *if $\{A_i\}_{i \in I}$ is a (possibly infinite) collection of convex sets, then $\bigcap_{i \in I} A_i$ is convex.*

If you will, we can thus say that convexity is robust scalar multiplication, set addition and intersections. You can try to prove points*(ii)* and *(iii)*, this is a good exercise. *(i)* is obvious from the convexity definition, and the proof for *(iv)* is given below.

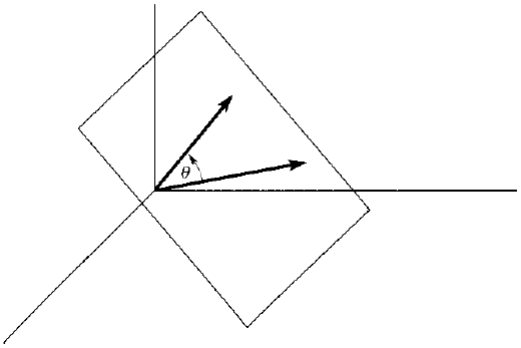
Proof. *(i)* and *(ii)* are self-study exercises. For *(iii)*, let $\{A_i\}_{i \in I}$ be a collection of convex sets. If $\bigcap_{i \in I} A_i = \varnothing$, then the proposition follows from *(i)* $\bigcap_{i \in I} A_i \neq \varnothing$, let $a_1, a_2 \in \bigcap_{i \in I} A_i$ and $\lambda \in [0,1]$. Note that this implies that $\forall i \in I : (a_1, a_2 \in A_i)$. By convexity of the sets, A_i, $\forall i \in I$: $\lambda a_1 + (1-\lambda)a_2 \in A_i$. Thus, $\lambda a_1 + (1-\lambda)a_2 \in \bigcap_{i \in I} A_i$. This yields that $\forall a_1, a_2 \in \bigcap_{i \in I} A_i \forall \lambda \in [0,1]$: $\lambda a_1 + (1-\lambda)a_2 \in \bigcap_{i \in I} A_i$. Thus, $\bigcap_{i \in I} A_i$ is convex. □

1.3.2 Hyperplanes and the Separating Hyperplane Theorem

Remember the definition of the scalar product earlier? It might have seemed to you to "fall from the sky" and that it its only reason to exist is that it is the only type of multiplication actually defined on vectors and presents a handy notation. Well, there is more to it, and now that we know the Euclidean norm, we are able to postulate the following theorem.

Theorem 18. *Let $u, v \in \mathbb{R}^n$, and consider the Euclidean space. In the plane spanned by the two vectors, let θ be the radian angle between them (see picture below). Then*

$$u \cdot v = \|u\| \cdot \|v\| cos(\theta).$$



Proof. See, e.g. SB, page 216. □

If u and v are different from the origin, that is, $u, v \neq \mathbf{0}$, then $\|u\|, \|v\| > 0$, and we can solve for the angle as

$$\theta = \arccos\left(\frac{u \cdot v}{\|u\| \cdot \|v\|}\right)$$

where the arccos-function is the inverse function of the cosine on $[0,\pi]$ and takes arguments from $[-1,1]$. Note that the radian angle satisfies $0° = 0\text{rad}, 45° = \pi/4\text{rad}, 90° = \pi/2\text{rad}$ etc. In this sense, if the vectors are orthogonal ($\theta = 90° = \pi/2$), $u \circ v = \|u\| \cdot \|v\| \cdot \cos(\pi/2) = 0$ since $\cos(\pi/2) = 0$. In this sense, this theorem indeed establishes that general vectors in Euclidean spaces are orthogonal if and only if their scalar product equals zero!

The importance of the scalar product brings us to the following concept:

Definition 31. *Let $\mathbb{X} \subseteq \mathbb{R}^n$. Then, a hyperplane of $\mathbb{X}$ is a set H of the form*

$$H(a,b) = \{x \in X : a \cdot x = b\} = \{x \in X : \sum_{i=1}^{n} a_i x_i = b\}$$

where $a \in \mathbb{R}^n$, $a \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Note that for any $x, y \in H(a,b)$, $a \cdot x = b = a \cdot y \Rightarrow a \cdot (x - y) = 0$, and any point on the hyperplane has the exact same inner product with a, namely b. b can be viewed as the hyperplane's intercept, so that the origin is included in the hyperplane if and only if $b = 0$.

First, let's look at $\mathbb{R}^2$. From high school, you know that a line has the form $x_2 = mx_1 + b$. That is, $(-m, 1) \cdot (x_1, x_2) = b$, which fits the definition of a hyperplane. At the stage, it might be wise to recall that there is also another way of describing the line, called the parametric representation. Here, one makes use of the fact that a line is completely determined by a point x_0 on the line and a direction vector v in which to move from x_0, that is a line can be represented by the equation:[27]

$$x(t) = x_0 + tv, \qquad t \in \mathbb{R}.$$

To see the equivalence, you can derive the first equation by deriving the parameterization of a line through the point $(0,b)$ in the direction $(1,m)$: $x(t) = (0,b) + t(1,m) = (t, b + tm)$. The parametric representation also works in higher dimension, and we will make use of this when talking about derivatives. For example, the line in $\mathbb{R}^3$ through the point $x_0 = (2,1,3)$ and in the direction $v = (4,-2,5)$ has the parameterization

$$x(t) = (2,1,3) + t(4,-2,5) = (2 + 4t, 1 - 2t, 3 + 5t).$$

As you can see, we have generalized a line as a function with a slope augmenting the magnitude t and an intercept vector, which will be the usual notion of lines in everything to follow. To make this discussion complete, note that a line is also defined by two points which are on the line. Suppose that x and y are on the line, which can then also be viewed as a line through x in the direction of $y - x$. Hence: $x(t) = x + t(y - x) = (1 - t)x + ty$.[28]

Let's add one dimension and look at planes. This time, we start with the parametric representation. A plane that passes through the "intercept" point $p \neq \mathbf{0}$ and is spanned by two

[27] By "represented", we mean formally that $x \in \mathbb{R}^2$ lies on the line if and only if there exists $t \in \mathbb{R}$ such that $x = x(t)$.

[28] If $0 \leq t \leq 1$, this defines the line segment between x and y. We have used this property in the characterization of convexity.

linearly independent direction vectors $v, w \in \mathbb{R}^3$ is represented by the equation

$$x(s,t) = p + sv + tw, \qquad s,t \in \mathbb{R}.$$

As two points on a line define it, three points p, q, r can define a plane that passes through p with direction vectors $p-q$ and $r-p$. Plugging it into the above equation yields $x = t_1 p + t_2 q + t_3 r$, $t_1, t_2, t_3 \in \mathbb{R}$. If we wish to consider a plane with these coordinates as "corner points", similar to the convex set, we may restrict $t_1, t_2, t_3 \geq 0$ and $t_1 + t_2 + t_3 = 1$. There is also a non-parametric equation of a plane, which comes from the fact that a plane is completely described by a point $p = (x_0, y_0, z_0)$ on the plane and its inclination. The latter is fixed by specifying a *normal vector* $n = (a, b, c)'$, which is orthogonal to the plane, i.e. $n \cdot x = 0$ for all points x on the plane. To see the representation, fix an arbitrary point $x = (x_1, x_2, x_3)$ on the plane, then $x - p$ is a vector in the plane and therefore orthogonal to n. Hence,

$$0 = n \cdot (x - p) = (a, b, c) \cdot (x - x_0, y - y_0, z - z_0) = a(x - x_0) + b(y - y_0) + c(z - z_0).$$

or $ax + by + cz = d$, which fits the form of a hyperplane.

To take away, hyperplanes can be thought of the generalization of lines, planes and higher-dimensional objects of the same structure in arbitrary spaces $\mathbb{R}^n$. They are very useful for the following central concept in Microeconomics. It is geometrically very intuitive, and the proof is way beyond the scope of this lecture. Therefore, we only illustrate the idea graphically.

Theorem 19. ***(Separating Hyperplane Theorem)*** *Let C and D be two convex and disjoint sets in a metric space $(\mathbb{X}, d)$ over the set X, i.e. $C \cap D = \emptyset$. Then, there exists $a \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and $b \in \mathbb{R}$ such that $\forall x \in C : a \cdot x \leq b$ and $\forall x \in D : a \cdot x \geq b$. The hyperplane $\{x \in X : a \cdot x = b\}$ is called a separating hyperplane for the sets C and D.*

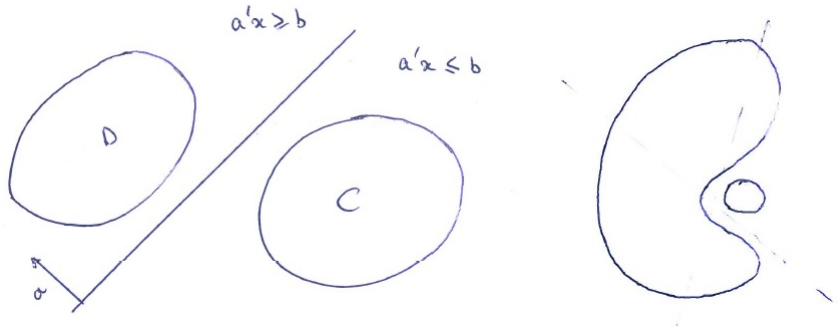


Figure 7: Separating Hyperplane Theorem: Graphical Illustration.

For the figure, note that the theorem is applicable in the left picture but not in the right, because while the sets are disjoint there, the bigger set is not convex.

1.4 Contents and Take-Aways

Chapter 1: Vector Spaces discusses

- how we define a vector space and the intuition of why this is useful
- how to measure mathematical distance using metrics and norms
- set properties in metric and normed spaces: closedness and openness, boundedness and compactness
- the argument for compactness of economic budget sets (with strictly positive prices)
- how to define convergence for general sequences in, and continuity for general functions mapping between metric spaces

Someone with profound knowledge of the contents of this chapter should

- be able to think intuitively about objects in vector spaces in terms of direction and magnitude
- know the central building blocks of a vector space definition: set, vector addition and scalar multiplication
- be able to give some examples for common vector spaces economists deal with
- be familiar with key concepts related to vector spaces, including the scalar product and linear independence
- know the definitions of a metric, a norm, and a norm-induced metric
- know that we prefer p-norm-induced metrics for economic applications, and why
- be able to illustrate the intuition of open and closed sets, as well as the concepts of interior, boundary and closure of a set, using a graph of a ball in the $\mathbb{R}^2$
- be familiar with a range of results that facilitate checking set openness and closedness (using e.g. the complement and unions/intersections, or limits of sequences in the set)
- know how to check set boundedness when using a norm-induced metric
- know the definitions of convergence and continuity in metric spaces, and the sequence characterization of continuity

and be able to answer a number of related questions, including

- Can you point out two issues of the baseline metric concept that motivate a norm-based approach to the mathematical distance definition?
- How is the Euclidean norm defined? Which p-norm does it correspond to?
- How can you use a norm to assess the length of a vector?
- What's the natural metric of the $\mathbb{R}$? How are p-norms a generalization of this natural metric to the $\mathbb{R}^n$?
- How can we use interior points, boundary points, and closure points to assess whether a set is open and/or closed?
- Can a set be both open and closed? Can it be neither?
- According to the Heine-Borel Theorem, which two conditions are equivalent to compactness of a subset of the $\mathbb{R}^n$?
- Is a budget set always closed? Under which condition is it compact?

1.5 Recap Questions

1. Addition and multiplication in $\mathbb{R}^n$: Let $(X, +, \cdot)$ be a real vector space, $x, y \in X$ and $\lambda, \mu \in \mathbb{R}$.
   (a) Is $x + y$ defined? If so, what concept does it refer to?
   (b) Is $\lambda + \mu$ defined? If so, what concept does it refer to?
   (c) Is $\lambda + x$ defined? If so, what concept does it refer to?
   (d) Is $\lambda \cdot x$ defined? If so, what concept does it refer to?
   (e) Is $x \cdot y$ defined? If so, what concept does it refer to?
2. Using Theorem 8, verify that if $B = \{b_1, \ldots, b_m\} \subseteq \mathbb{R}^n$, $m \leq n \in \mathbb{N}$ is a linearly independent set, then any subset $\tilde{B} \subseteq B$ is also linearly independent. (Try to come up with an intuition, admittedly, writing it down in a formally correct way may be a bit challenging.)
3. Think of three different bases for $\mathbb{R}^3$. Include $b_1 = (1, 1, 0)'$ in the second and $b_2 = (1, 0, 4)$ in the third.
4. Verify that the norm-induced metric as defined in Def. 17 is indeed a metric in the sense of Def. 15.
5. To be extended...

2 Matrix Algebra

Very frequently as economists, we have to deal with matrices because they are at the heart of (constrained) optimization problems where we have more than one variable, e.g. utility maximization given budget sets or cost-efficient production given an output target with multiple goods or deviation minimization in statistical analysis when considering multiple parameters (e.g. a standard regression model with two variables, or just one variable and an intercept). Indeed, the profession's heavy reliance on matrices is mainly what makes both empirical and theoretical economists prone to using MATLAB over alternative available computational software. To engage in such analysis, we need to know a variety of matrix properties, what operations we can apply to one or multiple matrices, and very importantly, how to use matrices to solve systems of linear equations.[1]

When considering such systems, a number of (frequently non-obvious) questions are

1. Does a solution exist?

2. If so, is the solution unique? Otherwise, how many solutions are there?

3. How can we (or our computer) efficiently derive the (set of) solutions?

We will shortly define how matrices can be multiplied with each other and with vectors formally. For now, let us confine to the intuition of why matrices are a useful concept in the context of linear equation systems. Consider the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 6 \\ x_2 - x_3 &= 0 \\ 5 \cdot x_1 + x_3 &= 1 \end{aligned}$$

which, by stacking the equations into a vector, we can re-write as

$$\begin{pmatrix} 1 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 \\ 0 \cdot x_1 + 1 \cdot x_2 + (-1) \cdot x_3 \\ 5 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ 1 \end{pmatrix} \quad \Leftrightarrow \quad Ax = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 5 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 0 \\ 1 \end{pmatrix} =: b$$

where $x = (x_1, x_2, x_3)'$ and A is the matrix on the LHS of the last equation. We will verify below that the equivalence holds after formally introducing matrix multiplication. Thus, a system of n equations in k variables has a matrix representation with a *coefficient matrix* A of dimension $n \times m$ and a *solution vector* $b \in \mathbb{R}^n$. You may be familiar with a few characterizations of A that help us to determine the number of solutions: Typically, if $n \geq k$, we can find at least one vector x that solves the system, and if $n > k$, then there are generally infinitely many solutions. However, there is much more we can say about the solutions from looking at A, and how exactly this works will be an important aspect of this chapter.

[1] For an economic application of linear equation systems, you can have a look at the Leontief closed exchange model, e.g. http://www.math.ucsd.edu/~math18/Fall/Lab2/Lab2.shtml. Depending on your current familiarity with matrices, it may be advisable to first read through this chapter to be able to easily follow this example.

2.1 THE VECTOR SPACE $M_{n\times m}$

Because we want to do mathematical analysis with matrices, a first crucial step is to make ourselves aware of the algebraic structure that we attribute to a set of matrices with given dimensions that allow to perform mathematical basis operations (addition and scalar multiplication), that serve as ground for any further analysis we will eventually engage in. As a first step, let us formally consider what a matrix is:

Definition 32. ***(Matrix of dimension*** $n\times m$***)*** *Let* $(a_{ij} : i \in \{1,\dots,n\}, j \in \{1,\dots,m\})$ *be a collection of elements from a basis set* X*, i.e.* $\forall i \in \{1,\dots,n\}, j \in \{1,\dots,m\} : a_{ij} \in X$*. Then, the matrix* A *of these elements is the ordered two-dimensional collection*

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

We call n *the row dimension of* A *and* m *the column dimension of* A*. We write* $A \in X^{n\times m}$*. As an alternative notation, one may write* $A = (a_{ij})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$*, or, if* $n = m$*,* $A = (a_{ij})_{i,j\in\{1,\dots,n\}}$*.*

In the following, we restrict attention to $X = \mathbb{R}$, so that the a_{ij} are real numbers. Note however that this need not be the case; indeed, an important concept in econometrics are so-called block-matrices $A = (A_{ij})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$ where the A_{ij} are matrices of real numbers, and for derivatives, we frequently consider matrices of (derivative) operators, that is, functions, as opposed to numbers.

To apply the vector space concept to matrices, note that matrices of real numbers can be viewed as a generalization of real vectors: a vector $x \in \mathbb{R}^n$ is simply a matrix of dimension $n\times 1$. We now consider objects that may have multiple columns, or respectively, stack multiple vectors in an ordered fashion. Thus, when $a_1,\dots,a_k \in \mathbb{R}^n$ are the columns of a matrix $A \in \mathbb{R}^{n\times m}$, we also write $A = (a_1, a_2, \dots, a_k)$.[2] Therefore, a natural way of defining addition and scalar multiplication for matrices is to apply the operators of the real vector context "element"-wise, i.e. for each column separately:[3]

Definition 33. ***(Addition of Matrices)*** *For two matrices* A *and* B *of* ***identical dimension****, i.e.* $A = (a_{i,j})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$ *and* $B = (b_{i,j})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$*, their sum is obtained from addition of their elements, that is*

$$A + B = (a_{i,j} + b_{i,j})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}.$$

Note that to conveniently define addition, we have to restrict attention to matrices of the *same* dimension! This already means that we will consider not the whole universe of matrices as a vector space, but each potential specific combination of dimensions separately.

[2] Similarly, an alternative, yet less frequently used notation stacks the rows of A on top of each other.

[3] Of course, vector addition and scalar multiplication in $\mathbb{R}^n$ work element-wise themselves, so that in fact, we are simply applying addition and scalar multiplication element-wise to the individual elements in A.

Definition 34. ***(Scalar Multiplication of Matrices)*** *Let $A = (a_{i,j})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$ and $\lambda \in \mathbb{R}$. Then, scalar multiplication of A with λ is defined element-wise, that is*

$$\lambda A := (\lambda a_{i,j})_{i=1,\dots,n,j=1,\dots,m}.$$

Theorem 20. ***(Vector Space of Matrices)*** *For any fixed $n, m \in \mathbb{N}$, the set of all $n \times m$ matrices, $M_{n\times m}$ together with the algebraic operations matrix addition and multiplication with a scalar as defined above defines a vector space.*

Proof. The proof is left as an exercise. □

As the definition sates, it is not the space of all matrices that constitutes a vector space, but rather, every set of matrices of *specific* dimensions $n\times m$, endowed with addition and scalar multiplication. Further, while we will not be concerned much with *distances* of matrices, defining them in accordance with the previous chapter is indeed possible: the matrix norm commonly used is very similar to the maximum-norm defined earlier:

$$\|A\|_\infty = \max\{|a_{ij}| : i \in \{1,\dots,n\}, j \in \{1,\dots,m\}\}.$$

The interested reader (or the one feeling to need some practice with the norm concept) may want to verify that this indeed constitutes a norm.

2.1.1 Important Matrices

Before we move to deeper analysis of matrices and their usefulness for the purpose of the economist, some review and introduction of important vocabulary is necessary. In this section, you can find a collection of the most central terminology for certain special characteristics matrices may have.

First, when characterizing matrices, it may be worthwhile to think about when we say that two matrices are equal.

Definition 35. ***(Matrix Equality)*** *The matrices $A = (a_{ij})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}}$ and $B = (b_{ij})_{i\in\{1,\dots,r\},j\in\{1,\dots,s\}}$ are said to be equal if and only if (i) they have the same dimension, i.e. $(n = r \wedge m = s)$, and (ii) all their elements are equal, i.e. $\forall i \in \{1,\dots,n\}, j \in \{1,\dots,m\} : a_{ij} = b_{ij}$.*

Note especially that it is thus not sufficient that e.g. all elements equal the same value, so that the matrices $\mathbf{0}_{2\times 2} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{0}_{2\times 3} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ are *not* equal.

Definition 36. ***(Transposed Matrix)*** *Let $A = (a_{ij})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}} \in \mathbb{R}^{n\times m}$. Then, the transpose of A is the matrix $A' = (a'_{ij})_{i\in\{1,\dots,n\},j\in\{1,\dots,m\}} \in \mathbb{R}^{m\times n}$ where $\forall i \in \{1,\dots,n\}, j \in \{1,\dots,m\}$, $a'_{ij} = a_{ji}$. Alternative notations are A^T or A^t.*

Thus, the transpose A' "inverts" the dimensions, or respectively, it stacks the columns of A in its rows (or equivalently, the rows of A in its columns). Note that dimensions "flip", i.e. that if A is of dimension $n \times m$, then A' is of dimension $m \times n$. Equation (3) below gives an example.

Before we proceed with the algebra, it is instructive to define and collect terminology for certain special characteristics matrices may have. You may already recognize the first two from our discussions above.

Definition 37. ***(Row and Column Vectors)*** *Let $A \in \mathbb{R}^{n\times m}$. If $n = 1$, A is called a row vector, and a column vector if $m = 1$. By convention, one also calls a column vector simply* ***vector****.*

According to this definition, as we also did before with vectors in $\mathbb{R}^n$, when you just read the word "vector", you should think of a column vector.

Definition 38. ***(Zero Matrices)*** *The zero matrix of dimension $n \times m$, denoted $\mathbf{0}_{n\times m}$, is the $n \times m$ matrix where every entry is equal to zero.*

Definition 39. ***(Square Matrix)*** *Let $A \in \mathbb{R}^{n\times m}$. Then, if $n = m$, A is said to be a square matrix.*

Definition 40. ***(Diagonal Matrix)*** *A* ***square*** *matrix $A = (a_{ij})_{i,j\in\{1,\dots,n\}} \in \mathbb{R}^{n\times n}$ is said to be a diagonal matrix if all of its off-diagonal elements are zero, i.e. $(i \neq j \Rightarrow a_{ij} = 0)$. We write $A = diag\{a_{11},\dots,a_{nn}\}$.*

Note that the diagonal elements a_{ii}, $i \in \{1,\dots,n\}$ need not be non-zero for A to be labeled "diagonal", and thus, e.g. the zero matrix is diagonal, and so is $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$.

Definition 41. ***(Upper and Lower Triangular Matrix)*** *A* ***square*** *matrix $A = (a_{ij})_{i,j\in\{1,\dots,n\}} \in \mathbb{R}^{n\times n}$ is said to be upper triangular if $(i > j \Rightarrow a_{ij} = 0)$, i.e. when the entry a_{ij} equals zero whenever it lies below the diagonal. Conversely, A is said to be lower triangular if A' is upper triangular, i.e. $(i < j \Rightarrow a_{ij} = 0)$.*

Rather than studying the definition, the concept may be more straightforward to grasp by just looking at an upper triangular matrix:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -4 & 3 \\ 0 & 0 & 0 & 2 \end{pmatrix} \quad \text{for which} \quad A' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & -4 & 0 \\ 4 & 0 & 3 & 2 \end{pmatrix}. \tag{3}$$

From its transpose, you can see why the transposition concept is used in the definition for the lower triangular matrix.

Definition 42. ***(Symmetric Matrix)*** *A* ***square*** *matrix $A = (a_{ij})_{i,j\in\{1,\dots,n\}} \in \mathbb{R}^{n\times n}$ is said to be symmetric if $\forall i, j \in \{1,\dots,n\} : a_{ij} = a_{ji}$.*

Equivalently, symmetry is characterized by coincidence of A and its transpose A', i.e. $A = A'$.

Definition 43. ***(Identity Matrices)*** *The $n \times n$ identity matrix, denoted $\mathbf{I_n}$, is a* ***diagonal*** *matrix with all its diagonal elements equal to 1.*

Again, the concept may be more quickly grasped visually:

$$\mathbf{I_n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

2.1.2 Calculus with Matrices

Now that we have laid the formal foundation by introducing the vector spaces of matrices of certain dimension and made ourselves familiar with a variety of important matrices, it is time to take a closer look on how we do calculus with matrices beyond the basis operations. Similar to the scalar product discussed for vectors, we first should know how to multiply *two elements* of the vector space with each other, rather than just one element with a scalar:

Definition 44. ***(Matrix Product)*** *Consider two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times k}$ so that the **column dimension of A is equal to the row dimension of** B. Then, the matrix $C \in \mathbb{R}^{n\times k}$ of **column dimension equal to the one of** A and **row dimension equal to the one of** B is called the product of A and B, denoted $C = A \cdot B$, if $\forall i \in \{1,\dots,n\}, j \in \{1,\dots,k\} : c_{ij} = \sum_{l=1}^{m} a_{il} b_{lj}$.*

As made clear by the bold text, matrix multiplication is subject to a compatibility condition, that slightly differs from the one discussed before for addition. Thus, not all matrices that can be multiplied with each other can also be added, and the converse is also true. To see just how closely this concept relates to the scalar product, write A in row notation and B in column notation, i.e. let $a_1,\dots,a_n \in \mathbb{R}^{1\times m}$ be row vectors such that $A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$[4] and $b_1,\dots,b_k \in \mathbb{R}^m$ column vectors so that $B = \begin{pmatrix} b_1 & \cdots & b_k \end{pmatrix}$. Then,

$$AB = \begin{pmatrix} a_1' \cdot b_1 & a_1' \cdot b_2 & \cdots & a_1' \cdot b_k \\ a_2' \cdot b_1 & a_2' \cdot b_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1}' \cdot b_k \\ a_n' \cdot b_1 & \cdots & a_n' \cdot b_{k-1} & a_n' \cdot b_k, \end{pmatrix}$$

and the matrix product emerges just as an ordered collection of the scalar products of A's rows with B's columns! Now, considering that we frequently leave out the multiplication dot, you can also see where the notation $a'b$ for the scalar product of a and b comes from.

Here is an example to help you practice seeing which element in the product matrix comes from which column and row in the multiplied matrices, and how precisely it is computed:

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 4 & 4 & 0 \\ -2 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 5 & -3 \end{pmatrix} = \begin{pmatrix} 1\cdot 1 + 2\cdot 0 + 3\cdot 5 & 1\cdot 0 + 2\cdot 1 + 3\cdot(-3) \\ 0\cdot 1 + 1\cdot 0 + 0\cdot 5 & 0\cdot 0 + 1\cdot 1 + 0\cdot(-3) \\ 4\cdot 1 + 4\cdot 0 + 0\cdot 5 & 4\cdot 0 + 4\cdot 1 + 0\cdot(-3) \\ (-2)\cdot 1 + 4\cdot 0 + 1\cdot 5 & (-2)\cdot 0 + 4\cdot 1 + 1\cdot(-3) \end{pmatrix} = \begin{pmatrix} 16 & -7 \\ 0 & 1 \\ 4 & 4 \\ 3 & 1 \end{pmatrix}$$

If you try to multiply this expression the other way round, you will quickly see that this doesn't work: recall that the "inner" dimensions needed to coincide, so if A is $n \times \underline{k}$, B must be $\underline{k} \times m$ for the product to exist. Thus, AB and BA exist only if the matrices are square and of equal dimension! And even then, it will generally **not** hold that $AB = BA$.

Besides its complicated look, matrix multiplication does have some desirable properties:

[4] If you prefer a unified representation of matrices in column notation, you can alternatively define $a_1',\dots,a_n'$ as the columns of A', so that $A' = (a_1',\dots,a_n')$, which gives the same matrix A.

Theorem 21. ***(Associativity and Distributivity of the Product)*** *Assuming that A, B, C are matrices of appropriate dimension, the product for matrices is*

(i) Associative: $(AB)C = A(BC)$

(ii) Distributive over matrix addition: $A(B+C) = AB + AC$ *and* $(A+B)C = AC + BC$

This means that standard rules related to addition and multiplication continue to hold for matrices, e.g. when A, B, C and D are matrices of appropriate dimension, then $(A+B)(C+D) = AC + BC + AD + BD$. An exception is of course that commutativity of multiplication is *not* guaranteed. It is noteworthy that the zero and identity element in the matrix space can be dealt with in a fashion very similar to the numbers 0 and 1 in $\mathbb{R}$:

Theorem 22. ***(Zero and Identity matrix)*** *Let $A \in \mathbb{R}^{n\times m}$. Then,*

(i) $A + \mathbf{0}_{n\times m} = A$.

(ii) For any $k \in \mathbb{N}$, $A \cdot \mathbf{0}_{m\times k} = \mathbf{0}_{n\times k}$ *and* $\mathbf{0}_{k\times n} \cdot A = \mathbf{0}_{k\times m}$.

(iii) For any $k \in \mathbb{N}$, $A \cdot \mathbf{I_m} = A$ *and* $\mathbf{I_n} \cdot A = A$.

Be sure to carefully think about where the dimensions of the zero and identity matrices come from, i.e. why they are chosen like this in the theorem! From this, take away that zero and identity matrix work in the way you would expect them to, and that there are no extraordinary features to be taken into account. Further useful properties of matrix operations are

Theorem 23. ***(Transposition, Sum, and Product)***

(i) Let $A, B \in \mathbb{R}^{n\times m}$. Then,

$$(A+B)' = A' + B'$$

(ii) Let $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{m\times k}$. Then:

$$(AB)' = B'A'$$

(iii) If $A \in \mathbb{R}^{1\times 1}$, then A is actually a scalar and $A' = A$.

While the former two points are more or less obviously useful, the third may appear odd; isn't this obvious?! Why should it be part of a theorem? Well, the practical use is that frequently, this can be used to achieve a more convenient representation of complex expressions. For instance, in econometrics, when β denotes a vector of linear regression model coefficients,[5] the squared errors in the model $y_i = x_i'\beta + u_i$ (y_i random variable, x_i random vector of length $k+1$) that are of interest to estimator choice are

$$u_i^2 = (y_i - x_i'\beta)^2 = y_i^2 - 2y_i x_i'\beta + (x_i'\beta)^2.$$

Now, when taking expectations of such an expression (summand-wise), we want the non-random parameters (here: β) to either be at the beginning or the end of an expression. For

[5] If you are not familiar with this expression, it's not too important what precisely this quantity is, just note that $\beta \in \mathbb{R}^{k+1}$, where k is the number of model variables.

the last one, this is not immediately the case: $(x_i'\beta)^2 = x_i'\beta \cdot x_i'\beta$. However, noting that $x_i'\beta$ is scalar, $x_i'\beta = (x_i'\beta)' = \beta' x_i$ (with *(ii)*), and thus, $(x_i'\beta)^2 = \beta' x_i x_i'\beta$.

As a final remark on notation, note that we can use matrices (and vectors as special examples of them) for more compact representations. Consider the sum of squared residuals in prediction of y_i using the information x_i and the prediction vector b, that the ordinary least squares (OLS) estimator minimizes over all b's:

$$SSR(b) = \sum_{i=1}^{n} (y_i - x_i'b)^2 = \sum_{i=1}^{n} u_i(b)^2.$$

When defining $u(b) = (u_1(b), \ldots, u_n(b))'$, we can simply write $SSR(b) = u(b)'u(b)$. Generally, note that the scalar product of a vector with itself is just the sum of squares: $\forall v = (v_1, \ldots, v_n)' \in \mathbb{R}^n : v'v = \sum_{j=1}^{n} v_j^2$. Similarly, when writing the model's design matrix as

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$$

with the individual observations stacked in the rows,[6] we can write the frequently considered expression, $\sum_{i=1}^{n} x_i x_i'$, a sum over vectors, as $X'X$. Make sure that you understand why in the matrix product, the former object is transposed while in the sum, the latter one is.

2.2 Matrices and Systems of Linear Equations

Re-consider the system of linear equations discussed earlier in this chapter. Here, you saw that stacking the equations into a vector, one could arrive at a matrix representation with just one equality sign, characterized by

$$Ax = b, \tag{4}$$

where A is a matrix of LHS coefficients of the variables $x_1, \ldots, x_n$ stacked in x, and b is the vector of RHS "result" coefficients. In the case of the example above, the specific system is

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 5 & 0 & 1 \end{pmatrix}}_{=A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{=x} = \underbrace{\begin{pmatrix} 6 \\ 0 \\ 1 \end{pmatrix}}_{=b}.$$

As an exercise of how matrix multiplication works, you can multiply out Ax in this example and verify that $Ax = b$ is indeed equivalent to the system of individual equations.

Recall that our central questions to this equation system were:

1. Does a solution exist?

[6] This notation may be quite counter-intuitive at first. However, it is based on the fact that most data matrices you see in statistical software store individuals in rows and characteristics in columns.

2. If so, is the solution unique? Otherwise, how many solutions are there?

3. How can we (or our computer) efficiently derive the (set of) solutions?

Thus, the issue at hand is to characterize the solution(s) for x, i.e. the vectors $x \in \mathbb{R}^n$ that satisfy equation (4), ideally in a computationally tractable way. If A, x and b were real numbers and A was unequal to zero, you would immediately know how to solve for x: just bring A to the other side by dividing by it. If instead $A = 0$ (i.e. A can not be inverted), you know that there is no solution for x if $b \neq 0$, but if $b = 0$, there are a great variety of solutions for x – indeed, every $x \in \mathbb{R}$ would solve the equation. The idea is very similar with matrix equations, we just need a slightly different or respectively more general notion of "dividing by" and "invertibility".

Similar to calculus with real numbers, we can define a multiplicative inverse for every $A \in \mathbb{R}^{n \times n}$:

Definition 45. ***(Inverse Matrices)*** *Consider two square matrices $A, M \in \mathbb{R}^{n \times n}$. Then, M is called the inverse of A if $MA = AM = \mathbf{I_n}$. We write $M = A^{-1}$ so that $A^{-1}A = AA^{-1} = \mathbf{I_n}$.*

As with real numbers, we can show that the multiplicative inverse is unique, i.e. that for every $A \in \mathbb{R}^{n \times n}$, there exists *at most* one inverse matrix A^{-1}:

Proposition 8. ***(Uniqueness of the Inverse Matrix)*** *Let $A \in \mathbb{R}^{n \times n}$ and suppose that the inverse of A exists. Then, the inverse of A is unique.*

Proof. Let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix, and consider two candidate inverse matrices $B, C \in \mathbb{R}^{n \times n}$, i.e. assume that $BA = AB = \mathbf{I_n}$ and $CA = AC = \mathbf{I_n}$. Then, it holds that

$$C = C\mathbf{I_n} = C(AB) = \underbrace{(CA)}_{=\mathbf{I_n}} B = \mathbf{I_n}B = B,$$

i.e. $C = B$. □

However, existence is not guaranteed, and in contrast to the real numbers, more than a single element (0) will be non-invertible in the space $R^{n \times n}$. Existence of the inverse is rigorously discussed below. For now, you should take away that the *easiest* case of a system of linear equations is one where the matrix A is invertible. Indeed, the following sufficient condition is what economists rely on most of the time when solving linear equation systems:

Proposition 9. *Consider a system of linear equations $Ax = b$ with unknowns $x \in \mathbb{R}^n$ and coefficients $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$. Then, if $m = n$ and A is invertible, the system has a unique solution given by $x^* = A^{-1}b$.*

Proof. Suppose A is invertible, and that $x^* \in \mathbb{R}^n$ solves the system. Then,

$$Ax^* = b \quad \Leftrightarrow \quad A^{-1}b = A^{-1}Ax^* = x^*.$$ □

As discussed above, if we can invert A, we can just bring it to the other side, this is exactly the same principle as with a single equation with real numbers. For this sufficient condition to be applicable, we **must have** that $m = n$ for A to be square, i.e. we must have as many equations

as unknowns. It may also be worthwhile to keep in mind that the converse of Proposition 9 is also true: If $Ax = b$ has a unique solution and A is square, then A is invertible. We return this fact in detail after discussing the Gauss-Algorithm and its relation to systems of equations.

2.3 Invertibility of Matrices

Above, we have introduced the most frequent form of a linear equation system economists study, and established Proposition 9, which tells us how to determine the unique solution, provided that it exists, i.e. provided that the matrix A is invertible. However, we do not yet know how we determine whether A can be inverted and if so, how to determine the inverse – these issues will be the subject of remaining discussion in this chapter.

First, some helpful relationships for inverse matrices are:

Proposition 10. *Suppose that $A, B \in \mathbb{R}^{n\times n}$ are invertible. Then,*

- *A' is invertible and $(A')^{-1} = (A^{-1})'$,*
- *AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$,*
- *$\forall \lambda \in \mathbb{R}$, $\lambda \neq 0$, λA is invertible and $(\lambda A)^{-1} = 1/\lambda A^{-1}$.*

Proof. The proof is rather simple. It is left to you to explore it as an exercise in the Recap Questions. You can also look it up in the review question solutions. □

An important corollary is obtained by iterating on *(ii)*:

Corollary 1. *For any $p \in \mathbb{N}$, if $A_1, \ldots, A_p \in \mathbb{R}^{n\times n}$ are invertible, then $A_1 \cdot \ldots \cdot A_p$ is invertible with inverse $(A_1 \cdot \ldots \cdot A_p)^{-1} = A_p^{-1} \cdot A_{p-1}^{-1} \cdot \ldots \cdot A_1^{-1}$.*

Proof. To estabilsh this corollary, we have to rely on a proof by induction.

Base case. We begin with the smallest number that we care about, here the product of two matrices, i.e. $p = 2$. Here, $A_1 \cdot \ldots \cdot A_p = A_1A_2$ is immediately invertible by Proposition 10 *(ii)* with inverse $(\Pi_1^p A)^{-1} = A_2^{-1}A_1^{-1}$.

Inductive Step. Now, we consider an arbitrary number $p-1$ and establish that if the statement holds for $p-1$, it does so for the next one (i.e. p) as well. Suppose that $p \in \mathbb{N}$, $p \geq 3$, and $A_1 \cdot \ldots \cdot A_{p-1}$ is invertible with inverse $A_{p-1}^{-1} \cdot \ldots \cdot A_1^{-1}$ (this is the inductive assumption). Then, since A_p is invertible, by Proposition 10 *(ii)*, $A_1 \cdot \ldots \cdot A_p$ is invertible and

$$(A_1 \cdot \ldots \cdot A_p)^{-1} = [(A_1 \cdot \ldots \cdot A_{p-1}) \cdot A_p]^{-1} = A_p^{-1}(A_1 \cdot \ldots \cdot A_{p-1})^{-1} = A_p^{-1} \cdot \ldots \cdot A_1^{-1}.$$

Thus, if the statement holds for $p-1$, it does so also for p.

Taken together, as in any proof by induction, the base case and the inductive step establish the corrolary: the base says that the statement holds for $p = 2$. The step concludes that it therefore holds for $p = 3$, and thus also for $p = 4$, and thus for $p = 5, \ldots$, i.e. that it holds for any fixed $p \in \mathbb{N}$! □

While this proposition is very helpful at times, it still assumes invertibility of some initial matrices, and therefore does not fundamentally answer when and how an inverse matrix can be computed. The deliberations in this section address the former issue: determining whether a

matrix can be inverted (without relying on invertibility of other matrices). Here, there are four common possible approaches: the determinant, the rank, the eigenvalues and the definiteness of the matrix. Before introducing and discussing these concepts, we turn to the elementary operations for matrices. Not only are they at the heart of solving systems of linear equations, but they also interact closely with all concepts discussed below.

2.3.1 Elementary Matrix Operations

Let's begin with the definition:

Definition 46. ***(Elementary Matrix Operations)*** *For a given matrix* $A = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ *with rows* $a_1', \dots, a_n' \in \mathbb{R}^n$*, consider an operation on* A *that changes the matrix to* $\tilde{A}$*, i.e.* $A \to \tilde{A}$ *where* $\tilde{A} = \begin{pmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_n \end{pmatrix}$*. The three elementary matrix operations are*

(E1) Interchange of two rows i, j*:* $\tilde{a}_i = a_j$*,* $\tilde{a}_j = a_i$ *and* $\tilde{a}_k = a_k$ *for all* $k \notin \{i, j\}$*,*

(E2) Multiplication of a row i *with a scalar* $\lambda \neq 0$*:* $\tilde{a}_i = \lambda a_i$ *and* $\tilde{a}_j = a_j$ *for all* $j \neq i$*,*

(E3) Addition of a multiple of row j *to row* i*:* $\tilde{a}_i = a_i + \lambda a_j$*,* $\lambda \in \mathbb{R}$*, and* $\tilde{a}_j = a_j$ *for all* $j \neq i$*.*

To increase tractability of what we do, the following is very helpful:

Proposition 11. ***(Matrix Representation of Elementary Operations)*** *Every elementary operation on a matrix* $A \in \mathbb{R}^{n \times m}$ *can be expressed as* ***left-multiplication*** *a* ***square*** *matrix* $E \in \mathbb{R}^{n \times n}$ *to* A*.*

(E1) The exchange of rows i *and* j *is represented by* $E^1 = (e^1_{kl})_{k,l \in \{1,\dots,m\}}$ *with* $e^1_{ij} = e^1_{ji} = 1$*,* $e^1_{kk} = 1$ *for all* $k \notin \{i, j\}$ *and* $e^1_{kl} = 0$ *else.*

(E2) Multiplication of a row i *with* $\lambda \neq 0$ *is represented by the diagonal matrix* $E^2 = (e^2_{kl})_{k,l \in \{1,\dots,m\}}$ *where* $e^2_{kl} = 0$ *for any* $k \neq l$ *(the definition of a diagonal matrix),* $e^2_{ii} = \lambda$ *and* $e^2_{jj} = 1$ *for* $j \neq i$*.*

(E3) Addition of a multiple $\lambda \in \mathbb{R}$ *of row* j *to row* i *is represented by the* ***triangular*** *matrix* $E^3 = (e^3_{kl})_{k,l \in \{1,\dots,m\}}$ *with* $e^3_{kk} = 1$ *for all* $k \in \{1, \dots, m\}$ *and* $e^3_{ij} = \lambda$*.*

To see these rather abstract characterizations in a specific example, consider a system with 4 rows and suppose that we use *(E1)* to exchange rows 1 and 3, *(E2)* to multiply the fourth row by 5 and *(E3)* to add row 2, multiplied by 2, to row 3. Then, the associated matrices are

$$E^1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad E^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}, \qquad E^3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The proof of this proposition is omitted because it is rather notation-intensive. However, it is extremely straightforward to verify, if you are motivated and have time, feel encouraged to

do so. As stated above, the practical value of this proposition lies in the fact that when we bring a matrix A to another matrix $\tilde{A}$ using the elementary operations $E_1, \ldots, E_k$, where the index j of E_j indicates that E_j was the j-th operation applied, then we can write

$$\tilde{A} = E_k \cdot E_{k-1} \cdot \ldots \cdot E_1 A.$$

This increases tractability of the process of going from A to $\tilde{A}$, a fact which we will repeatedly exploit below.

The key feature of the elementary operations is that one may use them to bring *any* matrix to (generalized) upper triangular form. This is helpful because, as will emerge, both the determinant and rank invertibility condition hold for an initial matrix if and only if they hold for a generalized upper triangular matrix obtained from applying elementary operations to A.

Definition 47. ***(Generalized Upper Triangular Matrix)*** *Let $A \in \mathbb{R}^{n\times m}$. Then, if $n \geq m$, we say that A has generalized upper triangular form if for a upper triangular matrix $A_T \in \mathbb{R}^{m\times m}$, $A = \begin{pmatrix} A_T \\ \mathbf{0_{n-m\times m}} \end{pmatrix}$. On the other hand, if $n \leq m$, we say that A has generalized upper triangular form if for a upper triangular matrix $A_T \in \mathbb{R}^{n\times n}$, $A = (A_T, X)$ where X is an arbitrary $n \times m - n$ matrix.*

Informally, the biggest possible upper square block of A must be upper triangular, and there must not be any zeros below it for A to have generalized upper triangular form. To see this more intuitively, consider the matrices

$$A = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 4 & 7 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 2 & 3 & 0 \\ 0 & 1 & 4 & 7 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then, A, B and D have generalized upper triangular form, but C does not.[7] The following will be central to all our discussions to follow:

Theorem 24. ***(Triangulizability of a Matrix)*** *Consider a matrix $A \in \mathbb{R}^{n\times m}$. Then, if $n = m$, A can be brought to upper triangular form applying only elementary operations to A. Generally, A can be brought to generalized upper triangular form.*

The proof is a bit tedious. It is given below, but it is fine if you just understand the result well. That being said, looking at it may help you train your understanding of induction proofs! Also, the triangularization algorithm may be quite helpful, but it will hopefully also become clear from the applications in class, so that the formal definition may not be too important.

Proof. The proof has two steps: We must show that...

1. any $n \times n$ (square) matrix can be triangularized (using the principle of induction).

2. regardless of n and m, we can reduce the problem to triangularizing a square matrix.

[7] Note that the block A_T in D corresponds to the left 3×3 block, and not to $\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ as with the other matrices!

Step 1. We apply the principle of induction. For the inductive base, consider $n = 2$,[8] and let $A \in \mathbb{R}^{2\times 2}$. Write $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. If $a = 0$, we can exchange rows 1 and 2 such that

$$\tilde{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ a & b \end{pmatrix}$$

is triangular. If instead $a \neq 0$, we can add $-c/a$ times row 1 to row 2 and obtain

$$\tilde{A} = \begin{pmatrix} 1 & 0 \\ -c/a & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & d - bc/a \end{pmatrix}$$

which is upper triangular. Thus, any 2×2 matrix can be brought to upper triangular form.
Inductive hypothesis. Suppose that for a fixed $n \geq 3$, any matrix of dimension $(n-1)\times(n-1)$ can be brought to upper triangular form.
Inductive step. Let $A \in \mathbb{R}^{n\times n}$, and write $A = \begin{pmatrix} A_{n-1} & a_{col} \\ a_{row} & a_{nn} \end{pmatrix}$, where A_{n-1} is the upper $(n-1)\times(n-1)$ block of A, a_{nn} is the (n,n)-element of a and a_{row}, a_{col} are the remaining row/column vectors in A. By the inductive hypothesis, A_{n-1} can be brought to upper triangular form. Denote the respective operations by $E_1,\dots,E_k$, and for $j \in \{1,\dots,k\}$ and the resulting upper triangular matrix as $\tilde{A}_{n-1} = (\tilde{a}_{ij})_{i,j\in\{1,\dots,n-1\}}$, and define $\tilde{E}_j = \begin{pmatrix} E_j & \mathbf{0}_{n-1\times 1} \\ \mathbf{0}_{1\times n-1} & 1 \end{pmatrix}$ as the transformation of A that only changes A_{n-1} according to E_j. Then,

$$\tilde{E}_k \cdot \tilde{E}_{k-1} \cdot \ldots \cdot \tilde{E}_1 A = \begin{pmatrix} \tilde{A}_{n-1} & \tilde{a}_{col} \\ a_{row} & a_{nn} \end{pmatrix}.$$

We now want to verify that we can use more elementary operations to arrive at upper triangular form. Note that the only thing in the way is $a_{row} := (a_{n1}, a_{n2}, \dots, a_{n,n-1})$. We move along the elements a_{nj}, $j \in \{1, 2 \dots, n-1\}$ with increasing index and ensure that we arrive at a zero using only elementary operations. Starting at $j = 1$, the following algorithm can be applied to arrive at triangular form:[9]

Algorithm 1. ***(Bringing the Last Row to Triangular Form)*** *Start at $j = 1$.*

1. *For the current j, eliminate the potentially non-zero entry $a_{nj} = 0$:*
   - *if $a_{nj} = 0$, no change needs to be applied.*
   - *if $a_{nj} \neq 0$ and $\tilde{a}_{jj} \neq 0$, add $-a_{nj}/\tilde{a}_{jj}$ times the j-th row to a_{nj}.*
   - *if $a_{nj} \neq 0$ and $\tilde{a}_{jj} = 0$, exchange rows n and j.*
2. *Increase j by 1.*
3. *Stop if $j > n-1$. Else, go to 1.*

[8] We do not start at $n = 1$ because by the definition of triangular matrices, it is ambiguous whether scalars are triangular – there are no elements below the diagonal!

[9] The algorithm will become much clearer from examples. Indeed, a first example is the inductive base, take the time to verify this. Further applications are found below and on the matrix problem set.

After applying the algorithm, we have a structure with an upper $(n-1)\times(n-1)$ block that is upper triangular and a last-row column vector with zeros up to index j. Therefore, the last row has become $(0,\dots,0,\tilde{a}_{nn})$ and the resulting matrix is upper triangular, which concludes the inductive step.

By the principle of induction, this establishes Step 1.

Step 2. If $n = m$, there is nothing to prove. If instead $m > n$, write $A = (A_{n\times n}, A_{n\times m-n})$ with blocks $A_{n\times n}$ of dimension $n\times n$ and $A_{n\times m-n}$ of dimension $n\times(n-m)$. By Step 1, we can triangularize $A_{n\times n}$ using operations $E_1,\dots,E_k$ to $\tilde{A}_{n\times n}$. Therefore,

$$E_k\cdot\ldots\cdot E_1A = (E_k\cdot\ldots\cdot E_1A_{n\times n}, E_k\cdot\ldots\cdot E_1A_{n\times n-m}) = (\tilde{A}_{n\times n}, E_k\cdot\ldots\cdot E_1A_{n\times n-m})$$

has generalized triangular form with arbitrary block $X = E_k\cdot\ldots\cdot E_1A_{n\times n-m}$.[10] Finally, if $n > m$, write A in row notation, i.e. $A = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$. Note that the $n > m$ rows have length n, and thus, at most n of them are linearly independent. Conversely, at least $n-m$ rows are linear combinations of other rows. Consider row a_j for which linear combination exists, and exchange rows a_j and a_n, so that in the new matrix A, we can write $a_n = \sum_{i=1}^{n-1}\lambda_i a_i$. Then, for any i such that $\lambda_i \neq 0$, subtract λ_i times row i from row n to arrive at $\tilde{A} = \begin{pmatrix} \tilde{a}_{11} & \dots & \tilde{a}_{1m} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{n-1,1} & \dots & \tilde{a}_{n-1,m} \\ 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \tilde{a}_1 \\ \vdots \\ \tilde{a}_{n-1} \\ \mathbf{0}_{1\times(n-1)} \end{pmatrix}$. If $n-1 = m$, the issue has reduced to triangularization of a square matrix. If instead $n-1 > m$, there must be at least one row $\tilde{a}_j$, $j\in\{1,\dots,n-1\}$ that linearly depends on the others, and we can re-apply the procedure above to get the next zero row. This can be repeated until the upper block in $\tilde{A}$ is square. □

Before moving on, it is instructive to establish the link of elementary operations to matrix *inversion* that we did not explicitly touch upon thus far. The connection is stunningly simple: suppose that we can bring a square $n\times n$ matrix A not only to triangular, but diagonal form with an all non-zero diagonal using the operations $E_1,\dots,E_k$, i.e.

$$E_k\cdot\ldots\cdot E_1A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_k \end{pmatrix} = \mathrm{diag}\{\lambda_1,\dots,\lambda_n\}$$

where $\forall j\in\{1,\dots,n\}:\lambda_j\neq 0$. Then, multiplying all columns j by $1/\lambda_j$, summarized by $E_{k+1} = \mathrm{diag}\{1/\lambda_1,\dots,1/\lambda_n\}$, with $E := E_{k+1}\cdot E_k\cdot\ldots\cdot E_1$, one has $EA = \mathbf{I_n}$. This is very convenient: the transformation matrix E is precisely what we call the inverse matrix of A, i.e. $E = A^{-1}$! Thus, we can summarize the following:

[10] The equation has used the concept of matrix block multiplication: Whenever the dimension conditions for multiplication are satisfied, i.e. the blocks are conformable, we can also apply matrix multiplication block-wise!

Proposition 12. ***(Invertibility and Identity Transformation)*** *If we can use elementary operations $E_1, \ldots, E_k$ to bring a matrix $A \in \mathbb{R}^{n\times n}$ to the identity matrix, then A is invertible with inverse $A^{-1} = E$ where $E := E_k \cdot \ldots \cdot E_1$.*

Indeed, we will establish later that the converse is also true. As you will see in the last section, the ensuing result is what we base upon our method of computing inverse matrices, the Gauß-Jordan algorithm: to compute the matrix E, note that $E = E \cdot \mathbf{I_n}$, so that when bringing an invertible matrix A to identity form, we just have to apply the *same operations* in the *same order* to the identity matrix to arrive at the inverse! Don't worry if this sounds technical, it's not, hopefully you will be convinced of this at the time you have seen some examples of the Gauss-Jordan method.

2.3.2 Determinant of a Square Matrix

Now, it is time to turn to our most common matrix invertibility check, the determinant. The very first thing to note about the determinant concept is that **ONLY square matrices have a determinant!!**, i.e. that for any non-square matrix, this quantity is NOT defined! To give some intuition on the determinant, it may be viewed as a scaling factor of a "base matrix" with determinant equal to one, similarly to the magnitude coefficient that augments the directionality in the case of vectors. This intuition is helpful because it turns out that as with real numbers, we can invert anything that is of non-zero magnitude – i.e. any matrix with non-zero determinant! However, note that unlike with a vector, a matrix with non-zero entries can have a zero determinant and thus have "zero magnitude", so that this reasoning should be viewed with some caution.

The general definition of the determinant is unnecessarily complex for our purposes.[11] We will define the determinant recursively here, that is, we first define it for simple 1×1 matrices, and then express the determinant of an $n\times n$ matrix as a function of that of several $(n-1)\times(n-1)$ matrices, which each can be expressed with the help of determinants of $(n-2)\times(n-2)$ matrices, and so on. It is conceptually more straightforward and sufficient for the use you will make of them.

Definition 48. ***(Determinant)*** *Let $A \in \mathbb{R}^{n\times n}$. Then, for the determinant of A, denoted* $\det(A)$ *or* $|A|$*, we define*

(i) if $n = 1$ and $A = (a)$ is a scalar, $\det(A) = \det(a) := a$.

(ii) for all $n \in \mathbb{N} : n \geq 2$, when $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$, $\det(A) := \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{-ij})$ *with*

[11] There does exist a general, direct formula to compute the determinant, but it makes reference to involved concepts we are not independently interested here, such as permutations and their associated parity.

$i = 1$ *and* A_{-ij} *as the matrix resulting from eliminating row* i *and column* j *from* A*, i.e.*

$$A_{-ij} = \begin{pmatrix} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{pmatrix}.$$

This definition allows to obtain the determinant for any arbitrary $n \times n$ matrix by decomposing the A_{-ij}'s into smaller matrices until we have just 1×1 matrices, i.e. scalars, where we can determine the determinant using *(i)*. The reason why there is the general index i in *(ii)* that is subsequently set to 1 is the following relationship:

Theorem 25. ***(Laplace Expansion)*** *For any* $i^*, j^* \in \{1, \dots, n\}$*, it holds that*

$$\det(A) = \sum_{j=1}^{n} (-1)^{i^*+j} a_{i^*j} \det(A_{-i^*j}) = \sum_{i=1}^{n} (-1)^{i+j^*} a_{ij^*} \det(A_{-ij^*}).$$

The definition deliberately makes use of stars for indices $i^*, j^* \in \{1, \dots, n\}$ to emphasize that the respective index is fixed and distinct from the running index of the sum. If we fix a row index i to calculate $\det(A)$, we call the computation method a Laplace expansion by the i-th row, if we fix the column index j, we call it a column expansion by the j-th column. **This method is the general way how we compute determinants.** However, it is quite computationally extensive, and luckily, most matrices that we deal with analytically (rather than with a computer who doesn't mind lengthy calculations) are of manageable size where we have formulas for the determinant.

Proposition 13. ***(Determinants of "small" Matrices)***

(i) If $n = 2$ *and* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$*, then* $\det(A) = ad - bc$*.*

(ii) If $n = 3$ *and* $A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$*, then* $\det(A) = aei + bfg + cdh - (ceg + bdi + afh)$*.*

Proof. For *(i)*, note that $A_{-11} = (d), A_{-12} = (c), A_{-21} = (b)$ and $A_{-22} = (a)$. Thus, applying Def. 48 *(ii)*, i.e. using the Laplace expansion by the first row, we have

$$\begin{aligned} \det(A) &= \sum_{j=1}^{2} (-1)^{1+j} a_{1j} \det(A_{-1j}) = (-1)^2 a_{11} \det(A_{-11}) + (-1)^3 a_{12} \det(A_{-12}) \\ &= 1 \cdot a \cdot \det(d) + (-1) \cdot b \cdot \det(c) = ad - bc. \end{aligned}$$

As an exercise and to convince you of the Laplace expansion, try expanding by the second column and verify that you get the same formula.

For $n = 3$, expansion for the first row gives

$$\det(A) = (-1)^2 a \det\begin{pmatrix} e & f \\ h & i \end{pmatrix} + (-1)^3 b \det\begin{pmatrix} d & f \\ g & i \end{pmatrix} + (-1)^4 c \det\begin{pmatrix} d & e \\ g & h \end{pmatrix}.$$

Applying the 2×2 formula we just proved, this yields

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg) = aei + bfg + cdh - ceg - bdi - afh. \qquad \square$$

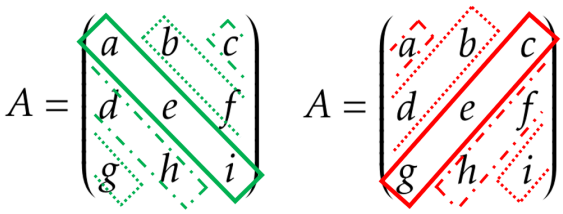


Figure 8: Computing a 3×3 determinant.

These two results are typically enough for most of our applications, and if not, they at least allow us to break the determinant down to 3×3 rather than 1×1 matrices using the Laplace expansion method, for which we can directly compute the determinants. Visually, for the 3×3 determinant, you can remember to add the product of all right-diagonals and subtract the one all left-diagonals (see Figure 8).

Equipped with these rules, two comments on the Laplace method deem worthwhile. First, when we have a row or column containing only one non-zero entry, we can reduce the dimension of determinant computation avoiding a sum: consider the lower-triangular matrix

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

The Laplace-expansion for the first *column* is $\det(A) = \sum_{i=1}^{4} (-1)^{i+1} a_{i1} \det(A_{-i1})$. However, for any $i \neq 1$, $a_{i1} = 0$, so that the expression reduces to $\det(A) = (-1)^2 a_{11} \det(A_{-11}) = 5 \det(A_{-11})$. Applying the 3×3-rule, it results that $\det(A) = 5 \cdot (-8) = -40$. Second, for triangular matrices generally, the determinant is given by the *trace*.

Definition 49. ***(Trace of a Square Matrix)*** *For a matrix $A \in \mathbb{R}^{n \times n}$, the trace $tr(A)$ is given by the product of diagonal elements, i.e. $tr(A) = \Pi_{i=1}^{n} a_{ii}$.*

Proposition 14. ***(Determinant of a Triangular Matrix)*** *The determinant of an upper or upper triangular matrix $A \in \mathbb{R}^{n \times n}$ is given by its trace, i.e.* $\det(A) = \mathrm{tr}(A)$.

Proof. The proposition follows by just iterating on what we have done in the example above

for a general matrix. Consider the upper triangular matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}.$$

Expanding iteratively by the first column, it follows that

$$\det(A) = a_{11} \det \begin{pmatrix} a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & a_{32} & \cdots & a_{3n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix} = a_{11} a_{22} \det \begin{pmatrix} a_{33} & a_{34} & \cdots & a_{3n} \\ 0 & a_{43} & \cdots & a_{4n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix} = \ldots = \Pi_{i=1}^{n} a_{ii} = \operatorname{tr}(A).$$

For the lower-triangular matrix, the procedure is analogous, here, we just have to expand for the first *row* instead of column iteratively. □

The two key take-aways are that when you have to compute the determinant of big matrices by hand, look for rows and/or columns with many zeros to apply the Laplace-expansion, or if you're lucky and face a lower or upper triangular matrix, you can directly compute the determinant from the diagonal.[12] The latter is nice especially because the elementary matrix operations introduced above affect the determinant in a tractable way, so that it is possible to avoid Laplace altogether:

Theorem 26. ***(Determinant and Elementary Operations)*** *Let $A \in \mathbb{R}^{n \times n}$ and $\tilde{A}$ the resulting matrix for the respective elementary operation. Then,*

(i) for operation (E1) (interchange of two rows), we have $\det(\tilde{A}) = -\det(A)$, i.e. the interchange of rows changes the sign of the determinant,

(ii) for operation (E2) (row multiplication with a scalar $\lambda \neq 0$), $\det(\tilde{A}) = \lambda \det(A)$,

(iii) for operation (E3) (addition of multiple of row to another row), $\det(\tilde{A}) = \det(A)$, i.e. (E3) does not change the determinant.

Proof. Parts *(ii)* and *(iii)* of this proposition are indeed a corollary of the proof of a later proposition, see Footnote 16. For part *(i)*, the intuition is to apply the Laplace expansion iteratively to E^1 of the referenced proof, and because every column and row contains only one non-zero element, you will reduce dimensionality in every step without summation and a factor $a_{ij} = 1$. Doing so, you face an expansion where $i + j$ is odd only once, namely at the column/row indicating the exchange of rows. The formal documentation of this procedure is left out here; feel free to write it down yourself for practice. □

Thus, if we properly document the operations that we apply to triangularize A, we can *directly obtain the determinant from the trace of the resulting upper triangular matrix*. You can easily check that especially, a non-zero determinant of A is equivalent to a non-zero determinant or trace of

[12] Note that this is of course especially true for diagonal matrices, since they are both upper and upper triangular.

the resulting triangular matrix. Thus, **the matrix A is invertible if and only if any associated triangular matrix has no zeros on the diagonal**. A formal explanation of what "associated" means can be found below.

Another important fact that will be helpful frequently is the following:

Theorem 27. ***(Determinant of the Product)*** *Let $A, B \in \mathbb{R}^{n\times n}$. Then,* $\det(AB) = \det(A)\det(B)$.

Note that in contrast to the product, for the sum, it does **not** hold in general that $\det(A+B) = \det(A) + \det(B)$. Now that we now how to compute a determinant, we care about its role in existence and determination of inverse matrices. As already stated above, the rule we will rely on is inherently simple:

Theorem 28. ***(Determinant and Invertibility)*** *Let $A \in \mathbb{R}^{n\times n}$. Then, A is invertible if and only if* $\det(A) \neq 0$.

The "only if" part is rather simple: suppose that A is invertible. Note that $\det(\mathbf{I_n}) = 1$ by Proposition 14. Then,

$$1 = \det(\mathbf{I_n}) = \det(AA^{-1}) = \det(A)\det(A^{-1}).$$

Therefore, $\det(A) \neq 0$. Moreover, this equation immediately establishes the following corollary:

Corollary 2. ***(Determinant of the Inverse Matrix)*** *Let $A \in \mathbb{R}^{n\times n}$ and suppose that A is invertible. Then,* $\det(A^{-1}) = 1/\det(A)$.

For the "if" part of Theorem 28, since we have not yet formally discussed matrix inversion and invertibility criteria (beyond the one stated in the theorem), we confine ourselves the intuition for the 2×2 case here. Consider the system of equations

$$\begin{aligned} ax_1 + bx_2 &= y_1 \\ cx_1 + dx_2 &= y_2 \end{aligned}$$

or, in matrix notation

$$Ax = y \quad \text{where} \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},\ x = (x_1, x_2)' \text{ and } y = (y_1, y_2)'.$$

One can solve this system multiplying the first equation by d, the second by $-b$, and adding the two, which yields:

$$(ad - cb)x_1 = dy_1 - by_2$$

Alternatively, multiplying the first equation by $-c$, the second by a, and adding, one gets:

$$(ad - cb)x_2 = ay_2 - cy_1$$

Thus, if the quantity $(ad - cb)$ is different from zero, these equations uniquely determine the value of the unknowns x and y, which, as we have argued earlier, implies invertibility of A. Recalling our 2×2 determinant formula, the invertibility condition $ad - cb \neq 0$ requires that $\det(A) \neq 0$! Indeed, this intuition can be generalized to any $n \times n$ system.

The key take-away is that invertibility is equivalent to a non-zero determinant. Consequently, because invertibility implies **unique existence of the solution**, so does a **non-zero determinant**. While the general determinant concept is a little notation-intensive, computation is easy for smaller matrices. Thus, the determinant criterion represents the **most common invertibility check** for "small" matrices or respectively, the most common unique solution check for "small" systems of linear equations when we have as many unknowns as equations.

2.3.3 Rank of a Matrix

Clearly, we don't always find ourselves in the comfortable situation that we have as many equations as unknowns, which motivates looking at different invertibility criteria. The next concept is closely linked to our discussion of linear dependence in Chapter 1. Let's again consider the general case of a system of m equations in n unknowns, i.e. equation (4) with $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ and potentially $m \neq n$. When writing A in column notation as $A = (a_1, a_2, \dots, a_n)$, $a_i \in \mathbb{R}^m$ for all i, it is straightforward to check that $Ax = \sum_{i=1}^{n} x_i a_i$.[13] Thus, the LHS of the system is nothing but a *linear combination of the columns of A with coefficients* $x = (x_1, \dots, x_n)'$! Therefore, the problem of solving the system of linear equations can be rephrased as as looking for the linear combination coefficients for the columns of our coefficient matrix A that yields the vector b! Let's see this abstract characterization in a practical example.

Consider the following system:

$$\begin{aligned} x_1 + 2x_2 &= 2 \\ x_1 - x_2 &= 0 \end{aligned}$$

with associated matrix form

$$Ax = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix} = b \quad \Leftrightarrow \quad \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \cdot x_1 + 2 \cdot x_2 \\ 1 \cdot x_1 + (-1) \cdot x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + x_2 \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

Recall that the linear combination coefficients can be viewed as combination magnitudes (or weights) of the vectors $a_1, \dots, a_n$. Thus, less formally, we are looking for the distance we need to go in every direction indicated by a column of A to arrive at the point b. Figure 9 shows how this can be illustrated geometrically.

Feel encouraged to repeat this graphical exercise for other points $b \in \mathbb{R}^2$; you should always arrive at a unique solution for the linear combination coefficients. The fundamental reason is that the columns of A, $(1,1)'$ and $(2,-1)'$ are linearly independent.[14] Think a second about what it would mean geometrically if they were linearly dependent before continuing. Done? Good! In case of linear dependence, the points lie on the *same* line through the origin, and either, b does not lie on this line and we never arrive there, i.e. there are no solutions, or it does, and an infinite number of combinations of the vectors can be used to arrive at b.

Before moving to the formal rank concept, let us make ourselves familiar the case with

[13] Try and verify it for the 3×3 case.

[14] Recall that this implies that they form a basis of the $\mathbb{R}^2$.

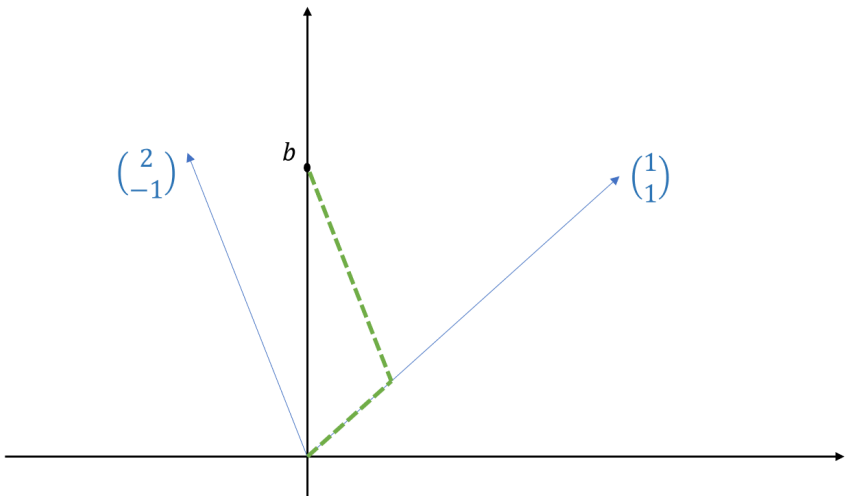


Figure 9: b as a linear combination of the columns.

more unknowns than equations explicitly: consider

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 1 & 4 & -3 \\ 2 & p & q \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + x_2 \begin{pmatrix} 4 \\ p \end{pmatrix} + x_3 \begin{pmatrix} -3 \\ q \end{pmatrix}$$

Now, we distinguish three scenarios: (i) $p = 8$ and $q = -6$, (ii) $p = 8$ or $q = -6$ but not both, or $p = -4/3q$ but $p \neq 8$, (iii) $p \neq -4/3q$, $p \neq 8$ and $q \neq -6$. In the first scenario, ALL column vectors of the matrix A lie on the same line, and despite there being more unknowns than equations, it may be that we have no solution at all, namely whenever b does not lie on this line. Otherwise, we have more than one direction vector and any point $b \in \mathbb{R}^2$ is accessible using linear combinations of the column vectors of A! Because there are more direction vectors than b has elements, there will generally be infinitely many solutions, and we can express the solutions in terms of one "free" variable, e.g. $x_1(x_3)$ and $x_2(x_3)$: When setting any $x_3 \in \mathbb{R}$, we can re-phrase the problem as moving from b to $x_3 a_3$, i.e. we may write $x_1 a_1 + x_2 a_2 = b - x_3 a_3$. However, in the second scenario, we are constrained by the fact that two columns (say 1 and 2) are linearly dependent, so that the free variable can not be the third one (say 3)! Otherwise, the LHS of the re-written problem is again represented by a line. In scenario 3, we can freely choose any of the variables. Keep in mind that the system with more equations than unknowns has (i) no solution if we cannot reach b using linear combinations of the columns of A and infinitely many solutions otherwise, i.e. there never is a unique solution!

It is time to formalize the idea of the "number of directions" that are reached by a matrix and whether "a point b can be reached combining the columns of A".

Definition 50. ***(Column Space of a Matrix)*** *Let $A \in \mathbb{R}^{m \times n}$ with columns $a_1, \ldots, a_n \in \mathbb{R}^m$, i.e. $A = (a_1, \ldots, a_n)$. Then, the column space $\mathrm{Co}(A)$ of A is the span of these columns, i.e. $\mathrm{Co}(A) = \mathrm{Span}(a_1, \ldots, a_n) = \{x \in \mathbb{R}^m : (\exists \lambda_1, \ldots, \lambda_m : x = \sum_{j=1}^m \lambda_j a_j)\}$ (equipped with matrix addition and scalar multiplication).*

Analogously, we define the row space as the space spanned by the rows of A.

Definition 51. ***(Rank of a Matrix)*** *Let $A \in \mathbb{R}^{n \times m}$. Then, the column (row) rank of A is the dimen-*

sion of the column (row) space of A. It coincides with the number of linearly independent columns (rows) of A and is denoted as $\mathrm{rk}(A)$, $\mathrm{rank}(A)$ *or* $\mathrm{rk}\, A$.

Recall that the dimension of a set's span was given by the number of linearly independent vectors in the set. This immediately gives the characterization of the rank given in the definition.[15] You may wonder why we use the same notation for the column and row rank, respectively. This is due to the following fact:

Theorem 29. ***(Column Rank = Row Rank)*** *Let $A \in \mathbb{R}^{n\times m}$. Then, the column rank and the row rank of A coincide.*

See e.g. https://en.wikipedia.org/wiki/Rank_(linear_algebra)#Proofs_that_column_rank_=_row_rank for ways to prove this theorem. Thus, "the rank" of A, $\mathrm{rk}(A)$ is a well-defined object, and it does not matter whether we compute it from the columns or rows.

Definition 52. ***(Full Rank)*** *Consider a matrix $A \in \mathbb{R}^{n\times m}$. Then, A has full row rank if* $\mathrm{rk}(A) = n$ *and A has full column rank if* $\mathrm{rk}(A) = m$. *If the matrix is square, A has full rank if* $\mathrm{rk}(A) = n = m$.

Corollary 3. ***(A Bound for the Rank)*** *Let $A \in \mathbb{R}^{n\times m}$. Then,* $\mathrm{rk}(A) \leq \min\{n, m\}$.

This follows immediately from Theorem 29 since at most n rows and m columns can be linearly independent.

From the definitions above, you should be able to observe that the rank captures the number of directions into which we can move using the columns of A, and that the column space is the set of all points we can reach using linear combinations of A's columns. Consequently, a solution will exist whenever $b \in \mathrm{Span}(A)$; a sufficient condition is that $\mathrm{Span}(A) = \mathbb{R}^n$ or respectively $\mathrm{rk}(A) = n$. Now, how does the rank help in solving systems of linear equations? Two relationships are particularly useful here:

Theorem 30. ***(Rank and Elementary Operations)*** *Let $A \in \mathbb{R}^{n\times m}$. Then,* $\mathrm{rk}(A)$ *is invariant to elementary operations, i.e. for any $\tilde{A}$ associated with operations (E1) to (E3),* $\mathrm{rk}(\tilde{A}) = \mathrm{rk}(A)$.

The proof is omitted here as it is not too insightful.

Theorem 31. ***(Rank Condition)*** *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then, the system $Ax = b$ has a unique solution if and only if $b \in \mathrm{Co}(A)$ and* $\mathrm{rk}(A) = n$.

Taken together with Corollary 3, Theorem 31 yields the general result that we have already seen in the example: if there are more unknowns than equations, that is, $n > m \geq \mathrm{rk}(A)$, then a unique solution cannot exist. To establish Theorem 31 (and also beyond this purpose), the following result for elementary operations is extremely helpful.

Proposition 15. ***(Elementary Operations and the Set of Solutions)*** *Consider a system $Ax = b$ of linear equations. Then, for an elementary operation characterized $\tilde{A} = EA$ with operation matrix E, the system $\tilde{A}x = \tilde{b}$ with $\tilde{b} = Eb$ is equivalent to $Ax = b$ in terms of the solutions x.*

[15] Formally, it is slightly sloppy to write a characterizing statement (rather than a defining one) into the definition. It was done here because what you should take away here is precisely that the rank is the number of linearly independent columns, and the rank definition via the column space is actually of subordinate importance.

Proof. The proposition sounds rather abstract and seems to fall from the sky. However, the intuition is fascinatingly simple. The key is, as already discussed earlier, that we can express elementary operations using a *square* matrix $E \in \mathbb{R}^{m\times m}$ and *left-multiply* it to A. Then, if E is invertible,

$$\tilde{A}x = \tilde{b} \quad \Leftrightarrow \quad EAx = Eb \quad \Rightarrow \quad E^{-1}EAx = E^{-1}Eb \quad \Leftrightarrow \quad Ax = b,$$

and the two equation systems are equivalent because $Ax = b$ clearly also implies $\tilde{A}x = \tilde{b}$ (just left-multiply E). Everything that needs to be established now is that the elementary operations can be represented by left-multiplication of *invertible* matrices to A. For what follows, recall the definitions of the matrices representing the elementary operations given in Proposition 11. We will rely on the determinant criterion and show invertibility by establishing that the operation matrices E^1, E^2 and E^3 have a non-zero determinant.

First, consider E^1 as the representation of the exchange of rows i and j. Note that exchanging rows i and j in E^1 gives $\mathbf{I_m}$, for which $\det(\mathbf{I_m}) = 1$! Recall that the exchange of rows implies multiplication of the determinant by -1; thus, $\det(E^1) = -\det(\mathbf{I_m}) = -1 \neq 0$.

Next, consider E^2 as the representation of multiplication of row i with $\lambda \neq 0$. Note that E^2 is diagonal and thus especially upper triangular. By Proposition 14, $\det(E^2) = \mathrm{tr}(E^2) = 1^{n-1}\lambda \neq 0$.

Finally, consider E^3 as the representation of addition of a multiple $\lambda \in \mathbb{R}$ of row j to row i. Note that E^3 is either lower or upper triangular, and that a lower triangular matrix can be brought to upper triangular form by inverting the order of the rows (i.e., repeated column swaps). Applying again Proposition 14, $\det(E^3) = \mathrm{tr}(E^3) = 1 \neq 0$.[16] □

As a side note, exchange of columns is achieved via *right-multiplication* of a matrix similar to E^1, and equivalence of the systems is not readily established. Thus, column swaps are not an elementary operation in the narrow sense, and don't generally preserve solutions![17] Note also that if the operations $E_1, \ldots, E_k$ have been used (again in order of the index) to transform A into the triangular matrix $\tilde{A}$, then $\tilde{A} = E_k \cdot E_{k-1} \cdot \ldots \cdot E_1 A$ and $\tilde{b} = E_k \cdot E_{k-1} \cdot \ldots \cdot E_1 b$. Thus, $\tilde{b}$ is an *invertible transformation* of b!

Now, we are ready to establish Theorem 31. The proof is given below to demonstrate that with our arguably limited toolbox which we have equipped ourselves with in this chapter is indeed sufficient to prove a powerful result such as Theorem 31. Moreover, in retracing the steps of proof, what becomes clear is the intuition of thinking about the rank as the *information content* of a system of linear equations or respectively, the number of equations that truly provide *independent* information and may therefore *not* be expressed using a (linear) combination of other equations. Thus, read the following lines partly as proof of Theorem 31, and partly as a verbal description of its interrelationship with Proposition 15.

The following is admittedly rather technical and lengthy. Try to superficially go over at least the intuition and Proposition 16, it is left to your motivation and time budget to decide how intensively you look at the rest. Another bold note will indicate where you should start to read more thoroughly again.

[16] Using the matrices E^2 and E^3, by the determinant product rule of Theorem 27, you can easily verify the respective determinant rules in Proposition 26!

[17] Of course, a column change is nothing but a re-labeling of variables, such that this is more of a technical comment rather than an actual issue.

Intuition. Suppose first that the system is *under-identified*, that is, that it has less equations than unknowns ($m < n$). Then, let $\tilde{A} = (\tilde{A}_T, X)$ be a matrix with upper triangular block A_T and an arbitrary block X that we arrive at by performing elementary operations on A.[18]

Then, let's add $n - m$ rows that just say $0 = 0$ (or more precisely, $\sum_{i=1}^{n} 0 \cdot x_i = 0$). This expands $\tilde{A}$ to the square matrix $\tilde{A}^* = \begin{pmatrix} A_T & X \\ \mathbf{0}_{n-m\times m} & \mathbf{0}_{n-m\times n-m} \end{pmatrix}$, and expand $\tilde{b}$ to $\tilde{b}^*$ by adding zeros: $\tilde{b}^* = \begin{pmatrix} \tilde{b} \\ \mathbf{0}_{n-m\times 1} \end{pmatrix}$. Then, the system $Ax = b$ is equivalent to $\tilde{A}x = \tilde{b}$ by Proposition 15, which in turn is equivalent to $\tilde{A}^*x = \tilde{b}^*$ because clearly, adding the "0 = 0" equations does not affect the solution(s) for x. However, this system is square, and by Proposition 14, $\det(\tilde{A}^*) = \text{tr}(\tilde{A}^*) = 0$ because at least one diagonal element of $\tilde{A}^*$ is equal to zero. Thus, by the determinant invertibility condition, $m < n$ is equivalent to non-existence of the unique solution.

Intuitively, if $m < n$, A does not contain enough information to uniquely pin down x. In the geometric interpretation discussed above, we will always have at least one column vector that captures a direction already spanned by the others, so that if there are solutions (that is, if $b \in \text{Co}(A)$), there will be infinitely many. This is because some of the x-coefficients necessarily augment the same direction, and thus, knowledge of the linear combination b is insufficiently informative about the individual values of a specific solution x.

On the other hand, for $m > n$, i.e. more equations than unknowns (where we say that the system is *over-identified*), we generally face a converse issue – the equations may impose too many restrictions on the variables x so that we may not have a solution: Here, $\text{rk}(A) = \dim(\text{Co}(A)) \leq n < m$, and $m - \dim(\text{Co}(A))$ directions of $\mathbb{R}^m$ are not reached. Thus, the vector $b \in \mathbb{R}^m$ may not be expressed as a linear combination of columns of A unless it satisfies a specific set of restrictions ensuring $b \in \text{Co}(A)$. While this sounds rather abstract, the intuition is very simple. Start from the initial system $Ax = b$. To facilitate the understanding, consider the concrete example

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ p \end{pmatrix}$$

where p is an unknown parameter. Because the rows of A have two entries but there are three rows, we know that not all rows can be linearly independent. Indeed, it is easily verified that

$$2 \cdot \begin{pmatrix} 1 \\ 2 \end{pmatrix} + (-7) \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} + (-1) \cdot \begin{pmatrix} 2 \\ -3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

i.e. there exist coefficients $\lambda_1 = 2$, $\lambda_2 = -7$ and $\lambda_3 = -1$ for which at least one $\lambda_j \neq 0$ and $\sum_{i=1}^{3} \lambda_i a_i = 0$. This proves linear dependence by establishing the contrapositive of Theorem 8. Thus, we can express one row as a linear combination of the other, e.g. $(2, -3) = 2 \cdot (1, 2) + (-7) \cdot (0, 1)$. The crucial aspect is what this means for our equation system: the first two rows already give information on $2x_1 - 3x_2$: when multiply the first row by 2, and subsequently subtract (-7)

[18] Recall Theorem 24 which tells us that the transformation $\tilde{A}$ always exists!

times the second row from it. Then, the equation system becomes

$$\begin{pmatrix} 2 & -3 \\ 0 & 1 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -8 \\ 2 \\ p \end{pmatrix} \tag{5}$$

Now, we have two cases: $p = -8$, where the third row is redundant because the information is also contained in the system when leaving it out, and $p \neq -8$, where the row provides information that conflicts with the remaining system. This reasoning applies more generally: with $m > n$ rows, at least $m-n$ rows are linear combinations of others, and the information contained in them is either redundant or inconsistent with the remaining system.

You will have picked up that whether the linearly dependent rows of A provide redundant or inconsistent information in the system depends on the vector b rather than the rows themselves. Indeed, whenever there is inconsistent information, $b \notin \mathrm{Co}(A)$ because then, we can not solve the system for x, i.e. we are unable to find a linear combination with coefficients $x_1, \ldots, x_n$ that reaches b. On the other hand, if $b \in \mathrm{Co}(A)$, all the linearly dependent equations are redundant. Let's investigate this case to see if and when we can find a *unique* solution.

Start again with the example above: the additional column is redundant if $p = -8$. Suppose this is the case, and subtract row 1 from row 3 in equation (5). Then, we get

$$\begin{pmatrix} 2 & -3 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -8 \\ 2 \\ 0 \end{pmatrix}.$$

Multiplying out the last row simply gives $0 = 0$, and the *redundancy of the third equation allows to eliminate the third row*. So long as there are no inconsistent rows and there are more rows than columns, we can continue to eliminate redundant information until we arrive at a square system! Let's express this a bit more formally for the general system $Ax = b$. Note that by Theorem 24 we may use the elementary operations to re-express $Ax = b$ as

$$\tilde{A}x = \tilde{b} \quad \text{with} \quad \tilde{A} = \begin{pmatrix} A_T \\ \mathbf{0}_{m-n \times n} \end{pmatrix}$$

where $A_T \in \mathbb{R}^{n \times n}$ is a upper triangular matrix (possibly containing zeros on the diagonal). Now, if $\tilde{b}$ has non-zero entries below the index n, there clearly is no solution for x, because the information system provides *conflicting information*. Otherwise, if $\tilde{b} = \begin{pmatrix} \tilde{b}_n \\ \mathbf{0}_{m-n \times 1} \end{pmatrix}$, multiplying out the last rows yields $0 = 0$, which is true, but contains no information of use in pinning down x. Thus, the system does not contain inconsistent but rather *redundant* information, which we can *eliminate* by excluding these last $m - n$ columns, and instead consider the new system $A_T x = b_n$ with *square* matrix A_T.

Thus, if a solution may exist because there is no conflicting information, eliminating redundant information yields a square equation system for which we know how to proceed: we can check the determinant of the coefficient matrix! Because A_T is upper triangular, we can find a

unique solution by inverting A_T if and only if $\det(A_T) = \text{tr}(A_T) \neq 0$, i.e. all diagonal elements of A_T are non-zero. This is equivalent to A_T having "full" rank $\text{rk}(A_T) = m$:

Proposition 16. ***(Rank of a Upper Triangular Matrix with strictly Non-zero Diagonal)*** *Consider the upper triangular matrix* $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & 0 & a_{nn} \end{pmatrix}$. *Then, all columns of A are linearly independent and* $rk(A) = n$ *if and only if* $\forall i \in \{1,\dots,n\} : a_{ii} \neq 0$.

Proof. "$\Leftarrow$ (if)" Suppose that $\forall i \in \{1,\dots,n\} : a_{ii} \neq 0$. We know that if A has n linearly independent columns, then $\text{rk}(A) = n$. Thus, it suffices that the columns $a_1,\dots,a_n$ of A are linearly independent. We apply our linear independence check of Theorem 8 to verify this: Suppose that $\sum_{i=1}^{n} \lambda_i a_i = \mathbf{0}$. From the last column, $\lambda_n a_{nn} = 0$, which implies $\lambda_n = 0$ because $a_{nn} \neq 0$ by assumption. This establishes our inductive base.[19] Suppose now that $\lambda_j = 0$ for all $j \in \{p+1, p+2, \dots n\}$. Then, from the p-th row of $\sum_{i=1}^{n} \lambda_i a_i = \mathbf{0}$, because of the triangular structure of A, we have that $0 = \sum_{i=p}^{n} \lambda_i a_{pi} = \lambda_p a_{pp}$ where the last equality follows from the inductive base. Thus, because $a_{pp} \neq 0$ by assumption, $\lambda_{pp} = 0$. This establishes the inductive step. By the principle of induction, $\lambda_j = 0$ holds for $j = n$, and thus for $j = n-1$, and thus for $n-2$, ..., and thus for $j = 1$, i.e. $\forall j \in \{1,\dots,n\} : \lambda_j = 0$. In consequence, by Theorem 8, the set $\{a_1,\dots,a_n\}$ is independent.
"$\Rightarrow$ (only if)" Here, we use the contrapositive method: suppose that $\neg(\forall i \in \{1,\dots,n\} : a_{ii} \neq 0)$, i.e. $\exists i \in \{1,\dots,n\} : a_{ii} = 0$. Note that if $\{a_1,\dots,a_i\}$ is not linearly independent, then especially $\{a_1,\dots,a_n\}$ is not, because we can find a linear combination of $a_1,\dots,a_n$ with coefficients $\lambda_j = 0$ for $j > i$ but $\lambda_j \neq 0$ for $1 \leq j \leq i$ so that $\sum_{k=1}^{n} \lambda_k a_k = 0$. Thus, it suffices to show that $\{a_1,\dots,a_i\}$ is not a linearly independent set. Note that because A is upper triangular, the last element of any a_j, $1 \leq j \leq i$ is zero, so that we can equivalently consider the set $\{\tilde{a}_1,\dots,\tilde{a}_i\}$ where $\tilde{a}_j$ eliminates the last element of a_j, i.e. $a_j = \begin{pmatrix} \tilde{a}_j \\ 0 \end{pmatrix}$ for $1 \leq j \leq i$. Because $\tilde{a}_j \in \mathbb{R}^{i-1}$, there can be at most $i-1$ linearly independent elements in $\{\tilde{a}_1,\dots,\tilde{a}_i\}$ and thus $\{a_1,\dots,a_i\}$, so that at $\{a_1,\dots,a_i\}$ is not a linearly independent set. In consequence, $\{a_1,\dots,a_n\}$ contains less than n independent vectors, so that $\text{rk}(A) = \dim(\text{Span}(\{a_1,\dots,a_n\} < n$. Thus, we have shown that

$$\neg(\forall i \in \{1,\dots,n\} : a_{ii} \neq 0) \Rightarrow \neg(rk(A) = n)$$

Which, by the contrapositive method ("negation flips direction of implication"), yields

$$rk(A) = n \Rightarrow \forall i \in \{1,\dots,n\} : a_{ii} \neq 0.$$ □

Showing that an upper triangular matrix has all non-zero diagonal elements if and only if it has full rank n is fully analogous! Make sure that you understand why.

[19] Remember the concept of proof by induction? Here comes another application, with the twist that we are now considering *decending* numbers! This works here because there is a biggest element.

Corollary 4. ***(Invertibility, Rank and Determinant of Square Matrices)*** *A square matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if its determinant is nonzero, if and only if its rank is equal to n, which is the case if and only if any associated triangular matrix has only non-zero diagonal entries.*

The first equivalence just re-states the respective result for the determinant. The reason why the rest follows is that the elementary operations do not alter the rank, and we can transform any square matrix into a upper triangular matrix using only these operations. Moreover, recall that the elementary operations do not affect invertibility as $\det(\cdot) \neq 0$ was maintained under all of them. Finally, recall that the determinant of a diagonal matrix is equal to the trace, which is non-zero if and only if all diagonal elements are non-zero or equivalently by the above proposition, if and only if it has full rank, i.e. its rank is equal to n.

This allows to conclude our investigation of the case $m \geq n$: If $m > n$, we necessarily have redundant information, and eliminating it yields a system $A_T x = \tilde{b}_n$. If $m = n$, the system $Ax = b$ is already characterized by a square matrix A. Here, we can apply Proposition 15 to bring the system in the upper triangular form. Either way, there exists a unique solution if and only if in the triangular system, there is no more redundant information, i.e. $\mathrm{rk}(A_T) = n$ or equivalently, $\det(A_T) \neq 0 \Leftrightarrow A_T$ has only non-zero diagonal entries. Note that $\mathrm{rk}(A_T) = n$ is equivalent to $\mathrm{Co}(A_T) = \mathbb{R}^n$, and thus, we certainly have that $b \in \mathrm{Co}(A_T)$.

After extensively having discussed the intuition, let's use our insight to give a brief formal proof of the rank condition stated above.

Proof of Theorem 31. "$\Rightarrow$" First, suppose that $Ax = b$ has a unique solution. Clearly, $n \geq m$, because we established that when $n < m$, no unique solution can exist. Because if $b \notin \mathrm{Co}(A)$ implies that there is no solution, it must be the case that $b \in \mathrm{Co}(A)$. Thus, we can reduce our system to $A_T x = \tilde{b}_n$ as above, and the unique solution exists if and only if A_T is invertible, or equivalently, has a non-zero determinant. By Proposition 16, this is equivalent to $\mathrm{rk}(A) = n$. To conclude, $b \in \mathrm{Co}(A)$ and $\mathrm{rk}(A) = n$.

"$\Leftarrow$" Now, suppose that $b \in \mathrm{Co}(A)$ and $\mathrm{rk}(A) = n$. Clearly, this implies that $m \geq n$, i.e. there are more equations than unknowns. Since $b \in \mathrm{Co}(A)$, there are only redundant but no inconsistent excess equations. Since the rank is unaffected by elementary operations, for $\tilde{A} = \begin{pmatrix} A_T \\ \mathbf{0}_{m-n\times n} \end{pmatrix}$, $\mathrm{rk}(\tilde{A}) = \mathrm{rk}(A)$. Because the zero columns are linearly dependent of the remaining columns, it follows that $\mathrm{rk}(A_T) = \mathrm{rk}(\tilde{A}) = n$. Thus, A_T is invertible and the system has a unique solution. □

If you have skipped the technical details, continue reading here, and have a look at the result in Corollary 4, which may give additional insight.

Let's briefly summarize the key insights from this rather lengthy investigation. As we have already established earlier, unique solutions exist in "square systems" with an invertible (square) matrix A. Now, we have seen more generally that **unless the system can be reduced to a square system with invertible matrix A, there can be no unique solution!** This follows from isolated consideration of the multiple cases of general equation systems: If we have **less equations than unknowns**, we can **never** have **enough information** to determine a unique solution, but it can still be that no solution exists at all, namely when the system contains contradictory information. Conversely, with **more equations than unknowns**, some of the LHS

expressions in the system $Ax = b$ will be linearly dependent, and the **linearly dependent expressions either contain contradictory or redundant information.** If they are contradictory, there is no solution, otherwise we can throw them out without altering the solution. **Without contradictory information**, we can always **reduce the system to a square system** that handled more easily, using the determinant criterion. We can **determine whether the system contains contradictory information by transforming it to generalized upper triangular form** and checking whether b contains non-zero entries below the column dimension n. For **square $n \times n$ matrices**, we have argued that **full rank** of A, i.e. $\text{rk}(A) = n$, **is equivalent to invertibility**, which in turn is equivalent to a non-zero determinant.

Thus, when in search of a unique solution as we frequently are in economics when looking for an equilibrium allocation or an estimator that optimizes a certain criterion, it is justified to look at square systems! Consequently, the following again restricts attention to such systems.

2.3.4 Eigenvalues, Eigenvectors and Definiteness of a Matrix

The concepts of eigenvalues and -vectors and matrix definiteness have a purpose far beyond the context of invertibility, and you will come across them frequently throughout your master studies. Their introduction here, however, restricts attention to their use for determining invertibility. Before getting started, as with the determinant, it is worthwhile to note that **only square matrices have eigenvalues and definiteness!** Thus, the only concept of this section that applies to more general matrices is the rank!

Definition 53. ***(Eigenvectors and -values)*** *Consider a **square** matrix $A \in \mathbb{R}^{n \times n}$. Then, $x \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is said to be an eigenvector of A if there exists a $\lambda \in \mathbb{R}$ such that $Ax = \lambda x$. λ is then called an eigenvalue of A.*

To practice again some quantifier notation, note that the definition of an eigenvalue here is equivalent to stating that $\lambda \in \mathbb{R}$ is an eigenvalue of $A \in \mathbb{R}^{n \times n}$ if $\exists x \in \mathbb{R}^n \backslash \{\mathbf{0}\} : (Ax = \lambda x)$. Moving away from technicalities, let's think about intuitively what it means if $Ax = \lambda x$: clearly, Ax and x are linearly dependent: $\lambda x + (-1)Ax = 0$, and thus, x and Ax lie on the *same line* through the origin! Note that if $x = \mathbf{0}$, then trivially, for any $\lambda \in \mathbb{R}$ it holds that $\lambda x = Ax = \mathbf{0}$, so that any $\lambda \in \mathbb{R}$ would constitute an eigenvalue and make the definition meaningless. This is why we require that $x \neq \mathbf{0}$. On the other hand, 0 can indeed be an eigenvalue, namely if there exists $x \neq \mathbf{0}$ so that $Ax = \mathbf{0}$, i.e. Ax is equal to the origin. The discussion to follow is interested in answering two questions: (i) how can we find eigenvalues (and associated eigenvectors)? and (ii) what do the eigenvalues tell us about invertibility of the matrix?

Starting with (i), we can re-write the search for an eigenvector x of A for an eigenvalue candidate $\lambda \in \mathbb{R}$ as a special system of linear equations: If x is an eigenvector of A for λ,

$$Ax = \lambda x = \lambda \cdot (\mathbf{I_n} x) \quad \Leftrightarrow \quad \mathbf{0} = Ax - \lambda \mathbf{I_n} x = (A - \lambda \mathbf{I_n}) x.$$

Thus, we have a *square* system of equations $C_\lambda x = b$ with coefficient matrix $C_\lambda = A - \lambda \mathbf{I_n}$ and solution vector $b = \mathbf{0}$. Now, how does this help? Note that if there is an eigenvector x of λ, it is not unique: if $Ax = \lambda x$, for any $c \neq 0$, $A(cx) = \lambda(cx)$, and $\tilde{x} = cx$ is also an eigenvector of A associated with λ! Thus, we are looking precisely for the situation where the square system

does *not* have a unique solution, i.e. where $\det(C_\lambda) = 0$ and C_λ is not invertible! This suggests that we can find the eigenvalues of A by solving

$$P(\lambda) = \det(A - \lambda \mathbf{I_n}) = 0,$$

i.e. by setting the *characteristic polynomial* of A to zero, or respectively, by finding its *roots*. To become more familiar with this method, let's consider an example. Let $A = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix}$. Then,

$$\begin{aligned} P(\lambda) = \det(A - \lambda I) = \det\left(\begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right) = \det\begin{pmatrix} 3-\lambda & 2 \\ 1 & 2-\lambda \end{pmatrix} \\ = (3-\lambda)(2-\lambda) - 2 \cdot 1 = 6 - 2\lambda - 3\lambda + \lambda^2 - 2 = \lambda^2 - 5\lambda + 4. \end{aligned}$$

Solving $P(\lambda) = 0$ can be done with the p-q-formula:

$$P(\lambda) = 0 \quad \Leftrightarrow \quad \lambda \in \left\{ -\frac{-5}{2} \pm \sqrt{\left(\frac{-5}{2}\right)^2 - 4} \right\} = \left\{ \frac{5}{2} \pm \sqrt{\frac{25-16}{4}} \right\} = \left\{ \frac{5}{2} \pm \frac{3}{2} \right\} = \{1, 4\}.$$

Consequently, our eigenvalue candidates are $\lambda_1 = 1$ and $\lambda_2 = 4$. To find the eigenvectors, we have to solve the equation system: for $\lambda_1 = 1$, $C_1 = A - 1 \cdot \mathbf{I_n} = \begin{pmatrix} 3-1 & 2 \\ 1 & 2-1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$. Clearly, you can see that this matrix does *not* have full rank and thus a multitude of solutions. $C_1 x = \mathbf{0}$ is equivalent to $x_1 + x_2 = 0$ or respectively, $x_1 = -x_2$. Thus, the eigenvectors of $\lambda_1 = 1$ are multiples of $(1, -1)'$. The set of all these vectors is $\left\{ c \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} : c \in \mathbb{R} \right\} = Span(\{(1,-1)'\})$ is the so-called *eigenspace* of $\lambda_1 = 1$. For $\lambda = 4, C_4 = A - 4 \cdot \mathbf{I_n} = \begin{pmatrix} 3-4 & 2 \\ 1 & 2-4 \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}$, the eigenvectors are multiples of $(2, 1)'$ and the eigenspace of $\lambda_1 = 4$ is $\left\{ c \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} : c \in \mathbb{R} \right\}$. Note that an eigenvalue may occur "more than once" and generally be associated with multiple linearly independent eigenvectors. In such a case, we still define the eigenspace as the span of the set containing all linearly independent eigenvectors associated with the eigenvalue.

To the second question, how do eigenvalues help in determining invertibility? This is very simple:

Proposition 17. ***(Eigenvalues and Invertibility)*** *Let $A \in \mathbb{R}^{n \times n}$. Then, A is invertible if and only if all eigenvalues of A are non-zero.*

Proof. A is invertible if and only if $0 \neq \det(A) = \det(A - 0 \cdot \mathbf{I_n})$, which is the case if and only if 0 is not an eigenvalue of A. □

The practical value is that sometimes, you may have already computed the eigenvalues of a matrix before investigating its invertibility. Then, this proposition can help you avoid the additional step of computing the determinant.

Now, coming to the last concept: definiteness. Let's first look at the definition:

Definition 54. ***(Definiteness of a Matrix)*** *A **symmetric** square matrix $A \in \mathbb{R}^{n \times n}$ is called*

- *positive semi-definite if* $\forall x \in \mathbb{R}^n : x'Ax \geq 0$
- *negative semi-definite if* $\forall x \in \mathbb{R}^n : x'Ax \leq 0$
- *positive definite if* $\forall x \in \mathbb{R}^n \backslash \{\mathbf{0}\} : x'Ax > 0$
- *negative definite if* $\forall x \in \mathbb{R}^n \backslash \{\mathbf{0}\} : x'Ax < 0$

Otherwise, it is called indefinite.

Note that the concept not only applies only to square matrices, but also requires them to be symmetric! Further, we exclude the zero vector from definiteness because $\mathbf{0}'A\mathbf{0} = 0$ for all matrices A. The concept's relation to invertibility is the given through the following characterization:

Proposition 18. ***(Definiteness and Eigenvalues)*** *A symmetric square matrix $A \in \mathbb{R}^{n \times n}$ is*

(i) positive (negative) definite if and only if all eigenvalues of A are strictly positive (negative).

(ii) positive (negative) semi-definite if and only if all eigenvalues of A are strictly non-negative (non-positive).

The full proof requires concepts that are beyond our scope, and is thus left out. The "only if ($\Rightarrow$)"-direction, however, is relatively straightforward, so that we will consider it for *(i)* for the case of positive definiteness explicitly; for the other points, this step is perfectly analogous using the respective inequality. Note that for $x \in \mathbb{R}^n$, $x'x = \sum_{i=1}^n x_i^2 = \|x\|_2^2$ where $\|\cdot\|_2$ is the Euclidean norm. Thus, for $x \neq \mathbf{0}$, by the norm properties, $x'x > 0$. For any eigenvalue λ with eigenvector $x_\lambda \in \mathbb{R}^n \backslash \{\mathbf{0}\}$,

$$0 < x_\lambda' A x_\lambda = x_\lambda'(\lambda x_\lambda) = \lambda \|x_\lambda\|_2^2.$$

Dividing this inequality by $\|x_\lambda\|_2^2$, it results that $\lambda > 0$.

Thus, with Propositions 17 and 18, the following corollary emerges:

Corollary 5. ***(Definiteness and Invertibility)*** *If $A \in \mathbb{R}^{n \times n}$ is symmetric and positive definite or negative definite, it is invertible.*

This follows because positive and negative definiteness rule out zero eigenvalues. Thus, positive and negative definiteness are *sufficient* conditions for invertibility! As an example where this may come in handy, consider the first order condition for the OLS estimator (in matrix notation):

$$X'Xb = X'y$$

where X is an $n \times (k+1)$ matrix where $n \geq k+1$ and $\text{rk}(X) = k+1$, i.e. X has full column rank (this is the so-called no-multi-collinearity condition) and y is a vector of length n. Now, we want to solve for *the* (unique!) vector b that satisfies this condition. Clearly, this requires that we can invert $X'X$. It turns out that this may be easiest to do with the definiteness criterion: take for granted that when multiplying a matrix with its transpose, the resulting matrix is symmetric. Then, for any $v \in \mathbb{R}^{k+1}$,

$$v'X'Xv = (Xv)'Xv = \|Xv\|_2^2 \geq 0.$$

Now, we just need to show that if $v \neq \mathbf{0}$, then $Xv \neq 0$, and we have proved that $X'X$ is positive definite and, by Corollary 5, also invertible! To do this, we proceed as follows: Suppose that $v \neq \mathbf{0}$. Then, what does $Xv = 0$ mean? We can re-write $X = (x_1, \ldots, x_{k+1})$ in column notation. Then, it is easily verified that $Xv = v_1 x_1 + \ldots + v_{k+1} x_{k+1}$. Thus, if $Xv = \mathbf{0}$ for $v \neq \mathbf{0}$, there exists a linear combination of the columns of X with non-zero coefficients that is equal to zero. Thus, some columns of X must be linearly dependent, i.e. $\text{rk}(X) < k+1$ – however, this is precisely what the no-multi-collinearity condition rules out! Thus, the matrix $X'X$ is invertible, and the unique solution to the first order condition is $\beta^{OLS} := (X'X)^{-1}X'y$.

2.4 Computing Inverse Matrices: the Gauss-Jordan Algorithm

After the extensive discussion of invertibility above, let's finally discuss how, if we have established invertibility, we can actually compute the inverse matrix. Our discussion of Proposition 12 above has already grasped at this issue: we can apply the same elementary operations that we use to transform an invertible matrix A to the identity matrix to an identity matrix and arrive at the inverse. What is left to do is to establish that whenever there exists an inverse, the procedure will identify it. Subsequently, we consider some examples how to practically apply this method.

Theorem 32. ***(Gauß-Jordan Algorithm Validity)*** *Suppose that $A \in \mathbb{R}^{n \times n}$ is an invertible matrix. Then, we can apply elementary operations $E_1, \ldots, E_k$ in ascending order of the index to A to arrive at the identity matrix $\mathbf{I_n}$, and the inverse can be determined as $A^{-1} = E_k \cdot \ldots \cdot E_1$.*

Proof. Because A is square, we can use elementary operations to arrive at the intermediate upper triangular matrix $\tilde{A}$. Because A is invertible, the diagonal of $\tilde{A}$ contains only non-zero elements (cf. Corollary 4). That is, $\tilde{A}$ is of the form

$$\tilde{A} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \cdots & \tilde{a}_{1n} \\ 0 & \tilde{a}_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{a}_{2,n-1} \\ 0 & \cdots & 0 & \tilde{a}_{nn} \end{pmatrix} \quad \text{where} \quad \forall j \in \{1, \ldots, n\} : \tilde{a}_{jj} \neq 0.$$

First, we can multiply all rows j by $1/\tilde{a}_{jj}$ to obtain the new matrix

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & \cdots & \hat{a}_{1n} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{a}_{2,n-1} \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Now, you may already see how we can arrive at the identity matrix: start with column $n-1$. If $\hat{a}_{2,n-1} = 0$, no transformation is needed. Else, subtract $\hat{a}_{2,n-1}$ times column n to arrive at the unit vector e'_{n-1} for the $n-1$-th column. For the $n-2$-th column, for any non-zero entry at index $j > n-2$, subtract $a_{n-2,j}$ times the j-th row to arrive at the unit vector e'_{n-2} for the $n-2$-th column. Iteratively repeat this until you arrive at column 1. Then, the resulting matrix is the identity matrix. □

In practice, to keep things tractable, we write the identity matrix next to the matrix to be inverted and perform operations iteratively. To convert A to the identity, you will first want to bring it to upper triangular form. Here, Algorithm 1 may be helpful - you will also see this in the example below. Then, apply the procedure described in the proof above to eliminate all remaining elements above the diagonal.[20]

As promised, here is an example: Start from the matrix

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

First, we want to know whether it is invertible – a quick check of the determinant criterion (that we can apply because the matrix is square), using e.g. our 3×3 formula (do it!), yields $\det(A) = 4 \neq 0$, so the matrix is indeed invertible. So, let's start the procedure by writing the matrix next to an identity matrix of appropriate dimension:

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Our first goal is to get a 2×2 triangular block in the upper left corner – this is *always* a good start. For this, because the (1,1) entry is non-zero, we add 1/2 times row 1 to row 2[21] to eliminate the -1 at position (2,1). Applying this transformation to *both* matrices gives

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Now, note that we want ones on the diagonal. Thus, we multiply rows 1 and 2 by 1/2 and 2/3, respectively:

$$\begin{pmatrix} 1 & -1/2 & 0 \\ 0 & 1 & -2/3 \\ 0 & -1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1/2 & 0 & 0 \\ 1/3 & 2/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

One more step to get to upper triangular form (here comes the application of Algorithm 1) – add row 2 to row 3:

$$\begin{pmatrix} 1 & -1/2 & 0 \\ 0 & 1 & -2/3 \\ 0 & 0 & 4/3 \end{pmatrix} \quad \begin{pmatrix} 1/2 & 0 & 0 \\ 1/3 & 2/3 & 0 \\ 1/3 & 2/3 & 1 \end{pmatrix}$$

Let's get our last one on the diagonal by multiplying the last column with 3/4:

[20] You can equivalently bring the matrix to upper triangular form and then eliminate the triangle below the diagonal, this is completely up to you.

[21] Otherwise, i.e. if the (1,1) entry was zero, we could simply exchange rows.

$$\begin{pmatrix} 1 & -1/2 & 0 \\ 0 & 1 & -2/3 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1/2 & 0 & 0 \\ 1/3 & 2/3 & 0 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}$$

Now, we're almost there. From now on, we proceed like in the proof of Theorem 32. First, get rid of the non-zero entry in position (2,3) by adding 2/3 times row 3 to row 2:

$$\begin{pmatrix} 1 & -1/2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1/2 & 0 & 0 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}$$

Finally, it remains to add 1/2 times row 2 to row 1:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}$$

Thus, our algorithm tells us that $A^{-1} = \begin{pmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}$. If you're suspicious and don't fully trust our abstract proofs, try and verify that $AA^{-1} = \mathbf{I_n}$! ;-)

The example given here was very extensive, usually, to save space and time, you would produce the zeros for the triangular form in the same step where you set the diagonal elements to one. If doing so, you may only need three steps, or even less, depending on how many transformations you manage to track in one step. Being less extensive will help you save time in exams or at problem sets, but when going fast you are also more prone to errors, so watch out!

Finally, when considering a 2×2 matrix, there is a rule that allows us to avoid the algorithm:

Proposition 19. ***(Inverse of a*** 2×2 ***Matrix)*** *Consider the matrix* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ *where* $ad \neq bc$. *Then,*

$$A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

You can try and prove it, it is straightforward using the algorithm. This fact gives the arguably quickest way to compute the inverse for any 2×2 matrix and it is worthwhile memorizing.

Before moving on to the last piece of information on matrices, let us summarize again what we have found:

- A *square* matrix $A \in \mathbb{R}^n$ is invertible if and only if either of the following equivalent conditions hold:

  1. $\det(A) \neq 0$,

2. $\text{rk}(A) = n$,
3. When transforming A to triangular form $\tilde{A}$, the diagonal of $\tilde{A}$ has only non-zero entries,
4. All eigenvalues of A are non-zero.

- Further, if A is symmetric, sufficient invertibility conditions are positive and negative definiteness.
- For a system $Ax = b$ of linear equations,
  1. if A is square, a unique solution exists if and only if A is invertible,
  2. Generally, there exists a solution if and only if $b \in \text{Co}(A)$ ("no conflicting information"),
  3. If there are more unknowns than equations, there is no unique solution ("insufficient information"),
  4. If there are more equations than unknowns and $b \in \text{Co}(A)$, some equations are linear combinations of others ("redundant information") and a unique solution exists if and only if the reduced square system has an invertible matrix.

→ When we know that a solution exists, investigating uniqueness is ultimately concerned with a square system, which are the easiest and most tractable system class!

- We can invert a matrix using the Gauß-Jordan algorithm or a rule if the matrix is 2×2.

2.5 Linear Functions

Now for the last matrix-related concept. In fact, a matrix of dimension $m \times n$ can also be thought of as a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let us start by defining what we mean by a "linear function" from $\mathbb{R}^n$ to $\mathbb{R}^m$. Similar to the way vector spaces preserve linear combinations of their elements, linearity of a function is also about preservation of linear combinations, but between two vector spaces, namely, the domain and the codomain. Less abstractly, the function f is linear if it maps any linear combination of x-values into the linear combination of y-values with the *same* coefficients. More formally:

Definition 55. ***(Linear Function)*** *Let $\mathbb{X}$ and $\mathbb{Y}$ be two vector spaces based on the sets X and Y, respectively. A function $f : X \to Y$ is said to be linear if*

$$\forall n \in \mathbb{N} \forall \lambda_1, \ldots, \lambda_n \in \mathbb{R} \forall x_1, \ldots, x_n \in X : \left(f\left(\sum_{i=1}^{n} \lambda_i x_i \right) = \sum_{i=1}^{n} \lambda_i f(x_i) \right)$$

As a comment, should you ever be asked to prove linearity of a function, here it suffices to check that (i) both the domain and codomain are vector spaces (i.e. the expressions on the left and right are well-defined) – this is mostly given by the setup when we consider functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ – and (ii) that $\forall \lambda \in \mathbb{R} \forall x_1, x_2 \in X : f(\lambda x_1 + x_2) = \lambda f(x_1) + f(x_2)$.[22]

[22] As an exercise, you can verify that this is indeed a necessary and sufficient condition for a linear function – or spend a couple minutes on Google to find a reference, should you need this for a problem set. ;-)

To give you some practice, it is easy to see that for any $a \in \mathbb{R}$ the function $f_1 : \mathbb{R} \to \mathbb{R}, x \mapsto a \cdot x$ is linear, and that the function $f : \mathbb{R} \to \mathbb{R}, x \mapsto a \cdot x^2$ is not. What about $x \mapsto ax + b$?

...This function is not linear in the sense of our definition - indeed, because we do not rule out that the λ's are zero, a linear function must necessarily satisfy $f(0) = 0$.[23]

As a preview of the next section, an important example is the function $\frac{\partial}{\partial x}$, which maps from the vector space $D^1(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R}, f \textit{ is differentiable}\}$ (equipped with appropriate basis operations, see below) of all differentiable functions over $\mathbb{R}$ into the vector space of all functions, i.e.,

$$\frac{\partial}{\partial x} : D^1(\mathbb{R}) \to \{f : \mathbb{R} \to \mathbb{R}\}, f \mapsto f'.$$

It is beyond our scope to prove this, but consider the typical basis operations on $D^1(\mathbb{R})$ defined by $f + g : ((f + g)(x) = f(x) + g(x) \forall x \in X)$ and $\lambda f : ((\lambda f)(x) = \lambda \cdot f(x) \forall x \in X)$. Then, you may indeed apply the "$\lambda x_1 + x_2$" rule to try and verify this, it's not that hard! Note that often, for easier communication, functions such as $\frac{\partial}{\partial x}$ which operate between vector spaces of functions are called *operators*.

An important case of linear functions are those from $\mathbb{R}^n$ to $\mathbb{R}^m$. If we consider the function that maps $x \in \mathbb{R}^n$ to its product with the matrix $A \in \mathbb{R}^{m \times n}$, i.e. $f : \mathbb{R}^n \mapsto \mathbb{R}^m, x \mapsto Ax$, then it is rather straightforward to verify that f is linear:

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) = A \cdot \left(\sum_{i=1}^{n} \lambda_i x_i\right) = \sum_{i=1}^{n} (A\lambda_i x_i) = \sum_{i=1}^{n} \lambda_i A x_i = \sum_{i=1}^{n} \lambda_i f(x_i).$$

In fact, the relation between the two types of objects is stronger than that, as the converse is also true: *All linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ are expressible via a matrix of dimension $m \times n$.*[24]

To conclude this chapter, let's combine invertibility and linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ by asking when we can find the *inverse function* of a linear function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$. As discussed above, there exists A so that $\forall x \in \mathbb{R}^n : f(x) = Ax$. Suppose we want to find the mapping $g : \mathbb{R}^m \mapsto \mathbb{R}^n$ that maps Ax onto x. If there is such a mapping, it may be represented by the matrix $B \in \mathbb{R}^{n \times m}$ for which $By = x$ for all $y = Ax \in \mathbb{R}^m$, $x \in \mathbb{R}^n$. Consequently, we must have that $BA = \mathbf{I_n}$. Clearly, a *sufficient* condition for invertibility of a linear function is that $n = m$ and the matrix A is invertible, where we can set $B = A^{-1}$ and define the mapping g by $y \mapsto A^{-1}y$.

[23] To see this, pick an arbitrary $x \in X$. Then, if f is linear, $f(0) = f(\sum_{i=1}^{n} 0 \cdot x_i) = \sum_{i=1}^{n} 0 \cdot f(x_i) = 0$.

[24] A proof can be found, for instance, in dlF chapter 3, theorem 3.5.

2.6 Contents and Take-Aways

Chapter 2: Matrix Algebra discusses

- the basics of the matrix concept, including addition and multiplication of matrices, and important matrix properties (symmetry, lower/upper triangular form, diagonality, etc.)
- matrices and linear equation systems, focused on the connection of invertability and existence of a unique solution in the context of "square" systems (as many unknowns as equations)
- criteria for matrix inversion and related concepts: determinants, rank, eigenvalues and -vectors, definiteness
- computation of inverse matrices: the 2x2-matrix formula and the Gauß-Jordan algorithm

Someone with profound knowledge of the contents of this chapter should

- know the dimension conformability conditions for matrix addition and multiplication, and how to compute the sum and product of conformable matrices
- be able to represent a system of n equations in k unknowns in matrix form
- know how invertability of a matrix A relates to existence of a unique solution in an associated equation system $Ax = b$
- be thoroughly familiar with the equivalent and sufficient conditions for matrix invertability as summarized at the end of the chapter (determinant, rank, eigenvalues, unique solution in associated system; definiteness)
- know the "sum-of-squares" property of the scalar product, i.e. $v'v = \sum_{j=1}^{n} v_j^2$ for $v \in \mathbb{R}^n$
- be aware of some inversion rules for "derived matrices", i.e. if A, B and C are invertible, how to invert e.g. A' and ABC
- be familiar with the Gauß-Jordan matrix inversion method, and have a rough idea of why it works
- know the definition of a column space of a matrix, and how it is useful when thinking about the existence of solutions in linear equation systems
- know how to find the eigenvalues of a matrix, and have a rough idea of why the method works

and be able to answer a number of related questions, including

- How are row and column rank defined? Can a matrix have a strictly greater column rank than row rank?
- What are the elementary matrix operations? How do they affect the rank and the determinant?
- Is the matrix product associative? Is it commutative?
- Provided that it exists for a given matrix A, is the inverse matrix A^{-1} always unique?
- Can any square matrix be brought to an upper triangular form using only elementary matrix operations? Which condition ensures that we can bring it to identity form?
- What characterizes an indefinite matrix? Is any positive definite matrix also positive semi-definite? Can it also be negative semi-definite?
- Consider the matrix

$$\begin{pmatrix} 1 & 4 & 0 \\ 2 & 3 & 1 \\ 0 & 0 & -2 \end{pmatrix}.$$

  What is its determinant? Is it invertible? If so, what is its inverse matrix?

2.7 Recap Questions

1. What is the difference between a square matrix and a diagonal matrix? How is the identity matrix defined? Is it diagonal, square, both, or neither?

2. Prove Proposition 10.

3 Multivariate Calculus

Chapters 0 and 1 have covered topics of fundamental concern to all mathematical disciplines, while the elaborations of Chapter 1 were already more closely linked to what mathematicians call *linear algebra*. Broadly speaking, this is because the basis operations allow to compute linear combinations, which are linear functions in the sense of Chapter 2, and the theory of vector spaces is concerned with characterizing sets related to these functions. Subsequently, Chapter 2 has discussed a central building block of linear algebra: characterizing and solving systems of linear equations. For the rest of this course, we want to move away from linear algebra and instead consider key issues in mathematical *analysis*, where we are concerned with analyzing mathematical objects, especially functions and related equations (say $\forall x \in X : f(x) = cf'(x)$), and investigate whether they are continuous, differentiable, invertible, have maxima and minima, and much more.

As an economist, you should care alot about these concepts: while you certainly know how to e.g. take derivatives of functions mapping from $\mathbb{R}$ to $\mathbb{R}$, being familiar with more general methods of (functional) analysis is invaluable because the typical functions we consider have more than one argument. To give a highly non-exhaustive list, you may, for instance, think about utility derived from a vector of goods (quantities), the welfare given quantities of a private and a public good or production cost with multiple inputs.

From Chapter 0, you should be thoroughly familiar with the concept of a function

$$f : X \mapsto Y, x \mapsto y = f(x)$$

with *domain* X, *codomain* Y, and *image* $f[X]$. There, we had also learned about the *graph* of f:

$$G(f) = \{(x, y) \in X \times Y : y = f(x)\} = \{(x, f(x)) : x \in X\}. \tag{6}$$

As you likely know, the most common way to geometrically represent a function is plotting the values of x against the y-dimension of the graph, which is feasible if x has at most two dimensions. There is some more related terminology that one should be familiar with before moving on, so let us turn to it here in a first step.

- If $X \subseteq \mathbb{R}$, we call f a univariate function.
- If $X \subseteq \mathbb{R}^n$ for $n > 1$, we call f a multivariate function.
- If $Y \subseteq \mathbb{R}$, we call f a real-valued function.
- If $Y \subseteq \mathbb{R}^m$ for $m > 1$, we call f a vector-valued function.
- If X and Y are sets of functions, we call f an operator.

In this chapter, we will first study how to deal with multivariate real-valued functions, i.e., functions $f : \mathbb{R}^n \mapsto \mathbb{R}$ where $n > 1$ but the codomain is still the standard real line. We have already seen many such functions, e.g. the norm of a vector, or the function $x'Ax$ of x when

contemplating definiteness of the matrix A, and turn to multivariate vector-valued functions $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, with $n > 1$ and $m > 1$ thereafter.[1]

Earlier versions of this course were more ambitious in linking this chapter's formal concepts to geometrical interpretations. This script deviates from the graphical approach due to the subjective impression that some of the graphical "intuition" was more abstract and complex than the math itself, and thus might be more confusing than helpful for some. If you believe you could benefit from a more thorough geometrical discussion, feel free to study the old course material, available at https://helmsmueller.wordpress.com/teaching/. That being said, of course, when a graphical interpretation is easily obtained, it is also discussed here.

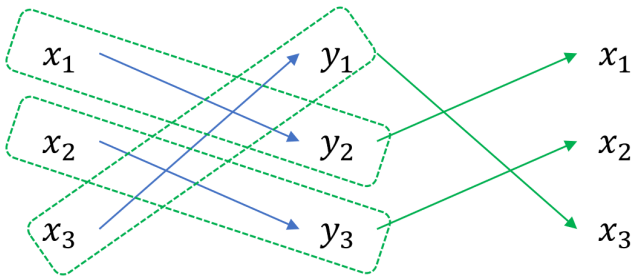
3.1 Basic Concepts

Let us begin with some basics. First, we will consider invertibility, a concept which translates in a one-to-one fashion from univariate real-valued functions, and then turn to the very important concepts of convexity and concavity.

3.1.1 Invertibility of Functions

Recall from Chapter 0 that we could invert a function $f : X \mapsto Y$, $X, Y \subseteq \mathbb{R}$ if and only if for any $y \in Y$, there exists *exactly* one $x \in X$ such that $f(x) = y$. This is the case because then, and only then, can we identify a unique element $x \in X$ that is mapped onto any $y \in Y$ – recall that when considering $g : Y \mapsto X$ as a candidate for the inverse function, we require g to be defined everywhere on Y, i.e. for all $y \in Y$, and by definition of g as a function, g must take exactly one value $x \in X$ for any $y \in Y$, rather than multiple values.[2] Indeed, this characterization is not specific to univariate, real-valued functions, but generally refers to any function $f : X \mapsto Y$ with domain $X \subseteq \mathbb{R}^n$ and codomain $Y \subseteq \mathbb{R}^m$, $n, m \in \mathbb{N}$.

To see these abstract elaborations graphically, let's consider two examples of functions where X and Y contain only three elements.

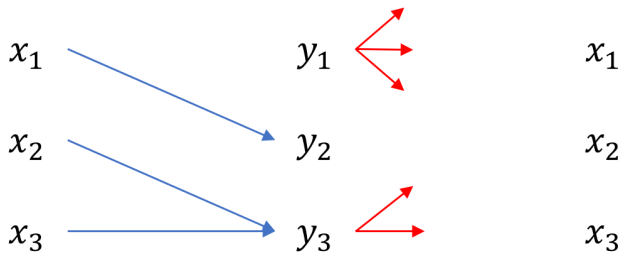


Here, the function f is represented by the blue arrows. As the green dotted boxes indicate, there are well-defined pairs (x, y), every one of which can be identified by knowing either the x– or the y-value. Hence, we can invert the rule of f and define the inverse function as the rule

[1] Although we could further generalize the concepts for infinite-dimensional spaces (e.g. function spaces), and this happens to be very close to the generalization for finite-dimensional spaces (e.g. $\mathbb{R}^n$, $n \in \mathbb{N}$), we restrict ourselves to finite-dimensional vector spaces. The reason is twofold: (i) you will not be expected to be able to work with infinite dimensional spaces in the master's curriculum, (ii) if you grasp the generalization for finite dimensional spaces, then you will easily grasp that for infinite dimensional spaces as it is presented in books.

[2] There exist more general *mappings* where it is allowed that a single element is mapped into a set, such mappings are called *correspondences*.

associating y's with x's to obtain the same pairs as we do under f, as indicated by the green arrows.



Here, again, f is represented by the blue arrows. However, now we have two sources of ambiguity in the attempt to invert the mapping of f: first, the value y_3 is associated with two x-values. By the definition of a function, however, we can map y_3 only onto one x value when considering a mapping $Y \mapsto X$, so that it is not possible to define a value of the inverse function at y_3. Secondly, the value y_1 is associated with no x-value at all, and there is no candidate value for the inverse function to take at y_1.

What should be especially clear from these examples is that the conditions under which the inverse function exists or not, respectively, are not specific to univariate, real-valued functions, but generally refer to any function $f : X \mapsto Y$ with arbitrary domain X and codomain Y.

We can define invertibility elegantly using the concepts of *injectivity* and *surjectivity*. As you saw in Chapter 1, in the case of a linear function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ that can be represented by the matrix A as $f(x) = Ax$, you can explicitly compute the inverse as $f^{-1}(y) = A^{-1}y$ if A is invertible. We will see later how we may characterize the inverse function more generally.

Definition 56. ***(Surjective Function)*** *Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ and $f : X \mapsto Y$. Then, f is said to be surjective if $\forall y \in Y \exists x \in X : (f(x) = y)$, i.e. for every y in the codomain of f, there exists at least one element x in the domain that is mapped onto it.*

Note that next to the mapping rule of f ($x \mapsto y = f(x)$), surjectivity crucially depends on the set Y we choose to define f. Consider e.g. $f(x) = x^2$ where $x \in \mathbb{R}$. Is f surjective? it depends! If we define $f : \mathbb{R} \mapsto \mathbb{R}$, i.e. $X = Y = \mathbb{R}$, then any $y \in Y, y < 0$ does *not* have an $x \in X$ for which $f(x) = y$, so that f is not surjective. On the other hand, if we set $Y = \mathbb{R}_+$, then for any y there exists $x = \sqrt{y} \in X : f(x) = y$, and f is surjective! This principle holds true more generally: given the domain X that we consider, we can simply define $Y = f[X]$ to "throw out" the values not mapped onto by f and ensure surjectivity. As we have seen, surjectivity is the first requirement satisfaction of which tells us that we can find elements in X to map $y \in Y$ onto when contemplating existence of the inverse function. Now, we just have to know whether the element in X is unique – enter injectivity:

Definition 57. ***(Injective Function)*** *Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ and $f : X \mapsto Y$. Then, f is said to be injective if $\forall x_1, x_2 \in X : (x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2))$, i.e. every two different elements in X have a different image under f.*

For the inverse function, injectivity rules out that for an $y \in Y$, we have two different elements $x_1, x_2 \in X$ so that $f(x_1) = f(x_2) = y$. Coming back to the example, is $f : \mathbb{R} \mapsto \mathbb{R}_+, x \mapsto x^2$

injective? Clearly not: e.g. $(-1)^2 = 1 = 1^2$, so that $f(-1) = f(1)$. Thus, it may also depend on the *domain* that we consider whether we can invert a given function – setting e.g. $f : \mathbb{R}_+ \mapsto \mathbb{R}_+, x \mapsto x^2$ achieves also injectivity because for $x_1, x_2 \in \mathbb{R}$, if $x_1 \neq x_2$ then also $x_1^2 \neq x_2^2$. Thus, if f is defined on $X = \mathbb{R}_+$ (rather than $X = \mathbb{R}$), we can invert f on $Y = \mathbb{R}_+$ (rather than $Y = \mathbb{R}$) as $f^{-1} : \mathbb{R}_+ \mapsto \mathbb{R}_+ : y \mapsto \sqrt{y}$. Then, indeed for any $x \in \mathbb{R}_+$, $f^{-1}(f(x)) = \sqrt{x^2} = x$.

In terms of language, sometimes, we also call surjective functions *onto*, because they map onto the whole space Y, and injective functions *one-to-one*, because they map every one element in X to one distinct element in Y. Before moving on, a last definition:

Definition 58. ***(Bijective Function)*** *Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ and $f : X \mapsto Y$. Then, f is said to be bijective if f is injective and surjective.*

Clearly, if we have inverted f to the function f^{-1}, then the function f^{-1} is also invertible with $(f^{-1})^{-1} = f$. This allows us to conclude:

Definition 59. ***(Inverse Function)*** *Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ and $f : X \mapsto Y$. Then, the function $g : Y \mapsto X, y \mapsto g(y)$ such that $\forall x \in X : g(f(x)) = x$ and $\forall y \in Y : f(g(y)) = y$ is called the inverse function of f. We write $g = f^{-1}$.*

Theorem 33. ***(Existence of the Inverse Function)*** *Let $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$ and $f : X \mapsto Y$. Then, the inverse function f^{-1} of f exists if and only if f is bijective.*

The proof is given below, mainly for completeness. Still, it is rather simple and it treats the relationship of inverse functions and bijectivity more formally than the elaborations above. Try to follow it if you have time.

Proof. "$\Rightarrow$" Suppose that f^{-1} exists. Then, for any $y \in Y$, there exists $x = f^{-1}(y)$ such that $f(x) = f(f^{-1}(y)) = y$, and f is surjective. Next, let $x_1, x_2 \in X$, $x_1 \neq x_2$, and denote $y_1 = f(x_1)$ and $y_2 = f(x_2)$. Now, if $y_1 = y_2 = y$, then $f^{-1}(y)$ is not defined, a contradiction to existence of f^{-1}. Thus, $f(x_1) \neq f(x_2)$ and f is injective. Because f is also surjective, f is bijective.
"$\Leftarrow$" Suppose that f is bijective. Let $y \in Y$. Then, by surjectivity, there exists $x \in X : f(x) = y$. By injectivity, this x is unique, and we conclude that $\forall y \in Y \exists! x(y) \in X : f(x(y)) = y$. Let $g : Y \mapsto X, y \mapsto x(y)$. Then, $\forall x \in X : g(f(x)) = x$, because x is the unique element of X that maps onto $f(x)$. □

Before moving on, two comments deem worthwhile. First, be reminded again to **not confuse the inverse function $f^{-1}(y)$ with the preimage of a set S, $f^{-1}[S]$!!** The latter quantity is always defined, but captures a fundamentally different concept. Secondly, we have seen that while surjectivity is more of a matter of properly defining the codomain, injectivity and thus invertibility strongly depends on the domain. Therefore, it is also common to consider invertibility not "globally" as we have done here, but rather "locally" in a small neighborhood (i.e., open ball) around some point of interest. That is, when starting from a function $f : X \mapsto Y$, for a point of interest x_0 and a small $\varepsilon > 0$, we consider the restricted function $f|_{B_\varepsilon(x_0)} : B_\varepsilon(x_0) \mapsto Y, x \mapsto f(x)$.

3.1.2 Convexity and Concavity of Multivariate Real-Valued Functions

In this subsection, we consider two elementary properties functions can have: convexity and concavity. We restrict attention to multivariate real-valued functions, i.e. those functions

$f : \mathbb{R}^n \mapsto \mathbb{R}$ that may take vectors as arguments but map into real numbers. The properties' importance stems from optimization and will thus be emphasized in the next chapter. For now, we proceed with the formal discussion.

Definition 60. ***(Convex and Concave Real Valued Function)*** *Let $X \subseteq \mathbb{R}^n$ be a* ***convex set****. A function $f : X \to \mathbb{R}$ is convex if for any $x, y \in X$ and $\lambda \in [0,1]$,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

Moreover, if for any $x, y \in X$ such that $y \neq x$ and $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

we say that f is strictly convex. Moreover, we say that f is (strictly) concave if $-f$ is (strictly) convex.

Note that the definition of a concave real-valued function also requires that the function be defined on a *convex* domain – i.e. a set X which satisfies $\forall x, y \in X \forall \lambda \in [0,1] : \lambda x + (1-\lambda)y \in X$. For the most frequent cases, $X = \mathbb{R}^n$ and $X = \mathbb{R}^n_+$, this is extremely straightforward to verify, and nothing you need to be scared of, but it should be kept in mind nonetheless. We require this in the definition because else, $f(\lambda x + (1-\lambda)y)$ is not always defined, and we can not judge on the inequality defining convexity/concavity. The definition of concavity using $-f$ may be a bit awkward, to check concavity, you can equvialently consider the defining inequalities

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) \qquad \forall x, y \in X \forall \lambda \in [0,1]$$

and for strict concavity

$$f(\lambda x + (1-\lambda)y) > \lambda f(x) + (1-\lambda)f(y) \qquad \forall x, y \in X \text{ so that } x \neq y \text{ and } \forall \lambda \in (0,1).$$

For univariate real-valued functions, you are likely familiar with the graphical representation of these concepts: note that all points $\lambda f(x) + (1-\lambda)f(y)$ with $\lambda \in [0,1]$ lie on the line segment connecting $f(x)$ and $f(y)$. Then, convexity (concavity) states that the graph of f must lie below (above) this line segment everywhere between x and y. This relationship is illustrated in Figure 10.

When considering functions with multiple arguments, the conceptual idea is similar, yet graphically more challenging to display. Let us have a look at a simple convex function defined in $X \subseteq \mathbb{R}^2$, say, $f(x_1, x_2) = x_1^2 + x_2^2$. Try and show convexity here – or consult the footnote.[3]

The graph of f as defined in equation (6) lies in $\mathbb{R}^3$ (see the left side of Figure 11). Recall that we consider real-valued functions that map to $\mathbb{R}$, and that the codomain of f corresponds to the third, or vertical dimension in the plot. Now, if one draws the segment line that connect two

[3] We take for granted that x^2 is strictly convex because the direct proof is unnecessarily inconvenient and we have not yet seen the derivative criterion. If you are interested, see https://math.stackexchange.com/questions/580856/proof-of-convexity-of-fx-x2. With this fact, for any $x, y \in \mathbb{R}^2$ so that $x \neq y$ and $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) = (\lambda x_1 + (1-\lambda)y_1)^2 + (\lambda x_2 + (1-\lambda)y_2)^2 < \lambda x_1^2 + (1-\lambda)y_1^2 + (\lambda x_2^2 + (1-\lambda)y_2^2 = \lambda f(x) + (1-\lambda)f(y).$$

The inequality follows by strict convexity of $(\cdot)^2$ and because either $x_1 \neq y_1$ or $x_2 \neq y_2$.

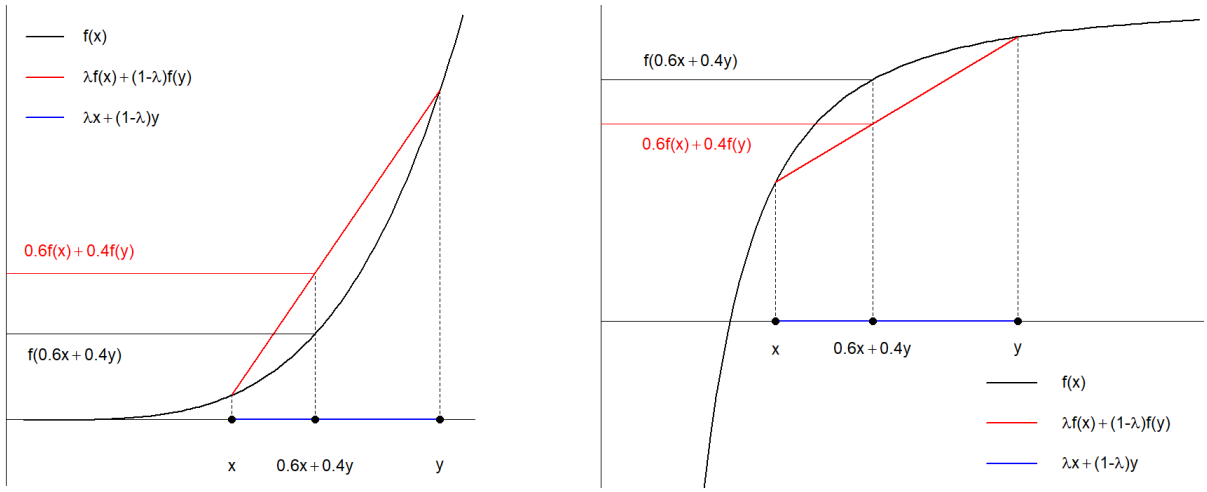


Figure 10: Convexity (left) and concavity (right) of univariate real-valued functions.

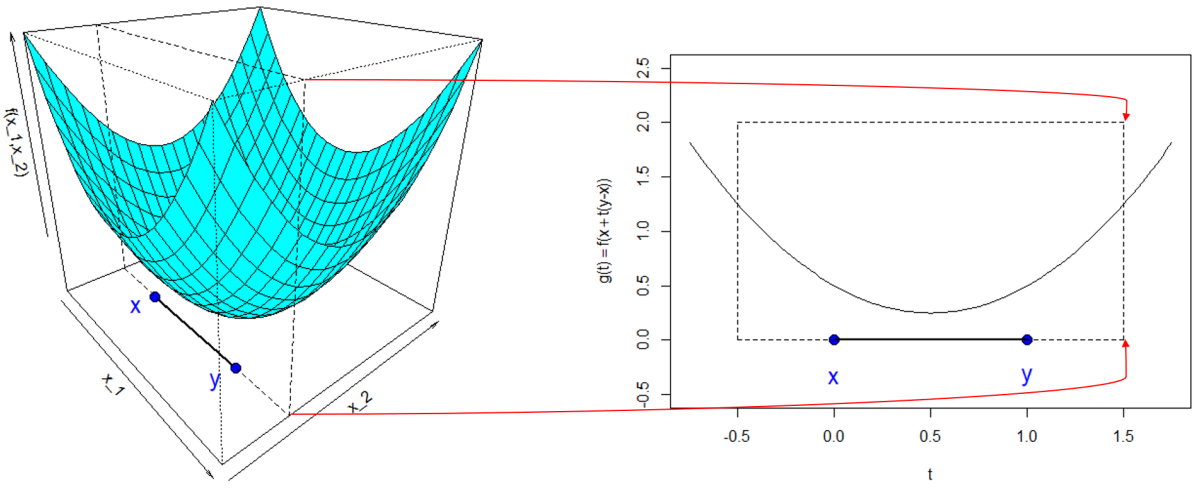


Figure 11: Lying "above" in 3 or more dimensions.

arbitrary values $f(x)$ and $f(y)$ of f as before, somehow, it does appear to lie "above" the graph of the function. Yet, a line exhausts only one out of three dimensions, and the notion of "above" isn't as clear as before in Figure 10 at first sight. However, it is upon closer investigation, as the figure already hints at: in the definition of convexity, we do not consider all points "between" x and y as in the univariate case, where between would e.g. refer to a rectangle with corners $x, y, (x_1, y_2)'$ and $(y_2, x_1)'$ but only points described by[4]

$$(1-\lambda)x + \lambda y = x + \lambda(y-x), \qquad \lambda \in [0,1]. \tag{7}$$

When not restricting λ to lie in $[0,1]$, the second expression looks very much like the hyperplane representation of a *line* we have discussed in Chapter 1! This line describes a subset of the $\mathbb{R}^2$, the domain of f, characterized by the *single* directionality $z := y - x$. Thus, when we interpret x as a new origin and the vector z as the directionality of the horizontal axis, as indicated by the blue dashed lines in Figure 11, we arrive at a two-dimensional graph again, as

[4]The coefficient $(1-\lambda)$ is assigned to x (rather than to y, as would be usual) to arrive at the line with "origin" x.

indicated in the drawn picture on the right.[5] The restricted graph drawn in this direction depends only on a univariate argument, namely t, because this is the variable our new horizontal axis (i.e. vectors that "start" at x, have directionality $z = y - x$ and magnitude $t \in \mathbb{R}$) moves along! Thus, the restricted graph is

$$G_R(f|x,z) = \{(t, f(x+tz)) : t \in \mathbb{R}\}$$

and describes the graph of the function $g : \mathbb{R} \mapsto \mathbb{R} : t \mapsto f(x+tz)$! This neatly highlights our dimensionality reduction to the graph of a univariate function.

Thus, it appears that our graphical intuition is preserved because even with arbitrary $n \in \mathbb{N}$, the logic of equation (7) applies, and the line connecting x and y has just a unidimensional diectionality! Indeed, because x and y were chosen arbitrarily, the property of (strict) convexity says that for *any* origin point $x \in X$ and *any* directionality $z \in X$, the restricted graph of f given x and z will be (strictly) convex! As the following theorem shows, this conclusion holds for arbitrary functions f and $n \in \mathbb{N}$:

Theorem 34. ***(Graphical Characterization of Convexity)*** *Let $X \subseteq \mathbb{R}^n$ be a* ***convex set*** *and $f : X \mapsto \mathbb{R}$. Then, f is (strictly) convex if and only if $\forall x, z \in X$ such that $z \neq \mathbf{0}$, the function $g : \mathbb{R} \mapsto \mathbb{R}, t \mapsto f(x+tz)$ is (strictly) convex.*

The formal proof gives you some opportunity to get more familiar with the formal structure of showing equivalence in two parts. But, if you conceptually understood the elaborations above, i.e. you know that z comes from picking an arbitrary y on top of x and setting $z = y - x$, the theorem should already be clear to you.

Proof. (strict convexity only – convexity is analogous)
"$\Rightarrow$" Suppose that f is strictly convex. Let $x, z \in X$ such that $z \neq \mathbf{0}$, and let $s, t \in \mathbb{R}$, $s \neq t$ and $\lambda \in (0,1)$. Then,

$$\begin{aligned} g(\lambda t + (1-\lambda)s) &= f(x + (\lambda t + (1-\lambda)s)z) = f(x(\lambda + 1 - \lambda) + \lambda tz + (1-\lambda)sz) \\ &= f(\lambda(x+tz) + (1-\lambda)(x+sz)) < \lambda f(x+tz) + (1-\lambda) f(x+sz) \\ &= \lambda g(t) + (1-\lambda) g(s). \end{aligned}$$

where the inequality follows because $s \neq t$ and $z \neq \mathbf{0}$. Thus, for any $x, z \in X$ such that $z \neq \mathbf{0}$ the function $g : \mathbb{R} \mapsto \mathbb{R}, t \mapsto f(x+tz)$ is strictly convex.

"$\Leftarrow$" Suppose that for any $x, z \in X$ such that $z \neq \mathbf{0}$ the function $g : \mathbb{R} \mapsto \mathbb{R}, t \mapsto f(x+tz)$ is strictly convex. Let $x, y \in X$ so that $x \neq y$ and let $\lambda \in (0,1)$. Let $z := y - x$. Because $y \neq x$, $z \neq \mathbf{0}$.[6] Then, the function $g(t) = f(x+tz)$ is strictly convex, and

$$\begin{aligned} f((1-\lambda)x + \lambda y) &= f(x + \lambda(y-x)) = f(x + (\lambda \cdot 1 + (1-\lambda)\cdot 0)z) = g(\lambda \cdot 1 + (1-\lambda)\cdot 0) \\ &< \lambda g(1) + (1-\lambda) g(0) = \lambda f(x + 1\cdot(y-x)) + (1-\lambda) f(x + 0\cdot(y-x)) \\ &= (1-\lambda) f(x) + \lambda f(y). \end{aligned}$$

Thus, f is strictly convex. □

[5] The hand-written set indicates the line that gives the horizontal direction in the new system.
[6] For concavity, if $y = x$, then trivially, $f(\lambda x + (1-\lambda)y) = f(x) = \lambda f(x) + (1-\lambda) f(y)$.

Now that we have a proper idea of how convexity (and concavity as its opposite) looks like in more general vector spaces or respectively, for general multivariate real-valued functions $f : \mathbb{R}^n \mapsto \mathbb{R}$, we move to some related but weaker concept: *quasi-convexity*, with the natural opposite quasi-concavity. The reason is that for many applications, requiring convexity in the narrow sense as discussed above is too restrictive – e.g. when considering monotonic transformations[7] of an initially convex function, it is not guaranteed that the resulting function will also be convex. As such, the narrow range of functions convexity (and concavity) applies to restricts our ability to perform general functional analysis. The appealing aspect of considering quasi-convexity instead is that while applying to a much broader class of functions, it preserves most of the convenient properties of convex functions that we are interested in.

As you will see in the next chapter, the convexity of the *upper-level set* (for concave functions) and convexity of the *lower-level set* (for convex functions) are the specific characteristics of concave and convex one would wish to preserve. As multivariate convexity and concavity can be reduced to univariate ones, let me illustrate these concepts for the univariate case. For what follows, recall that a subset $S \subseteq \mathbb{R}$ of the real line is convex if and only if S is an interval, i.e. if there are $-\infty \leq a \leq b \leq \infty$ such that $S = [a, b]$, $S = (a, b]$, $S = [a, b)$ or $S = (a, b)$.

If one considers a convex function and draws a horizontal line (a "level" line), the set of elements x in the domain with an image below this line is called a lower-level set of the function and is convex. Similarly, if one considers a concave function, and draw an horizontal line (a "level" line) through it, the set of elements x in the domain with an image above this line is called an upper-level set of the function and is convex.

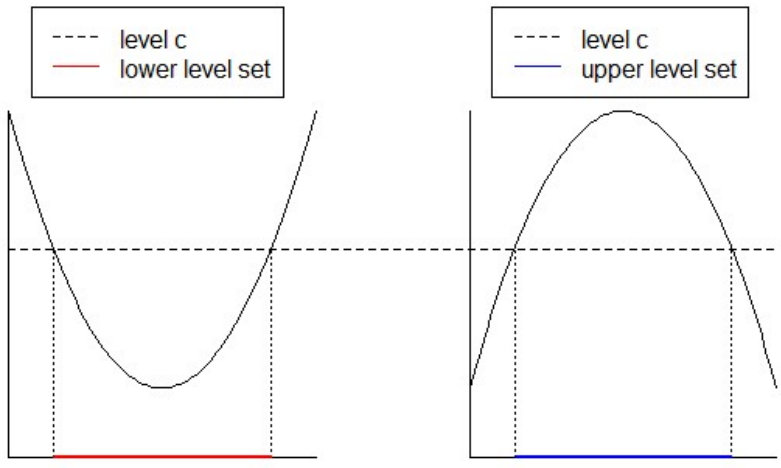


Figure 12: Convexity and Concavity via lower- and upper-level sets.

Quasiconvexity and quasiconcavity are precisely defined so as to preserve these two characteristic properties:

Definition 61. ***(Lower and Upper Level Set of a Function)*** *Let $X \subseteq \mathbb{R}^n$ be a convex set and*

[7] An increasing transformation of f is $(g \circ f)(x) = g(f(x))$ such that the function $g(y)$ is increasing, i.e. $y_1 \geq y_2 \Rightarrow g(y_1) \geq y_2$. A decreasing transformation is the opposite, where $y_1 \geq y_2 \Rightarrow g(y_1) \leq y_2$. Strict versions with strict inequalities also exist. See also the definition of a monotonic function in the introductory chapter.

$f : X \to \mathbb{R}$ be a real-valued function. Then, for $c \in \mathbb{R}$, the set

$$L_c^- := \{x | x \in X, f(x) \leq c\},$$

is called the lower-level set of f at c, and

$$L_c^+ := \{x | x \in X, f(x) \geq c\}$$

is called the upper-level set of f at c.

Definition 62. ***(Quasiconvexity, Quasiconcavity)*** *Let $X \subseteq \mathbb{R}^n$ be a convex set. A real-valued function $f : X \to \mathbb{R}$ is called quasiconvex if for all $c \in \mathbb{R}$, the lower-level set of f at c is convex. Alternatively, f is called quasiconcave if for all $c \in \mathbb{R}$, the upper-level set of f at c is convex.*

The following is an often more workable characterization:

Theorem 35. *(**Quasiconvexity, Quasiconcavity**) Let $X \subseteq \mathbb{R}^n$ be a convex set. A real-valued function $f : X \to \mathbb{R}$ is quasiconvex if and only if*

$$\forall x, y \in X \forall \lambda \in [0,1] : f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$$

Conversely, f is quasiconcave if and only if

$$\forall x, y \in X \forall \lambda \in [0,1] : (\lambda x + (1-\lambda)y) \geq \min\{f(x), f(y)\}$$

In the spirit of the definitions above, we further have the following characterizations that can sometimes be helpful:

Definition 63. ***(Strict Quasiconvexity, Strict Quasiconcavity)*** *Let $X \subseteq \mathbb{R}^n$ be a convex set. A real-valued function $f : X \to \mathbb{R}$ is called strictly quasiconvex if*

$$\forall x, y \in X \text{ such that } x \neq y \text{ and } \forall \lambda \in (0,1) : f(\lambda x + (1-\lambda)y) < \max\{f(x), f(y)\}$$

Conversely, f is strictly quasiconcave if

$$\forall x, y \in X \text{ such that } x \neq y \text{ and } \forall \lambda \in (0,1) : f(\lambda x + (1-\lambda)y) > \min\{f(x), f(y)\}$$

This generalization allows us to consider also a variety some non-convex and non-concave functions while ruling out only "too messy" functions, such as "camel backs" (see Figure 13). To see that we are dealing with a strict broadening of concepts i emphasize that *all convex functions are quasi-convex*, and *all concave functions are quasi-concave.*[8] Linear functions are both quasi-concave and quasi-convex (and thus *quasi-linear*), but they are not the only functions with this property – Monotonic functions are another instance of quasi-linear functions, you can convince yourself by looking at the graph of $f : \mathbb{R}_+ \mapsto \mathbb{R}_+, x \mapsto x^2$, which is strictly monotonically increasing (note that we take only positive arguments), and look at the lower-level and upper-level sets.

[8] This stems from the fact that we have defined the concepts from a *characteristic feature* of convex or respectively,

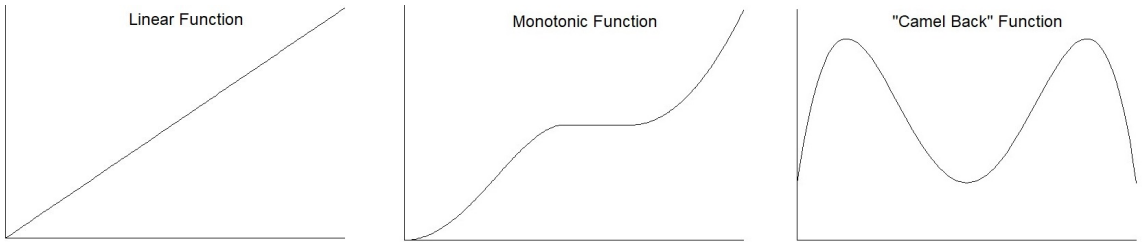


Figure 13: Quasiconvexity and Quasiconcavity. For more pictures, see https://en.wikipedia.org/wiki/Quasiconvex_function.

As a final note of caution before moving on, an established result is that convex and concave functions are continuous. This was *not* the property we wanted to maintain when coming up with our definitions of quasi-convexity, and indeed, there are quasi-convex or quasi-concave functions that are discontinuous, e.g. indicator functions such as $\mathbb{1}[x > 0]$.

3.2 Multivariate Differential Calculus

Wikipedia provides a good explanation of what calculus actually is about:[9]

> "Calculus [...] is the *mathematical study of continuous change*, in the same way that geometry is the study of shape and algebra is the study of generalizations of arithmetic operations. It has two major branches, differential calculus and integral calculus. Differential calculus concerns instantaneous rates of change and the slopes of curves. Integral calculus concerns accumulation of quantities and the areas under and between curves. These two branches are related to each other by the fundamental theorem of calculus [stating that differentiation and integration are inverse operations]."

The remainder of this chapter is concerned with introducing you to both of these branches in the context of multivariate functions, with greater emphasis on differential calculus, which surpasses integral calculus in importance in most economic disciplines (an exception is theoretical econometrics) any may indeed be one of, if not the most important mathematical concept an economist should be well-familiar with. This is because not only is differentiation at the heart of our favorite exercise, namely constrained optimization (as discussed in the next chapter), but also, fundamentally important concepts such as marginal utility or marginal cost are based upon the derivative.

3.2.1 Basics and Review of Univariate Differential Calculus

As repeatedly done before, let's start from the most simple case – and the one we are hopefully all at least somewhat familiar with: univariate real-valued functions $f : \mathbb{R} \mapsto \mathbb{R}$. In the next step, we will again be concerned with generalization of this concept. Now, when asked what the *instantaneous rate of change*, or *slope*, of f at $x_0 \in X$ is, how would you go about and answer this? Here, we don't mean to simply give a rule how to determine the slope, but rather, to

concave functions.

[9] https://en.wikipedia.org/wiki/Calculus, accessed August 03, 2019.

conceptually and formally describe the concept of the slope for a general function f at an arbitrary point x_0! One common characterization is that the slope tells us how sensitive f is to changes in x, i.e. how much $f(x)$ varies relative to the variation in x. You may have also heard (from Wikipedia) that the slope at x_0 corresponds to the rate of change in f associated with an infinitely small change in in the argument – the *marginal* rate of change in f. But how do we write this down formally? Let's consider $h := x - x_0$ for a *fixed* $x \in X$, $x \neq x_0$. Then clearly, h is equal to a fixed real number and $|h| > 0$, so that h is not "infinitely small". But this consideration is very helpful because it allows to characterize the *relative change* of f, i.e. the ratio of the variation in f and the one in the argument when moving from x_0 to x:

$$\frac{\Delta f(x)}{\Delta x} := \frac{f(x) - f(x_0)}{x - x_0} = \frac{f(x_0 + h) - f(x_0)}{h}. \tag{8}$$

Now, we know the relative change for any *fixed* change $h \in \mathbb{R}$. This suggests that, when concerned with finding the relative change induced by a *marginal*, i.e. infinitely small variation in x, we should be able to derive it from letting h go to zero in equation (8)! Indeed, this is exactly how we proceed to define the derivative – we just have to be careful about one detail: the expression in equation (8) is always well-defined for fixed $h > 0$; a limit, on the other hand, is not guaranteed to exist.

Definition 64. ***(Differentiability and Derivative of a Univariate Real-Valued Function)*** *Let $X \subseteq \mathbb{R}$ and consider the function $f : X \mapsto \mathbb{R}$. Let $x_0 \in X$. If*

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

*exists, f is said to be differentiable at x_0, and we call this limit the **derivative of** f **at** x_0, denoted by $f'(x_0)$. If for all $x_0 \in X$, f is differentiable at x_0, f is said to be differentiable over X or differentiable. If f is differentiable, the function $f' : X \mapsto \mathbb{R}, x \mapsto f'(x)$ is called the **derivative** of f.*

Note the following crucial distinction: the derivative of f at x_0, $f'(x_0)$, is a limit and takes a value in $\mathbb{R}$, i.e. it is a real number. On the other hand, the derivative of f, f', like f is a *function* that maps from X to $\mathbb{R}$! To take away, in words, we define the derivative by first looking at a fixed change h and then study what happens to $\Delta f(x)/h$ if h becomes infinitely small. If (and only if) we arrive at the same, well-defined rate $f'(x_0)$ regardless of how we let h approach 0, this rate of marginal change is unique and well-defined, and we can use it to infer on the function's behavior at x_0. This concept is extremely helpful because it allows us to study the *local* behavior of functions (i.e. in small neighborhoods around fixed points x_0) even if we can not graphically represent them anymore – hence, it lets symbols and equations become our eyes when we can no longer draw the objects we are interested in!

3.2.2 Notation and Conceptual Foundations of Differential Calculus

Above, we had already seen two levels of concepts at the heart of differential calculus: a derivative of a function at a point in the domain (*real number*) and a derivative *function* mapping points onto the derivative of the function at them. The third, and highest level concept is the **derivative operator**, a mapping between function spaces, associating the function f with its

derivative f'. As the derivative operator can associate a derivative function only with functions f that are indeed differentiable, its domain corresponds to the set of once differentiable functions:

Definition 65. ***(Set of k times Differentiable Functions)*** *Let $X \subseteq \mathbb{R}$. Consider a differentiable function $f : X \mapsto \mathbb{R}$. If its derivative f' is also differentiable, we say that f is two times differentiable, and call the derivative f'' of f' the second derivative of f. In analogy, we define $f^{(k)}$, the k-th derivative of f, recursively as the derivative of $f^{(k-1)}$, the $k-1$-st derivative of f. If $f^{(k)}$ exists, we call f k times differentiable. For any $k \in \mathbb{N}$, we define*

$$D^k(X, \mathbb{R}) = \{f : X \mapsto \mathbb{R} : f \text{ is } k \text{ times differentiable over } X\}$$

as the set of univariate real-valued functions with domain X that are k times differentiable. Moreover, we define

$$C^k(X, \mathbb{R}) = \{f \in D^k(X, \mathbb{R}) : f^{(k)} \text{ is continuous}\}$$

as the set of ***k times continuously differentiable functions****, i.e. k times differentiable functions with continuous k-th derivative $f^{(k)}$.*
If $X = \mathbb{R}$, we write $D^k(X, \mathbb{R}) = D^k(\mathbb{R})$ and $C^k(X, \mathbb{R}) = C^k(\mathbb{R})$.

Now we know the domain of the differential operator. By the way, $f \in D^1(X, \mathbb{R})$ is a convenient way of writing that f is a differentiable function mapping from X to $\mathbb{R}$ – if domain and codomain coincide, e.g. $X = Y = \mathbb{R}$, we write $f \in D^1(\mathbb{R})$. This notation is used more generally: e.g. if f is twice differentiable, i.e. the derivative of f is also differentiable, then we write $f \in D^2(X, \mathbb{R})$. Note that any twice differentiable function is especially once differentiable, such that $\forall f \in D^2(X, \mathbb{R}) : f \in D^1(X, \mathbb{R})$! A similar notation $C^n(X, \mathbb{R})$ is used to indicate that f is not only n times differentiable, but that the n-th derivative is continuous.

As for the codomain of differential operator, i.e. the set containing the derivative functions, we impose no restrictions but its function property, so that we simply consider the space of all functions mapping from X to $\mathbb{R}$, which we denote as F_X:

$$F_X := \{f : X \mapsto \mathbb{R}\}.$$

This gives us everything we need to define the differential operator:

Definition 66. ***(Differential Operator for Univariate Real-Valued Functions)*** *Let $X \subseteq \mathbb{R}$. Then, the differential operator is defined as the function*

$$\frac{d}{dx} : D^1(X, \mathbb{R}) \mapsto F_X, f \mapsto f'$$

where f' denotes the derivative of $f \in D^1(X, \mathbb{R})$.

Take the time to appreciate what this means. While the definition appears rather straightforward, it encompasses two details that are frequently missed even in advanced textbooks and university lecture material, but that anyone claiming to have a proper command of mathematics should be well-aware of. First, the derivative and the differential operator are *not* the same

thing, indeed, they are fundamentally different functions, because one maps between function spaces and the other between real vector spaces.[10] The precise relationship is that **the derivative of a specific function** f (in the domain $D^1(X, \mathbb{R})$ of the differential operator) **is a specific value (in the codomain) of the differential operator!**

Secondly, you frequently see the expressions

$$\frac{df}{dx}(x) \quad \text{and} \quad \frac{df(x)}{dx} \tag{9}$$

Without further explanation, note that these quantities are not defined! So what should we make of them? Despite their frequent use, things are actually a bit tricky; the deliberations to follow give a thorough discussion. If you care less about these technical details, it is fine if you just memorize the take-away, as summarized below.

The formally correct way of writing the mapping rule of the derivative function, i.e. the rule $x \mapsto y = f'(x)$ of the function f', is:

$$f'(x) = \left[\frac{d}{dx}(f)\right](x).$$

This states that we first evaluate the differential operator at f to obtain the derivative f': $f' = \frac{d}{dx}(f)$; the resulting function is then evaluated at x. Because this looks a bit weird, one commonly writes $f' = \frac{df}{dx}$, so that the first expression in equation (9) represents as a justified notational convention for evaluating the derivative function at specific points $x \in X$, or respectively, when x is interpreted as a variable argument, as the mapping rule $x \mapsto y = f'(x)$ of the derivative function.

The second expression in equation (9) is supposed to refer to the same object. However, this is arguably problematic: $f(x)$ is not a function (like f), but rather a specific value in $\mathbb{R}$, or alternatively, a description of the mapping rule $x \mapsto y = f(x)$ of f, as e.g. in $\frac{d(x^2+sin(x))}{dx}$), and thus *not an element in the domain of the differential operator!* [11] As such, this notation blurs the lines between the concepts of functions and mapping rules, and may easily cause conceptual misunderstandings. Hence, this course will not make use of this notation.

Still, in concrete applications, it may be more convenient to work with the mapping rules directly as a reduced form representation of the function f, especially when it is clear what the domain f is. For instance, suppose that we want to compute the derivative of $f : \mathbb{R} \mapsto \mathbb{R}, x \mapsto x^2 + \sin(x)$. Here, we know that the derivative of f will, like f, have domain $\mathbb{R}$, and to fully characterize it, we only need to compute its mapping rule. We write this rule as $\frac{d}{dx}f(x)$, where $f(x)$ is the mapping rule of f. Then, it is formally correct to write:

$$\frac{d}{dx}f(x) = \frac{d}{dx}(x^2 + \sin(x)) = \frac{d}{dx}x^2 + \frac{d}{dx}\sin(x) = 2x + \cos(x)$$

To summarize, in practice,

[10] Of course, the domain need not always be a space, e.g. when we deal with functions $\mathbb{N} \mapsto \mathbb{R}$. More precisely, one should only talk about sets here.

[11] Even if we think of $f(x)$ as the mapping rule with variable argument, it is still not clear what the domain of the resulting object should be.

1. It is justified to pull the *function* onto the numerator of the derivative operator, as in $\frac{df}{dx}(x)$.

2. It is imprecise and not advisable to pull the mapping rule onto the numerator of the derivative operator, as in $\frac{df(x)}{dx}$, e.g. $\frac{d(x^2+\sin(x))}{dx}$.

3. When working with mapping rules as a reduced form representation of the function, as e.g. in $f(x) = x^2 + \sin(x)$, the correct way to express the derivative's mapping rule is $\frac{d}{dx}f(x) = \frac{d}{dx}(x^2 + \sin(x))$.

Generally, to express yourself as unambiguously as possible, you are well-advised to omit the argument x completely if possible, and in any case to not write it in the numerator of the derivative expression.

To get used to this (perhaps less familiar, but in advanced texts more prominent!) way of handling derivatives, as an exercise, let us re-state the central rules for derivatives of univariate real-valued functions we considered in Chapter 0, Table 6. Before doing so, because this is the first time we formally deal with *function spaces*, we first need to define the concept formally and introduce the basis operations we will use.[12]

Theorem 36. ***(Basis Operations in Spaces of Vector Functions)*** *Let F be a set of functions with domain $X \subseteq \mathbb{R}^n$ and codomain $Y \subseteq \mathbb{R}^m$, $n, m \in \mathbb{N}$. Then, we usually consider the operations "+" and "$\cdot$" which are such that for any $f, g \in F$ and any $\lambda \in \mathbb{R}$, $f + g : ((f + g)(x) = f(x) + g(x) \forall x \in X)$ and $\lambda f : ((\lambda f)(x) = \lambda \cdot f(x) \forall x \in X)$. If F is closed under these operations, then $\mathbb{F} := (F, +, \cdot)$ constitutes a vector space.*

This is a theorem rather than a definition because it asserts that the operations satisfy properties *(ii)* to *(vii)* of Definition 5 of a real vector space. You can verify this as an exercise, it's rather straightforward. Further, in this space, *vector multiplication* is defined as $f \cdot g : ((fg)(x) = f(x)g(x) \forall x \in X)$, and *vector division* as $f/g : ((f/g)(x) = f(x)/g(x) \forall x \in X)$, where the quotient f/g exists at x_0 (generally) if $g(x_0) \neq 0$ $(\forall x_0 \in X)$. Now, for the derivative rules:

Theorem 37. ***(Rules for Univariate Derivatives)*** *Let $X \subseteq \mathbb{R}$, $f, g \in D^1(X, \mathbb{R})$ and $\lambda, \mu \in \mathbb{R}$. Then,*

(i) (Linearity) $\lambda f + \mu g$ is differentiable and $\frac{d}{dx}(\lambda f + \mu g) = \lambda \frac{df}{dx} + \mu \frac{dg}{dx}$,

(ii) (Product Rule) The product fg is differentiable and $\frac{d}{dx}(fg) = \frac{df}{dx} \cdot g + f \cdot \frac{dg}{dx}$

(iii) (Quotient Rule) If $\forall x \in X$, $g(x) \neq 0$, the quotient f/g is differentiable and $\frac{d}{dx}(f/g) = \frac{\frac{df}{dx} \cdot g - f \cdot \frac{dg}{dx}}{g \cdot g}$

(iv) (Chain Rule) If all the following expressions are well-defined, then $g \circ f$ is differentiable and $\frac{d}{dx}(g \circ f) = \left(\frac{dg}{dx} \circ f\right) \cdot \frac{df}{dx}$.

Note that **all expressions** in Theorem 37 are **sums, products and compositions of functions** and therefore functions themselves! To make this circumstance even more clear, compare the statement of Theorem 37 to the respective relationships for **derivatives *at specific points***, which are no longer functions, but **values in** $\mathbb{R}$, given below in Theorem 38 (pay attention to where we put the argument x: never in the numerator of the differential expression!).

[12] If you have carefully read the previous chapter, you should already be familiar with them.

Theorem 38. ***(Rules for Univariate Derivatives at Specific Points)*** *Let $X \subseteq \mathbb{R}$, $f, g : X \mapsto \mathbb{R}$ and $\lambda, \mu \in \mathbb{R}$. Let $x_0 \in X$ and suppose that f and g are differentiable at x_0. Then,*

(i) (Linearity) $\lambda f + \mu g$ is differentiable at x_0 and $\frac{d(\lambda f + \mu g)}{dx}(x_0) = \lambda \frac{df}{dx}(x_0) + \mu \frac{dg}{dx}(x_0)$,

(ii) (Product Rule) The product fg is differentiable at x_0 and $\frac{dfg}{dx}(x_0) = \frac{df}{dx}(x_0) \cdot g(x_0) + f(x_0) \cdot \frac{dg}{dx}(x_0)$,

(iii) (Quotient Rule) If $g(x_0) \neq 0$, then quotient f/g is differentiable at x_0 and $\frac{df/g}{dx}(x_0) = \frac{\frac{df}{dx}(x_0) \cdot g(x_0) - f(x_0) \cdot \frac{dg}{dx}(x_0)}{g(x_0)^2}$,

(iv) (Chain Rule) if all the following expressions are well-defined, then $g \circ f$ is differentiable at x_0 and $\frac{d(g \circ f)}{dx}(x_0) = \left[\frac{dg}{dx} \circ f\right](x_0) \cdot \frac{df}{dx}(x_0)$.

Before moving on, make sure that you are thoroughly familiar with the three conceptual levels of differential calculus and know the differences between them (here summarized from highest to lowest level):

1. *Derivative Operator* $\frac{d}{dx}$: Function that maps between function spaces; maps differentiable functions onto their derivative function,
2. *Derivative Function* $\frac{df}{dx}$: Function that maps between spaces of real vectors: maps arguments x of a function f onto the derivative of f evaluated at x,
3. *Derivative at a Point* $\frac{df}{dx}(x_0)$: Element of a real vector space (e.g. $\mathbb{R}^n$): concrete value of the derivative function at a specific point.

3.2.3 From Univariate to Multivariate Derivatives

With the conceptual foundation laid, we are now prepared to have a look at how precisely we can derive insight about the local behavior of a function from computing the derivative. For what follows, let f be a function with domain $X \subseteq \mathbb{R}$ and codomain $\mathbb{R}$.

The existence and value of the derivative of a function f gives us three important pieces of information about f:

1. If f is differentiable at $x_0 \in X$, then f is also continuous at x_0.

Proof. Suppose that f is differentiable at $x_0 \in X$. Recall that in Chapter 0, we characterized continuity at x_0 by $\lim_{x \to x_0} f(x) = f(x_0)$, so this is what we want to show. Now, consider the derivative of f at x_0,

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

This gives

$$0 = 0 \cdot f'(x_0) = \lim_{x \to x_0} (x - x_0) \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \left((x - x_0) \frac{f(x) - f(x_0)}{x - x_0} \right) = \lim_{x \to x_0} [f(x) - f(x_0)].$$

where third equality follows from the product rule of the limit. Because $f(x_0)$ is a constant, $\lim_{x \to x_0} [f(x) - f(x_0)] = [\lim_{x \to x_0} f(x)] - f(x_0)$, and the equation above becomes

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

□

2. If f is differentiable at $x_0 \in X$, then there exists a "good" linear approximation to f around x_0, called the Taylor Approximation.

We like linear functions because they are simple and we know how they work. Unfortunately, it is not likely that the functions involved in our applications be linear. If a given function is too complex to handle but differentiable around a point of interest, a good solution is often to work with a local linear approximation to the function rather than the function itself, and rely on the following result that ensures that when x is "close enough" to x_0, the linear approximation based on the derivative is "sufficiently good": let's make no further restrictions on f than assuming differentiability and else allow it to be any (arbitrarily complex and potentially erratically-behaving) function. Consider the following *approximation to f at* $x_0 \in X$:

$$T_{1,x_0}(x) = f(x_0) + f'(x_0)(x - x_0),$$

the so-called Taylor-Approximation to f at x_0 of order 1 (because we only take the first derivative). This expression is a linear function with the *known values* $f(x_0)$ as intercept and $f'(x_0)$ as slope, with the difference to the point of investigation, $x - x_0$, as the variable argument. Now, what do we mean when we say that this is a "good approximation around x_0"? Denote by $\varepsilon_1(x) := T_{1,x_0}(x) - f(x)$ the error we make when approximating f using T_{1,x_0} at $x \in X$. Because f is an arbitrary function, when x is far away from x_0, this error may be quite large – however, as we approach x_0, the error becomes negligibly small relative to the distance $x - x_0$! Formally:

$$\frac{\varepsilon_1(x)}{x - x_0} = \frac{f(x_0) + f'(x_0)(x - x_0) - f(x)}{(x - x_0)} = f'(x_0) - \frac{f(x) - f(x_0)}{(x - x_0)} \quad \overset{x \to x_0}{\to} \quad 0.$$

The graphical intuition is illustrated in Figure 14, where the point x_1 is too far away from x_0 for approximation quality to be decent, but x_2 is close enough to x_0 such that the Taylor approximation and the true function almost coincide – note especially that $|\varepsilon_1(x_2)|$ is much smaller than $|x_2 - x_0|$.

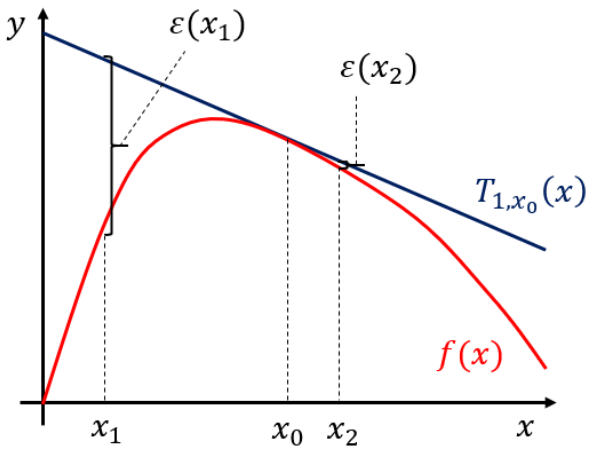


Figure 14: Quality of the Taylor Approximation.

In practice, however, we usually don't know how close is close enough – the Taylor approximation is just a limit statement for moving *infinitely close* to the point x_0, and for specific functions, even at small but fixed distances $x - x_0 = 0.0000001$, the difference may be quite

large, so treat this result with a grain of caution.

Note also that, if f is more than once differentiable, we can get an even better approximation by taking higher order derivatives, and computing the Taylor approximation as a polynomial of higher degree – the higher the polynomial order, the more flexible the function and the better the approximation.[13] Because this concept will be helpful repeatedly, let's take the time to consider the formal definition.

Theorem 39. ***(Taylor Expansion for Univariate Functions)*** *Let $X \subseteq \mathbb{R}$ and $f \in D^k(X, \mathbb{R})$ where $k \in \mathbb{N} \cup \{\infty\}$, i.e. f is k times differentiable. Then, for $N \in \mathbb{N} \cup \{\infty\}$, $N \leq k$, the Taylor expansion of order N for f at $x_0 \in X$ is*

$$f(x) = T_{N,x_0}(x) + \varepsilon_N(x) = f(x_0) + \sum_{n=1}^{N} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \varepsilon_N(x),$$

*where $\varepsilon_N(x)$ is the approximation error of T_{N,x_0} for f at $x \in X$, and $n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$ denotes the **faculty** of n. Then, the approximation quality satisfies $\lim_{h \to 0} \varepsilon_N(x_0 + h)/h^N = 0$. Further, if f is $N+1$ times differentiable, there exists a $\lambda \in (0,1)$ such that*

$$\varepsilon_N(x_0 + h) = \frac{f^{(N+1)}(x_0 + \lambda h)}{(N+1)!} h^{N+1}.$$

Some remarks are worth noting: (i) in contrast to the Taylor *approximation* $T_{N,x_0}(x)$ for f at x_0, the Taylor *expansion* is always equal to the function value $f(x)$ because it encompasses the approximation error as an unspecified object, (ii) when considering small deviations from x_0 rather than arbitrary points x, it may at times be more convenient to express the expansion directly in terms of the deviation $h = x - x_0$ rather than x:

$$f(x_0 + h) = T_{N,x_0}(x_0 + h) + \varepsilon_N(x_0 + h) = f(x_0) + \sum_{n=1}^{N} \frac{f^{(n)}(x_0)}{n!} h^n + \varepsilon_N(x_0 + h),$$

and (iii) $\lim_{h \to 0} \varepsilon_N(x_0 + h)/h^N = 0$, says that higher N yield better approximations because h^N goes "faster" to zero the larger h is. To see this, consider a small h, e.g. $h = 0.01$, and compute h^1, $h^2 = 0.0001$, $h^4 = 10^{-8}$ etc. Thus, when considering larger N, we can divide the error by ever smaller numbers and still have convergence to zero – indeed, as $N \to \infty$, because $\lim_{N \to \infty} h^N = 0$ for any $h < 1$, the Taylor approximation of infinite order $N = \infty$ yields perfect approximation so that $\varepsilon_\infty(x) = 0 \ \forall x \in X$ (one can also show this more formally)!

An immediate corollary, and nonetheless a very useful one, of the Taylor expansion theorem is the following:

Corollary 6. ***(Mean Value Theorem)*** *Let $X \subseteq \mathbb{R}$ and $f \in D^1(X, \mathbb{R})$, i.e. f is a differentiable function. Then, for any $x_1, x_2 \in X$ such that $x_2 > x_1$, there exists $x^* \in (x_1, x_2)$ such that*

$$f'(x^*) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

[13] For instance, our graphical example looks rather close to $-cx^2 + d$, so that a second order approximation should fare much better on a wider neighborhood of x_0.

Proof. Consider the Taylor expansion of order 0 at x_1, evaluated at x_2:

$$f(x_2) = f(x_1 + (x_2 - x_1)) = f(x_1) + \varepsilon_{0,x_1}(x_1 + (x_2 - x_1)).$$

Because f is differentiable, there exists $\lambda \in (0,1)$ such that $\varepsilon_{0,x_1}(x_1 + (x_2 - x_1)) = f'(x^*)(x_2 - x_1)$ with $x^* = x_1 + \lambda(x_2 - x_1) \in (x_1, x_2)$. Subtracting $f(x_1)$ on both sides on the above equation and plugging in the expression for $\varepsilon_{0,x_1}(x_1 + (x_2 - x_1))$,

$$f(x_2) - f(x_1) = f'(x^*)(x_2 - x_1) \quad \Leftrightarrow \quad f'(x^*) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$ □

We will see later that this theorem is incredibly helpful for establishing existence of (local) maxima and minima that satisfy a first order condition $f'(x) = 0$ – indeed, you can already see here that all we need is two different points in X with the same value for the *differentiable* function f.

Excursion. The second derivative and proving the Taylor Expansion Theorem for $N = 2$. The reasoning above has established Taylor's Theorem for $N = 1$. Establishing the general theorem requires an inductive approach; insight into this procedure can be obtained from considering the proof for the second order expansion. Following this proof may also help you to more formally understand higher order derivatives of univariate functions, so it may be worthwhile looking at. You may know that the second derivative can be obtained from differentiating the first derivative – this is precisely how we define it formally: Let $f \in D^2(X, \mathbb{R})$, $X \subseteq \mathbb{R}$. Then, the second derivative of f at $x_0 \in X$ is

$$f''(x_0) = \lim_{h \to 0} \frac{f'(x_0 + h) - f'(x_0)}{h}.$$

The following characterization will be helpful for understanding why the Taylor theorem holds for $N = 2$:

$$\lim_{h \to 0} \left\{ \frac{1}{h} \left(\frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) \right) \right\} = \frac{1}{2} f''(x_0). \tag{10}$$

To see why this holds, note that the LHS limit satisfies L'Hôpital's rule: both h and the expression in round brackets go to zero as $h \to 0$. Thus, we take the derivative of both expressions to obtain the limit – the denominator becomes one, and for the numerator, we apply the quotient rule to the first summand (the second summand does not depend on h and drops out):[a]

$$\begin{aligned}
\frac{d}{dh} \left(\frac{f(x_0 + h) - f(x_0)}{h} \right) &= \frac{f'(x_0 + h)h - (f(x_0 + h) - f(x_0))}{h^2} \\
&= \frac{f'(x_0 + h) - f'(x_0)}{h} + \frac{f'(x_0)}{h} - \frac{(f(x_0 + h) - f(x_0))}{h^2} \\
&= \frac{f'(x_0 + h) - f'(x_0)}{h} - \frac{1}{h} \left(\frac{(f(x_0 + h) - f(x_0))}{h} - f'(x_0) \right).
\end{aligned}$$

Therefore, we obtain

$$\lim_{h\to 0}\left\{\frac{1}{h}\left(\frac{f(x_0+h)-f(x_0)}{h}-f'(x_0)\right)\right\}=\lim_{h\to 0}\left\{\frac{f'(x_0+h)-f'(x_0)}{h}-\frac{1}{h}\left(\frac{(f(x_0+h)-f(x_0))}{h}-f'(x_0)\right)\right\}$$

Adding $\lim_{h\to 0}\left\{\frac{1}{h}\left(\frac{f(x_0+h)-f(x_0)}{h}-f'(x_0)\right)\right\}$ on both sides,

$$2\lim_{h\to 0}\left\{\frac{1}{h}\left(\frac{f(x_0+h)-f(x_0)}{h}-f'(x_0)\right)\right\}=\lim_{h\to 0}\frac{f'(x_0+h)-f'(x_0)}{h}=f''(x_0).$$

Dividing by 2, this gives equation (10).

Now, coming to the Taylor expansion theorem:

$$\begin{aligned}\frac{\varepsilon_2(x_0+h)}{h^2}&=\frac{T_{2,x_0}(x_0+h)-f(x_0+h)}{h^2}=\frac{f(x_0)+f'(x_0)h+\frac{f''(x_0)}{2}h^2-f(x_0+h)}{h^2}\\&=\frac{f''(x_0)}{2}-\frac{1}{h}\left(\frac{f(x_0+h)-f(x_0)}{h}-f'(x_0)\right)\end{aligned}$$

By the sums rule of the limit and equation (10), it follows that $\lim_{h\to 0}\frac{\varepsilon_2(x_0+h)}{h^2}=0$.

[a] The first equality follows from chain rule: $\frac{d}{dh}f(x_0+h)=f'(x_0+h)\cdot\left(\frac{d}{dh}\{x_0+h\}\right)=f'(x_0+h)\cdot 1=f'(x_0+h)$.

3. If f is differentiable on the interval $(a,b)\subseteq\mathbb{R}$, then

(i) f **is constant on** (a,b) **if and only if** $\forall x_0\in(a,b): f'(x_0)=0$**,**

(ii) f **is monotonically increasing (decreasing) on** (a,b) **if and only if** $\forall x_0\in(a,b): f'(x_0)\geq 0$ $(f'(x_0)\leq 0)$**,**

(iii) If $\forall x_0\in(a,b): f'(x_0)>0$ $(f'(x_0)<0)$**, then** f **is strictly monotonically increasing (decreasing) on** (a,b)**.**

Proof for (i). "$\Rightarrow$" Suppose that f is constant on (a,b), i.e. $\forall x,y\in(a,b): f(x)=f(y)=c$. Then, because $x_0+h\in(a,b)$ for h small enough, for any $x_0\in X$:

$$f'(x_0)=\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}=\lim_{h\to 0}\frac{c-c}{h}==\lim_{h\to 0}0=0.$$

"$\Leftarrow$" Suppose that $\forall x_0\in X: f'(x_0)=0$. Note the Taylor representation theorem: if $f\in D^\infty(X,\mathbb{R})$ then $\forall x\in(a,b)$, $f(x)=f(x_0)+\sum_{n=1}^{n}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$.[14] However, because $f'(x_0)=f^{(1)}(x_0)=0$, it follows that $f^{(n)}(x_0)=0\forall n\in\mathbb{N}$, and $\forall x\in(a,b)$,

$$f(x)=f(x_0)+\sum_{n=1}^{n}\frac{0}{n!}(x-x_0)^n=f(x_0).$$ □

You can try to prove the remaining points on your own or consult your favorite mathematical textbook or website, but really, the fundamental message lies in the statements of 3., rather in the details of why they are true.

[14] Don't worry that this may look rather complex - verbally, this is easy to grasp: it just says that if it exists, a Taylor approximation of infinite order perfectly approximates the function.

Note that 3.*(iii)* only provides a sufficient condition for strict monotonicity. As we will also see in the next chapter, there are strictly monotonic functions (e.g. $f(x) = x^3$ on $\mathbb{R}$) that have a zero derivative at some points in their domain.

Now that we have convinced ourselves that derivatives of univariate functions are extremely helpful in characterizing them when they exist, we are concerned with transferring these concepts and their intuitions to multivariate functions. In fact, this is all we will do in the remainder of this section – next to generalizing the Taylor theorem to higher orders and multivariate functions!

3.2.4 Partial Derivatives and the Gradient

Now, lets move away from univariate real-valued functions but maintain $\mathbb{R}$ as the co-domain – that is, let us consider multivariate real-valued functions of the form $f : \mathbb{R}^n \mapsto \mathbb{R}$. You are already familiar with many examples, e.g. the scalar product, any norm, and any matrix function of the form $x \mapsto x'Ax$ with $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Now, how can we generalize the concept of the derivative? The fundamental issue is that unlike with the $\mathbb{R}$, when considering the $\mathbb{R}^n$, we can move away from x_0 in multiple directions, and it is no longer clear what precisely we mean by a "small change". Still, you will see shortly that this issue can be resolved with only a slight conceptional twist.

To approach this issue, let's first stick to what we know: when fixing a point $x_0 = (x_{0,1}, \ldots, x_{0,n})' \in \mathbb{R}^n$, for any position $j \in \{1, \ldots, n\}$, we can consider the function

$$\tilde{f}_{j,x_0}(t) = f(x_{0,1}, \ldots, x_{0,j-1}, x_{0,j} + t, x_{0,j+1}, \ldots, x_{0,n})$$

that takes the point x_0 as given and only varies f by varying the j-th entry of the argument. The appeal is that for a given $x_0 \in \mathbb{R}^n$, this function maps $\mathbb{R}$ into $\mathbb{R}$, i.e. it is a univariate real-valued function that we know how to take the derivative of! Indeed, the derivative of this function is of crucial importance of everything that follows, which is why we take the time to formally define it and give it an explicit name:

Definition 67. ***(Partial Derivative)*** *Consider a function $f : X \mapsto \mathbb{R}$ where $X \subseteq \mathbb{R}^n$, and let $x_0 \in X$. Then, if for $j \in \{1, \ldots, n\}$, the function $\tilde{f}_{j,x_0} : \mathbb{R} \mapsto \mathbb{R}, t \mapsto f(x_{0,1}, \ldots, x_{0,j-1}, x_{0,j} + t, x_{0,j+1}, \ldots, x_{0,n})$ is differentiable at $t = 0$, we say that f is partially differentiable at x_0 with respect to (the j-th argument) x_j, and we call $\frac{d\tilde{f}_{j,x_0}}{dt}(0)$ the partial derivative of f at x_0 with respect to x_j, denoted by $\frac{\partial f}{\partial x_j}(x_0)$ or $f_j(x_0)$.*

A few comments on this definition are worthwhile considering. First, as with the univariate derivative, also the partial derivative is described by application of an operator mapping between function spaces, namely

$$\frac{\partial}{\partial x_j} : D_j^1(X, \mathbb{R}) \mapsto F_X, f \mapsto \frac{\partial f}{\partial x_j}$$

where $D_j^1(X, \mathbb{R})$ is the set of real-valued functions with domain X that are partially differentiable with respect to the j-th argument.[15]

[15] To bring across the conceptual distinction between operator, function and value as clearly as possible, the

As with the derivative of the univariate function, note that $\frac{\partial}{\partial x_j}$ is an *operator*, $\frac{\partial f}{\partial x_j}$ is a *real-valued function* and $\frac{\partial f}{\partial x_j}(x_0)$ is a *real number*. Next, clearly, when $n = 1$, then the partial derivative and the derivative defined for univariate functions coincide, because there is only one direction x_j. Otherwise, if $n > 1$, the derivative defined previously is no longer applicable, while the partial derivative is – this is precisely why we introduced the concept.[16]

Less formally, the j-th partial derivative is the derivative you take of f when you treat all x_l, $l \neq j$ as *constants* rather than variables, and differentiate the function as if it had only one variable argument, namely x_j. To get some feeling for this concept, let's consider some examples: $f^1(x_1, x_2) = x_1 + x_2$, $f^2(x_1, x_2) = x_1 x_2$ and $f^3(x_1, x_2) = x_1 x_2^2 + \cos(x_1)$. Then, the partial derivatives are (switching between the two possible notations to get you used to both of them):

$$f_1^1(x_1, x_2) = 1, \qquad f_2^1(x_1, x_2) = 1,$$

$$\frac{\partial f^2}{\partial x_1}(x_1, x_2) = x_2, \qquad \frac{\partial f^2}{\partial x_2}(x_1, x_2) = x_1,$$

$$f_1^3(x_1, x_2) = x_2^2 - \sin(x_1), \qquad f_2^3(x_1, x_2) = 2x_1 x_2.$$

As you see, the partial derivatives can include none, some, or all of the components of the vector x! Generally, the point to be made is that simply because we are taking the derivative into one direction (say x_1, the other components (here: x_2, more generally, $x_2, x_3, \ldots, x_n$) do not drop out because they may interact *non-linearly* in f! Only if terms containing x_j are *strictly linearly separable* (e.g. f^1 or $\cos(x_1)$ in f^3), they will drop out when taking the partial derivative with respect to a different x_l, $l \neq j$.

An object that you will see frequently when concerned with differential calculus is the so-called gradient. It is nothing but the *ordered collection* of all partial derivatives of f at $x_0 \in X$:

Definition 68. ***(Gradient)*** *Consider a function $f : X \mapsto \mathbb{R}$ where $X \subseteq \mathbb{R}^n$, and let $x_0 \in X$. Then, if for all $j \in \{1, \ldots, n\}$, f is partially differentiable with respect to x_j at x_0, then we call the* ***row vector***

$$\nabla f(x_0) = (f_1(x_0), f_2(x_0), \ldots, f_n(x_0))$$

the gradient of f at x_0. If for all $x_0 \in X$ and for all f is partially differentiable with respect to x_j at x_0, then we call the ***function*** $\nabla f : \mathbb{R}^n \mapsto \mathbb{R}^{1 \times n}, x_0 \mapsto \nabla f(x_0)$ *the gradient of f.*

Can you already guess the first comment on this definition? If not, look at the words in bold in the definition and try to guess again.

$\ldots \nabla f(x_0)$ **is a real (row) vector and** ∇f **is a function!** The function exists under much stronger conditions than the vector – it requires all partial derivatives to exist at all points, whereas for a specific point x_0, existence of the gradient at this point requires only existence of

elaborations below also always define the respective operator of interest. Unfortunately, the operator's domain will continue to look somewhat ugly – don't worry about this too much, though, just remember that the domain is always a set of functions that requires the operator to be well-defined and ignore the unwieldy notation.

[16] At https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivatives/v/partial-derivatives-and-graphs, you can find a video which provides a graphic illustration of partial derivatives. You might find this helpful for developing an intuition for the concept, but note that a graphic illustration is again only possible for low-dimensional function, whereas it was our aim to generalize the concept of differentials to higher dimensional spaces.

all partial derivatives at this specific point. Finally, we can again go one step further and define the gradient **operator**

$$\nabla : D^1_{partial}(X,\mathbb{R}) \mapsto F_X, f \mapsto \nabla f, \text{ where } D^1_{partial}(X,\mathbb{R}) = \{f : X \mapsto \mathbb{R} : (\nabla f(x_0) \text{ exists } \forall x_0 \in X)\}.$$

Before discussing the ultimate punchline of this subsection, let's go one more step further and consider a multivariate function that is not necessarily real-valued, but more generally maps into $\mathbb{R}^m$ with $m \in \mathbb{N}$, i.e. $f : \mathbb{R}^n \mapsto \mathbb{R}^m$. To see how we extend the collection of partial derivatives to this function, note that we may write it as

$$f = \begin{pmatrix} f^1 \\ f^2 \\ \vdots \\ f^m \end{pmatrix} \text{ so that } \forall x \in \mathbb{R}^n : f(x) = \begin{pmatrix} f^1(x) \\ f^2(x) \\ \vdots \\ f^m(x) \end{pmatrix} \tag{11}$$

where for any $i \in \{1,\dots,m\}$, $f^i : \mathbb{R}^n \mapsto \mathbb{R}$ is a multivariate real-valued function. Can you guess the comment on this equation?[17] Let's see how we stack the partial derivatives of all these functions into a collecting object:

Definition 69. ***(Jacobian)*** *Let $n,m \in \mathbb{R}^n$, $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}^m$ and for $i \in \{1,\dots,m\}$, let $f^i : X \mapsto \mathbb{R}$ such that $f = (f^1,\dots,f^m)'$. Let $x_0 \in X$. Then, if at x_0, $\forall i \in \{1,\dots,m\}$, f^i is partially differentiable with respect to any x_j, $j \in \{1,\dots,n\}$, we call*

$$J_f(x_0) = \begin{pmatrix} \nabla f^1(x_0) \\ \nabla f^2(x_0) \\ \vdots \\ \nabla f^m(x_0) \end{pmatrix} = \begin{pmatrix} f^1_1(x_0) & f^1_2(x_0) & \dots & f^1_n(x_0) \\ f^2_1(x_0) & f^2_2(x_0) & \dots & f^2_n(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ f^m_1(x_0) & f^m_2(x_0) & \dots & f^m_n(x_0) \end{pmatrix}$$

the Jacobian of f at x_0. If the above holds at any $x_0 \in X$, we call the mapping $J_f : \mathbb{R}^n \mapsto \mathbb{R}^{n\times m}, x_0 \mapsto J_f(x_0)$ the Jacobian of f.

For the last time, the usual comment: make sure that you understand that the Jacobian of f at x_0 is a matrix, and that the Jacobian of f is a function mapping $\mathbb{R}^n$ into the matrix space $\mathbb{R}^{n\times m}$. As before, we can also define the Jacobian operator

$$J : D^1_{partial}(X,\mathbb{R}^m) \mapsto F_X, f \mapsto J_f, \text{ where } D^1_{partial}(X,\mathbb{R}^m) = \{f : X \mapsto \mathbb{R}^m : (J_f(x_0) \text{ exists } \forall x_0 \in X)\}.$$

Next, from Definition 69 of the Jacobian, you can see why we defined the gradient as a row vector. Alternatively, we could have defined the Jacobian as the row vector of column-vector gradients, but we definitely need to define one as a column and the other as a row in order to arrive at a matrix rather than a super-long vector. Finally, let me stress that while the gradient and the Jacobian may look intimidating at first, they are nothing but mere collections of partial derivatives, i.e. they describe rules for how we order them when presenting them together.

[17] The LHS object is a function, the RHS object a specific evaluation point and thus a vector in $\mathbb{R}^m$!

This means that once you have understood firmly what a partial derivative is, these concepts are indeed rather straightforward to grasp!

3.2.5 Differentiability of Real-Valued Functions

Before moving on, a comment on what follows in this subsection: the way of generalizing the derivative to multivariate functions is somewhat technical and non-intuitive at first. After defining everything formally, we will arrive at the fundamental punchline that **the collections of partial derivatives defined above, i.e. the gradient or the Jacobian, are in fact equal to the derivative whenever it exists, and it exists whenever all partial derivatives are continuous!** This fact is the key message you should take away from what follows, understanding everything beyond this is a plus.[18]

Now that we come to multivariate differentiability, a disclaimer: This script adopts the usual textbook notation $D_f(x_0)$ for the derivative of a multivariate function f at x_0 in its domain. However, you may think of this object as the same thing as $\frac{df}{dx}(x_0)$, it is the *exact same concept*! The reason why textbooks may hesitate to write the multivariate differential operator as $\frac{d}{dx}$ is because changes dx in the denominator refers to instantaneous variation in a multivariate object, and we don't know how to divide by vectors. Then again, as we will see below, the notation $\frac{\partial f}{\partial \bar{x}}(x_0)$ for taking the partial derivatives with respect to multiple elements $\bar{x}$ of x (but not all, i.e. $\bar{x} \neq x$, thus the "partial" operator ∂), is widely accepted. Long story short, if the new notation $D_f(x_0)$ confuses you, be clear that it is nothing else but (the generalization of) the regular derivative of f at x_0, $\frac{df}{dx}(x_0)$.

Recall that our fundamental issue in generalizing the derivative to multivariate functions was that the arguments $x \in \mathbb{R}^n$ can move in "multiple" directions and it is ex-ante not clear what a marginal change should refer to in this context – in other words, there is some ambiguity with respect to *the directionality of variation in* x. Before coming to the general concept of the multivariate derivative (or simply: "derivative") of a multivariate function, we briefly turn to directional derivatives to show you the link of partial derivatives, as defined above, to the directionality of variation in x.

Recall that we obtained the partial derivative of $f : X \mapsto \mathbb{R}$ at $x_0 \in X$ from the function

$$\tilde{f}_{j,x_0}(t) = f(x_{0,1}, \ldots, x_{0,j-1}, x_{0,j} + t, x_{0,j+1}, \ldots, x_{0,n}).$$

Now, if we consider the j-th unit vector $e_j \in \mathbb{R}^n$ with $e_{ji} = 1$ if $i = j$ and $e_{ji} = 0$ else, we can also write

$$\tilde{f}_{j,x_0}(t) = f(x_0 + te_j).$$

Taking the derivative of this function with respect to t, at $t = 0$, we get the change of f associated with varying only x_j, or respectively, of moving into direction $(0, \ldots, 0, 1, 0, \ldots, 0)'$ with the 1 at position j. Thus, e_j is the direction of variation that we study when considering the partial derivative with respect to x_j! Now, rather than moving in only "fundamental directions" e_j,

[18] Note that "it exists whenever all partial derivatives are continuous" is a sufficient condition rather than an equivalent one, so that differentiable functions with non-continuous partial derivatives may exist. An example is given at https://mathinsight.org/differentiable_function_discontinuous_partial_derivatives. To understand it, you may want to read through the deliberations on multivariate derivatives below first.

$j \in \{1,\dots,n\}$, we can generalize this concept to *unidirectional variation* in arbitrary directions $z \in \mathbb{R}^n$! For this to be well-defined, we need to normalize the direction in a fashion similar to we have done before, as you will see from the following definition:

Definition 70. ***(Directional Derivative)*** *Consider the normed vector space $(\mathbb{R}^n, \|\cdot\|)$, let $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}$. Further, let $z \in \mathbb{R}^n$ such that $\|z\| = 1$, and define*

$$\tilde{f}_{z,x_0}(t) = f(x_0 + tz).$$

If $\tilde{f}_{z,x_0}(t)$ is differentiable at $t = 0$, we call

$$D_z f(x_0) := \frac{d\tilde{f}_{z,x_0}}{dt}(0)$$

the directional derivative of f in direction z.

The key take-away from this definition is that the gradient of f, ∇f, collects all the directional derivatives in the fundamental directions of the $\mathbb{R}^n$, i.e. $e_1, e_2, \dots, e_n$, and therefore tells us what happens if we vary f alongside any of the horizontal axes in isolation. At this point, it may already be at least plausible from an intuitive point of view why we may call the gradient the derivative: it is a complete summary of a function's variation along all fundamental directions in its domain!

Now we know how to consider changes in isolated directions more generally, it is time to move on to the real deal – how do we define a multivariate derivative where we allow simultaneous variation in arbitrary directions? To narrow down the issue, recall that we call d^* the derivative of f at x_0 in the domain of f if

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} = d^*. \tag{12}$$

Our conceptual problem now is that when the domain is multivariate, i.e. $X \subseteq \mathbb{R}^n$ and $n > 1$, then there is no clear notion of what we mean with "$\lim_{h\to 0}$". So, how can we rephrase this statement to something that does generalize to the $\mathbb{R}^n$? Note that we can re-write the equation above as

$$0 = \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} - d^* = \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0) - d^* \cdot h}{h}. \tag{13}$$

A key fact that we can use now is the following:

Proposition 20. ***(Continuity of the Norm)*** *Consider the normed vector space $(\mathbb{X}, \|\cdot\|)$ where $\mathbb{X} = (X, +, \cdot)$ is a real vector space. Then, the norm $\|\cdot\|$ is continuous.*

Proof. Upon a closer look, this is a direct implication of a definition of continuity: what we have to show for any function $f : X \mapsto \mathbb{R}$ is that

$$\forall x_0 \in X \forall \varepsilon > 0 \exists \delta > 0 : (\|x - x_0\| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon).$$

Now, the function we are interested in is $f(x) = \|x\|$, so that the requirement becomes

$$\forall x_0 \in X \forall \varepsilon > 0 \exists \delta > 0 : (\|x - x_0\| < \delta \Rightarrow \big|\|x\| - \|x_0\|\big| < \varepsilon).$$

If $\|x - x_0\| < \delta$, then by the inverse triangle inequality (Proposition 3):

$$\big|\|x\| - \|x_0\|\big| \leq \|x - x_0\| < \delta.$$

Thus, for any $x_0 \in X$ and any $\varepsilon > 0$, we can just choose $\delta = \varepsilon$ to establish continuity of $\|\cdot\|$. □

While sounding rather abstract, this fact is actually strikingly intuitive: continuity says that if two arguments x and x_0 considered don't lie far apart, the function values don't lie far apart either. Consider now the scenario where $f(x) = \|x\|$. Clearly, for two vectors x and x_0 to lie "close" to each other, a *necessary condition* is that they are of similar length - which is precisely what is measured by the norm of the vectors, which need to be "close" to each other as well in consequence!

To see how this helps us, recall that we can pull the limit in and out of any continuous function. Thus, the proposition above tells us that it can especially be pulled in and out of norms. Since we are still dealing with a univariate function f, we use the absolute value as our norm as induces the natural metric of the $\mathbb{R}$, $d(x, y) = |x - y|$. Applying it on both sides, we see that equation (13) implies

$$\begin{aligned} &\left|\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0) - d^* h}{h}\right| = |0| = 0 \\ \Leftrightarrow \quad &\lim_{h\to 0} \frac{|f(x_0 + h) - f(x_0) - d^* h|}{|h|} = 0. \end{aligned} \tag{14}$$

The equivalence follows because *(i)* we can pull the limit out of the absolute value and *(ii)* because for any $a, b \in \mathbb{R}$ such that $b \neq 0$, it holds that $|a/b| = |a|/|b|$.

The next step to generalization is the insight that the previous expression is *equivalent* to

$$\lim_{h\to 0} \frac{{}_a\|f(x_0 + h) - f(x_0) - d^* h\|}{{}_b\|h\|} = 0 \tag{15}$$

as the characterization of the univariate derivative d^* of f at x_0 in its domain, where ${}_a\|\cdot\|$ and ${}_b\|\cdot\|$ are any two norms on $\mathbb{R}$. To see why this equivalence holds for any two arbitrary norms on $\mathbb{R}$, note that for a norm $\|\cdot\|$ on $\mathbb{R}$, because $x \in \mathbb{R}$, by absolute homogeneity, $\|x\| = \|x \cdot 1\| = |x| \cdot \|1\|$, so that for any norm $\|\cdot\|$ on $\mathbb{R}$, there exists $c = \|1\| > 0$ such that $\|\cdot\| = c \cdot |\cdot|$. Then, when $c_a, c_b > 0$ are such that $\forall x \in \mathbb{R} : ({}_a\|x\| = c_a|x| \wedge {}_b\|x\| = c_b|x|)$, it holds that

$$\lim_{h\to 0} \frac{{}_a\|f(x_0 + h) - f(x_0) - d^* h\|}{{}_b\|h\|} = \frac{c_a}{c_b} \cdot \lim_{h\to 0} \frac{|f(x_0 + h) - f(x_0) - d^* h|}{|h|} \overset{(14)}{=} \frac{c_a}{c_b} \cdot 0 = 0,$$

and equation (14) implies equation (15). Conversely, starting from equation (15), one obtains

$$\lim_{h\to 0} \frac{|f(x_0 + h) - f(x_0) - d^* h|}{|h|} = \frac{c_b}{c_a} \cdot \lim_{h\to 0} \frac{{}_a\|f(x_0 + h) - f(x_0) - d^* h\|}{{}_b\|h\|} = \frac{c_b}{c_a} \cdot 0 = 0.$$

What is more, equation (15) is actually equivalent to our initial definition of the univariate derivative in equation (12): we have seen that equation (15) is implied by equation (12), and that it is equivalent to equation (14). Hence, to establish equivalence to equation (12), it suffices to show that equation (14) also implies (12). This is easily done: Note that *(i)* $|a|/|b| = |a/b|$ for

any $a, b \in \mathbb{R}$, and that $|\cdot|$ defines a norm on $\mathbb{R}$[19], which implies that *(ii)* $|x| = 0 \Leftrightarrow x = 0$ for $x \in \mathbb{R}$ and *(iii)* it is continuous in the metric space $(\mathbb{R}, |\cdot|)$. Hence, starting from equation (14),

$$0 = \lim_{h\to 0} \frac{|f(x_0+h) - f(x_0) - d^*h|}{|h|} \overset{(i)}{=} \lim_{h\to 0} \left| \frac{f(x_0+h) - f(x_0) - d^*h}{h} \right| \overset{(iii)}{=} \left| \lim_{h\to 0} \frac{f(x_0+h) - f(x_0) - d^*h}{h} \right|.$$

Hence, with *(ii)*, equation (12) follows.

In the previous expressions, formally, generalization of the derivative concept to the $\mathbb{R}^n$ was limited by the fact that we don't know how to divide by the direction h if h is a vector. Now, with equation (15) we have arrived at an *equivalent* expression in the *norm* of h, that equals a real number regardless of whether h is scalar or a vector! This inspires the following definition:[20]

Definition 71. ***(Multivariate Derivative of Real-valued Functions)*** *Let $X \subseteq \mathbb{R}^n$, $f : X \mapsto \mathbb{R}$ and $\|\cdot\|$ a norm on $\mathbb{R}^n$. Further, let $x_0 \in int(X)$ be an interior point of X. Then, f is differentiable at x_0 if there exists $d^* \in \mathbb{R}^{1\times n}$ such that*

$$\lim_{\|h\|\to 0} \frac{|f(x_0+h) - f(x_0) - d^*h|}{\|h\|} = 0.$$

Then, we call d^* the derivative of f at x_0, denoted $\frac{df}{dx}(x_0)$ or $D_f(x_0)$. If f is differentiable at any $x_0 \in X$, we say that f is differentiable, and we call $D_f : X \mapsto \mathbb{R}^n, x \mapsto D_f(x)$ the derivative of f.

Like the derivative at x_0, we can alternatively denote the derivative D_f as $\frac{df}{dx}$. Note that the derivative is defined as a row vector in order to have compatible dimension for the multiplication d^*h, and that in the defining equation, the numerator uses the natural norm of the $\mathbb{R}$, while the denominator norm is an arbitrary norm for $\mathbb{R}^n$. Further, besides the usual comment that $D_f(x_0)$ is a vector in $\mathbb{R}^n$, D_f a function and with $D^1(X, \mathbb{R})$ as the set of differentiable functions from $X \to \mathbb{R}$, the differential *operator* is

$$D : D^1(X, \mathbb{R}) \mapsto F_X, f \mapsto D_f$$

and maps differentiable functions onto their derivative function, one needs to pay attention what precisely "$\lim_{\|h\|\to 0}$" means. Verbally, we say that we approach the zero vector in an arbitrary fashion, i.e. not restricted to any specific direction, and if the limit exists, regardless of how we approach zero, the rate of marginal change as captured by the derivative will always be the same. Formally, we call g_0 the limit $\lim_{\|h\|\to 0} g(h)$ of a function $g : X \mapsto \mathbb{R}$ if

$$\forall \varepsilon > 0 \exists \delta > 0 : (h \in B_\delta(\mathbf{0}) \Rightarrow |g(h) - g_0| < \varepsilon).$$

That is, we consider a ball of radius δ around $\mathbf{0}$, the *only* element of $\mathbb{R}^n$ with $\|\cdot\| = 0$, and study the function's behavior on this potentially very small ball, which however covers vectors in *any* of $\mathbb{R}^n$'s directions – this solves the issue of approaching x_0 only from a single direction! Moreover, as a more technical note, you can see from the definition that the derivative is only

[19] In fact, it is the norm that induces $\mathbb{R}$'s natural metric!

[20] To see how this emerges from equation (15), note that for real-valued functions, for any $x \in X$, $f(x_0+h) - f(x_0) - d^*h \in \mathbb{R}$, so that the absolute value is an appropriate norm to use in the numerator.

defined on the interior of X. The simple reason is that for any $x_0 \notin \text{int}(X)$, for every $\delta > 0$, there exist points in $B_\delta(x_0)$ where f is not defined and the limit can not exist by definition.

Now, for the fundamental insight of this subsection that was already hinted to at the beginning: while formally, the generalization of the derivative to the multivariate case is logical and arguably quite clever, it is of little help in actually computing the derivative – how should we even begin looking for vectors d^* that satisfy the limit condition of the definition? Luckily, we have the following very powerful result, stated here for real-valued (rather than vector-valued) functions:

Theorem 40. ***(The Gradient and the Derivative)*** *Let $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}$. Further, let $x_0 \in \text{int}(X)$ be an interior point of X, and suppose that f is differentiable at x_0. Then, all partial derivatives of f at x_0 exist, and $D_f(x_0) = \nabla f(x_0)$.*

In words, $D_f(x_0)$, the true generalization of the univariate derivative, is nothing else than the vector of partial derivatives of f at x_0, i.e., the gradient $\nabla f(x_0)$ of f at x_0! That is, the instantaneous rate of change of f in *all* directions coincides with the collection of unidimensional rates of change when varying x along the *fundamental* directions of $\mathbb{R}^n$! More intuitively, the derivative, equal to the gradient, collects all the possible fundamental effects varying x into *any arbitrary* direction can have. Before looking at the proof, let me again stress the following:

This theorem is the main take-away from this subsection. It tells you that, once you know how to compute a partial derivative as a specific example of the derivative of a standard univariate function, ***given that a function is differentiable****, you can simply stack these objects in a vector to differentiate a multidimensional function, and thus, despite its complex look,* ***multivariate differentiation is nothing but a rule for ordering and stacking univariate partial derivatives!!*** *This is the central detail that you* ***must*** *understand here – any formal detail or conceptual insight beyond that will also be helpful, but comes nowhere close in practical importance.*

Below the proof of Theorem 40, we turn to the condition that ensures that a function is differentiable. But first, let's look at some applications – re-consider the functions from before (with slightly altered f^3): $f^1(x_1, x_2) = x_1 + x_2$, $f^2(x_1, x_2) = x_1 x_2$ and $f^3(x_1, x_2, x_3) = x_1 x_2^2 + \cos(x_1)e^{x_3}$ which are all differentiable. What are the derivatives of these functions? (Think first, then read on.) Applying the theorem and noting that we had computed the partial derivatives for f^1 and f^2 already above, we get $D_{f^1}(x) = \nabla f^1(x) = (1, 1)$ and $D_{f^2}(x) = \nabla f^2(x) = (x_2, x_1)$. Finally, for f^3, it is left to you to verify that $D_{f^3}(x) = \nabla f^3(x) = (x_2^2 - \sin(x_1)e^{x_3}, 2x_1 x_2, \cos(x_1)e^{x_3})$. Now, let's prove the theorem:

Proof of Theorem 40. The proof relies heavily on the following fact that we discussed earlier (recall Theorem 17): If the limit $\lim_{x \to x_0} g(x) = g_0$ exists, then for any sequence $\{x_n\}_{n \in \mathbb{N}}$ such that $\lim_{n \to \infty} x_n = x$, it holds that $\lim_{n \to \infty} g(x_n) = g_0$. More generally, this implies that $g(x)$ arrives at g_0 when x approaches x_0 in an *arbitrary* fashion, it does so especially if we let x approach x_0 in a *specific* fashion! This implies for example that if $\lim_{\|h\| \to 0} g(h) = g_0$ exists, then especially e.g. $\lim_{h_1 \to 0} g(h_1, 0, \dots, 0) = g_0$ or $\lim_{t \to 0} g(t \cdot z) = g_0$ for a fixed vector z.

As a roadmap, we will now want to show that any j-th partial derivative of f at x_0 exists, and that it is indeed equal to the j-th entry of the derivative $D_f(x_0)$. Recall that the j-th partial

derivative of f at x_0 was defined as

$$f_j(x_0) = \lim_{t\to 0} \frac{f(x_0 + te_j) - f(x_0)}{t} \quad \Leftrightarrow \quad f_j(x_0) = d_j^* \text{ with } \lim_{t\to 0} \left| \frac{f(x_0 + te_j) - f(x_0)}{t} - d_j^* \right|, \tag{16}$$

where e_j is the j-th unit vector, i.e. the vector with a one in position j and zeros everywhere else.

Suppose that f is differentiable at $x_0 \in X$ with derivative $D_f(x_0)$. Then, by definition,

$$\lim_{\|h\|\to 0} \frac{|f(x_0 + h) - f(x_0) - D_f(x_0)h|}{\|h\|} = 0. \tag{17}$$

Let $j \in \{1, \ldots, n\}$, and let e_j be the j-th unit vector. Then, equation (17) implies

$$0 = 0 \cdot \|e_j\| = \|e_j\| \cdot \lim_{t\to 0} \frac{|f(x_0 + te_j) - f(x_0) - D_f(x_0)te_j|}{\|te_j\|} = \lim_{t\to 0} \frac{|f(x_0 + te_j) - f(x_0) - D_f(x_0)te_j|}{|t|}$$

Now, note that $|a|/|b| = |a/b|$, and thus

$$\frac{|f(x_0 + te_j) - f(x_0) - D_f(x_0)te_j|}{|t|} = \left| \frac{f(x_0 + te_j) - f(x_0) - D_f(x_0)te_j}{t} \right| = \left| \frac{f(x_0 + te_j) - f(x_0)}{t} - D_f(x_0)e_j \right|.$$

Therefore (cf. equation (16)),

$$\lim_{t\to 0} \frac{f(x_0 + te_j) - f(x_0)}{t} = D_f(x_0)e_j,$$

i.e. the LHS limit exists, which establishes existence of the partial derivative. Further, the j-th partial derivative at x_0 is equal to this limit, i.e. $\frac{\partial f}{\partial x_j} = D_f(x_0)e_j$, the j-th entry of $D_f(x_0)$.[21] Thus, $\nabla f(x_0)$ exists and $D_f(x_0) = \nabla f(x_0)$. □

To establish the converse, i.e. to see how we get from existence of the partial derivatives to differentiability, note the following:

Theorem 41. ***(Partial Differentiablility and Differentiability)*** *Let $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}$. Further, let $x_0 \in \text{int}(X)$ be an interior point of X. If all the partial derivatives of f at x_0 exist and are continuous, then f is differentiable at x_0.*

For a proof, see e.g. dlF, Chapter 4, Theorem 3.4. This tells us that when concerned with differentiating a function, we have to (i) verify that all partial derivative exist, (ii) that they are continuous and if so, (iii) stack them into the gradient to obtain the derivative. Note however that continuity of all partial derivatives is a *sufficient* condition for differentiability, so that if it is violated, if in doubt, we have to check the technical definition directly.

3.2.6 Differentiability of Vector-Valued Functions

Generalization of the differentiability concept to functions with codomain $\mathbb{R}^m$ is straightforward – you may even be able to tell how it works without reading this subsection if you have

[21] Note that for $a = (a_1, \ldots, a_n)' \in \mathbb{R}^n$, $a'e_j = \sum_{i=1}^n a_i e_{ji} = 0a_1 + 0a_2 + \ldots + 0a_{j-1} + 1a_j + 0a_{j+1} + \ldots + 0a_n = a_j$, such that $a'e_j$ picks out the j-th element of a.

properly understood the previous one. The reason is equation (11), which shows us that a function $f : X \mapsto \mathbb{R}^n$ is just an ordered collection of functions $f^j : X \mapsto \mathbb{R}$, that we know how to take the derivative of. Indeed, maintaining this order, we can generalize every insight from the previous subsection to vector-valued functions, we just need to adjust notation slightly.

First, in the definition of the derivative, because f dos no longer map onto $\mathbb{R}$, in accordance with our "generalization equation" (15), we now need to consider a norm of the $\mathbb{R}^m$ in the numerator (rather than $\|\cdot\|$ as a norm of $\mathbb{R}$, the codomain of real-valued functions):

Definition 72. ***(Multivariate Derivative of Vector-valued Functions)*** *Let $X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}^m$. Further, let $x_0 \in \text{int}(X)$ be an interior point of X. Denote ${}_k\|\cdot\|$ as a norm of $\mathbb{R}^k$, $k \in \{n, m\}$. Then, f is differentiable at x_0 if there exists* ***a matrix*** *$D^* \in \mathbb{R}^{m\times n}$ such that*

$$\lim_{{}_n\|h\|\to 0} \frac{{}_m\|f(x_0+h) - f(x_0) - D^*h\|}{{}_n\|h\|} = 0,$$

Then, we call D^* the derivative of f at x_0, denoted $D_f(x_0)$ or $\frac{df}{dx}(x_0)$. If f is differentiable at any $x_0 \in X$, we say that f is differentiable, and we call $D_f : X \mapsto \mathbb{R}^{m\times n}, x \mapsto D_f(x)$ the derivative of f.

Our key theorem continues to hold as well:

Theorem 42. ***(The Jacobian and the Derivative)*** *Let $X \subseteq \mathbb{R}^n$, $f : X \mapsto \mathbb{R}^n$ and $f^1, \dots, f^n : X \mapsto \mathbb{R}$ such that equation (11) holds. Further, let $x_0 \in \text{int}(X)$ be an interior point of X, and suppose that f is differentiable at x_0. Then, for any f^i, $i \in \{1, \dots, m\}$, all partial derivatives of f^i at x_0 exist, and $D_f(x_0) = J_f(x_0)$.*

Proof. Very similar to before. Consider again $j \in \{1, \dots, n\}$ and $e_j \in \mathbb{R}^n$ as the j-th unit vector of length n.[22] Let $x_0 \in \text{int}(X)$, and suppose that $D_f(x_0)$ exists. Then by definition,

$$\lim_{\|h\|\to 0} \frac{{}_m\|f(x_0+h) - f(x_0) - D_f(x_0)h\|}{{}_n\|h\|} = 0,$$

and in analogy to before, this implies

$$0 = \lim_{t\to 0} \frac{{}_m\|f(x_0+te_j) - f(x_0) - D_f(x_0)te_j\|}{|t|} = \lim_{t\to 0} {}_m\| \frac{f(x_0+te_j) - f(x_0)}{t} - D_f(x_0)e_j\|$$

or respectively,

$$D_f(x_0)e_j = \lim_{t\to\infty} \frac{f(x_0+te_j) - f(x_0)}{t}.$$

Note that this is a *vector equation*, i.e. that similar to before, $D_f(x_0)e_j$ picks out the j-th *column* of $D_f(x_0)$, and the RHS is based on f which is vector-valued by definition. Noting that the limit of a vector is defined element-wise[23] we obtain

$$e_i' Df(x_0)e_j = \lim_{t\to\infty} e_i' \frac{f(x_0+te_j) - f(x_0)}{t} = \lim_{t\to\infty} \frac{f^i(x_0+te_j) - f^i(x_0)}{t} = \frac{\partial f^i}{\partial x_j}(x_0) \qquad \forall i \in \{1, \dots, m\}.$$

[22] Like before, the index j is chosen for objects in the domain $X \subseteq R^n$. Objects in the codomain will from now on be denoted by i, as in f^i.

[23] This means that if we have a sequence $\{x_k\}_{k\in\mathbb{N}}$, where for all $k \in \mathbb{N}$, $x_k = (x_{k1}, \dots, x_{kn})' \in \mathbb{R}^n$, then the limit, if it exists, is given by $\lim_{k\to\infty} x_k = (\lim_{k\to\infty} x_{k1}, \dots, \lim_{k\to\infty} x_{kn})'$.

Thus, all directional derivatives exist, and the j-th directional derivative of f^i is equal to $e_i' D_f(x_0) e_j$, the (i,j) element of $D_f(x_0)$. □

If you have trouble following the notation above, it may be a good time to take pen and paper and verify that $e_i A e_j$ picks out the (i,j)-element of a matrix $A \in \mathbb{R}^{n\times m}$ for specific examples, this should help your understanding a lot. Finally, also our differentiability condition continues to go through:

Theorem 43. ***(Partial Differentiablility and Differentiability)*** *Let $X \subseteq \mathbb{R}^n$, $f : X \mapsto \mathbb{R}^n$ and $f^1, \dots, f^n : X \mapsto \mathbb{R}$ such that $f = (f^1, \dots, f^m)'$. Further, let $x_0 \in \text{int}(X)$ be an interior point of X. If for any $i \in \{1, \dots, m\}$, all the partial derivatives of f^i at x_0 exist and are continuous, then f is differentiable at x_0.*

To conclude our discussion of multivariate differentiability, let me highlight that **the rules you are well-familiar with (linearity of the derivative, product rule and chain rule) go through also for the multi-dimensional case.** An exception is the quotient rule, because the quotient of two vectors is not a well-defined object. That being said, we need to **apply some care to ensure concerning the order of derivatives,** because unlike with real-valued functions where e.g. $f'(x)g(x) = g(x)f'(x)$, recall that matrix multiplication is *not* commutative! Thus, be sure to respect the order in which the differential expressions appear in the following theorem:

Theorem 44. ***(Rules for General Multivariate Derivatives)*** *Let $X \subseteq \mathbb{R}^n$, $f, g : X \mapsto \mathbb{R}^m$ and $h : \mathbb{R}^m \mapsto \mathbb{R}^k$. Suppose that f, g and h are differentiable functions. Then,*

1. *(Linearity) For all $\lambda, \mu \in \mathbb{R}$, $\lambda f + \mu g$ is differentiable and $D_{\lambda f + \mu g} = \lambda D_f + \mu D_g$.*
2. *(Product Rule) $f' \cdot g$ is differentiable and $D_{f'g} = f' \cdot D_g + g' \cdot D_f$.*
3. *(Chain Rule) $h \circ f$ is differentiable and $D_{h\circ f} = (D_h \circ f) \cdot D_f$.*

As an example for the chain rule, consider $v(x) = x'A'Ax$ where $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$. Then, what is $D_v(x)$? Note that we can write $v(x) = h(f(x))$ where $f(x) = Ax$ and $h(y) = y'y$. Albeit somewhat tedious, taking the partial derivatives of f is rather straightforward, and you should arrive at $D_f(x) = A$. For h, you should obtain $D_h(y) = 2y'$. Then, the chain rule tells us that

$$D_v(x) = D_h(f(x)) \cdot D_f(x) = 2(Ax)' \cdot A = 2x'A'A.$$

This has helped us to compute the derivative of the quite complex-looking function $x'A'Ax$ without multiplying everything out and considering squares and cross-terms, so this should convince you that this formula may come in extremely handy at times – you should definitely have it in the back of your mind whenever you come across something that looks too tedious to handle otherwise!

Excursion. An Economic Example of the Chain Rule. Consider the utility $u(c,l)$ over consumption c and leisure l, where consumption is determined in its simplest form and is equal to labor income, i.e. $c = w(1-l)$ where w is the wage which is exogenous (i.e., invariant to c and l). Then, c is implicitly determined by l, such that $c = g(l)$. Indeed, u is then really just a function of leisure: $U(l) = u(g(l), l)$. This function, we call the implicit

utility of leisure. To understand how it changes with leisure, we can consult the chain rule and the *partial* derivatives of u with respect to its arguments, denoted $u_c = \frac{\partial u}{\partial c}$ and $u_l = \frac{\partial u}{\partial l}$ and typically referred to as the marginal utilities of consumption and leisure, respectively. Then, when we apply the chain rule with $f(l) = (g(l), l)$ and $h(y) = u(y)$, we have

$$\frac{du}{dl} = D_h(f(l))f'(l) = \left(\frac{\partial u}{\partial c}, \frac{\partial u}{\partial l}\right)\begin{pmatrix}\frac{dg}{dl}\\ 1\end{pmatrix} = \frac{\partial u}{\partial c}\frac{dg}{dl} + \frac{\partial u}{\partial l} = u_c\frac{dg}{dl} + u_l.$$

This expression will be helpful, amongst others, to determine the optimal value $l^* \in [0,1]$ of leisure that sets this derivative to 0, given that this point is indeed the global maximizer, as discussed in the next chapter.

While you don't necessarily need to read the excursion, the more general fact is worthwhile noting: If a function $f(x_1(t), x_2(t))$ in arguments x_1 and x_2 is driven by a latent factor t, we can apply the chain rule to compute the derivative of f with respect to t from the derivatives of x_1 and x_2 with respect to t and the partial derivatives of f:

$$\frac{df}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt}.$$

You will likely see this equation alot in the course of your master studies, so make sure that you understand it well, and perhaps also try to prove it! (For inspiration, see the excursion, the proof works in a perfectly analogous fashion.)

Before moving on, a last piece of notation: the definitions have told us that $\frac{df}{dx}$ is the notation for the (multivariate) derivative of f with respect to its argument, and $\frac{\partial f}{\partial x_j}$ the one for the partial derivative of f with respect to x_j, the j-th component of its argument. Something that you will at times see that falls in neither of these to categories is $\frac{\partial f}{\partial x}$ (with $x \in \mathbb{R}^n$). When you see this object, we are mostly interested in functions of the form $f(x, y)$ where y itself may be a function of x, i.e. $y = g(x) \in \mathbb{R}^m$, potentially also with $m = 1$. When we write $\frac{\partial f}{\partial x}$, we explicitly state that we consider only the derivative of f with respect to its first n arguments, and abstract from the potential dependence of y through x. Generally, $\frac{\partial f}{\partial x}$ is established as shorthand notation for taking the derivative with respect to the collection of arguments x while holding constant the others if the function has more arguments than the vector x, e.g. when $f : \mathbb{R}^3 \mapsto \mathbb{R}$, then $\frac{\partial f}{\partial (x_2, x_3)} = \left(\frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}\right)$ (in contrast to $\frac{df}{dx} = \nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}\right)$). However, **when x is the vector of all arguments** of f (and thus also when f has just a single argument, i.e. $f : \mathbb{R} \mapsto \mathbb{R}$) it makes **no sense to write $\frac{\partial f}{\partial x}$!!**

3.2.7 Higher Order Derivatives, Taylor Approximations and Total Derivatives

As the last part of our investigations into differentiation, we turn to derivatives of higher order for general vector functions, and discuss how to generalize the Taylor approximation and relate it to the total derivative. You may have grasped that when starting from a function $f : X \mapsto \mathbb{R}$ where $X \subseteq \mathbb{R}^n$, then taking the derivative comes with an increase in dimension: while for an $x_0 \in X$, $f(x_0) \in \mathbb{R}$, $\frac{df}{dx}(x_0)' = \nabla f(x_0)' \in \mathbb{R}^n$, and for $f : X \mapsto \mathbb{R}^m$ where $f(x_0) \in \mathbb{R}^m$, the derivative $\frac{df}{dx}(x_0) = J_f(x_0)$ is already a matrix in $\mathbb{R}^{m\times n}$. Because there is no need to touch

multi-dimensional matrices (i.e. spaces of the form $\mathbb{R}^{n_1 \times n_2 \times \ldots \times n_k}$), which you indeed never come across in regular economic studies, this puts a natural limit to the derivatives we consider here: the first derivative for $f : X \mapsto \mathbb{R}^m$, which you already know from the last subsection, and the second derivative for $f : X \mapsto \mathbb{R}$. Like with univariate functions, it can be obtained from taking the derivative of the first derivative, provided that it exists.

Definition 73. ***(Second Order Partial Derivative)*** *Let $X \subseteq \mathbb{R}^n$ be an open set and $f : \mathbb{R}^n \mapsto \mathbb{R}$. Further, let $x_0 \in X$, and suppose that f is differentiable at x_0. Then, if the i-th partial derivative of f, $f_i = \frac{\partial f}{\partial x_i}$ is differentiable at x_0, then we call its j-th partial derivative at x_0 the (i,j)-second order partial derivative at x_0, denoted $f_{i,j}(x_0) = \frac{\partial f_i}{\partial x_j}(x_0) = \frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)$.*

Higher order partial derivatives are defined in the exact same way, so that e.g. the (i, j, k, l) fourth order derivative of f at x_0 is $\frac{\partial^4 f}{\partial x_i \partial x_j \partial x_k \partial x_l}(x_0)$. By requiring X to be an open set, we ensure $X = \text{int}(X)$ so that it has only interior points. Recall that like the function f, the partial derivatives $\frac{\partial f}{\partial x_i}$ are functions from $X \subseteq \mathbb{R}^n$ to $\mathbb{R}$, so it makes indeed sense to think about their *partial* derivatives. As an example, re-consider our function $f^3(x_1, x_2, x_3) = x_1 x_2^2 + \cos(x_1)e^{x_3}$ with gradient $\nabla f^3(x) = (x_2^2 - \sin(x_1)e^{x_3}, 2x_1 x_2, \cos(x_1)e^{x_3})$. Before computing the second order partial derivatives of this function, we look at how we order them to obtain a second derivative, and introduce a very powerful rule in computing them.

Definition 74. ***(Hessian or Hessian Matrix)*** *Let $X \subseteq \mathbb{R}^n$ be an open set and $f : X \mapsto \mathbb{R}$. Further, let $x_0 \in X$, and suppose that f is differentiable at x_0 and that all second order partial derivatives of f at x_0 exist. Then, the matrix*

$$H_f(x_0) = \begin{pmatrix} \nabla f_1(x_0) \\ \nabla f_2(x_0) \\ \vdots \\ \nabla f_n(x_0) \end{pmatrix} = \begin{pmatrix} f_{1,1}(x_0) & f_{1,2}(x_0) & \cdots & f_{1,n}(x_0) \\ f_{2,1}(x_0) & f_{2,2}(x_0) & \cdots & f_{2,n}(x_0) \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1}(x_0) & f_{n,2}(x_0) & \cdots & f_{n,n}(x_0) \end{pmatrix}$$

is called the Hessian of f at x_0.

Requiring X to be open ensures that $x_0 \in \text{int}(X)$, which is the case for any differentiable function.[24] Now, remember the set $C^k(X, \mathbb{R})$ that we defined for $X \subseteq \mathbb{R}$ to indicate that f is k times differentiable with continuous k-th derivative? In the more general context of $X \subseteq \mathbb{R}^n$, this *functional class* is defined as the set $C^k(X, \mathbb{R})$ of functions $f : X \mapsto \mathbb{R}$ where f is k times differentiable and any k-th order partial derivative is continuous. Alternatively, one may write $C^k(X)$ where it is implicit that only real-valued functions are considered, or C^k if the domain of interest is clear from the context, and/or equal to $\mathbb{R}$. Typically, you will read $f \in C^k(\mathbb{R}^n)$ to indicate a k times differentiable function with continuous partial k-th derivatives, and $f \in C^k$ for a n times differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ with continuous n-th derivative. This concept is useful in the given context because:

Theorem 45. ***(Schwarz's Theorem/Young's Theorem)*** *Let $X \subseteq \mathbb{R}^n$ be an open set and $f : \mathbb{R}^n \mapsto \mathbb{R}$. If $f \in C^k(X)$, then the order in which the derivatives up to order k are taken can be permuted.*

[24] Differentiability is only defined at interior points of X, but differentiability requires that f is differentiable at any $x_0 \in X$. Therefore, it is necessary (but by no means sufficient) that for any $x_0 \in X$, it holds that $x_0 \in \text{int}(X)$, i.e. that X is open.

A proof can be found e.g. in dlF, Chapter 4, Theorem 2.6. Here, *permuted* means simply to interchange in order, so that e.g. $f_{1,2}(x) = f_{2,1}(x)$. You can assume that the functions we are typically concerned with are sufficiently well-behaved such that their partial derivatives we consider are continuous, and the order is interchangeable! You'll be convinced of this fact in a second when we come to our example f^3. For now, note the following corollary of our theorems in the previous subsection:

Corollary 7. ***(Hessian and Gradient)*** *Let $X \subseteq \mathbb{R}^n$ and $f \in C^2(X)$. Then, the Hessian is symmetric and corresponds to the derivative, i.e. the Jacobian of the transposed gradient $(\nabla f)'$: $H_f = J_{(\nabla f)'}$.*

From a technical side, note that $(\nabla f)'$ is the function that maps x onto the *column* vector $(\nabla f(x))'$. The corollary holds because if the second order partial derivatives, i.e. the partial derivatives of the functions in the gradient, are all continuous, then we can take the derivative of the (transposed) gradient $(\nabla f)' : \mathbb{R}^n \mapsto \mathbb{R}^n$. Because the (transposed) gradient is nothing but a vector-valued function, its derivative will coincide with its Jacobian $J_{(\nabla f)'}$. However, from the way the second order partial derivatives are arranged in the Hessian H_f, it follows that these two objects are precisely the same! Note, however, that **the Hessian is certainly equal to second derivative only if** $f \in C^2(X)$ because otherwise, the second derivative may not even be defined! Also note that the Hessian is always a Jacobian (of the transposed gradient), but not every Jacobian is a Hessian – be sure to know the distinction between these two concepts.

Finally, let's put all this knowledge to work. For $f^3(x_1, x_2, x_3) = x_1 x_2^2 + \cos(x_1)e^{x_3}$, computing the second order partial derivatives, at the point (x_1, x_2, x_3), we arrive at the Hessian (try it on your own first)

$$H_{f^3}(x_0) = \begin{pmatrix} -\cos(x_1)e^{x_3} & 2x_2 & -\sin(x_1)e^{x_3} \\ 2x_2 & 2x_1 & 0 \\ -\sin(x_1)e^{x_3} & 0 & \cos(x_1)e^{x_3} \end{pmatrix}.$$

We see rather easily that all second order partial derivatives are composed only of continuous functions, such that the Hessian should be symmetric - and it is! Indeed, this also tells us that the second derivative of f^3 will correspond to the Hessian.

Next to its importance in optimization where it determines the nature of an extreme value, where it tells us whether a solution to the first order condition it is a maximum, a minimum, or neither, the second derivative, equal to the Hessian matrix, can also be used to compute a second order Taylor approximation for a multivariate function:

Theorem 46. ***(Second Order Multivariate Taylor Approximation)*** *Let $X \subseteq \mathbb{R}^n$ be an open set and consider $f \in C^2(X)$. Let $x_0 \in X$. Then, the second order Taylor approximation to f at $x_0 \in X$ is*

$$T_{2,x_0}(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)' \cdot H_f(x_0) \cdot (x - x_0).$$

The error $\varepsilon_{2,x_0}(x) = T_{2,x_0}(x) - f(x)$ approaches 0 at a faster rate than $\|x - x_0\|^2$, i.e. $\lim_{\|h\| \to 0} \frac{\varepsilon_{2,x_0}(x+h)}{\|h\|^2} = 0$.

Again, this theorem tells us how around x_0, we can arrive at a "good" functional approximation of an arbitrary $f \in C^2(X)$. The proof is similar to the one presented in the excursion

after defining the Taylor expansion for the univariate case above, it is just a bit more notation-intensive and thus omitted here. Of course, the first-order Taylor expansion can also be obtained for $f \in C^1(X)$, and is defined in an analogous way as $T_{1,x_0}(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0)$. The multivariate Taylor *expansion*, adding the error to the Taylor *approximation*, is also defined analogy to the univariate case, however, for the second order approximation, a formula due to the dimensional complication is harder to come by; for the first order approximation, if $f \in C^2(X)$, it is simply equal to

$$\varepsilon_{1,x_0}(x_0 + h) = \frac{1}{2} h' \cdot H_f(x_0 + \lambda h) \cdot h$$

for a $\lambda \in (0,1)$, and thus a direct generalization of the univariate case; the error formula also holds for an approximation of order 0 if $f \in C^1(X)$.

Recall that in the univariate case, we had the mean value theorem as a corollary of Taylor's theorem. What about the multivariate case? In analogy to before, for a real-valued function f, we may arrive at

$$f(x_2) - f(x_1) = \nabla f(x^*)(x_2 - x_1).$$

Now, the issue arises that the RHS is a scalar product, and we can not solve for $\nabla f(x^*)$. Thus and unfortunately, a multivariate generalization of the mean value theorem does not exist, such that we can not as easily derive a sufficient condition for existence of a vector x^* that sets the gradient to zero using the Taylor approach.

The Total Derivative. An object frequently used in economics is the *total derivative*. It is instructive to discuss it here, since it is closely linked to Taylor's theorem. Typically, you will read the two-variable version like this:

$$df = \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2. \tag{18}$$

So, how do we make sense of this? And how can we use it to derive insights on the function f? The purpose of this expression is to capture the **instantaneous rate of change of** f as the the **vector of arguments marginally varies in a specific direction** $dx = (dx_1, dx_2)'$. That is, it characterizes $df(x_0) = \lim_{t \to 0} \Delta f(x_0)/t$ when considering changes of the form $\Delta x = t \cdot dx$ with a **fixed vector** dx **of relative variation of elements in** x**.** Accordingly, df is a function of x_0 and moreover of the direction components dx_1 and dx_2! As such, a more explicit way of writing this concept is

$$df(x_0, dx) = \frac{\partial f}{\partial x_1}(x_0) dx_1 + \frac{\partial f}{\partial x_2}(x_0) dx_2 = \nabla f(x_0) \cdot \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix}.$$

Why should we care about fixing specific ratios for the variation in the argument's components? After all, we already have the gradient which tells us about the variation along all the fundamental directions of the $\mathbb{R}^2$, isn't that enough? Well, yes and no. In economics, we are frequently concerned with *trade-offs* so that increasing one argument (e.g. consumption of the first good) can not go without decreasing the other (e.g. consumption of the second good, leisure, etc.), and the exchange ratio is usually exogenously given, at least when fixing

the starting point of variation x_0. The total derivative tells us more directly how *instantaneous variations in light of such trade-offs* look like! Indeed, it tells us that *computing this variation is as simple as multiplying the respective vector of relative variations to the gradient*.

Let us go over the concept more carefully to develop a thorough understanding of it, and see why the relationship asserted by the total derivative holds. Our first objective is to express the change in f, denoted Δf, resulting from variation of x_1 and x_2 by Δx_1 and Δx_2, respectively, collected in the vector $\Delta x = (\Delta x_1, \Delta x_2)'$. To connect this to the previous notation, starting from a specific point of variation x_0, we can consider $x = x_0 + \Delta x$, or simply put, variation of the argument of f by $h = \Delta x$, and the associated variation in f we are looking for is $\Delta f(x_0) = f(x_0 + \Delta x) - f(x_0)$. Then, the Taylor *expansion* of first order is

$$f(x) = T_{1,x_0}(x_0 + \Delta x) + \varepsilon_1(x_0 + \Delta x) = f(x_0) + \nabla f(x_0)\Delta x + \varepsilon_1(x_0 + \Delta x).$$

Recall that $\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right)$. Multiplying out $\nabla f(x_0)\Delta x$ and bringing $f(x_0)$ to the other side,

$$\Delta f(x_0) = f(x) - f(x_0) = \frac{\partial f}{\partial x_1}(x_0)\Delta x_1 + \frac{\partial f}{\partial x_2}(x_0)\Delta x_2 + \varepsilon_1(x_0 + \Delta x).$$

Next, we want to characterize $df(x_0, dx) = \lim_{t\to 0} \Delta f(x_0)/t$ when considering changes of the form $\Delta x = t \cdot dx$. The equation above gives

$$\begin{aligned}\frac{\Delta f(x_0)}{t} &= \frac{\partial f}{\partial x_1}(x_0)\frac{dx_1 t}{t} + \frac{\partial f}{\partial x_2}(x_0)\frac{dx_2 t}{t} + \varepsilon_1(x_0 + t \cdot dx) \\ &= \frac{\partial f}{\partial x_1}(x_0)dx_1 + \frac{\partial f}{\partial x_2}(x_0)dx_2 + \varepsilon_1(x_0 + t \cdot dx)\end{aligned}$$

The Taylor theorem tells us that the first order approximation is "good", i.e. $\lim_{h\to 0} \frac{\varepsilon_1(x_0+h)}{h} = 0$. This especially implies $\lim_{h\to 0} \varepsilon_1(x_0+h) = 0$ and thus also $\lim_{t\to 0} \varepsilon_1(x_0+t\cdot dx) = 0$, i.e. the Taylor error vanishes as we consider instantaneous variation in the argument x. Therefore, applying the limit $t \to 0$ in the equation above gives

$$df(x_0, dx) = \lim_{t\to 0} \frac{\Delta f(x_0)}{t} = \frac{\partial f}{\partial x_1}(x_0)dx_1 + \frac{\partial f}{\partial x_2}(x_0)dx_2.$$

This is precisely the statement of the total derivative. In fact, you may wonder - the argument varying from x_0 into a specific direction, isn't this precisely the definition of the directional derivative? Indeed, you are right: note that by the chain rule, when $g(t) = x_0 + t \cdot dx$,

$$\frac{df \circ g}{dt} = \left[\frac{df}{dx} \circ g\right]\frac{dg}{dt} = [\nabla f \circ g] \cdot dx$$

so that for $x_0 \in \mathbb{R}^2$,

$$\frac{df \circ g}{dt}(x_0)\bigg|_{t=0} = \nabla f(x_0 + 0 \cdot dx) \cdot dx = \nabla f(x_0) \cdot dx = \frac{\partial f}{\partial x_1}(x_0)dx_1 + \frac{\partial f}{\partial x_2}(x_0)dx_2.$$

This shows that the total derivative and the directional derivative of f in direction dx coincide. However, the total derivative is arguably more convenient to represent and thus to remember,

as it directly tells us that the "directional slope" we are interested in can just be computed by multiplying the direction of change to the gradient. Therefore, this label is more frequently used in practice. However, the only formal difference between the two concepts is that the direction of change is a *variable* in the total derivative, and a *fixed vector* for any concrete directional derivative!

To see this rather abstract concept in action, consider the following scenario: suppose you care only about studying and sleeping, so that your utility function may be written as

$$u(p,s) = 5\sqrt{p} - (8-s)^2$$

where p is the number of pages you read in your favorite economics textbook in a day, and s is the number of hours of sleep you are getting per night. Suppose you are currently getting 6 hours of sleep and reading 36 pages. You are thinking about reading *just a bit* more at the expense of sleeping. Assuming that you can read 4 pages in 1 hour, the vector characterizing how you marginally exchange pages for sleep is $(dp, ds) = (4, -1)$. So, how does your utility change as you move towards reading more and sleeping less? Let us consult the total derivative:

$$du(p_0,s_0) = \frac{5}{2\sqrt{p_0}}dp + 2(8-s_0)ds.$$

Plugging in your current schedule $(p_0, s_0) = (36, 6)$ and the variation vector (dp, ds), you get

$$du(36,6) = \frac{5}{2\sqrt{36}} \cdot 4 + 2 \cdot (8-6) \cdot (-1) = \frac{5}{3} - 4 = -\frac{7}{3}.$$

Thus, your utility is decreasing, and you should in fact be reading less and sleeping more!

If, on the other hand, you currently manage to read the 36 pages and still get 7 hours of sleep, and you are more efficient at reading, managing 6 pages per hour, things look different:

$$du(36,7) = \frac{5}{2\sqrt{36}} \cdot 6 + 2 \cdot (8-7) \cdot (-1) = \frac{5}{2} - 2 = \frac{1}{2}.$$

The idea we saw above for two arguments can, of course, be generalized to arbitrary functions $f : \mathbb{R}^n \mapsto \mathbb{R}$. Defining the total derivative as a *function* of the considered location x_0 and the direction of change $dx = (dx_1, \ldots, dx_n)$, we write

$$df : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}, (x_0, dx) \mapsto df(x_0, dx) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)dx_i$$

or more compactly

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}dx_i = \nabla f \cdot dx. \tag{19}$$

In economics, these considerations are valuable in theoretical models when we are doing *comparative statics*, i.e. we consider how some equilibrium state (corresponding to x_0) and an economic output quantity, e.g. $GDP(x_0)$, marginally responds to exogenous impulses that change economic quantities in fixed ratios, e.g. technology shocks that increase capital pro-

ductivity twice as much as labor productivity. Oftentimes, as the example just given already hints at, we will not choose these the ratio of changes ad hoc, but assume that they are driven by some background variable, such as technology shocks. If z denotes the level of technology in the economy, we consider $f = GDP$ as the outcome, and x_i are all relevant determinants of GDP, to highlight that the ratios are driven by the technology variable and endogenously depend upon it, we would write accordingly:

$$\frac{df}{dz} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dz} = \nabla f \cdot \frac{dx}{dz}. \tag{20}$$

Note that this relationship also directly follows from the chain rule: x_i varies with z, such that x can be written as a function of z: $x = x(z)$. Then, the outcome we consider is actually a composite function $f(x(z)) = (f \circ x)(z)$, and the chain rule gives

$$\begin{aligned}\frac{df \circ x}{dz} &= \left[\frac{df}{dx} \circ x\right] \cdot \frac{dx}{dz} = \left[\left(\frac{\partial f}{\partial x_1} \circ x, \ldots, \frac{\partial f}{\partial x_n} \circ x\right)\right] \cdot \begin{pmatrix} \frac{dx_1}{dz} \\ \vdots \\ \frac{dx_n}{dz} \end{pmatrix} \\ &= \sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i} \circ x\right] \cdot \frac{dx_i}{dz}\end{aligned}$$

so that with $x_0 = x(z_0)$:

$$\frac{df \circ x}{dz}(z_0) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) \cdot \frac{dx_i}{dz}(z_0).$$

Before moving on, a note of caution: equations (19) and (20) look quite similar, indeed, the naive mathematician could think that the latter could be obtained from just dividing the former by "dz". Of course, this is in no way a well-defined operation, as dz is not a well-defined mathematical object, but arises only from our notational convention for the derivative of x with respect to z, $\frac{dx}{dz}$. Thus, remember that the total derivative works also when the direction of change is determined implicitly through a background variable z, but that this result does **not** follow by dividing the total derivative by the change dz! Moreover, **do not extrapolate the total derivative to non-marginal changes!** The concept, by its definition, captures an *instantaneous* variation relying on the Taylor approximation, and for non-marginal changes, the linear approximation is by no means guaranteed to fare well. To illustrate correct and false interpretation in an example, re-consider our reading-and-sleeping utility: here, we saw that moving *marginally* in direction $(6, -1)$ (exchange rate 6 units of pages for one unit of sleep) when starting from the status quo of $(36, 7)$ could increase utility *locally around this point*. When considering instead the non-marginal change of reading 6 more pages at the expense of sleeping one less hour, one gets

$$\Delta u(p, s) = u(42, 6) - u(36, 7) = 5(\sqrt{42} - \sqrt{36}) - (8-6)^2 + (8-7)^2 \approx 5(6.48 - 6) - 4 + 1 = -0.6,$$

and thus a *loss* in utility.

You should take away from this:

1. The standard form of the total derivative,

$$df(x_0, dx) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) dx_i,$$

corresponds to the **instantaneous variation** of f from x_0 in the *relative* change direction $dx = (dx_1, \ldots, dx_n)'$, and is nothing but a *directional derivative* with variable direction.

2. The total derivative with a "background variable" in the denominator,

$$\frac{df}{dz} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dz},$$

captures the same concept and follows directly from the chain rule.

3. Some less formal courses derive 2. from 1. by dividing the equation by dz. This is in no way well-defined and should *not* be how you memorize these concepts!

4. The arguments dx_i, $i \in \{1, \ldots, n\}$, express how argument i changes *marginally relative to the other arguments* x_j, $j \neq i$, and does **not** represent an absolute change in the i-th argument!

3.2.8 Differentiability and Continuity

Let us consider for a moment what we summarized for the univariate derivative's properties: while we saw the linear approximation generalization in the previous subsection, we have so far not addressed the other two points. Thus, let's address continuity here. The insight from the univariate context that we want to start from is that if $X \subseteq \mathbb{R}$ and $f : X \mapsto \mathbb{R}$ is differentiable, then f is continuous.

The first critical point here is that *existence of all partial derivatives is NOT sufficient to guarantee continuity* of a multivariate function f at a point x_0. To see this, look at the following function:

$$f : \mathbb{R}^2 \mapsto \mathbb{R}, x \mapsto f(x, y) = \begin{cases} 0 & \text{if } xy = 0 \\ 1 & \text{else} \end{cases}.$$

It is partially differentiable with respect to x and y with partial derivative $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = 0$, but it is not continuous in $(0,0)$.

However, our usual logic goes through! You can check from the definition of continuity that if $\lim_{\|h\| \to 0} \|f(x_0 + h) - f(x_0)\| = 0$ is equivalent to continuity.[25]

In the way we have defined continuity,

$$0 = \lim_{{}_n\|h\| \to 0} \frac{{}_m\|f(x_0 + h) - f(x_0) - D_f(x_0)h\|}{{}_n\|h\|}.$$

[25] If $\lim_{\|h\| \to 0} \|f(x_0+h) - f(x_0)\| = 0$, then $\forall \varepsilon > 0 \exists \delta > 0$ such that $\forall h \in B_\delta(\mathbf{0}) : \|f(x_0+h) - f(x_0)\| < \varepsilon$, or equivalently, $\forall x \in B_\delta(\mathbf{x_0}) : \|f(x_0 + h) - f(x_0)\| < \varepsilon$.

This is helpful because the inverse triangle inequality of the norm tells us that

$$\frac{{}_m\|f(x_0+h)-f(x_0)-D_f(x_0)h\|}{{}_n\|h\|} \geq \frac{\left|{}_m\|f(x_0+h)-f(x_0)\| - {}_m\|D_f(x_0)h\|\right|}{{}_n\|h\|} \geq 0.$$

Now, as ${}_n\|h\| \to 0$, trivially the RHS expression, 0, "goes to" (rather: stays at) zero, and the left expression goes to zero by definition of the derivative. Here, we can use the sandwich theorem that we have used before, that tells us that because the middle expression (if you will, the toppings of the sandwich) is restricted on both sides by terms ("the bread") that go to zero, it will do so as well. By continuity of the absolute value and because $|x| = 0$ implies $x = 0$, one obtains

$$\lim_{{}_n\|h\|\to 0}\left(\frac{{}_m\|f(x_0+h)-f(x_0)\|}{{}_n\|h\|} - \frac{{}_m\|D_f(x_0)h\|}{{}_n\|h\|}\right) = 0.$$

By linearity of the limit, this gives

$$\lim_{{}_n\|h\|\to 0}\frac{{}_m\|f(x_0+h)-f(x_0)\|}{{}_n\|h\|} = \lim_{{}_n\|h\|\to 0}\frac{{}_m\|D_f(x_0)h\|}{{}_n\|h\|}.$$

Thus, the limit $\lim_{{}_n\|h\|\to 0}\frac{{}_m\|f(x_0+h)-f(x_0)\|}{{}_n\|h\|}$ exists! Because the denominator goes to zero, the numerator **must** so as well, otherwise, the limit is not defined (cf. L'Hôpital's rule). This gives

$$\lim_{{}_n\|h\|\to 0}{}_m\|f(x_0+h)-f(x_0)\| = 0,$$

which shows that f is continuous at x_0.

To summarize, **our intuition that differentiability implies continuity extends to the multivariate case.** However, **Watch out that mere existence of all partial derivatives is not sufficient for continuity!** On the other hand, we had shown that differentiability is implied by continuity of all partial derivatives. Thus, **a multivariate function is continuous if all partial derivatives exist and are continuous!**

3.2.9 Derivatives and Convexity

Above, it was already argued that derivatives are at the heart of optimization, an issue we are frequently concerned with in economics, and that important economic quantities are based on derivatives. For the last subsection on derivatives, we turn to a further aspect of derivatives which makes them a powerful tool in our box of mathematical skills: if the second derivatives exist for a function f, they can be used to check convexity (and thus also concavity) of said function!

First, why should we care? Recall the third important feature (and not yet generalized) of univariate derivatives: they told us whether a given function was increasing, decreasing or constant on some interval. For multivariate functions, this characterization is no longer too meaningful – if the gradient is zero everywhere, then the function will be "constant" in the same sense as a univariate function, but such functions are typically not too interesting. On the other hand, the concept of monotonicity is difficult to transfer because how f evolves along one dimension depends on the positions in the other dimension: e.g. $f(x_1, x_2) = x_1 x_2$ is

monotonically increasing in x_1 if and only if $x_2 \geq 0$. Thus, the **more convenient concept** to characterize multivariate functions is convexity, which, as we have seen, can be (more or less) easily generalized to the $\mathbb{R}^n$ case!

For the test of convexity using the second derivative, note our geometrically-motivated characterization that we derived earlier: the *multivariate function* f is convex if and only if for any $x, z \in \mathbb{R}^n$ with $z \neq 0$, the restriction of f to the *unidimensional* line $L(x,z) = \{x + tz : t \in \mathbb{R}\}$ is convex (recall Figure 11 and the elaborations in the corresponding paragraphs). Denote by $f|_{L(x,z)}$ the restriction to this line, i.e.

$$f|_{L(x,z)}(t) = f(x + tz) \qquad \forall t \in \mathbb{R}.$$

Now, if $f \in C^2(\mathbb{R}^n)$ is twice differentiable with continuous partial derivatives, then $f|_{L(x,z)}(t)$ is twice differentiable. The second derivative is obtained from the chain rule, with "outer" function f and "inner" function $t \mapsto x + tz$ and derivatives ∇f and z, respectively. We get

$$f|''_{L(x,z)}(t) = \frac{d}{dt}\left(\frac{d}{dt} f(x + tz)\right) = \frac{d}{dt}\left(((\nabla f(x + tz))')'z\right) = (H_f(x + tz)z)'z = z' H_f(x + tz)z. \tag{21}$$

At the third equality, note that we applied the chain rule to the gradient, and used that its derivative is equal to the Hessian because f is twice differentiable. Since the Hessian is the derivative of the *transposed* gradient, we get $\frac{d}{dt}(\nabla f(x+tz))' = H_f(x+tz)z$, but the outer transpose remains, which gives the third equality.[26] The last equality uses symmetry of the Hessian, given by $f \in C^2(\mathbb{R}^n)$. Once you know that you have to apply chain rule, this computation is straightforward, but note that we are taking vector derivatives here, which should make us pay attention to the order in which the isolated derivatives are combined!

Now, the following result for univariate functions will be immensely helpful:

Proposition 21. ***(Convexity of Twice Differentiable Univariate Functions)*** *Let $X \subseteq \mathbb{R}$ be a **convex**, open subset of $\mathbb{R}$ and suppose that $f \in C^2(X)$, i.e. f is a twice differentiable univariate function such that $f''(x)$ is continuous. Then, f is convex if and only if $\forall x \in X$: $f''(x) \geq 0$.*

The proof is admittedly tedious, go through it if you wish to improve your understanding of proofs and/or how to show convexity. However, it is most important that you know the result and can apply it, understanding everything more here is a plus but perhaps not too essential.

Proof. "$\Rightarrow$" Suppose that f is convex. Let $x \in X$. Note the first order Taylor expansion

$$f(x - h) = f(x) - f'(x)h + \frac{h^2}{2} f''(x + \lambda h) \quad \Leftrightarrow \quad f'(x)h = f(x) - f(x - h) + \frac{h^2}{2} f''(x + \lambda h)$$

for a $\lambda \in (0,1)$. Note that by Taylor's theorem, $\lim_{h \to 0} \frac{\varepsilon_1(x-h)}{h} = 0$. Thus, with equation (10), one

[26] This holds as $\frac{d}{dt}(g') = \left(\frac{dg}{dt}\right)'$ for any differentiable function g, i.e. the order in which we apply the derivative and transposition operator does not matter.

obtains

$$\begin{aligned} f''(x) &= 2\lim_{h\to 0}\left\{\frac{1}{h}\left(\frac{f(x+h)-f(x)-f'(x)h}{h}\right)\right\} \\ &= \lim_{h\to 0}\left\{\frac{f(x+h)+f(x-h)-2f(x)-1/2f''(x+\lambda h)h^2}{h^2}\right\} \\ &= 2\lim_{h\to 0}\left\{\frac{f(x+h)+f(x-h)-2f(x)}{h^2}\right\} - \lim_{h\to 0} f''(x+\lambda h). \end{aligned}$$

By continuity of f'', $\lim_{h\to 0} f''(x+\lambda h) = f''(\lim_{h\to 0} x+\lambda h) = f''(x)$. Adding $f''(x)$ on both sides and dividing the equation by two,

$$f''(x) = \lim_{h\to 0}\left\{\frac{f(x+h)+f(x-h)-2f(x)}{h^2}\right\}.$$

By convexity,

$$f(x) = f\left(\frac{1}{2}(x+h)+\frac{1}{2}(x-h)\right) \leq \frac{1}{2}f(x+h)+\frac{1}{2}f(x-h).$$

Multiplying by 2, it results that $f(x+h)+f(x-h)-2f(x)\geq 0$. Because also $h^2\geq 0$, for any $h\in\mathbb{R}\backslash\{0\}$, it results that $\frac{f(x+h)+f(x-h)-2f(x)}{h^2}\geq 0$. Because weak inequalities are preserved under the limit, $f''(x)\geq 0$.

"$\Leftarrow$" Now for the more interesting direction establishing our convexity check. Suppose that $\forall x\in X : f''(x)\geq 0$. Let $x,y\in X$, $\lambda\in[0,1]$, and without loss of generality suppose that $x\leq y$ (else re-label variables and set $\tilde{\lambda}=1-\lambda$). First, if $x=y$ or $\lambda\in\{0,1\}$, then it trivially holds that $f(\lambda x+(1-\lambda)y)\geq \lambda f(x)+(1-\lambda)f(y)$ with equality. Thus, suppose that $x\neq y$ and $\lambda\notin\{0,1\}$. Then,

$$f(\lambda x+(1-\lambda)y) = f(x)+(1-\lambda)(y-x)) = f(y-\lambda(x-y)).$$

The Taylor expansion gives

$$f(\lambda x+(1-\lambda)y) = f(x)+f'(x)(1-\lambda)(y-x)+\varepsilon_1^x = f(y)-f'(y)\lambda(y-x)+\varepsilon_1^y.$$

Because f is twice differentiable, there exist $\mu_x,\mu_y\in[0,1]$ such that $\varepsilon_1^x = [((1-\lambda)(y-x))^2/2]\cdot f''(x+\mu_x(1-\lambda)(y-x))$ and $\varepsilon_1^y = [(\lambda(y-x))^2/2]\cdot f''(y-\mu_y\lambda(y-x))$. Note that $x+\mu_x(1-\lambda)(y-x)\in int(X)$ and $y-\mu_y\lambda(y-x)\in int(X)$. Because $f''(\cdot)\geq 0$, $\varepsilon_1^x\geq 0$ and $\varepsilon_1^y\geq 0$. Combining the two Taylor expansions,

$$\begin{aligned} f(\lambda x+(1-\lambda)y) &= \lambda\left(f(x)+f'(x)(1-\lambda)(y-x)+\varepsilon_1^x\right)+\left(f(y)-f'(y)\lambda(y-x)+\varepsilon_1^y\right) \\ &\geq \lambda\left(f(x)+f'(x)(1-\lambda)(y-x)\right)+\left(f(y)-f'(y)\lambda(y-x)\right) \\ &= \lambda f(x)+(1-\lambda)f(y)+\lambda(1-\lambda)(y-x)(f'(y)-f'(x)) \\ &\geq \lambda f(x)+(1-\lambda)f(y). \end{aligned}$$

The last inequality follows because $y\geq x$ and $f'(y)\geq f'(x)$, which results from the fact that $f''(\cdot)$ is the slope of f', and $f''(\cdot)\geq 0$ implies that $f'(\cdot)$ is monotonically increasing. □

Corollary 8. ***(Strict Convexity of Univariate Functions)*** *Let $X\subseteq\mathbb{R}$ be a **convex** subset of $\mathbb{R}$ and*

suppose that $f \in C^2(X)$. Then, if for any $x \in \text{int}(X) : f''(x) > 0$, f is strictly convex.

This follows directly from the proof above by noting that if $x \neq y$ and $\lambda \in (0,1)$, then the errors ε_1^x and ε_1^y are *strictly* positive, which gives a strict inequality in the second line. The converse is not necessarily true, however, because strict inequalities are not preserved under the limit. Take away from this that **a sufficient condition for (strict) convexity is $f''(x) \geq 0$ ($f''(x) > 0$)!** Finally, let me comment on the restriction that X is an open set: this simply ensures that f is differentiable everywhere, i.e. that $\forall x \in X$: $x \in int(X)$. However, as you can see from the alternative formulation in the corollary, boundary points are not an issue for these considerations, so that to establish convexity of a function with closed domain, it suffices to show the derivative condition for any interior point.

As promised earlier in the chapter, we will now establish that with this corollary it is easy to see that a variety of univariate functions are strictly convex. Consider for instance $f(x) = x^2$ and $g(x) = -\sqrt{x}$ for $x \geq 0$: $f''(x) = 2 > 0$, $g''(x) = \frac{1}{4x^{3/2}} > 0$ for any $x > 0$. Thus, both functions are strictly convex, and that $\sqrt{x} = -g(x)$ is strictly concave.[27]

Now, back to our study of multivariate functions. Re-consider equation (21). How does Proposition 21 help? Well, it tells us that $f|_{L(x,z)}$ is convex if and only if $f|''_{L(x,z)} \geq 0$ and strictly convex if $f|''_{L(x,z)} > 0$. The requirement that this holds can be written as

$$\forall t \in \mathbb{R} : f|''_{L(x,z)}(t) \geq 0 \ (> 0) \quad \Leftrightarrow \quad \forall t \in \mathbb{R} : z' H_f(x + tz) z \geq 0 \ (> 0).$$

As a final step, note that we said this must hold for any $x, z \in \mathbb{R}^n$ such that $z \neq \mathbf{0}$. Note that for fixed z, because we can arbitrarily vary $x \in \mathbb{R}^n$ (and $t \in \mathbb{R}$) and thus reach *any* $y \in \mathbb{R}^n$ as $y = x + tz$, requiring $z' H_f(x + tz) z \geq 0 (> 0)$ $\forall x \in \mathbb{R}^n, t \in \mathbb{R}$ is equivalent to requiring $z' H_f(y) z \geq 0$ (> 0).[28]

Because the condition must hold for all $z \neq \mathbf{0}$, *(strict) multivariate convexity is equivalent to*

$$\forall y \in \mathbb{R}^n : \Big(\forall z \in \mathbb{R}^n \backslash \{\mathbf{0}\} : z' H_f(y) z \geq 0 \ (> 0) \Big).$$

Does the inner expression look familiar? This is precisely what we require for positive semi-definiteness (or definiteness in the case of strict inequality) of $H_f(y)$![29] Thus, we get the following proposition as the result:

Proposition 22. ***(Multivariate Convexity)*** *Let X be a **convex** subset of $\mathbb{R}^n$ and $f \in C^2(X)$. Then, f is convex if and only if, for all $x \in \text{int}(X)$, $H_f(x)$ is positive semi-definite. Further, if for all $x \in \text{int}(X)$, $H_f(x)$ is positive definite, then f is strictly convex.*

Proof. Established above. □

Note that we again focus only on interior points where the derivative exists, and that **positive semi-definiteness (positive definiteness) of the Hessian is always a sufficient condition for convexity (strict convexity)!** In this sense, the definiteness of the Hessian can be viewed

[27] Note that our interior restriction requires us to consider only points $x \in \text{int}([0, \infty)$, i.e. $x \in (0, \infty)$ or $x > 0$.

[28] We saw that if for fixed z, the condition holds for any (x, t), then it holds for any y. Equivalence holds as when the condition holds for any $y \in \mathbb{R}^n$, then of course it will for any $x + tz$, because this object is also an element of $\mathbb{R}^n$.

[29] For semi-definiteness, we usually don't exclude the zero vector, but this doesn't matter, because $\mathbf{0}' H_f(y) \mathbf{0} = 0 \geq 0$ is always trivially satisfied.

as the sign of the second derivative when generalizing our insights from univariate differential calculus.

3.3 Integral Theory

For the last bit on multivariate calculus, we turn to integration – albeit far less extensively as differentiation. The conceptual perspective here is quite the opposite as with differentiation: while thus far, we were interested in the *marginal change* of a function f, we now care about its *accumulation* in the codomain. Actually, it is quite intuitive that we should consider integration and differentiation as "inverse" operations also in a narrow sense. This is because f is the instantaneous change of its accumulation, i.e. the rate at which the area under it accumulates! Accumulation is also an issue of frequent interest to economists, e.g. when we care about aggregation of (the choices of) individual firms/households to a national economy, or when forming expectations about outcomes, where we aggregate all possible events and weight them by their probability.

Indeed, also more formally, the idea is to construct the integral as an, in an appropriate sense, inverse operator to the differential operator. Recall that the derivative, D, is an *operator* on the space of differentiable functions, mapping functions f onto their derivative $D(f)$, or, in our notation, D_f. Now, we ask ourselves: does this operator have an inverse, i.e. can we find D^{-1} such that $D^{-1}(D_f) = D(D^{-1}(f)) = f$, or respectively, can we, for any function f, find a *unique* function F such that $D(F) = D_F = f$? If we again restrict attention to univariate functions for the moment, you are likely aware that this is not possible, for the reason that constants vanish when taking derivative. Thus, $F_1(x) = x^2 + 5$ and $F_2(x) = x^2 - 2$, such that the function $f(x) = 2x$ has more than one function characterized by the feature we are looking for. In other words, the derivative is not injective, and thus, as we discussed earlier, we can not invert it!

However, similar to the non-invertible (because non-injective) function $f(x) = x^2$, we can of course define the preimage of f under the differential operator, $D^{-1}[\{f\}] = \{F : D_F = f\}$ of functions that have f as their derivative, just like we can define $f^{-1}[\{y\}] : \{x \in \mathbb{R} : y = x^2\}$ as the pre-image of any value y of $f(x) = x^2$. For the case of univariate functions, you likely know the following characterization:

$$\int f(x)dx := F(x) + C, \qquad C \in \mathbb{R}$$

where $F(x)$ is the *stem function* of f. Recall that the reason for ambiguity in the inverse derivative, or *antiderivative*, was that constants vanish. Thus, up to said constant, we should be able to uniquely pin down the antiderivative through the function $F(x)$ that does *not contain a constant*... and we indeed can! The object in this equation is called the **indefinite integral** of f and, in some generalized sense, describes a "function".[30] Note, however that the expression is a notational simplification for the pre-image of f under the differential operator, and that we describe a set here, rather than an equation.

Before moving on to the well-defined definite integral, check that you are familiar with the following rules for indefinite integrals:

[30] Since the value it maps to is not unique, we do not call it a function – recall that uniqueness of the image is part of what defines a function! Instead, we would call the indefinite integral a *correspondence*, the mathematical term for a mapping of individual elements onto sets.

Theorem 47. ***(Rules for Indefinite Integrals)*** *Let f, g be two* integrable *functions*[31] *and let $a, b \in \mathbb{R}$ be constants, $n \in \mathbb{N}$. Then*

- $\int (af(x) + g(x))dx = a\int f(x)dx + \int g(x)dx$,
- $\int x^n dx = \frac{x^{n+1}}{n+1} + C$ *if* $n \neq -1$ *and* $\int \frac{1}{x}dx = \ln(x) + C$,
- $\int e^x dx = e^x + C$ *and* $\int e^{f(x)}f'(x)dx = e^{f(x)} + C$,
- $\int (f(x))^n f'(x)dx = \frac{1}{n+1}(f(x))^{n+1} + C$ *if* $n \neq -1$ *and* $\int \frac{f(x)}{f'(x)}dx = \ln(f(x)) + C$.

Another important rule, which can be thought of as the reverse of the product rule, is integration by parts:

Theorem 48. ***(Integration by parts)*** *Let u, v be two differentiable functions. Then,*

$$\int u(x)v'(x)dx = u(x)v(x) - \int u'(x)v(x)dx.$$

3.3.1 Definite Integrals and the Fundamental Theorem of Calculus

Remaining with univariate functions for now, we know that a unique function F can be attributed to $f : X \mapsto \mathbb{R}$ such that $DF = f$ if we require that F does *not* contain a constant. For simplicity, let's focus on convex sets X, i.e. intervals. So, how do we compute F? The idea is quite easy: note that while the antiderivative is not well-defined in general because of the constants C, for any function $\tilde{F} \in D^{-1}[\{f\}]$, i.e. any function that satisfies $\tilde{F}(x) = F(x) + C$, for specific values $x, y \in X$, we have that $\tilde{F}(y) - \tilde{F}(x) = F(y) + C - (F(x) + C) = F(y) - F(x)$. Supposing that $y \geq x$, this can be used to compute the **uniquely defined definite intergral** that tells us the accumulation of f from x to y, that is, on the interval (x, y),[32]

Definition 75. ***(Definite Integral)*** *Let $X \subseteq \mathbb{R}$ and consider an integrable function $f : X \mapsto \mathbb{R}$. Then, the definite integral of f from $x \in X$ to $y \in X$, is*

$$\int_x^y f(t)dt = \tilde{F}(y) - \tilde{F}(x), \quad \textit{where} \quad \tilde{F}(x) \in D^{-1}[\{f\}].$$

This gives us the usual rule that you are likely familiar with: to compute $\int_x^y f(t)dt$, compute the stem function F, and take the difference $F(y) - F(x)$. For instance, the stem function of $f(x) = 4x^3$ is x^4, such that $\int_{-1}^3 4x^3 dx = 3^4 - (-1)^4 = 81 - 1 = 80$.

Before moving on, keep the following in mind: **the inverse of the differential operator D is not generally well-defined. However, any function $\tilde{F}$ in the preimage of a function f under D is characterized by a uniquely defined accumulation between any two points x and y, called the definitive integral of f.** Because we like uniquely defined quantities, we mostly restrict attention to this object – indeed, when we care about accumulation as we mostly do when considering antiderivatives, it's all we need.

[31] Although there is a more formal definition, let it suffice here to understand an integrable function as being a functions whose integral you can compute with the usual rules.

[32] It does not matter if we include x and y to this interval or not, because at single points, there is no accumulation. Thus, to be more general and allow that $x, y \in \{\pm\infty\}$, we consider the open interval.

Definition 76. ***(Infimum and Supremum of a Set)*** *Let $X \subseteq \mathbb{R}$. Then, the infimum* $\inf(X)$ *of X is the largest value smaller than any element of X, i.e.* $\inf(X) = \max\{a \in \mathbb{R} : \forall x \in X : x \geq a\}$*, and the supremum* $\sup(X)$ *of X is the smallest value larger than any element of X, i.e.* $\sup(X) = \min\{b \in \mathbb{R} : \forall x \in X : x \leq b\}$.

These concepts are a helpful generalization of maximum and minimum, and exist under much more general conditions. For instance, for an interval (a, b), there is no maximum or minimum, but infimum and supremum exist and are equal to a and b, respectively. We need these concepts here the theorem below to ensure that a always defines an interval X as $X = (a, b)$, $X = [a, b)$, $X = (a, b]$ or $X = [a, b]$), regardless of whether the lower bound is open or closed. Note that we may have $a = -\infty$.

Theorem 49. ***(Fundamental Theorem of Calculus)*** *Let X be an interval in $\mathbb{R}$ and $f : X \mapsto \mathbb{R}$. Let $a = \inf(X)$, suppose that f is integrable, and define $F := X \mapsto \mathbb{R}, x \mapsto \int_a^x f(t)dt$. Then, F is differentiable, and*

$$\forall x \in X : F'(x) = D_F(x) = f(x).$$

Proof. (slightly informal) We take for granted that (i) if $f(x) \leq g(x)\ \forall x \in X$, then $\int_a^x f(t)dt \leq \int_a^x g(t)dt\ \forall x \in X$. Let $x \in X$. Then, observe that for $h \geq 0$,

$$h \min\{f(s) : s \in [x, x+h]\} = \int_x^{x+h} \min\{f(s) : s \in [x, x+h]\}dt \leq h \max\{f(s) : s \in [x, x+h]\}.$$

The converse can be shown for $h \leq 0$, where by definition of the indefinite integral, $\int_x^{x+h} f(t)dt = -\int_{x-|h|}^x f(t)dt$. In any case, for $h \neq 0$, one arrives at

$$\min\{f(s) : s \in \bar{B}_{h/2}(x + h/2)\} \leq \frac{1}{h}\int_x^{x+h} f(x)dt \leq \max\{f(s) : s \in \bar{B}_{h/2}(x + h/2)\}$$

where $B_{h/2}(x+h/2)$ denotes the closed ball of radius $h/2$ around $x+h/2$.[33] We can show (omitted) that under some regularity ensuring integrability of f, both the LHS and RHS quantities go to $f(x)$ as $h \to 0$, intuitively because x is the only element that remains in $\bar{B}_{h/2}(x + h/2)$ as $h \to 0$. Thus, by the sandwich theorem, also the middle expression does, i.e.

$$f(x) = \lim_{h\to 0} \frac{1}{h}\int_x^{x+h} f(x)dt = \lim_{h\to 0} \frac{1}{h}\left(\int_a^{x+h} f(x)dt - \int_a^x f(x)dt\right) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = D_F(x). \quad \square$$

3.3.2 Multivariate Integrals

As we did with the derivatives, let us extend the notion of the integral to the multivariate case by first looking at a function mapping from $\mathbb{R}^2$ to $\mathbb{R}$. If in the univariate case, the definite integral was measuring an area under the graph, it is now only natural to require the definite integral to measure the volume under the graph. In higher dimensions, we have to go on without graphic illustrations, but the concept of "summing up the function values over infinitely

[33] This is more convenient notation because the relevant set is $[x, x+h]$ if $h \geq 0$ but $[x+h, x]$ if $h < 0$, and the closed ball expresses both these quantities. We consider the closed ball so as to have x in there, which facilitates the consideration $h \to 0$.

small areas of the domain" remains valid. Also, indefinite integrals should still be considered the antiderivative, but now to the multivariate derivative, and again intuition might fail us here. Luckily, for probably all intends and purposes that you will come across integrals in your master courses, the following theorem will be of practical help:

Theorem 50. ***(Fubini's theorem)*** *Let X and Y be two intervals in $\mathbb{R}$, let $f : X \times Y \to \mathbb{R}$ and suppose that f is continuous. Then, for any $I = I_x \times I_y \subseteq X \times Y$ with intervals $I_x \subseteq X$ and $I_y \subseteq Y$,*

$$\int_I f(x,y)d(x,y) = \int_{I_x}\left(\int_{I_y} f(x,y)dy\right)dx,$$

and all the integrals on the right-hand side are well-defined.

It tells us that when concerned with a multi-dimensional integral, we can integrate with respect to each dimension (or fundamental direction) "in isolation" or rather, integrate in an arbitrary succession with respect to all the single variables. The theorem is pretty powerful as it only needs continuity of the function as a prerequisite, and then allows you to reduce a multivariate integral to a lower-dimensional one! You can also apply the theorem repeatedly if you are faced with higher dimensional integrals, so that

$$\int_I f(x_1,\dots,x_n)d(x_1,\dots,x_n) = \int_{I_1}\left(\cdots\left(\int_{I_n} f(x_1,\dots,x_n)dx_n\right)\cdots\right)dx_1.$$

Thus, a scheme applies that is very similar to what we have seen for differentiation of multivariate functions: if the operation can be performed, that is, here if we can integrate the function f, **so long as f satisfies a continuity condition, then the multivariate version of the operation can be computed by repeatedly applying the univariate concept subject to a certain scheme of ordering!**

For a final property, note that linearity of the integral implies especially that we can pull constants with respect to the integrating variable x, i.e. any expression $c(z)$ that may depend on arbitrary variables z but not on x, out of the integral, so that $\int c(z)f(x)dx = c(z)\int f(x)dx$. Thus, we obtain the following corollary of Fubini's theorem:

Corollary 9. ***(Integration of Multiplicatively Separable Functions)*** *Let $X_f \in \mathbb{R}^n, X_b \in \mathbb{R}^m$, $f : X_f \to \mathbb{R}$, $g : X_b \to \mathbb{R}$ continuous functions. Then, for any intervals $A \subseteq X_f$, $B \subseteq X_g$,*

$$\int_{A\times B} f(x)g(y)d(x,y) = \left(\int_A f(x)dx\right)\left(\int_B g(y)dy\right).$$

Proof. Define $h : X_f \times X_g \to \mathbb{R}$ with $h(x,y) = f(x)g(y)$. Then, this function is continuous and we can apply Fubini's theorem:

$$\begin{aligned}\int_{A\times B} f(x)g(y)d(x,y) &= \int_A\left(\int_B f(x)g(y)dy\right)dx\\ &= \int_A f(x)\left(\int_B g(y)dy\right)dx\\ &= \left(\int_A f(x)dx\right)\left(\int_B g(y)dy\right)\end{aligned}$$

where the second equality follows because $f(x)$ is a constant in terms of integration with respect to y and the third because $\int_B g(y)dy$ is a constant in terms of integration with respect to x. □

Note that f and g can be multivariate, so that whenever you can separate a function into two factors that depend on a disjoint subset of variables, you can multiplicatively separate integration! An important economic example is for instance the Cobb-Douglas production function: Suppose that firms' stock of capital k and labor l are independently and uniformly distributed on $[0,1]$,[34] so that individual level output is $y = f(k,l) = Ak^\alpha l^{1-\alpha}$. Here, the theorem for multiplicatively separable variables can help us determine the aggregated output of the whole economy Y as

$$\begin{aligned} Y &= \int_{[0,1]\times[0,1]} f(k,l)d(k,l) = \int_{[0,1]\times[0,1]} Ak^\alpha l^{1-\alpha} d(k,l) = A\left(\int_{[0,1]} k^\alpha dk\right)\left(\int_{[0,1]} l^{1-\alpha} dl\right) \\ &= A\left(\left[\frac{x^{2-\alpha}}{(2-\alpha)}\right]_{x=0}^{x=1}\right)\left(\left[\frac{x^{1+\alpha}}{(1+\alpha)}\right]_{x=0}^{x=1}\right) = \frac{A}{(2-\alpha)(1+\alpha)}. \end{aligned}$$

To conclude this section on integrals, take away that (i) the differential operator can generally not be inverted, (ii) the definite integral, referring to the accumulation of a function between to points, can be well-defined nonetheless, and corresponds to the usual integral you are familiar with, and (iii) that like differentiation, we can handle multivariate integration by applying techniques for univariate functions according to a certain scheme of ordering, which applies under rather general conditions.

[34] It is not too important what this means here, it just ensures that the first equality below holds.

3.4 Contents and Take-Aways

Chapter 3: Multivariate Calculus discusses the fundamental concepts of functional analysis in metric spaces, especially spaces of vectors of real numbers, including

- invertibility and convexity and concavity
- differentiation: the approach to generalizing arbitrary derivatives from the ones of univariate functions and how multivariate derivatives are defined and how they can be computed
- Taylor approximations, Taylor expansions and total derivatives of multivariate functions
- integration: conceptual basics and rules for computing multivariate integrals

Someone with profound knowledge of the contents of this chapter should

- be able to perform common operations on functions (addition, multiplication, composition)
- be familiar with terminology related to vector-valued functions, and know e.g. the conceptual difference between a multivariate real-valued function and a univariate vector-valued function
- have a graphical intuition for how convexity generalizes from univariate to multivariate functions
- know the concepts of quasi-convexity and quasi-concavity and how to investigate whether a certain function satisfies them
- know the formal definition of a function's derivative, and how the definition of multivariate derivatives is motivated and generalized from the one of univariate derivatives
- be aware of the 3 conceptual levels of objects in differential calculus: operators, functions, values
- be familiar with gradients, Jacobians and Hessians, and their relation to multivariate derivatives
- know how the total derivative can be used to study economic trade-offs and indirect effects
- be able to investigate multivariate convexity using the second derivative
- be familiar with the definitions of a set's infimum and supremum
- know how to compute multivariate integrals using Fubini's Theorem (= iterative integration)

and be able to answer a number of related questions, including

- How do injectivity and surjectivity relate to invertibility? What about bijectivity?
- When and how can a "matrix function" ($f : \mathbb{R}^n \mapsto \mathbb{R}^n, x \mapsto Ax$ where $A \in \mathbb{R}^{n \times n}$) be inverted?
- Which criterion must a function satisfy to be an element of the set $D^k(X, \mathbb{R})$? What about $C^k(X, \mathbb{R})$?
- Does a multivariate version of the chain rule for the derivative exist? If so, does the order in which the derivative's elements are multiplied with each other matter?
- What is the difference between a Taylor expansion and a Taylor approximation? Is either one of these concepts always equal to the underlying function?
- Why, mathematically, is the Taylor approximation for f at x_0 is a "good" around x_0?
- Is any norm $\|\cdot\|$ continuous in the metric space $(\mathbb{X}, \|\cdot\|)$? What's the intuition for this?
- How is the limit $g_0 = \lim_{\|h\| \to 0} g(h)$ formally defined (ε/δ statement) when g has domain $\mathbb{R}^n$, $n > 1$?
- Is it sufficient for differentiability of f that all partial derivatives of f exist?
- Is every Hessian a Jacobian? Is every Jacobian a Hessian?
- Why is the differential operator not invertible? Roughly, how does the definite integral circumvent this issue?

3.5 Recap Questions

1. Apply point 3. of the first section to the functions $f_1(x) = x^2$, $f_2(x) = \min\{x, 5\}$ and $f_3(x) = \sin(x)$. On which intervals are these functions (i) constant, (ii) increasing and (iii) strictly increasing?
2. Define the gradient, the Jacobian and the Hessian, and pay special attention to the type of function regarding the domain and codomain, they concern. When, if ever, is the gradient equal to the Jacobian?
3. Discuss the nature of the concepts $\frac{\partial}{\partial x_j}$, $\frac{\partial f}{\partial x_j}$, $\frac{\partial f}{\partial x_j}(x_0)$ and $\frac{\partial f(x_0)}{\partial x_j}$, especially the difference between them and how they relate to each other.
4. How is the total differential of a function $f(x_1, x_2, x_3)$ defined? What does it tell us? How does it relate to Taylor's theorem?
5. Integrate $f(x_1, x_2) = \sin(x_1)e^{2x_2}$ over $[0, \pi/2] \times [0, \ln(2)]$. (Hint: think about the corollary from Fubini.)

4 Optimization

Disclaimer: This script focuses on studying the Lagrangian technique, its formal justification and intuition, and its generalization to inequality constraints and the Kuhn-Tucker method, which, at least for the coursework in all Economics Master tracks in Mannheim, will be sufficient for solving most if not all optimization problems. Earlier versions of this chapter have, as was also the case with the last chapter, focused more on graphical representations of formal details, and also discussed convex optimization, which is left out here. These notes are available at https://helmsmueller.wordpress.com/teaching/.

There is no need to emphasize the importance of optimization in economics. If you have studied economics, you already have experienced it. If not, you will in the coming months. This chapter focuses on multivariate real-valued functions for two reasons: they (i) represent the kind of functions you will mostly be working with, and (ii) allow us not to dig too far into order theory. It is important to keep in mind, however, that the methods we will use do not depend specifically on this fact.[1]

After introducing a last pinch of vocabulary, we proceed with optimization techniques, which are an application of the concepts we have studied in the previous chapter. We will start with the unconstrained case, which should make it easier for you to relate the unfamiliar multivariate case to the more familiar univariate case. We will then consider two important classes of constrained optimization, namely, convex programming with inequality constraints and convex programming with equality constraints. Throughout the entire chapter, we will most often consider maximization problems. That will not seem very natural if you studied mathematics or engineering, but it seems most natural to economists. You may consider the material as understood if you feel confident rephrasing everything in terms of minimization problems.

As a first step to solving it, let us consider how to write down a constrained optimization problem mathematically.

$$(\mathcal{P}_{min}) \qquad \begin{aligned} &\underset{x\in \text{dom}(f)}{\text{minimize}} && f(x) \\ &\text{subject to} && g_i(x) = 0,\ i = 1,\dots,m. \\ & && h_i(x) \leq 0,\ i = 1,\dots,k. \end{aligned}$$

$$(\mathcal{P}_{max}) \qquad \begin{aligned} &\underset{x\in \text{dom}(f)}{\text{maximize}} && f(x) \\ &\text{subject to} && g_i(x) = 0,\ i = 1,\dots,m. \\ & && h_i(x) \leq 0,\ i = 1,\dots,k. \end{aligned}$$

Here, f, g_i, $i \in \{1,\dots,m\}$ and h_i, $i \in \{1,\dots,k\}$ are functions mapping from the set of x's, the domain $\text{dom}(f)$ (previously denoted by X), to $\mathbb{R}$ (due to our focus on real-valued functions). We call f our *objective function* and x the *choice variables*. In words, we are looking for the x in the domain of f that yields either the smallest (or largest) value f can possibly attain when we require that the functions g_i must attain the value 0 and h_i can not lie strictly above 0 – we want to minimize (or maximize) f subject to the *equality constraints* given by the g_i and the *inequality constraints* by h_i. You may wonder if it is not a restrictive formulation to have only equality constraints requiring a function to be equal to zero, and

[1] In case you're interested, the methods actually depend on the fact that we are working in a vector space with a specific structure, namely a Hilbert space, i.e. a complete vector space with an inner product. Even the requirement of the existence of an inner product may be partially relaxed, and if you are to face such problems, then simple generalizations of our methods are available and nicely described in books such as Luenberger (1969). Optimization by vector space methods. John Wiley & Sons, 1969. You may actually hear about duality theory during your master's years. This theory is specifically concerned about the extension of the methods we present here to vector spaces that cannot be equipped with an inner product. It is also useful in Hilbert spaces, as it captures some of the geometrical intuition behind our techniques.

"less or equal" inequalities. Indeed, it is not: note that we can always re-write a condition $\tilde{g}(x) = c$ as $g(x) := \tilde{g}(x) - c = 0$ and $\tilde{h}(x) \geq c$ as $h(x) := c - \tilde{h}(x) \leq 0$. Lastly, if $m = k = 0$, we say that the problem is *unconstrained*.

As you may have noticed until now, this script puts great emphasis on characterizing any mathematical expression or object by its type, recall our discussions of differential operators (functions between function spaces), derivatives (functions with the same domain as the function of interest) and derivatives evaluated at certain elements in the domain (real numbers, vectors or matrices). Also here, please take a second to appreciate the mathematical nature of the optimization problem – and especially that it does not constitute any structure that we have introduced so far. It is neither equal to its solution(s), which may or may not exist and may or may not be unique, nor to the maximum value f may attain under the restrictions – these quantities will be defined separately shortly! As such, we need a new label, and we will use the one you are already familiar with: the optimization problem. Thus, when you read "optimization problem" in the following, note that it is a distinct mathematical object that we have now defined properly.

4.1 Some last Vocabulary and Basic Results

Before starting the whole optimization process, it is wise to (i) formally define what we are looking for (i.e. maximum or minimum) and (ii) evaluate how likely it is that we will find it. The second task is generally achieved by looking at specific properties of the objective function and of its domain. So let us proceed!

4.1.1 Definitions

Definition 77. ***(Extremum: Minimum and Maximum)*** *Let $X \subseteq \mathbb{R}$. Then, $\bar{x} \in \mathbb{R}$ is called the maximum of X, denoted $\bar{x} = \max X$, if $\bar{x} = \sup(X)$ and $\bar{x} \in X$. Conversely, $\underline{x} \in \mathbb{R}$ is called the minimum of X, denoted $\underline{x} = \min X$, if $\underline{x} = \inf(X)$ and $\underline{x} \in X$. $x \in \mathbb{R}$ is called an extremum of X if $x = \max X$ or $x = \min X$.*

Verbally, x is the maximum of X if x is the smallest number greater or equal than all elements of X, and it is also contained in X. More simply put, it is the largest value in the set if there is any such value. This need not be the case, recall that e.g. open intervals (a, b) neither have a maximum nor a minimum. However, for sets of real numbers (since sets don't contain duplicates and there is a strict ordering of the real numbers, such that for any $x, y \in \mathbb{R}$ with $x \neq y$, either $x > y$ or $x < y$), if there is a minimum or a maximum, it (i) is unique and (ii) coincides with the infimum or the supremum, respectively.

Already at this stage one can get a hint at why real-valuedness of the function makes things easier. How would you define inequalities of vectors, i.e. when would you say that for $x, y \in \mathbb{R}^m$, $x \geq y$? And how would you find the maximum of a set $X \subseteq \mathbb{R}^m$?[2]

In terms of our optimization problem, we will be looking for the maximum of the set of attainable values of f under the constraints of the problem $\mathcal{P}$:

$$A_{\mathcal{P}}(f) = \{f(x) : x \in \text{dom}(f),\ g_i(x) = 0,\ h_j(x) \leq 0\ \forall i \in \{1, \dots, m\} \forall j \in \{1, \dots, k\}\}.$$

Now, next to the maximum attainable value, we are frequently also (most of the time: even more) interested in the (set of) solution(s), i.e. the arguments x that maximize f under the constraints of $\mathcal{P}$. So, let's define them in a next step:

Definition 78. ***(Local and Global Maximizers)*** *Let $X \subseteq \mathbb{R}^n$, $f : X \mapsto \mathbb{R}$. Then, $x_0 \in X$ is*

[2] There exist orderings for the $\mathbb{R}^m$ as well, for instance the relation (recall Chapter 0) on $R^m \times R^m$ defined by $\forall (x, y) \in \mathbb{R}^m \times R^m : (x \geq y \Leftrightarrow x_i \geq y_i \forall i \in \{1, \dots, m\})$, or the strict version where $x > y$ if $x \geq y$ and one equality $x_i \geq y_i$ is strict. But those lack the very basic property of completeness, such that there exist $x, y \in \mathbb{R}^m$ for which neither $x \geq y$ or $y \geq x$, which immensely complicates things like search for extreme values.

- *a global maximizer for f if* $\forall x \in X : f(x_0) \geq f(x)$
- *a strict global maximizer for f if* $\forall x \in X \backslash \{x_0\} : f(x_0) > f(x)$.
- *a local maximizer for f if there exists $\varepsilon > 0$ such that* $\forall x \in X \cap B_\varepsilon(x_0) : f(x_0) \geq f(x)$
- *a strict local maximizer for f if there exists $\varepsilon > 0$ such that* $\forall x \in X \cap B_\varepsilon(x_0) \backslash \{x_0\} : f(x_0) > f(x)$

Verbally, local means that there must be a neighborhood (i.e. an open ball $B_\varepsilon(x_0)$ around x_0) such that x_0 maximizes f in this neighborhood, or respectively, it maximizes the restricted function

$$f|_{B_\varepsilon(x_0) \cap X} : B_\varepsilon(x_0) \cap X \mapsto \mathbb{R}, x \mapsto f(x).$$

Strict means that x_0 is the unique maximizer of f (local: when restricted to a neighborhood), such that all other x (in this neighborhood) yield strictly lower values for f if f is defined in x, i.e. $x \in X$. Note that local implies global, because if all elements of X yield smaller values for f than x_0, then so do especially all those that lie in neighborhoods of x_0. Note that the local maximizer concept must restrict the ball $B_\varepsilon(x_0)$ to the set X to ensure that the global maximizer concept is strictly more powerful![3] That being said, this is a rather technical comment – the take-away is clearly that global maximizers are always local maximizers, but the converse is not always true. In optimization, infrequently care about the whole set X, rather, we are interested in the restriction $f|_C : C \mapsto \mathbb{R}, x \mapsto f(x)$ where $C \subseteq X$, because not all $x \in X$ satisfy the constraints of our problem. Thus, we need some more definitions (the first of which was already used repeatedly above):

Definition 79. ***(Restriction of a Function)*** *Let X, Y be sets and $f : X \mapsto Y$ a function. Then, for $S \subseteq X$, the function $f|_S : S \mapsto Y, x \mapsto f(x)$ is called the restriction of f on S.*

Having defined the restriction, the deliberations below will no longer always spell it out explicitly in the following, so be sure that you know what is meant when you see an expression like $f|_S$, i.e. f followed by a large vertical line and a subscript!

Definition 80. ***(Constraint Set)*** *Consider an optimization problem $\mathcal{P}$ with objective function $f : X \mapsto \mathbb{R}$, $X \subseteq \mathbb{R}^n$, equality constraints $g_i(x) = 0 \; \forall i \in \{1, \ldots, m\}$ and inequality constraints $h_j(x) \leq 0 \; \forall j \in \{1, \ldots, k\}$. Then, the set*

$$C(\mathcal{P}) := \{x \in X : ((\forall i \in \{1, \ldots, m\} : g_i(x) = 0) \wedge (\forall j \in \{1, \ldots, k\} : h_j(x) \leq 0))\}$$

is called the constraint set of $\mathcal{P}$.

The constraint set of a problem $\mathcal{P}$ defines the restriction $f|_{C(P)}$. Beyond now being able to write optimization problems in one line as

$$\underset{x \in C(\mathcal{P})}{\text{maximize}} f(x) \quad \text{with} \quad C(\mathcal{P}) = \{x \in X : ((\forall i \in \{1, \ldots, m\} : g_i(x) = 0) \wedge (\forall j \in \{1, \ldots, k\} : h_j(x) \leq 0)),$$

the value of having defined the constraint set is that it more formally narrows down what we are indeed after when considering the problem $\mathcal{P}$ mathematically: finding the maximizers of the restricted function $f|_{C(\mathcal{P})}$! Note that for an unconstrained problem, we simply have $C(\mathcal{P}) = X$ so that $f|_{C(\mathcal{P})} = f$. Beyond the general maximizers defined above, this allows us to define the maximizers we care about in optimization problems:

[3] Else, boundary points could be local but not global maximizers because there exists no neighborhood around them on which f is defined. For instance, the function $f : [0, \infty) \mapsto \mathbb{R}, x \mapsto -\sqrt{x}$ clearly has a global maximizer at $x = 0$, but in every neighborhood $(-\varepsilon, \varepsilon)$ of this point with $\varepsilon > 0$, there exist points $x < 0$ where f is not defined!

Definition 81. ***(Solutions, Maximizing Arguments)*** *Consider an optimization problem $\mathcal{P}$ with constraint set $C(\mathcal{P})$. The* ***solutions*** *to $\mathcal{P}$ are given by the* ***set*** *of global maximizers of $f|_{C(\mathcal{P})}$,*

$$\arg\max_{x\in C(\mathcal{P})} f(x) := \{x_0 \in C(\mathcal{P}) : (\forall x \in C(\mathcal{P}) : f(x_0) \geq f(x)\}.$$

We call $\arg\max_{x\in C(\mathcal{P})} f(x)$ the maximizing arguments or the arg max of the problem $\mathcal{P}$. If this set contains only a single element x^*, we also write $x^* = \arg\max_{x\in C(\mathcal{P})} f(x)$.

As indicated by the bold words, it is crucial to note that the solutions typically constitute a set that may contain arbitrarily many elements or none at all (consider e.g. $\arg\max_{x\in\mathbb{R}} x^2 = \emptyset$ or $\arg\max_{x\in\mathbb{R}} \cos(x) = \{2\pi n : n \in \mathbb{N}\}$). Only if there is a unique maximizer, by convention, the arg max *also* refers to a number or a vector (but still to the set as well)![4] Finally, note that regardless of whether the problem $\mathcal{P}$ has solutions, and even regardless of whether there are any values that satisfy the constraints (i.e., whether $C(\mathcal{P}) \neq \emptyset$)), the arg max is always defined!

This definition concludes our introductory vocabulary section. Thus, it is a good time to make sure that we have our notation straight. Perhaps one of the most important notational relationships in the optimization context is the distinction between the following two objects that you frequently see across all economic fields, and applications of mathematics more generally:

$$\max_{x\in C(\mathcal{P})} f(x) \quad \text{and} \quad \arg\max_{x\in C(\mathcal{P})} f(x).$$

Hopefully, the "arg max" has become clear from the elaborations above, so let me focus on the first object. First, note that it must be a convention, because with our definitions thus far, we are not able to narrow down what precisely this is supposed to mean! Indeed, it is "shorthand" notation for $\max\{f(x) : x \in C(\mathcal{P})\}$, i.e. the maximum of the set of values that f can attain on the constraint set $C(\mathcal{P})$. So, it is a *real number* rather than a set, in contrast to the general definition of the arg max! Further, unlike the arg max, it need not always exist, recall that many sets, even very simple ones, do not have a maximum. A further distinction is that if existent, $\max_{x\in C(\mathcal{P})} f(x) \in \mathbb{R}$, i.e. **this object is an object of the codomain** of f, whereas $\arg\max_{x\in C(\mathcal{P})} f(x) \subseteq X$, i.e. **the arg max is a subset of the domain of X!** Despite all these differences, there is one key relationship that, if you understand it, you know everything there is to take away here:

$$\forall x^* \in \arg\max_{x\in C(\mathcal{P})} f(x) : (f(x^*) = \max_{x\in C(\mathcal{P})} f(x)).$$

Before moving on to the conditions for existence and uniqueness of solutions, a comment on maximization and minimization problems: **any minimization problem $\mathcal{P}_{min}$ with objective f and constraint set $C(\mathcal{P}_{min})$ can be written as the maximization problem $\mathcal{P}_{max}$ with objective $-f$ and constraint set $C(\mathcal{P}_{max}) = C(\mathcal{P}_{min})$!** The solutions are maintained, now called the arg min or minimizing arguments, just be careful to flip the sign of the objective at the solution(s)! This fact can directly be seen from the definition of the problems, and does not require proof. Thus, our convention here to study maximization problems is without loss of generality. Because of this "equivalence" but the slight difference in notation, as noted in the introduction, re-stating all concepts to follow in terms of a minimization problem may be a good test of your understanding of them.[5]

4.1.2 Characterizing the Set of Solutions

[4] While this is slightly weird notation at first sight, it is arguably intuitive since when you have a unique maximizer, say e.g. $x_0 = 0 = \arg\max_{x\in\mathbb{R}} -x^2$, you would be typically more inclined to call $x_0 = 0$ the solution to maximizing $f(x) = -x^2$ over $C(\mathcal{P})$ ($\mathbb{R}$), rather than the set $\{x_0\} = \{0\}$.

[5] You may also derive value from this if you like econometrics, because really, what is the purpose of maximizing error functions?! ;-)

For this subsection, make sure to recall the concepts of (i) a bounded set, (ii) a closed set, (iii) supremum and infimum, and (iv) the Heine-Borel theorem for characterization of compactness in $\mathbb{R}^n$ that have been introduced throughout the earlier chapters of this script.

First, we need one more definition:

Definition 82. ***(Bounded Function)*** *Let $f : X \mapsto \mathbb{R}$ be a real-valued function. If $\text{im}(X)$, the image of X under f (or range of f) is bounded, we say that f is a bounded function. Moreover, for $S \subseteq X$, we say that f is bounded on S if $f|_S$ is bounded.*

Recall that the discussions of Heine-Borel highlighted the value of compact subsets of $\mathbb{R}$ for optimization: a continuous function will *necessarily* assume both a (global) maximum and minimum on them, such that compactness of the set is a *sufficient* condition for existence of solutions to optimization problems! To transfer this intuition to the $\mathbb{R}^n$, we need to review what Heine-Borel precisely says about compact sets: they are closed and bounded. Boundedness of a set $X \subseteq \mathbb{R}^n$, defined as X being contained in some ball $B_r(x_0)$ where $x_0 \in \mathbb{R}^n$ and $0 \leq r < \infty$ is immensely helpful because it restricts the degree to which the arguments x of a function $f : X \mapsto \mathbb{R}$ can extend into any direction in X – if they move too far away from x_0, then boundedness tells us that they can no longer be arguments of f! Similar to the univariate case where bounded sets are intervals with finite bounds, this prevents solution-breakers like limits $x_k \to \infty$, as e.g. in $f(x_1, x_2) = x_1 + x_2$ where if x is allowed to infinitely expand e.g. in direction $e_2 = (0,1)'$, then it is allowed to diverge to $+\infty$ as $x_2 \to \infty$. As with univariate functions, we can prevent this by defining f on a bounded set. Next, why do we need closedness? For univariate functions, this is also rather straightforward: if e.g. $f : (a,b) \mapsto \mathbb{R}, x \mapsto 2x + 3$, then the set is bounded and f can not diverge due to "too large" arguments, but clearly, for any $x, y \in (a,b) : (x < y \Rightarrow f(x) < f(y))$, and since the set (a,b) does not have extrama (only infimum and supremum), so does its range $f[(a,b)]$![6] The more general issue is that when starting from $f : [a,b] \mapsto \mathbb{R}$ where the maximizer lies on the boundary $\{a,b\}$, then considering any set that does not include the whole boundary may omit the maximizer, and we may be able to move arbitrarily close to the maximizer but never reach it on the new set, preventing us to find a solution. This logic can be extended to the multivariate case in a one-to-one fashion.

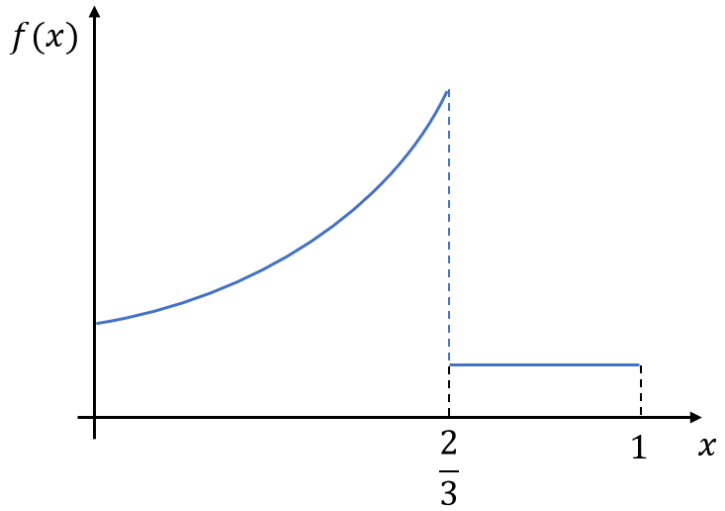


Figure 15: Graph of $f : [0,1] \mapsto \mathbb{R}, x \mapsto \mathbb{1}[x < 2/3](x^2 + 2) + \mathbb{1}[x \geq 2/3]$.

[6] That is, when e.g. looking for the maximum, consider any point $x \in (a,b)$ and let $\varepsilon := b - x > 0$. Then, for $y = b - \varepsilon/2 \in (a,b)$, $y > x$ and thus $f(y) > f(x)$, and x can not be a maximizer of f. Consequently, no point in $x \in (a,b)$ is a maximizer of f, no matter how close it lies to b, i.e. no matter how small $\varepsilon > 0$.

Finally, why do we need continuity? To see this, consider the function

$$f : [0,1] \mapsto \mathbb{R}, x \mapsto f(x) = \begin{cases} x^2 + 2 & x < 2/3 \\ 1 & 1 \geq 2/3 \end{cases}$$

the graph of which is illustrated in Figure 15. What is the maximizer of f? Clearly, it is not an element of $[2/3, 1]$, so we can restrict our search to $[0, 2/3)$. But this set is no longer closed, and we may run into our boundary problem again – and the way f is defined, we indeed do! More generally, if f is discontinuous, the function can "jump", i.e. it can increase or decrease steadily into one direction, but just before reaching the high/low point, attain an entirely different value. In other words, when f approaches $x_0 \in X$, it approaches the level $\lim_{x \to x_0} f(x)$. If this level would be a maximum/minimum, we need to ensure that it lies in the range of f by requiring $f(x_0) = \lim_{x \to x_0} f(x)$, which is precisely the definition of continuity.

It is possible to show formally that the intuitive line of reasoning above goes through formally:

Theorem 51. ***(Weierstrass Extreme Value Theorem)*** *Suppose that $X \subseteq \mathbb{R}^n$ is compact, and that $f : X \mapsto \mathbb{R}$ is continuous, then, f assumes its maximum and minimum on X, such that $\arg\max_{x \in X} f(x) \neq \emptyset$ and $\arg\min_{x \in X} f(x) \neq \emptyset$.*

The proof is omitted as it relies on sequence theory which we have not touched in this class.[7]

4.2 Unconstrained Optimization

As per our tradition of moving from easier to harder problems, we begin the study of optimization problems with those that are not subject to any constraints, but just seek to maximize a function f over its domain, i.e.

$$(\mathcal{P}) \qquad \underset{x \in \text{dom}(f)}{\text{maximize}} \; f(x)$$

where $\text{dom}(f) = X \subseteq \mathbb{R}^n$ and $f : X \mapsto \mathbb{R}$ is a real-valued function. Note that when considering constrained problems, we may equivalently consider the "unconstrained" problem with objective $f|_{C(\mathcal{P})}$, but the key complication is that continuity of f and appealing properties of the domain do not transfer easily to this issue, such that we need to consider it later on its own right. We now discuss the necessary and sufficient conditions for a value $x \in X$ being a maximizer of f.

4.2.1 First and Second Order Necessary Conditions

For univariate functions $f : X \mapsto \mathbb{R}$, $X \subseteq \mathbb{R}$, you may know how to approach the problem of unconstrained optimization: set the first derivative of f to zero, and check that the second derivative is smaller than 0 – any point that satisfies these conditions is a candidate for an *interior* solution. Next to the interior solutions, you will have to consider *border* solutions: points of non-differentiability and points on the boundary of the support X of f. There are, of course, applications where you need not worry about border solutions: if f is differentiable everywhere (especially: $f \in C^2(X)$) or the boundary points are either non-existent ($X = \mathbb{R}$) or excluded from the domain ($X = (a, b)$). Either way, you evaluate f at all candidate points that you found, and the one yielding the highest value is your global maximizer. Below, we will (i) formally justify this procedure, while (ii) extending it to the multivariate context.

First of all, it is noteworthy that we restrict attention to *local* maximizers in our analytical search, and only subsequently identify the global maximum from the rather crude and inelegant technique of

[7] The interested reader can consult https://en.wikipedia.org/wiki/Extreme_value_theorem, which gives several variants of the proof.

comparing the value of the objective at all candidates. To find necessary conditions, **we restrict attention to the interior, the (finite set of) boundary points are considered in isolation subsequently!**. Here, we imagine for the moment that there is a point $x^* \in int(X)$ that maximizes f locally, and search for some (ideally strong) characteristic features x^* must *necessarily* have. Starting from the univariate case, suppose x^* is a local maximizer, such that for the ε-ball restricted to X with $\varepsilon > 0$, there are no values of f above $f(x^*)$, i.e. $\forall x \in B_\varepsilon(x^*) : f(x) \leq f(x^*)$. Then, how does f look like around x^*, provided that it is (twice) differentiable and thus especially continuous? Figure 16 illustrates the possible shapes.

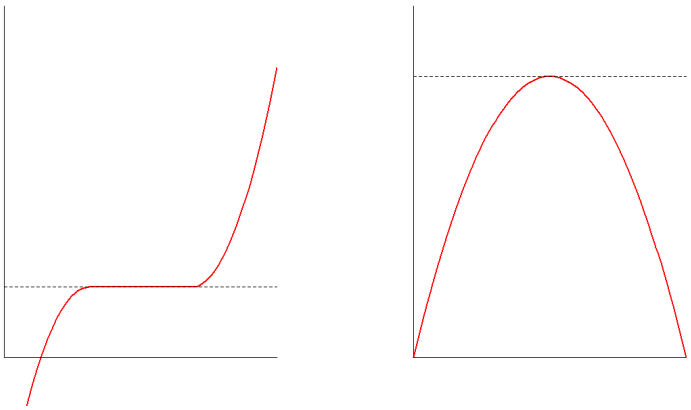


Figure 16: Shapes of f around a local maximizer.

Thus, f must either be flat around x^*, or constitute a "hill" with peak x^* (a mixture also exists, with flatness on one side and a "downhill" part on the other). What does this mean? As illustrated, intuitively, we should have a zero slope of f at x^*, or for a multivariate function f, a zero gradient, i.e. a zero slope into any direction (generalizing the concepts of flatness and hilltop to the $\mathbb{R}^n$). But how to come about this formally? Let's directly consider the multivariate case. It is a rather technical proof, the intuition is simple; it is discussed once the result is established.

Theorem 52. ***(Unconstrained Interior Maximum - First Order Necessary Condition)*** *Let $X \subseteq \mathbb{R}^n$ and $f \in C^1(X)$. Suppose that $x^* \in \text{int}(X)$ is a local maximizer of f. Then, $\nabla f(x^*) = \mathbf{0}$.*

Proof. By contradiction: suppose that $x^* \in int(X)$ is a local maximizer of f, but that $\nabla f(x^*) \neq \mathbf{0}$. The goal is to show that a non-zero entry of $\nabla f(x^*)$ breaks x^*'s property as a maximizer. Thus, let $j \in \{1, \dots, n\}$ such that $f_j(x^*) = \frac{\partial f}{\partial x_j}(x^*) \neq 0$. Suppose first that $f_j(x^*) > 0$. Let $\varepsilon > 0$ small such that $\forall x \in B_\varepsilon(x^*)$, $f_j(x) > 0$, or equivalently, $\forall h \in B_\varepsilon(\mathbf{0})$, $f_j(x^* + h) > 0$.[8] By Taylor's theorem, there exists $\lambda \in (0,1)$ such that for $h \in B_\varepsilon(\mathbf{0})$,

$$f(x^* + h) = f(x^*) + \nabla f(x^* + \lambda h)h.$$

Because $\|\lambda h\| = |\lambda| \cdot \|h\| < \|h\|$, $\lambda h \in B_\varepsilon(\mathbf{0})$, and thus $f_j(x^* + \lambda h) \neq 0$. Now, consider $\delta > 0$ small so that $\delta e_j \in B_\varepsilon(\mathbf{0})$. Then, the above yields

$$f(x^* + \delta e_j) = f(x^*) + \nabla f(x^* + \lambda \delta e_j)\delta e_j = f(x^*) + \delta \underbrace{f_j(x^* + \lambda \delta e_j)}_{>0} > f(x^*),$$

i.e. there exists $x = x^* + \delta e_j \in B_\varepsilon(x^*)$ such that $f(x) > f(x^*)$.

If $f_j(x^*) < 0$, we proceed in an analogous fashion but consider the vector $x = x^* - \delta e_j$, for which $f(x) > f(x^*)$. Because $\varepsilon > 0$ was chosen randomly, it follows that x^* is not a local maximizer of f, a contradiction. Thus, it must hold that $\nabla f(x^*) = \mathbf{0}$. □

[8] The existence of such small $\varepsilon > 0$ is ensured by continuity of f_j. We restrict attention to any such small balls around x^*, because any larger ball contains them, and if x^* is not a local maximizer with respect to the small ball, then especially not with respect to larger balls.

The intuition of this admittedly abstract proof is easy to see from the univariate case: recall that a differentiable function f was strictly increasing (decreasing) in x_0 if and only if $f'(x_0) > 0$ ($f'(x_0) < 0$), and because f is twice differentiable, f' is differentiable and thus especially continuous. Thus, if $f'(x^*) > 0$, ($f'(x^*) < 0$), by continuity, $f'(x) > 0$ ($f'(x) < 0$) for points "close enough" to x^*, and thus, there lie points slightly to the right (left) of x^* with strictly larger values, and x^* can not be a maximizer! This suggests that $\neg(f'(x^*) = 0) \Rightarrow \neg(x^* \text{ is local maximizer})$, and by our negation-flipping-rule, we get $(f'(x^*) = 0) \Leftarrow (x^* \text{ is local maximizer})$. Indeed, this idea is exactly what we generalize for the multivariate case where we have to ensure that such behavior is precluded in any (fundamental) direction as collected by the gradient, the rest is just notation.

This suggests that points with a zero gradient are important. Indeed, they deserve an own label:

Definition 83. ***(Critical Point or Stationary Point)*** *Let $X \subseteq \mathbb{R}^n$, $f : X \mapsto \mathbb{R}$ and $x^* \in X$. Then, if f is differentiable at x^* and $\nabla f(x^*) = \mathbf{0}$, we call x^* a critical point of f or a stationary point of f.*

It is crucial to note that just because we have defined generally, it need not exist in every specific scenario! There are a broad variety of functions that have a non-zero gradient everywhere, e.g. $f(x_1, x_2) = x_1 + x_2$, just like there are univariate functions with a globally non-zero derivative. Also, it is easily seen that this condition is **not sufficient** for a local maximizer, because $\nabla f(x^*) = \mathbf{0}$ applies also to local minima that are not local maxima and so-called "saddle points", as we discuss more thoroughly later (see also Figure 17).

As has just emerged, the distinction between local maxima and minima lies beyond the first derivative. **Thus, it is generally not sufficient to set the first derivative to zero when looking for a maximum!** But as you know, we can consult the second derivative to tell maximizers and minimizers apart. Similar to before, we consult the Taylor expansion around a candidate x^* for a local maximizer. Consider $\varepsilon > 0$ small (justified as in the proof above), and suppose that f is twice continuously differentiable on the open ball around x^* with radius ε, i.e. $f|_{B_\varepsilon(x^*)} \in C^2(B_\varepsilon(x^*))$. Then, the Taylor theorem tells us that for all $h \in B_\varepsilon(\text{ꟾ})$, we can find a $\lambda \in (0,1)$ for which

$$f(x^* + h) = f(x^*) + \nabla f(x^*)h + 1/2h'H_f(x^* + \lambda h)h.$$

If x^* satisfies the *necessary condition* of being a critical value, $\nabla f(x^*) = 0$, and we can re-arrange

$$f(x^* + h) - f(x^*) = 1/2h'H_f(x^* + \lambda h)h.$$

For x^* to be a local maximizer, we must thus have

$$\forall h \in B_\varepsilon(x^*): \qquad 1/2h'H_f(x^* + \lambda h)h \leq 0.$$

This already looks very much like a definiteness condition, the only issues are that $H_f(x^* + \lambda h)$ depends variably on h (and λ), and that h is chosen from the ε-open ball, rather than arbitrarily from $\mathbb{R}^n$. The proof below formally demonstrates how to resolve this issue.

Theorem 53. ***(Unconstrained Interior Maximum – Second Order Necessary Condition)*** *Let $X \subseteq \mathbb{R}^n$ and $f \in C^1(X)$. Suppose that $x^* \in \text{int}(X)$ is a local maximizer of f. Then, if f is twice continuously differentiable at x^*, $H_f(x^*)$ is negative semi-definite.*

Proof. Again by contradiction. Suppose that $x^* \in \text{int}(X)$ is a local maximizer of f, such that f is twice continuously differentiable at x^*. For contradiction, suppose that $H_f(x^*)$ is not negative semi-definite, i.e. that there exists $v \in \mathbb{R}^n$: $v'H_f(x^*)v > 0$.

Note that the Hessian is continuous at x^* because all second order partial derivatives of f are continuous at x^* by assumption. Note further that similar to our lines of reasoning in the previous chapter, this gives continuity of the univariate *directional* function $\delta \mapsto H_f(x^* + \delta \cdot (\lambda v))$ at $\delta = 0$.

What follows relies on a result frequently exploited in economics: continuity is preserved under functional composition! This means that if g, h are continuous functions, then so is $g \circ h$. In the given context, because $\delta \mapsto H_f(x^* + \delta \cdot (\lambda v))$ is continuous at $\delta = 0$ and left- and right-multiplication of the fixed vector v can be viewed as generalization of multiplication with a constant, i.e. a continuous operation, the function $\delta \mapsto v' H_f(x^* + \delta \lambda v) v$ is continuous at $\delta = 0$.

Verbally, note that because this function takes a value above zero at $\delta = 0$, by continuity, it will lie strictly above zero everywhere in small enough environments around $\delta = 0$. Formally, let $r > 0$ small such that $r \leq \varepsilon / \|v\|$[9] and $\forall \delta \in (0, r)$, with $\tilde{v} = \delta v$:

$$0 < v' H_f(x^* + \lambda \tilde{v}) v \quad \Leftrightarrow \quad 0 < \delta^2 v' H_f(x^* + \lambda \tilde{v}) v = \tilde{v}' H_f(x^* + \lambda \tilde{v}) \tilde{v}.$$

Because $\delta > 0$, by absolute homogeneity of the norm, $\|\tilde{v}\| = \delta \|v\| < r \|v\| = \frac{\varepsilon}{\|v\|} \|v\| = \varepsilon$, and we obtain $\tilde{v} \in B_\varepsilon(\mathbf{0})$. Thus, there exists $x = x^* + \tilde{v} \in B_\varepsilon(x^*)$ such that

$$f(x) - f(x^*) = f(x^* + \tilde{v}) - f(x^*) = \tilde{v} H_f(x^* + \lambda \tilde{v}) \tilde{v} > 0,$$

and x^* is not the maximizer of $f|_{B_\varepsilon(x^*)}$. Because $\varepsilon > 0$ was chosen arbitrarily small, x^* is not a local maximizer of f, a contradiction. □

The equivalent necessary condition for a minimum is positive definiteness – try to prove it using the proof above as help!

One can nicely illustrate the grasp of this necessary condition by illustrating critical values of bivariate functions it rules out: consider $f(x_1, x_2) = x_1^2 - x_2^2$. The necessary condition of a critical point tells us that any candidate $x^* = (x_1^*, x_2^*)$ must satisfy

$$\nabla f(x_1^*, x_2^*) = (2x_1^*, -2x_2^*) = \mathbf{0} \quad \Leftrightarrow \quad (x_1^*, x_2^*) = \mathbf{0}.$$

Thus, there is a unique candidate... But is it also a maximizer (or minimizer)? Let's consult the Hessian – it is straightforward to verify that for all $x = (x_1, x_2) \in \mathbb{R}^2$,

$$H_f(x) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}.$$

What about its definiteness? Pick $z = (z_1, z_2)' \in \mathbb{R}^2$. Then,

$$z' H_f(x) z = (z_1, z_2) \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = (z_1, z_2) \begin{pmatrix} 2z_1 \\ -2z_2 \end{pmatrix} = 2(z_1^2 - z_2^2).$$

Thus, there exist $z = (1, 0)$ and $\tilde{z} = (0, 1)$ such that

$$z' H_f(x) z > 0 > \tilde{z}' H_f(x) \tilde{z}$$

and $H_f(x)$ is neither positive nor negative semi-definite, i.e. it is *indefinite*. Thus, the solution $x^* = \mathbf{0}$ can neither constitute a local maximum nor a local minimum: we call this a *saddle point*. To see why, consult Figure 17, and think how this would fit on a horse (or your preferred animal to ride).

The crucial thing to take away from saddle points is that the critical value constitutes a maximum in one direction (here: $e_2 = (0, 1)'$, i.e. moving along the x_2-axis) but a minimum in the other ($e_1 = (1, 0)'$). Note that with some abuse of concept, we called matrix definiteness the "sign" of the matrix. This allows to neatly illustrate how this necessary condition fails to be sufficient using the univariate case, where

[9] You will see in a second why this is useful – we ensure that $\tilde{v} = \delta v$ breaks the local maximizer property of x^* on the ε-open ball we consider, for which $\tilde{v}$ must of course be contained in this ball.

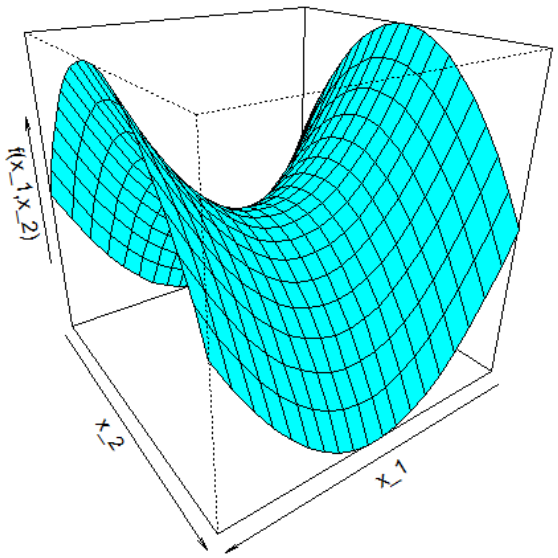


Figure 17: Plot of $f(x_1, x_2) = x_1^2 - x_2^2$: A saddle point.

the definiteness of $f''(x)$ indeed corresponds to the sign also in the narrow sense. Consider $f(x) = x^3$ (for the graph see Figure 18), which has a saddle point at $x = 0$. Since $f'(x) = 3x^2$ and $f''(x) = 6x$, we get the unique critical value at $x = 0$ with second derivative $f''(x) = 0$, such that the negative semi-definiteness condition for the local maximum holds. However, so does the positive semi-definiteness condition for the local minimum. The graph will ensure you that the point is neither: $f'(x) > 0$ for any $x \neq 0$, the function strictly increases (decreases) in value for infinitely small deviations to the right (left) of $x = 0$. Thus, both the first and second order necessary conditions hold, but we don't have a maximum (or a minimum). The interested reader can check the first pages of SB's Chapter 16, which introduces quadratic forms, an insightful topic in these matters.

4.2.2 Sufficient Conditions

Let's see how *strengthen* the necessary conditions to sufficient ones, because after all, that's what we're after: deriving rules that we can check to arrive at our solutions.

Here, intuitively, we want to rule out cases like $f(x) = x^3$ as seen before. Why did our conditions fail us there? well, the Hessian was both positive and negative semi-definite, i.e. failure occured due to definiteness not being "strict"! Thus, can we arrive at sufficient conditions by simply requiring definiteness in the second order condition? Luckily, this is indeed all we need to do.

Theorem 54. ***(Unconstrained Interior Local Maximum – Sufficient Conditions)*** *Let $X \subseteq \mathbb{R}^n$, $f \in C^2(X)$ and $x^* \in \text{int}(X)$. Suppose that x^* is a critical point of f, and that $H_f(x^*)$ is negative definite. Then, x^* is a strict local maximizer of f.*

Proof.[10]. For $h \in \mathbb{R}^n$, the Taylor expansion around x^* of first order gives for a $\lambda \in (0, 1)$

$$f(x^* + h) = f(x^*) + \nabla f(x^*)h + 1/2h' H_f(x^* + \lambda h)h \quad \Leftrightarrow \quad f(x^* + h) - f(x^*) = 1/2h' H_f(x^* + \lambda h)h$$

where equivalence follows because x^* is a critical point and thus, $\nabla f(x^*) = 0$. By continuity of H_f (implied by $f \in C^2(X)$), one can show that because $H_f(x_0)$ is positive definite, H_f is negative definite on

[10] This proof is conceptually very similar to our considerations before. One might consider it easier, because we can go the "direct" way and need not rely on a proof by contradiction/contrapositive.

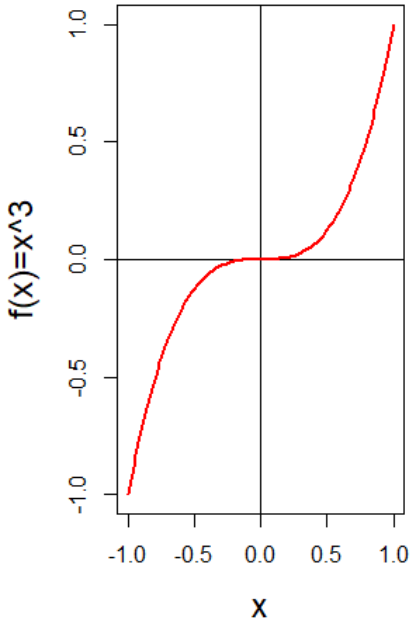


Figure 18: Plot of $f(x) = x^3$: A univariate saddle point.

a neighborhood $B_\varepsilon(x^*)$ of x^* with $\varepsilon > 0$. The intuition is the one of definiteness and the sign: if a function g is continuous and $g(x) > 0$, then $g > 0$ in a neighborhood of x. Formally, this is a bit tricky. The details are given in the next paragraph, but feel free to skip it if you have understood the intuition well.

Recall that a matrix A is negative definite if and only if all eigenvalues λ of A are < 0. The eigenvalues set the characteristic polynomial to zero: $\mathcal{P}(\lambda) = \det(H_f(x^*) - \lambda I) = 0$. Now, note that the mapping $(x^*, \lambda) \mapsto \det(H_f(x^*) - \lambda I)$ is a composition of continuous functions, because H_f is continuous by assumption and the determinant is a large product which is also continuous. Now, if $\lambda < 0$ and $\|h\|$ is sufficiently small, the new solution $\tilde{\lambda}$ to $\det(H_f(x^* + h) - \tilde{\lambda} I) = 0$ will still satisfy $\tilde{\lambda} < 0$.[11] So, for any eigenvalue λ of $H_f(x^*)$ there exists $\varepsilon_\lambda > 0$ such that for all $h \in B_{\varepsilon_\lambda}(\mathbf{0})$ the associated eigenvalue $\tilde{\lambda}(h)$ of $H_f(x + h)$ satisfies $\tilde{\lambda}(h) < 0$. Because the number of eigenvalues is bounded by the dimension of X, n, it is finite and we can set $\varepsilon = \min_\lambda \varepsilon_\lambda$ so that for all $h \in B_\varepsilon(\mathbf{0})$, all eigenvalues of $H_f(x + h)$ are negative, i.e. $H_f(x + h)$ is negative definite.

Thus, we can find $\varepsilon > 0$ such that for all $h \in B_\varepsilon(\mathbf{0})$, $H_f(x + h)$ is negative definite. Note especially that this implies that for any $h \in B_\varepsilon(\mathbf{0})$ and any $\lambda \in (0, 1)$, $H_f(x + \lambda h)$ is negative definite, because $\|\lambda h\| = |\lambda| \|h\| < \lambda\varepsilon < \varepsilon$ and $\lambda h \in B_\varepsilon(\mathbf{0})$. By negative definiteness, for $\mathbf{0} \neq h \in B_\varepsilon(\mathbf{0})$,

$$h' H_f(x + \lambda h) h < 0.$$

From the Taylor expansion, it results that for any $h \in B_\varepsilon(\mathbf{0})$ so that $h \neq \mathbf{0}$,

$$f(x^* + h) - f(x^*) = 1/2 h' H_f(x^* + \lambda h) h < 0$$

so that for any $x \in B_\varepsilon(\mathbf{x}^*)$, if $x \neq x^*$ then $f(x^*) > f(x)$.

Thus, x^* is indeed a *strict* local maximizer of f. □

Now, we have established the sufficient condition for a local interior maximum. Let us take a mo-

[11] This can be justified more formally using the implicit function theorem discussed later.

ment to think about what to make of the results thus far. When concerned with *interior* maxima, we can restrict the candidate set to critical values, because all potential maxima are *necessarily* critical values! Then, if f is "sufficiently smooth"[12] we can check the Hessian and determine its definiteness. If at a critical value x^*, H_f is negative (positive) definite, we have *sufficient* evidence that x^* is a local maximum (minimum)! That may not be necessary – but we can rule out any value x^* where H_f is not at least negative (positive) semi-definite, because they violate the necessary conditions. Thus, we are generally happy if all critical values are associated with either a positive/negative definite Hessian or an indefinite one – but we have to beware semi-definite Hessians, for which no theorem tells us whether or not such points are extrema! As a further note of caution, all theorems thus far have concerned *interior* points – for boundary points, no theorem exists, and we must either rule them out somehow from the specific nature of the function, or simply consider f with an open set $\text{dom}(f)$ such that the boundary issue is avoided altogether.[13]

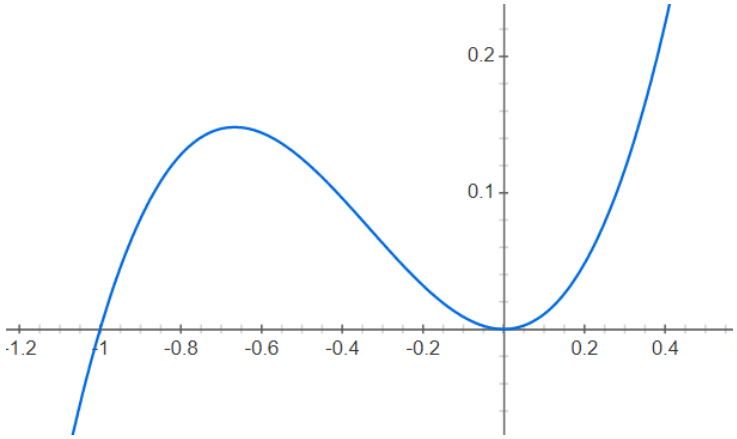


Figure 19: Plot of $f(x) = x^2 + x^3$ (image from Google search).

Finally, let's see how we can go from local to global: As you may know, just because there are local extrema does not mean that global extrema exist, consider e.g. the function $f(x) = x^2 + x^3$ defined on $\mathbb{R}$ as plotted in Figure 19. However, when using the concept of *convexity*, one can establish a sufficient condition:

Theorem 55. ***(Sufficiency for the Global Unconstrained Maximum)*** *Let $X \subseteq \mathbb{R}^n$ be a **convex** set, and $f \in C^2(X)$. Then, if f is concave and for $x^* \in \text{int}(X)$, it holds that $\nabla f(x^*) = \mathbf{0}$, then x^* is a global maximizer of f.*

Proof. Let $x^* \in \text{int}(X)$ and suppose that $\nabla f(x^*) = \mathbf{0}$. Recall Proposition 22 which told us that f is concave if and only if $H_f(x)$ is negative semi-definite $\forall x \in int(X)$. This can be combined with the Taylor expansion of second order: For any $x \in X$, there exists $\lambda \in (0,1)$ such that

$$f(x) = f(x^*) + \nabla f(x^*)(x - x^*) + 1/2(x - x^*)' H_f(x^* + \lambda(x - x^*))(x - x^*).$$

Because $\lambda \in (0,1)$, we have $x^* + \lambda(x - x^*) \in \text{int}(X)$, and by negative semi-definiteness of H_f at any interior point, $(x - x^*)' H_f(x^* + \lambda(x - x^*))(x - x^*) \leq 0$. Further, because $\nabla f(x^*) = \mathbf{0}$ by assumption, we obtain

$$\forall x \in X : f(x) - f(x^*) = 1/2(x - x^*)' H_f(x^* + \lambda(x - x^*))(x - x^*) \leq 0 \quad \Leftrightarrow \quad \forall x \in X : f(x) \leq f(x^*). \qquad \square$$

[12] I.e., it satisfies all required differentiability conditions – here: being twice continuously differentiable at all critical values.

[13] However, if we wish to use the Weierstrass extreme value theorem to ensure existence in a first step, we have to consider closed sets (which include the boundary)!

This is extremely powerful: when we have a concave objective function (and many classical economic functions are concave, e.g. most standard utility or production functions), then the critical point criterion (before only *necessary* for a *local* maximizer) is now **sufficient for a global maximizer**!

4.2.3 Limit Behavior

The previous subsection has given one example of how we can conclude from local on global extrema. However, we may not always be as lucky and have a convex/concave objective. Hence, it is instructive to think about when it is the case that the largest/smallest local extrema are also constitute global extrema. For this, it may be useful to recall the discussion of the Weierstrass theorem arguing why we need compactness and continuity to ensure existence of global extrema.

First, why may this break down? On one hand, the function may have points of discontinuity which may cause the function to jump above local maxima or below local minima. Because we usually deal with continuously differentiable functions, we disregard this point for now. Secondly, the Weierstrass theorem needed a compact domain. If the domain indefinitely expands along some direction(s), we can run into a limit problem similar to the one of Figure 19. Fortunately, this can be addressed in a separate investigation.

Let us again start with the easy case of a univariate function f. Here, the asymptotic behavior is easily studied, and we can establish a neat result that simplifies the analysis.

Proposition 23. ***(Limit Behavior: Univariate Functions)*** *Consider a continuously differentiable function on $\mathbb{R}$, $f \in C^1(\mathbb{R})$. Let x_c^+ and x_c^- be critical points of f with $f(x_c^+) \geq f(x_c)$ and $f(x_c^-) \leq f(x_c)$ for all critical points x_c of f (i.e., the candidates for global extremizers). Then,*

1. *If $f(x_c^+) \geq \max\{\lim_{x\to\infty} f(x), \lim_{x\to-\infty} f(x)\}$, x_c^+ is a global maximizer of f.*
2. *If $f(x_c^-) \leq \min\{\lim_{x\to\infty} f(x), \lim_{x\to-\infty} f(x)\}$, x_c^- is a global minimizer of f.*

Proof. We only consider the global maximizer as the proof for the minimizer is entirely analogous. The proof is by contradiction. Suppose that $f(x_c^+) \geq \max\{\lim_{x\to\infty} f(x), \lim_{x\to-\infty} f(x)\}$ and that x_c^+ is not a global maximizer of f. Then, there exists $\bar{x} \in \mathbb{R}$ such that $f(\bar{x}) > f(x_c^+)$. Because $f(x_c^+) \geq f(x_c)$ for all critical points x_c of f, $\bar{x}$ is not a critical point of f. Furthermore, because

$$f(\bar{x}) > f(x_c^+) \geq \max\{\lim_{x\to\infty} f(x), \lim_{x\to-\infty} f(x)\}$$

there exists $c_1 > 0$ such that $f(\bar{x}) > f(x)$ for all $x \geq c_1$ and $c_2 > 0$ such that $f(\bar{x}) > f(x)$ for all $x \leq -c_2$. Because $[-c_2, c_1]$ is compact, there exists a global maximizer x_g^* of $f|_{[-c_2,c_1]}$ (by the Weierstrass theorem). Because $f(x) < f(\bar{x})$ for all $x \in \mathbb{R}\backslash[-c_2, c_1]$, any global maximizer of $f|_{[-c_2,c_1]}$ is also a global maximizer of f. Because $f(-c_2) < f(\bar{x})$ and $f(c_1) < f(\bar{x})$, the global maximizer x_g^* lies in the interior of $[-c_2, c_1]$, i.e. in $(-c_2, c_1)$. By the first order necessary condition, x_g^* is a critical point of f. Hence,

$$f(x_g^*) \overset{x_g^* \text{ glob. max.}}{\geq} f(\bar{x}) > f(x_c^+) \overset{x_g^* \text{ crit. point}}{\geq} f(x_g^*),$$

a contradiction. Thus, when $f(x_c^+) \geq \max\{\lim_{x\to\infty} f(x), \lim_{x\to-\infty} f(x)\}$, it can not be that x_c^+ is not a global maximizer of f, which establishes the proposition. □

Verbally, this theorem says that limits only "break" the global maximizer property for the largest local maximizer x_c^+ if there is a limit that **strictly** exceeds the value f attains at x_c^+.

For multivariate limits "to infinity", there are two potential concepts. Recall from the last chapter that we generalized the standard limit $\lim_{h\to0} f(x)$ as $\lim_{\|h\|\to0} f(x)$ where we wrote that $c = \lim_{\|h\|\to0} f(x)$ if

$$\forall\varepsilon > 0 \exists\delta > 0 : (\forall h \in B_\delta(\mathbf{0}) : |f(h) - c| < \varepsilon)$$

Accordingly, this suggests the generalization $c = \lim_{\|x\|\to\infty} f(x)$ when

$$\forall \varepsilon > 0 \exists \delta > 0 : (\forall x \in \mathbb{R}^n \text{ such that } \|x\| \geq \delta : |f(x) - c| < \varepsilon)$$

or for "divergent" functions, $\lim_{\|x\|\to\infty} f(x) = \infty$ $(= -\infty)$ if

$$\forall C > 0 \exists \delta > 0 : (\forall x \in \mathbb{R}^n \text{ such that } \|x\| \geq \delta : f(x) > C (f(x) < -C)).$$

In practical applications, you can either "see" what happens as $\|x\|$ becomes large, e.g. for $f(x_1, x_2) = -x_1^2 - x_2^2$ where clearly, $f(x) \to -\infty$ as $\|x\| \to \infty$. In such cases, you don't need a formal investigation to determine the limit, and fortunately, many practical examples we consider in economics are similarly obvious.

For the case of maximization (minimization) problems, if $\lim_{\|x\|\to\infty} f(x) = -\infty$ $(\lim_{\|x\|\to\infty} f(x) = \infty)$, we do not need to worry about limit behavior:

Proposition 24. ***(Limit Behavior: Vanishing Multivariate Functions)*** *Consider a continuously differentiable function on $\mathbb{R}^n$, $f \in C^1(\mathbb{R}^n, \mathbb{R})$. Let x_c^+ and x_c^- be critical points of f with $f(x_c^+) \geq f(x_c)$ and $f(x_c^-) \leq f(x_c)$ for all critical points x_c of f (i.e., the candidates for global extremizers). If $\lim_{\|x\|\to\infty} f(x) = -\infty$, then x_c^+ is a global maximizer of f, and if $\lim_{\|x\|\to\infty} f(x) = \infty$, then x_c^- is a global minimizer of f.*

Proof. We only consider the global maximizer as the proof for the minimizer is entirely analogous. Let $\delta > 0$ such that $\forall x \in \mathbb{R}^n$ such that $\|x\| \geq \delta : f(x) < f(x_c^+)$ (δ exists by definition of $\lim_{\|x\|\to\infty} f(x) = -\infty$; if $f(x_c^+) < 0$ set $C = -f(x_c^+)$, otherwise set $C = 0$). Further, let $\delta^* = \max\{\delta, \|x_c^+\| + 1\}$ such that x_c^+ is an interior point of $\bar{B}_{\delta^*}(\mathbf{0})$. Because this ball is closed, it is compact (the norm is bounded by δ^*), and by the Weierstrass theorem, f attains a global maximum and minimum on $\bar{B}_{\delta^*}(\mathbf{0})$.

For $x \in \mathbb{R} \backslash \bar{B}_{\delta^*}(\mathbf{0})$, because $\|x\| > \delta^* \geq \delta$, we have $f(x) < f(x_c^+)$, and because $x_c^+ \in \bar{B}_{\delta^*}(\mathbf{0})$, the global maximizer of $f|_{\bar{B}_{\delta^*}(\mathbf{0})}$ is also the global maximizer of f.

Next, on the boundary of $\bar{B}_{\delta^*}(\mathbf{0})$, $\|x\| = \delta^* \geq \delta$ so that $f(x) < f(x_c^+)$ for x on the boundary of $\bar{B}_{\delta^*}(\mathbf{0})$. Hence, the global maximum of f is an interior point of $\bar{B}_{\delta^*}(\mathbf{0})$, i.e. a local maximum of f on $\bar{B}_{\delta^*}(\mathbf{0})$.

Because x_c^+ is the local maximum yielding the largest value of f on $\mathbb{R}$, it is especially the local maximum yielding the largest value of f on $\bar{B}_{\delta^*}(\mathbf{0})$. Hence, x_c^+ is a global maximizer of f. □

To take away, if a function vanishes asymptotically according to the generalized limit into the "opposite" direction of optimization (e.g. to $-\infty$ for maximization), limit behavior can be neglected.

If the quantity can either not be guessed or $\lim_{\|h\|\to\infty} f(x)$ does not exist, we have to pursue a more thorough approach. Non-existence may, in fact, be a very common issue as the generalized limit concept is somewhat crude: there is just one limit $\lim_{\|h\|\to\infty} f(x)$ for *every direction* of the $\mathbb{R}^n$ – even for univariate functions, we had already considered two potential axis of infinite expansion: $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$, or equivalently, $\lim_{\lambda\to\infty} f(\lambda \cdot v)$ for $v \in \{-1, 1\}$.

For maximization problems, $\lim_{\|x\|\to\infty} f(x) = -\infty$ can be established by showing that $\max_{v\in D(n)} f(\lambda v)$ diverges to $-\infty$ as $\lambda \to \pm\infty$, where $D(n) = \{v \in \mathbb{R}^n : \|v\| = 1\}$ is the set of "directions" of the $\mathbb{R}^n$, normalized to unit length.[14] If this argument applies, then Proposition 24 applies and limit behavior can be disregarded.

Note that $\max_{v\in D(n)} f(\lambda v)$ is guaranteed to have a solution because $D(n)$ is compact: it is clearly bounded as for all $v \in D(n)$, $\|v\| \leq C$ for any $C \geq 1$, and furthermore closed by continuity of the norm: if $\{v_n\}_{n\in\mathbb{N}}$ is a sequence over $D(n)$ with limit $v \in \mathbb{R}^n$, then

$$\|v\| = \|\lim_{n\to\infty} v_n\| = \lim_{n\to\infty} \|v_n\| = \lim_{n\to\infty} 1 = 1$$

[14] For minimization, you would show that $\min_{v\in D(n)} f(\lambda v)$ diverges to ∞ as $\lambda \to \pm\infty$.

so that $v \in D(n)$. Thus, $D(n)$ is compact and any continuous function g will attain a global maximum and minimum $D(n)$.[15]

If we can not establish $\lim_{\|h\|\to\infty} f(x) = -\infty$, we need a different approach. Here, the problem $\max_{v\in D(n)} f(\lambda v)$ is helpful again: by solving for $\arg\max_{v\in D(n)} f(\lambda v)$, one obtains an indirect value function $V(\lambda)$ that indicates the maximum value obtained in $\max_{v\in D(n)} f(\lambda v)$ for fixed λ, i.e. $V(\lambda) = f(\lambda v^*)$ for $v^* \in \arg\max_{v\in D(n)} f(\lambda v)$. If $V(\lambda)$ is continuously differentiable, we have successfully reduced the multivariate problem to an univariate one: here, judging the limit issue is easy again, and finding $\lambda^* \in \arg\max V(\lambda)$ yields the global maximizer $x^* = \lambda^* v^*(\lambda^*)$ where $v^*(\lambda)$ is the solution of $\max_{v\in D(n)} f(\lambda v)$ that was used to compute $V(\lambda)$.

As a take-away, in terms of limit behavior, your approach to unconstrained multivariate optimization should be to check whether you can neglect the limit because the function vanishes into the "opposite" direction of optimization, either informally or considering the problem $\max_{v\in D(n)} f(\lambda v)$. If this is not the case, you need to reduce the problem to a univariate one in a first step by finding the optimal directionality conditional on the magnitude of the solution.

4.3 Optimization with Equality Constraints

In the previous section, we have given a very formal and granular treatment to unconstrained optimization, a technique which you were likely (roughly) familiar with, at least for the univariate case. We have thoroughly justified the method, and considered it for general, multivariate problems. Now, we move to the next stage: problems in which constraints can be expressed by *equality of some functions to zero*, i.e.

$$(\mathcal{P}) \qquad \underset{x\in C(\mathcal{P})}{\text{maximize}}\ f(x) \quad \text{where} \quad C(\mathcal{P}) = \{x \in \text{dom}(f) : g_i(x) = 0\ \forall i \in \{1,\dots,m\}\}.$$

Alternatively, you may read the problem in forms like

$$\max f(x) \quad \text{subject to} \quad g_i(x) = 0\ \forall i \in \{1,\dots,n\}.$$

Perhaps, you are familiar with the Lagrangian method – but unless you have had some lectures explicitly devoted to mathematical analysis in your undergraduate studies, you may not be familiar with its formal justification, and perhaps also not how we may apply it for multivariate functions. Thus, the following addresses precisely these issues.

Before moving on, note that the loss in generality from omitting inequality constraints from the problem formulation, depending on the application in mind, may not be too severe: consider for instance utility maximization with the budget constraint $p_1x_1 + p_2x_2 \leq y$, meaning that the consumer can not spend more for consumption than his income. Now, if there is no time dimension (so that the consumer does not want to save) and he has a utility function $u(x_1, x_2)$ which strictly increases in both arguments, as we typically assume ("strictly more is strictly better"), the consumer will never find it optimal to not spend all his income, so that the problem is entirely equivalent to one with the constraint $p_1x_1 + p_2x_2 = y$, i.e. we can expect the problem with the inequality constraint and the one with the equality constraint to yield the exact same solution and associated value of u, such that we can equivalently solve the equality-constrained problem![16]

4.3.1 Implicit Functions and the Lagrangian Method for One Constraint

[15] $g(v) = f(\lambda v)$ is continuous whenever f is continuous.

[16] Once we have completed the formal discussion of the optimization methods, we will explicitly take the time to look at how problems can be re-written in a simplifying fashion more generally. There, you can also find a more thorough justification of why we may sometimes replace inequality constraints with equality constraints.

Do you recall our definition of the lower- and upper-level sets when we discussed the generalization of convexity to the multivariate case? Here, we need a closely related concept: the *level set*, which is both a lower and the upper level set given a certain value!

Definition 84. ***(Level Set)*** *Let $X \subseteq \mathbb{R}^n$, $g : X \mapsto \mathbb{R}$, and $c \in \mathbb{R}$. Then, we call $L_c(g) = \{x \in X : g(x) = c\}$ the c-level set of g.*

This definition may look indeed familiar, in the sense that you should be able to describe $L_c(g)$ with vocabulary that has been introduced very early on: think about what we typically would call a set $\{x \in X : g(x) = c\}$ when c is a value in the codomain of g. As a hint, we denote it by $g^{-1}[\{c\}]$. ...if you have practiced your notation, you will know from this that the **level set is nothing but the pre-image of $\{c\}$ under g!** To see why we care about level sets is analytically straightforward: in our optimization problem, when there is only one constraint g, we are looking for solutions on the level set $L_0(g)$, i.e. $C(\mathcal{P}) = L_0(g)$, in words, the constraint set is nothing but the zero-level set of g! With multiple constraints, note that

$$\begin{aligned} C(\mathcal{P}) &= \{x \in \text{dom}(f) : g_i(x) = 0 \; \forall i \in \{1, \dots, m\}\} \\ &= \{x \in \text{dom}(f) : (g_1(x) = 0 \wedge g_2(x) = 0 \wedge \dots \wedge g_m(x) = 0)\} \\ &= \{x \in \text{dom}(f) : g_1(x) = 0\} \cap \{x \in \text{dom}(f) : g_2(x) = 0\} \cap \dots \cap \{x \in \text{dom}(f) : g_m(x) = 0\} \\ &= \bigcap_{i=1}^{m} L_0(g_i), \end{aligned}$$

and that we are looking for solutions in the *intersections* of the level sets.

From a geometrical perspective, level sets are oftentimes useful to map three-dimensional objects onto a two-dimensional space. The most famous practical example is perhaps the one of contour lines as used for geographical maps, as shown in Figure 20

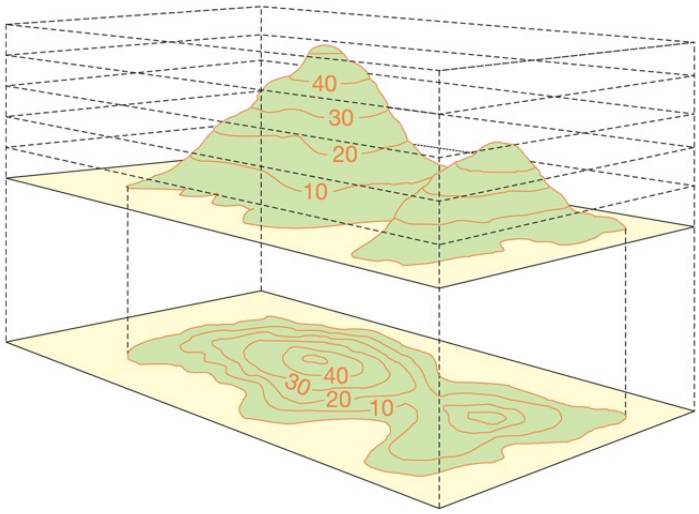


Figure 20: Level Sets in Geography (from http://canebrake13.com/fieldcraft/map_compass.php).

Also in economics, we have various important examples: for instance budget sets (with the restriction that all money is spent) where the level is equal to the disposable income y, and utility functions with level sets $L_{\bar{u}}(u) = \{x \in \mathbb{R}^2 : u(x) = \bar{u}\}$ for the $\mathbb{R}^2$. For utility, we usually call sets $L_{\bar{u}}(u)$ indifference sets, and their graphical representation in the $\mathbb{R}^2$ the *indifference curve*, which represents all points $(x_1, x_2) \in X$ that yield the same utility level $\bar{u}$. For the Cobb-Douglas utility with equal coefficients, i.e. $f(x) = \sqrt{x_1 x_2}$, this relationship is shown in Figure 21 (note the application-oriented labeling of axes).

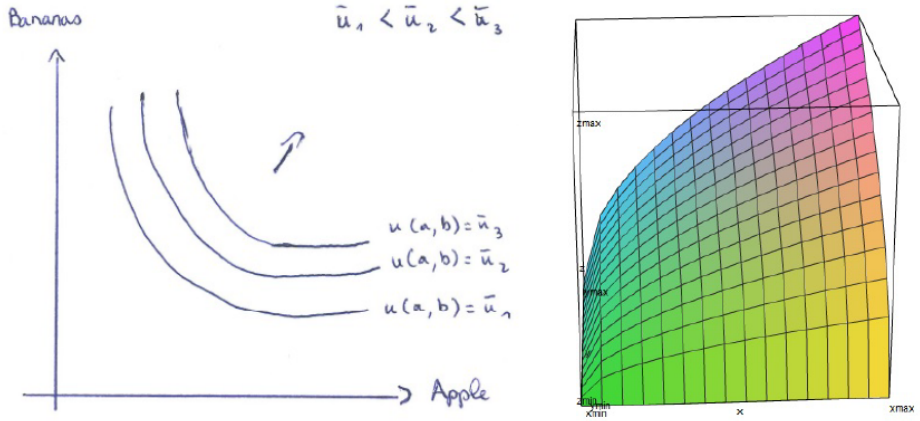


Figure 21: Indifference Map of a Cobb-Douglas Utility Function.

Coming back to our pre-image interpretation, let me stress again that one fixes a value of interest, c, that the function may take, for instance either the altitude of a coordinate point, or the utility level associated with a consumption choice. Then, the level set is the collection of *all* the points x that yield this value c when evaluating g at x, i.e. $g(x) = c$. Note that level sets may just consist of one element, be empty altogether, but conversely may also encompass even the whole domain of g, namely when g is a "flat" function.

Now that we have (hopefully) convinced ourselves that equality-constrained maximization is nothing but maximization on level sets, let us think about whether we can use this insight to reduce our new issue of solving the equality-constrained problem to the old one that we already know, the unconstrained problem. There is good news and bad news. The first piece of good news is that precisely this is indeed possible! The (full set of) bad news, however, is that things will get quite heavy in notation. To somewhat mitigate this problem, the following restricts attention to the scenario of *one* equality constraint – but it is important to keep in mind that this is just to have simpler notation, and as we will convince ourselves later, the generalization to multiple equality constraints works in a one to one fashion! Finally, the last piece of good news, and perhaps the most re-assuring fact is that while mathematically somewhat tedious to write down, the intuition of the approach is really simple.

Start from a function that is at least bivariate, i.e. $X \subseteq \mathbb{R}^n, n \geq 2$ and $f, g : X \mapsto \mathbb{R}$.[17] Suppose that f and g are "smooth" in an appropriate sense, for now at least once continuously differentiable. Suppose our problem is somewhat meaningful, such that $C(\mathcal{P}) = L_0(g) \neq \emptyset$ (in words: there are values for x that satisfy the restriction), and let $x_0 \in L_0(g)$. Now, suppose that we start from x_0 and only marginally vary the components at positions $2, 3, \ldots, n$ of x_0, but not the first component, say to $\tilde{x} = x_0 + \Delta x$ where $\Delta x_1 = 0$ and $\|\Delta x\| < \varepsilon$ where $\varepsilon > 0$ small. Because g is continuous, $g(\tilde{x}) = g(x_{0,1}, x_{0,2} + \Delta x_2, \ldots x_{0,n} + \Delta x_n)$ should still lie "really close" to $g(x_0) = 0$.[18] However, by continuous differentiability of g, the first partial derivative of g is continuous, so should we not also be able to vary the first argument by Δx_1 to move $g(x)$ back to zero, i.e. cancel any potential change in g induced by the marginal variation in the other arguments? That is, can we not find a function $\tilde{h}(\tilde{x} - x_0)$ such that

$$g(x_{0,1} + \tilde{h}(\tilde{x} - x_0), x_{0,2} + \Delta x_2, \ldots x_{0,n} + \Delta x_n) = g(x_0) = 0?$$

[17] Constrained problems in one variable are typically not too interesting. This is because already a single constraint will typically fundamentally restrict the set of x that are even possible, i.e. the zero-level sets of non-redundant functions $g(x)$ contain only a very small number of elements (think about any function $g(x)$ that is not constant at 0 and does not jump discontinuously), such that we need not apply sophisticated analytic techniques.

[18] $g(x_0) = 0$ holds because $x_0 \in L_0(g)$!

It turns out that so long as g actually moves if we vary the first argument (i.e., the first partial derivative is non-zero at x_0), the answer to this question is yes! To get closer to the formal theorem telling us about this fact, note that because the function $\tilde{h}$ takes $\tilde{x} - x_0$ as its argument and x_1 has not changed from x_0 to $\tilde{x}$, it depends only of the variables with index above one! Thus, more formally, when setting $h = x_0 + \tilde{h}$, the function h we are looking for would satisfy

$$g(h(x_2,\dots,x_n),x_2,\dots,x_n) = 0$$

for all $(x_2,\dots,x_n)$ **in a neighborhood of** $(x_{0,2},\dots,x_{0,n})$, since we were considering only marginal changes. Of course, we typically need not pick the first element to do this, so that we may instead simply decompose $x = (y,z)$ where y is the univariate variable that we want to adjust, and z are all the other variables that are allowed to move freely.[19] Then, the crucial step is that if a point $x^* = (y^*,z^*)$ is a local maximizer in the problem

$$\max_{(y,z)\in\mathbb{R}^n} f(y,z) \quad \text{subject to} \quad g(y,z) = 0,$$

because $x^* \in L_0(g)$, we can find our *implicit function* $h(z)$ such that **for any** z **in a neighborhood** U **of** z^***,** $g(h(z),z) = 0$**, and** z^* **maximizes the unconstrained problem**

$$\max_{z\in U} f(h(z),z).$$

But because we already know how to solve unconstrained problems, from here, things should be simple, right? Well, almost, because we have to ensure that we can sufficiently narrow down $h(z)$ to determine the optimal z^* and $y^* = h(z^*)$.

If you are suspicious why we picked out the first variable, note that we can do this with any other variable with non-zero partial derivative to study what happens when x_1 is part of the freely moving vector. If no other variables have non-zero partial derivatives, then they don't move g around x_0 anyways, so that we need only consider variation in x_0.

Let us formalize this idea. We begin by a result the proof of which is far beyond the scope of this course, if you took some mathematics classes in your undergraduate, you may have had the "pleasure" of finding out that it is indeed not too straightforward to establish. So let's just state it here:[20]

Theorem 56. ***(Univariate Implicit Function Theorem)*** *Let $X_1 \subseteq \mathbb{R}$, $X_2 \subseteq \mathbb{R}^{n-1}$ and $X := X_1 \times X_2$, and $g : X \mapsto \mathbb{R}$. Suppose that $g \in C^1(X)$, and that for a $(y^*,z^*) \in X_1 \times X_2$, $g(y^*,z^*) = 0$. Then, if $\frac{\partial g}{\partial y}(y^*,z^*) \neq 0$, there exists an open set $U \subseteq \mathbb{R}^{n-1}$ such that $z^* \in U$ and $h : U \mapsto \mathbb{R}$ for which $y^* = h(z^*)$ and $\forall z \in U : g(h(z),z) = 0$. Moreover, it holds that $h \in C^1(U)$ with derivative*

$$\nabla h(z) = -\left(\frac{\partial g}{\partial y}(h(z),z)\right)^{-1} \frac{\partial g}{\partial z}(h(z),z) \qquad \forall z \in U.$$

NOTE THAT $\frac{\partial g}{\partial z}$ REFERS TO A COLLECTION OF PARTIAL DERIVATIVES OF LENGTH $n-1$. Indeed, it is a notational convention we have discussed before, namely the one for taking partial derivatives with respect to more than one, but not all, arguments of a function!

Hopefully, the elaborations above have made the theorem somewhat accessible. Indeed, until the last line, everything just re-states what we said above, with the exception of the partial derivative condition $\frac{\partial g}{\partial y}(y^*,z^*) \neq 0$. This will tell us in practice which component y of $(x_1,\dots,x_n)$ we should choose

[19] We follow an established mathematical convention here and slightly abuse notation: although we write $x = (y,z)$, we don't require y to be the first element of x! You may e.g. have seen this in decompositions like $x = (x_i, x_{-i})$ that says that x is composed of the i-th element and all other elements.

[20] You may note that our notation slightly differs from other instances of the theorem. This is for the purpose of specifically connecting the definition to our elaborations before and the optimization notation.

to apply our implicit representation for the adjustment. A special case is $\nabla g(y^*, z^*) = \mathbf{0}$ where there is no direction for us to adjust alongside. **Such points are called singularities of g and require isolated investigation!** Apart from this, the thing that is new is that we learn from the theorem that h is continuously differentiable, and that we can indeed find a formula for the derivative. At first sight, this expression is not too helpful, because it depends of h itself, and the function h is still unknown, however, we know that $h(z^*) = y^*$, so that

$$h'(z^*) = -\left(\frac{\partial g}{\partial y}(y^*, z^*)\right)^{-1} \frac{\partial g}{\partial z}(y^*, z^*),$$

and this is an expression we know how to compute once we know y^* and z^*! Now, you also see why we needed the condition $\frac{\partial g}{\partial y}(y^*, z^*) \neq 0$, as otherwise, $h'(z^*)$ does not exist.

Still, we have thus far not discussed how we put this to use. Let's change this: suppose that $x^* = (y^*, z^*)$ is a local maximizer of the constrained problem, i.e. a local maximizer of $f|_{L_0(g)}$. Then, there exists our environment U such that z^* is a maximizer of $\max_{z \in B} f(h(z^*), z^*)$. Because the problem is unconstrained and U is open (i.e. there is no boundary, so no boundary points that may be an issue), we know that z^* is a critical point of $f(h(z^*), z^*)$, i.e.

$$\frac{d}{dz} f(y^*, z^*) = \frac{d}{dz} f(h(z^*), z^*) = \mathbf{0}.$$

Applying chain rule, we obtain

$$\frac{df}{dz} = \frac{\partial f}{\partial y}\frac{dh}{dz} + \frac{\partial f}{\partial z},$$

and because $\frac{dh(z)}{dz} = h'(z)$ is given by the implicit function theorem,

$$0 = \frac{\partial f}{\partial y}(y^*, z^*)\left(-\left(\frac{\partial g}{\partial y}(y^*, z^*)\right)^{-1} \frac{\partial g}{\partial z}(y^*, z^*)\right) + \frac{\partial f}{\partial z}(y^*, z^*).$$

Consequently, it results that

$$\frac{\partial f}{\partial y}(y^*, z^*)\left(\frac{\partial g}{\partial y}(y^*, z^*)\right)^{-1} \frac{\partial g}{\partial z}(y^*, z^*) = \frac{\partial f}{\partial z}(y^*, z^*). \tag{22}$$

Define

$$\lambda := \frac{\partial f}{\partial y}(y^*, z^*)\left(\frac{\partial g}{\partial y}(y^*, z^*)\right)^{-1},$$

and as a small spoiler (to see that we're almost there), call it the *Lagrange multiplier*. Then by definition, we clearly have

$$\frac{\partial f}{\partial y}(y^*, z^*) = \lambda \frac{\partial g}{\partial y}(y^*, z^*),$$

and by plugging the definition of λ into equation (22), we also get

$$\frac{\partial f}{\partial z}(y^*, z^*) = \lambda \frac{\partial g}{\partial z}(y^*, z^*).$$

Combining these expressions, this tells us that

$$\nabla f(x^*) = \nabla f(y^*, z^*) = \left(\frac{\partial f}{\partial y}(y^*, z^*), \frac{\partial f}{\partial z}(y^*, z^*)\right) = \left(\lambda \frac{\partial g}{\partial y}(y^*, z^*), \lambda \frac{\partial g}{\partial z}(y^*, z^*)\right) = \lambda \nabla g(y^*, z^*) = \lambda \nabla g(x^*).$$

This is indeed what we were after: a characterization of how the derivative of f relates to a derivative of g at a specific point x^* that is a local maximizer of the constrained problem. Note that this result is an

implication of x^* being a local maximizer, such that it is a *necessary condition* for a local maximum. Let us write this down:

Theorem 57. ***(Lagrange's Necessary First Order Condition)*** *Consider the constrained problem* $\max_{x \in L_0(g)} f(x)$ *where* $X \subseteq \mathbb{R}^n$ *and* $f, g \in C^1(X)$*. Let* $x^* \in L_0(g)$ *and suppose that* $\nabla g(x^*) \neq \mathbf{0}$*. Then,* x^* *is a local maximizer of the constrained problem only if there exists* $\lambda \in \mathbb{R} : \nabla f(x^*) = \lambda \nabla g(x^*)$*. If such* $\lambda \in \mathbb{R}$ *exists, we call it the Lagrange multiplier associated with* x^**.*

As in the unconstrained problem, the necessary conditions allow us to quite substantially narrow down the set of candidates that we have to consider: we moved from "the solution could be any point that satisfies the constraint" to "any solution x^* satisfies $\nabla f(x^*) = \lambda \nabla g(x^*)$", and typically, this step corresponds to a reduction from an infinitely large set to a set of two or three values! However, we still need to take care of other **candidates uncaptured by the theorem: boundary points of the support of f (provided that they satisfy the constraint) and singularities of the level set, i.e. points x^s where $\nabla g(x^s) = 0$.**

Let's take a moment to consider and formally justify the Lagrangian function, or simply, the Lagrangian. The necessary condition is "$\exists \lambda \in \mathbb{R} : \nabla f(x^*) = \lambda \nabla g(x^*)$". Thus, for said λ, it is an easy exercise to verify that evaluated at x^*, the gradient of the following function is equal to $\mathbf{0}$:

$$\mathcal{L}(\lambda, x) = f(x) - \lambda g(x) = f(x) + \lambda(0 - g(x)).$$

We call this function the Lagrangian. Importantly, from our considerations in the previous section, the necessary condition for a solution to unconstrained maximization of the Lagrangian is $\nabla \mathcal{L}(\lambda, x) = 0$, which is however equivalent to our necessary condition for the constrained problem we had just derived. Finally, note that the derivative of $\mathcal{L}(\lambda, x)$ with respect to λ is just $-g(x)$, and requiring $g(x) = 0$ is thus equivalent to requiring $\frac{\partial \mathcal{L}}{\partial \lambda}(\lambda, x) = 0$. Thus, **any candidate for an interior solution of the constrained problem has a $\lambda \in \mathbb{R}$ so that (λ, x^*) is a critical point of the (unconstrained) Lagrangian maximization problem!**[21] This gives the usual Lagrangian first order conditions for a solution x^* that you may be familiar with:

$$\begin{aligned} [x]: &\quad \frac{\partial \mathcal{L}}{\partial x}(x^*) = 0 \quad \Leftrightarrow \quad \nabla f(x^*) = \lambda \nabla g(x^*) \\ [\lambda]: &\quad \frac{\partial \mathcal{L}}{\partial \lambda}(x^*) = 0 \quad \Leftrightarrow \quad g(x^*) = 0 \end{aligned}$$

This gives a system of $n+1$ equations is $n+1$ unknowns,[22] representing our starting point for searching our solution candidates of any specific equality-constrained problem with one constraint. So, we can search for candidates with a (typically rather strong) necessary condition, and have a few values to compare. But can we also establish sufficiency in a fashion similar to the unconstrained problem? Yes and no. The issue is that **while every local extremum necessarily is also a critical point of the (unconstrained) Lagrangian problem, it is not necessarily an extremum of the Lagrangian!** To make sense of this, recall that being a critical point was necessary, but not sufficient for a maximum (or a minimum). Indeed, when considering $f, g \in C^2(X)$ and construct the Hessian of the Lagrangian, the so-called *bordered Hessian*, we obtain for any $(\lambda, x) \in X \times \mathbb{R}$:

$$H_{\mathcal{L}}(\lambda, x) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \lambda^2}(\lambda, x) & \frac{\partial^2 \mathcal{L}}{\partial \lambda \partial x}(\lambda, x) \\ \frac{\partial^2 \mathcal{L}}{\partial x \partial \lambda}(\lambda, x) & \frac{\partial^2 \mathcal{L}}{\partial x^2}(\lambda, x) \end{pmatrix} = \begin{pmatrix} 0 & -\nabla g(x) \\ -(\nabla g(x))' & H_f(x) - \lambda H_g(x) \end{pmatrix}$$

which results from $\frac{\partial^2 \mathcal{L}}{\partial \lambda^2} = \frac{\partial}{\partial \lambda}\left(\frac{\partial \mathcal{L}}{\partial \lambda}\right) = \frac{\partial}{\partial \lambda}(-g) = F_0$, where $F_0 : \mathbb{R} \mapsto \mathbb{R}, t \mapsto 0$ denotes the zero function, $\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial x} = \frac{\partial}{\partial \lambda}\left(\frac{\partial \mathcal{L}}{\partial x}\right) = \frac{\partial}{\partial \lambda}(\nabla f - \lambda \nabla g) = -\nabla g$, $\frac{\partial^2 \mathcal{L}}{\partial x \partial \lambda} = \left(\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial x}\right)' = (-\nabla g)'$ from symmetry, and $\frac{\partial^2 \mathcal{L}}{\partial x^2} = \frac{\partial \mathcal{L}}{\partial \lambda \partial x}(\nabla f - \lambda \nabla g) = H_f - \lambda H_g$.

[21] As formally shown below, any Lagrangian has only saddle points and no maximum or minimum, so do not call x^* a local maximum of the Lagrangian!!

[22] If it is linear, what do we require to obtain a unique solution candidate?

As shown below, the Lagrangian function does not have maximizers or minimizers, because **all its critical values violate the second order necessary condition**, that is, the matrix $H_{\mathcal{L}}$ can never be positive or negative semi-definite! For this, note that we can multiply matrices block-wise given conformable dimensions, the elaborations below comment on this more thoroughly. Now, consider an arbitrary point $(\lambda, x) \in X \times \mathbb{R}$, and pick any vector $v = (v_\lambda, v_x')' \in \mathbb{R}^{n+1}$ with $x_\lambda \in \mathbb{R}$ and $v_x \in \mathbb{R}^n$. Then,

$$v' H_{\mathcal{L}}(\lambda, x) v = (v_\lambda, v_x') \begin{pmatrix} 0 & -\nabla g(x) \\ -(\nabla g(x))' & H_f(x) - \lambda H_g(x) \end{pmatrix} \begin{pmatrix} v_\lambda \\ v_x \end{pmatrix}.$$

now, consider the dimension of the elements in the RHS expression, and see if the dimension conditions hold for multiplying everything out "block-wise": First, consider multiplication of the matrix with the right vector, i.e.

$$\begin{pmatrix} 1\times 1 & 1\times n \\ n\times 1 & n\times n \end{pmatrix} \begin{pmatrix} 1\times 1 \\ n\times 1 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} (1\times 1)\cdot(1\times 1) + (1\times n)\cdot(n\times 1) \\ (n\times 1)\cdot(1\times 1) + (n\times n)\cdot(n\times 1) \end{pmatrix} = \begin{pmatrix} 1\times 1 \\ n\times 1 \end{pmatrix}.$$

As we see, all products are of conformable dimension, and we also only add elements of same dimension, such that block multiplication works here. Note that the resulting object is a vector of length $n+1$ with an upper 1×1 block, and thus, this object is also conformable with $v' = (v_\lambda, v_x')$ in terms of block multiplication. So, multiplying everything out, we obtain

$$\begin{aligned} v' H_{\mathcal{L}}(\lambda, x) v &= (v_\lambda, v_x') \begin{pmatrix} 0 v_\lambda - \nabla g(x) v_x \\ -(\nabla g(x))' v_\lambda + (H_f(x) - \lambda H_g(x)) v_x \end{pmatrix} \\ &= v_\lambda(-\nabla g(x) v_x) + v_x'(-(\nabla g(x))' v_\lambda + (H_f(x) - \lambda H_g(x)) v_x) \\ &= -2 v_\lambda \nabla g(x) v_x + v_x' H_f(x) - \lambda H_g(x)) v_x \end{aligned}$$

where the last line has used that $v_x' \nabla g(x)'$ is scalar, and we can apply the transpose without changing the object to obtain $(v_x' \nabla g(x)')' = \nabla g(x) v_x$. Now, what does this tell us? Recall that, in order for $H_{\mathcal{L}}(\lambda, x)$ to not be indefinite, we need that

$$\forall v = (v_\lambda, v_x')' \in \mathbb{R}^{1+n} : (v' H_{\mathcal{L}}(\lambda, x) v \geq 0 \quad \vee \quad v' H_{\mathcal{L}}(\lambda, x) v \leq 0).$$

But **unless the gradient $\nabla g(x) = 0$, a case which we initially ruled out as a singularity we consider in separation**, this may never hold: for any fixed $v_x \in \mathbb{R}^n$ with $\nabla g(x) \neq 0$, for

$$v_\lambda^* = \frac{v_x' H_f(x) - \lambda H_g(x)) v_x}{2 \nabla g(x) v_x},$$

we have $v' H_{\mathcal{L}}(\lambda, x) v = 0$, and because $v' H_{\mathcal{L}}(\lambda, x) v$ is strictly monotonic in v_λ^*, if $\nabla g(x) v_x < 0$ (> 0), then $v' H_{\mathcal{L}}(\lambda, x) v > 0$ (< 0) for all $v_\lambda > v_\lambda^*$ but $v' H_{\mathcal{L}}(\lambda, x) v < 0$ (> 0) for all $v_\lambda < v_\lambda^*$!

Long story short, **it is not possible to establish that constrained maximization is equivalent to unconstrained Lagrangian maximization because the Lagrangian has only saddle points as critical values**! Thus, we need a different approach to second order conditions. Unfortunately, they will fall from the sky at this point, but their proof is beyond the scope of this class. We will need a further concept:

Definition 85. ***(Leading Principal Minor)*** *Consider a symmetric matrix $A = (a_{ij})_{i,j\in\{1,\dots,n\}} \in \mathbb{R}^{n\times n}$. Then, for $k \leq n$, the k-th leading principal minor of A, or the leading principal minor of A of order k is the matrix obtained from eliminating all rows and columns with index above k from A, i.e. the matrix $M_k^A = (a_{ij})_{i,j\in\{1,\dots,k\}} \in \mathbb{R}^{n\times n}$.*

In words, the k-th leading principal minor of A corresponds to the upper $k\times k$ block of A. Let's

consider an example. Let

$$A = \begin{pmatrix} 1 & 4 & 3 & 2 \\ 2 & 0 & 0 & 0 \\ 3 & 4 & -1 & -2 \\ 0 & 1 & e & \pi \end{pmatrix}.$$

Then, the leading principal minors of order $1, 2, 3$ and 4, respectively, are

$$M_1^A = (1), \qquad M_2^A = \begin{pmatrix} 1 & 4 \\ 2 & 0 \end{pmatrix}, \qquad M_3^A = \begin{pmatrix} 1 & 4 & 3 \\ 2 & 0 & 0 \\ 3 & 4 & -1 \end{pmatrix}, \qquad M_4^A = A.$$

Because it will important for the theorem, answer the following: what are the last two leading principal minors of A?[23]

Theorem 58. ***(Sufficient Conditions for the Constrained Problem)*** *Consider the constrained problem* $\max_{x \in L_0(g)} f(x)$ *where* $X \subseteq \mathbb{R}^n$ *and* $f, g \in C^2(X)$. *Let* $x^* \in L_0(g)$ *and* $\lambda^* \in \mathbb{R}$ *such that* (λ^*, x^*) *is a critical point of the Lagrangian function, i.e.* $\nabla f(x^*) = \lambda^* \nabla g(x^*)$ *and* $g(x^*) = 0$. *If* $m = 1$ *is the number of equality constraints, denote by* $M_{n-m+1}^{H_{\mathcal{L}}}(\lambda^*, x^*), \ldots, M_n^{H_{\mathcal{L}}}(\lambda^*, x^*)$ *the last* $n - m$ *principal minors of* $H_{\mathcal{L}}(\lambda^*, x^*)$. *If*

- $\forall j \in \{n - m + 1, \ldots, n\} : sgn(\det(M_j^{H_{\mathcal{L}}})) = (-1)^m$, *then* x *is a local minimizer of the constrained problem.*
- $\forall j \in \{n - m + 1, \ldots, n\} : sgn(\det(M_j^{H_{\mathcal{L}}})) = (-1)^j$, *then* x *is a local maximizer of the constrained problem.*

Here, $\text{sgn}(x)$ denotes the *sign function* equal to -1 if $x < 0$, to 0 for $x = 0$ and to 1 else. If your experience with this definition is like mine, then you will find it very unwieldy to consider in general. Still, the theorem above has already introduced the variable $m = 1$ to indicate the generalization to multiple constraints; on one hand to preview that things get as bad as this, but at least they don't get worse if we move to more constraints, and on the other, to not risk you memorizing the "wrong" conditions for multiple constraints. We will shortly consider an example to convince you that despite its ugly look, applying this theorem is rather straightforward when the problem we consider is not "too big". Further, note that (i) here, we only have a second order sufficient, but not a second order necessary condition, so our ability to rule out critical values based on failure of a necessary condition is comparatively limited (this is not too bad, usually, there are not too many critical values anyway) and (ii) as with the second order condition in the unconstrained case, we require the functions f and g to be C^2, i.e. twice continuously differentiable, such that **again, the second order condition is subject to a stronger smoothness regularity**. Typically in economic problems, this regularity assumption will be satisfied.

Finally, note that the search for global maxima, because the level set $L_0(g)$ is usually not convex, the theorem for the unconstrained case does not transfer, such that **a concave objective is not sufficient for the global maximum.** However, as we will see in the last section, if we can express our problem in *only convex inequality constraints*, then the level set will be convex, and an extreme value for a concave objective will indeed constitute a global maximum!

4.3.2 Lagrangian with One Equality Constraint: Example and Interpretation

Let us study an example to get a feeling for this method.Consider the problem

$$\max_{x \in \mathbb{R}^2} -\|x\|_2 \quad \text{subject to} \quad x_1 + x_2 = 1.$$

[23] M_3^A and M_4^A.

That is, we seek the vector with minimum Euclidean length $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$ in the $\mathbb{R}^2$ that satisfies $x_1 + x_2 = 1$ (for economic context, you may think of the length as a cost function and 1 as a production target). How do we approach this issue? First, a trick to get rid of the unwieldy square root; let's consider the equivalent problem

$$\max_{x \in \mathbb{R}^2} -(x_1^2 + x_2^2) \quad \text{subject to} \quad x_1 + x_2 = 1.$$

This may always be worthwhile to think about – can you re-write the objective to a simpler expression without changing the solutions? Just remember to plug the solutions into the original objective in the end, then this approach works just fine and may save you a lot of work!

In **step zero, we think about points that the Lagrangian method won't find: boundary points and singularities!** However, because $\mathbb{R}^2$ is an open set, boundary points are not an issue. Next, at any $x \in \mathbb{R}^2$, the gradient of the constraint function $g(x) = x_1 + x_2 - 1$ is $\nabla g(x) = (1,1)' \neq \mathbf{0}$, such that there are no singularities. Thus, we can restrict attention to critical values as found by the Lagrangian method. **Make sure that you always think about this step!**

In the first step, we look for our candidates using the first order necessary conditions, or the Lagrangian first order conditions (FOCs), given by

$$\begin{aligned} [x]: &\quad f(x^*) = \lambda g(x^*) \\ [\lambda]: &\quad g(x^*) = 0 \end{aligned}$$

So, let's compute the gradient of $f(x) = -\|x\|_2^2$ (the one of g, we already took care of above):

$$\nabla f = -(2x_1, 2x_2).$$

Thus, the Lagrangian FOC for x^* tells us that

$$\begin{pmatrix} 2x_1^* \\ 2x_2^* \end{pmatrix} = \lambda^* \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

This gives either $\lambda^* = x_1^* = x_2^* = 0$ or $x_1^* = x_2^* = \lambda^*/2 \neq 0$. Considering the FOC for λ, i.e. the constraint, the former case can be ruled out; plugging in the latter gives

$$\lambda^*/2 + \lambda^*/2 = 1 \quad \Rightarrow \quad \lambda^* = 1,\ x_1^* = x_2^* = \frac{1}{2}.$$

Hence, there is just a unique candidate to consider: $x^* = (1/2, 1/2)'$. Let's consult the second order sufficient condition and see whether we can decide upon the type of extremum: recall the form of the Bordered Hessian above. Computing it for an arbitrary (λ, x) gives

$$H_{\mathcal{L}}(\lambda, x) = \begin{pmatrix} 0 & -1 & -1 \\ -1 & -2 & 0 \\ -1 & 0 & -2 \end{pmatrix}.$$

Thus, it is the same irrepsective of x and λ, and we don't need to plug in our candidate solution.[24] Now, let's apply the theorem. We have $n = 2$ variables and $m = 1$ constraints, so $n - m = 1$, and indeed, we only have to compute the determinant of the *last* leading principal minor, so the Bordered Hessian itself. Before doing so, let's form expectations: the point is (i) a maximum if the determinant's sign is $(-1)^n = 1$, i.e. $\det(H_{\mathcal{L}}(\lambda^*, x^*)) > 0$, (ii) a minimum if the determinant's sign is $(-1)^m = -1$, i.e. $\det(H_{\mathcal{L}}(\lambda^*, x^*)) < 0$, and (iii) may be either of the two, or a saddle point if $\det(H_{\mathcal{L}}(\lambda^*, x^*)) = 0$.

[24] Otherwise, you would have to plug in the candidate – including the associated multiplier, which may differ across candidates!

Applying our 3×3 determinant rule, all the right-diagonals yield a zero product, while we twice get $(-1)(-1)(-2) = -2$ on the left-diagonal, such that $\det(H_{\mathcal{L}}(\lambda^*, x^*)) = 4 > 0$, and we have indeed found a maximum!

To conclude, $x^* = (1/2, 1/2)'$ minimizes the Euclidean length of vectors in $\mathbb{R}^2$ that satisfy $x_1 + x_2 = 1$, and it has length $\|x^*\| = \sqrt{(1/2)^2 + (1/2)^2} = \frac{\sqrt{2}}{2}$.

To conclude our investigations on the case of one equality constraint, some intuition and interpretation of the method. First, for our weird second order sufficient condition that more or less fell from the sky, while it is very abstract, there is something intuitive to note about it: the number of principal minors to consider. Recall that the gradient of g, ∇g, was ruled out to have a "row rank deficit", which for row vectors is a fancy way of saying that they are not zero (generally, a matrix $A \in \mathbb{R}^{n \times m}$ has a row rank deficit if its rank is below n; this concept will be important in the multivariate extension below, so we already mention it here). In general, this means that all constraints in g move in *independent directions* at the solution, in our case of just one constraint, it is sufficient that the function g moves at all with variation in x around the solution, i.e. we have not included a nonsense constraint such as $0 = 0$, or even worse, $-5 = 0$.[25] As such, the constraint restricts $m = 1$ directions of the $\mathbb{R}^n$ in maximization of f, but it leaves free the remaining $n - m$, which is exactly the number of leading principal minors that we have to consider! Thus, for every direction we can freely choose alongside, there is exactly one condition telling us whether a point is a maximum along this dimension, and only if all maximum conditions of the individual directions are satisfied at our solution, we have indeed found a maximum! Recall also our discussions of the saddle point graph – exactly like there, the conditions here rule out that we have a minimum in one and a maximum into the other direction.

The second, and perhaps far more important piece of interpretation concerns the Lagrangian multiplier. Have you ever heard the expression "shadow price of the budget constraint"? Then you may know that we associate it with the Lagrangian multiplier in constrained utility maximization subject to the budget constraint. But what does it mean? Let's consider a general constraint $\tilde{g}(x) = y$, but to ease interpretation, keep in mind that y may be the budget and we may have $\tilde{g}(x) = \sum_{i=1}^{n} p_i x_i$ as the sum of expenditures on individual goods. Note that given $\tilde{g}$, we can express the optimal solution (assuming for now that it is unique) as $x^*(y)$, i.e. it will depend only on the value of y (the budget available) that constrains the problem. Denote $F(y) = f(x^*(y))$ as the objective function's value at the optimum (the implicit utility function $U(y) = u(x^*(y))$), and let $G(y) = \tilde{g}(x^*(y))$ be the implicit constraint function (expressing expenditures at the optimum, $E(y) = \sum_{i=1}^{n} p_i x_i^*(y)$). Then, note that by the chain rule,

$$\frac{d}{dy}F(y) = \nabla f(x^*(y))\frac{dx^*}{dy}(y) = \lambda^* \nabla g(x^*(y))\frac{dx^*}{dy}(y) = \lambda^* \frac{d}{dy}G(y)$$

and thus

$$\lambda = \frac{\frac{d}{dy}F(y)}{\frac{d}{dy}G(y)} \qquad \left(\lambda = \frac{\frac{d}{dy}U(y)}{\frac{d}{dy}E(y)}\right).$$

In case of the utility maximization problem, λ tells us the *ratio of change in utility and change in expenditures*. Because the household always spends all his income (provided that the utility function is strictly increasing in consumption), we have $\frac{d}{dy}E(y) = 1$ so that $\lambda = \frac{d}{dy}U(y)$, i.e. it resembles the utility's responsiveness to increases in y. Hence, we can conversely interpret it as the *marginal utility cost of being budget-constrained*! For the general problem, the intuition is similar, because if we continue to require $\tilde{g}(x) = y$, then trivially $\frac{d}{dy}G(y) = 1$, and we can interpret λ as the marginal increase in the objective function associated with marginally relaxing the constraint. **Note that this means that the multiplier**

[25] The function need not be constant on the whole domain for this issue to occur, also the seemingly smart attempt to re-write inequality constraints using indicator functions, e.g. $x'x - 5 \leq 0$ as $\mathbb{1}[x'x - 5 > 0] = 0$ falls victim to this issue at every point where $\mathbb{1}[x'x - 5 > 0]$ is differentiable.

MUST BE non-negative, i.e. $\lambda \geq 0$**!** If $\lambda = 0$, we are in an interesting special case: here, the constraint does not matter, in the sense that $x^*(y)$ is the solution to the unconstrained problem as well.

Indeed, this intuition will be crucial when we generalize our insights to inequality constraints: if an inequality constraint "matters", in the set of x such that $h_j(x) \leq 0$, it constrains our optimization opportunities, then it must be the case that it has a non-zero "cost" in terms of the objective, that is, it is associated with a multiplier unequal to zero, and the constraint is *binding* at x^* in the sense that $h_j(x^*) = 0$. Otherwise, if the $h_j(x^*) < 0$, we say that the constraint *is slack*, we could deviate marginally from x^* and still satisfy the constraint – it thus seems not to have a value cost for the objective, and must be associated with a multiplier = 0. This suggests that when μ_j is the multiplier associated with the constraint $h_j \leq 0$, we should have an equality of the form $\mu_j h_j(x^*) = 0$, called a *complementary slackness* condition (think about why this equality encompasses the two cases before). Indeed, next to the constraint holding, this is *almost* all we need to generalize the method to inequality constraints!

Consider again the Lagrangian (with constraint $g(x) - y = 0$):

$$\mathcal{L}(\lambda, x) = f(x) + \lambda(y - \tilde{g}(x)).$$

Suppose now that we marginally relax the (budget) constraint, that is, assume that y marginally increases (the household gets more money). Then, note that

$$\frac{\partial \mathcal{L}}{\partial y} = \lambda.$$

4.3.3 Lagrangian Optimization: Multiple Constraints

Here, we turn to the results that ensure generalization. There will not be much discussion, all formal ideas and the intuition transfer from the case of one constraint. The only tricky part in generalization here is the rank condition for the Jacobian of the constraint function $g = (g_1, \ldots, g_m)' : \mathbb{R} \supseteq X \mapsto \mathbb{R}^m$, the intuition of which was already discussed above. This is because when starting from a problem of the form

$$\max_{x \in C(\mathcal{P})} f(x), \qquad C(\mathcal{P}) = \{x \in \mathbb{R}^n : (\forall i \in \{1, \ldots, m\} : g_i(x) = 0)\},$$

we can always define g as above so that

$$C(\mathcal{P}) = \{x \in \mathbb{R}^n : g(x) = \mathbf{0}\}$$

which looks very much like the constraint set of the one-constraint problem, only that we are now dealing with a *vector*-valued function that must be equal to the zero *vector*. But, as you may recall from Chapter 1, vector spaces are endowed with very similar algebraic structures as real numbers, that in many cases, including this one, allow for convenient and relatively straightforward generalizations. Indeed, we may rely on the multivariate implicit function theorem:

Theorem 59. ***(Multivariate Implicit Function Theorem)*** *Let* $X_1 \subseteq \mathbb{R}^m$, $X_2 \subseteq \mathbb{R}^{n-m}$ *and* $X := X_1 \times X_2$, *and* $g : X \mapsto \mathbb{R}^m$. *Suppose that* $g \in C^1(X, \mathbb{R}^m)$, *and that for a* $(y^*, z^*) \in X_1 \times X_2$, $g(y^*, z^*) = \mathbf{0}$. *Then, if* $\text{rk}\left(\frac{\partial g}{\partial y}(y^*, z^*)\right) = m$, *there exists an open set* $U \subseteq \mathbb{R}^{n-m}$ *such that* $z^* \in U$ *and* $h : U \mapsto \mathbb{R}^m$ *for which* $y^* = h(z^*)$ *and* $\forall z \in U : g(h(z), z) = \mathbf{0}$. *Moreover, it holds that* $h \in C^1(U, \mathbb{R}^m)$ *with derivative*

$$J_h(z) = -\left(\frac{\partial g}{\partial y}(h(z), z)\right)^{-1} \frac{\partial g}{\partial z}(h(z), z) \qquad \forall z \in U.$$

The important changes from the univariate theorem are indicated with red color. You can see that the only thing that has really changed is the rank condition, which, as has been discussed in the pre-

vious section, encompasses the univariate case, because when $m = 1$, $\text{rk}\left(\frac{\partial g}{\partial y}(y^*, z^*)\right) = m$ is equivalent to $\frac{\partial g}{\partial y}(y^*, z^*) \neq 0$. Note that in the general case, $\frac{\partial g}{\partial y}$ is the matrix-valued Jacobian. The intuition is the same as before: for every constraint $g_i(x) = 0$, we need to vary g_i independently from the other constraints so as to ensure that it holds in a neighborhood of the point of interest. As in the univariate case, we can not have $m > n$, i.e. more (meaningful) constraints than directions of variation.

In analogy to above, this theorem allows us to derive the necessary first order condition that gives us the set of "interior" candidates for extrema:

Theorem 60. ***(Lagrange's Necessary First Order Conditions for Multiple Constraints)*** *Consider the constrained problem* $\max_{x \in L_{\mathbf{0}}(g)} f(x)$ *where* $X \subseteq \mathbb{R}^n$ *and* $f \in C^1(X), g \in C^1(X, \mathbb{R}^m)$*. Let* $x^* \in L_{\mathbf{0}}(g) = \cap_{i \in \{1,\dots,m\}} L_0(g_i)$ *and suppose that* $\text{rk}(J_g(x^*)) = m$*. Then,* x^* *is a local maximizer of the constrained problem only if there exists* $\Lambda = (\lambda_1, \dots, \lambda_m)' \in \mathbb{R}^m : \nabla f(x^*) = \Lambda' J_g(x^*)$*. If such* $\Lambda \in \mathbb{R}$ *exists, we call* λ_i *the Lagrange multiplier associated with* x^* *for the* i*-th constraint.*

Here, instead of $\nabla f(x^*) = \Lambda' J_g(x^*)$, you may be more familiar with the condition

$$\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*).$$

You can check that these two characterizations are equivalent using the definition of the Jacobian (and the block-multiplication rule). As before, this theorem gives us all the candidates for local extrema that we have to consider when searching for the global maximum/minimum. As with the univariate case, the i-th multiplier can be interpreted as the (non-negative) value cost of the i-th constraint at x^*.

Ruling out extreme values with the second order necessary condition is exactly analogous to before (because we had already formulated the second order condition quite generally earlier):

Theorem 61. ***(Second Order Sufficient Conditions for Multiple Constraints)*** *Consider the constrained problem* $\max_{x \in L_{\mathbf{0}}(g)} f(x)$ *where* $X \subseteq \mathbb{R}^n$ *and* $f \in C^2(X), g \in C^2(X, \mathbb{R}^m)$*. Let* $x^* \in L_{\mathbf{0}}(g) = \cap_{i \in \{1,\dots,m\}} L_0(g_i)$ *and* $\Lambda^* \in \mathbb{R}$ *such that* (Λ^*, x^*) *is a critical point of the Lagrangian function, i.e.* $\nabla f(x^*) = (\Lambda^*)' J_g(x^*)$ *and* $g(x^*) = \mathbf{0}$*. Denote by* $M_{n-m+1}^{H_{\mathcal{L}}}(\lambda^*, x^*), \dots, M_n^{H_{\mathcal{L}}}(\lambda^*, x^*)$ *the last* $n - m$ *principal minors of* $H_{\mathcal{L}}(\lambda^*, x^*)$*. If*

- $\forall j \in \{n - m + 1, \dots, n\} : sgn(\det(M_j^{H_{\mathcal{L}}})) = (-1)^m$*, then* x *is a local minimizer of the constrained problem.*
- $\forall j \in \{n - m + 1, \dots, n\} : sgn(\det(M_j^{H_{\mathcal{L}}})) = (-1)^j$*, then* x *is a local maximizer of the constrained problem.*

Next to the slightly different first order condition, the only difference is the form of the Bordered Hessian. It still corresponds to the Hessian of the Lagrangian function, however, due to the multitude of constraints, this Hessian now looks slightly different:

$$H_{\mathcal{L}}(\Lambda, x) = H_{\mathcal{L}}(\lambda_1, \dots, \lambda_m, x) = \begin{pmatrix} \mathbf{0}_{m \times m} & (J_g(x))' \\ J_g(x) & H_f(x) - \sum_{i=1}^{m} \lambda_i H_{g_i}(x) \end{pmatrix}.$$

You can try to compute this yourself using the definition of the Hessian and appreciating the vector structure of Λ. Note again that we need a stronger smoothness assumption for the second order necessary condition, as f and g need to be C^2.

4.3.4 Equality Constraints and the Lagrangian: A Recipe

We have done quite some math to establish our approach to equality-constrained problems using the Lagrangian function. Having a good understanding of this analytic justification will greatly aid your general knowledge in econ-related math, however, it may be a bit overwhelming, such that the clear path to approach a specific problem may not be crystal clear just yet. Thus, in our last discussion of

problems with equality constraints, let us outline a "cookbook recipe" to approach such problems for you to follow when confronted with specific applications.

Starting Point. Consider a problem

$$\max_{x \in \text{dom}(f)} f(x) \quad \text{subject to} \quad \forall i \in \{1, \ldots, m\} : g_i(x) = 0$$

and suppose that f and $g = (g_1, \ldots, g_m)'$ are *sufficiently smooth*, i.e. $f \in C^2(X)$ where $X = \text{supp}(f) \subseteq \mathbb{R}^n$ and $g \in C^2(X, \mathbb{R}^m)$. This ensures that *both* our first and second order conditions are applicable.

Analytical Steps.

0. Collect all candidate points that can not be identified via the "interior" (local extremum) method, i.e. **boundary points** of the support X of f (if any), and **singularities** of the level set $L_g(x)$, i.e. x^* with $\text{rk}(J_g(x^*)) < m$. Next, turn to the interior candidates.
1. ("Apply First Order Necessary Condition"/"FOC") Compute the set of interior candidates, i.e. the critical points of the Lagrangian function, denoted for now as $\{x_1^*, \ldots, x_K^*\}$, and the (vectors of) associated multipliers $\{\Lambda_1^*, \ldots, \Lambda_K^*\}$.
2. ("Apply Second Order Sufficient Condition"/"SOC") Can we identify the type of all candidates using the Bordered Hessian Criterion? Rule out Minima!
3. (in case of multiple candidates) For all candidates x^* that remain until here (including those of step 0), compute $f(x^*)$ and determine which point(s) remain(s) as candidates for the global minimizer/maximizer.

Note that step 3 may, as in our example above, not be necessary if only one candidate remains after step 2. Finally, as a technical point, note also that we have to ensure somehow that maximum and minimum exist, because we are only dealing with *necessary* conditions here, and identifying a single plausible extreme value candidate that satisfies them is only sufficient for identifying the extreme value of interest if its existence is ensured in the first place! In practice, this is ensured by compactness of the constraint set, as demonstrated earlier e.g. for the budget set.[26] **Nonetheless, you could think of step -1 as ensuring existence of the extreme value of interest!** Accordingly, the procedure may be compactly summarized as follows:

$$\begin{aligned} \text{Existence} \quad &\rightarrow \quad \text{Freak Points} \quad \rightarrow \quad \text{Regular Points: FOC} \\ &\rightarrow \quad \text{Regular Points: SOC} \quad \rightarrow \quad \text{Compare values} \end{aligned}$$

4.4 Inequality Constraints

Finally, it is time to turn to our approach to inequality constraints. The logic we employ to solve problems featuring also inequality constraints is the one outlined above: a solution candidate will be required to satisfy the Lagrangian conditions for all equality constraints, and further, for the inequality constraints we will impose (i) feasibility, i.e. being contained in the constraint set for a solution candidate, and (ii) our complementary slackness condition. The result we rely on is the Karush-Kuhn-Tucker theorem that tells us the necessary and sufficient conditions for an optimum. Here, we only state it, as the application procedure is more or less analogous to the "Lagrangian cookbook formula", and the central take-away from this Chapter should be the general foundation of unconstrained optimization and how we formally/intuitively justify the Lagrangian approach.

[26] When using the budget set in line with an equality-constrained problem, i.e. $B(y|p) = \{x : p'x = y\}$, we can establish compactness as follows. Note that $B(y|p)$ is the complement of the union of the sets $\{x : p'x < y\}$ and $\{x : p'x > y\}$ which are both open. Thus, their union is open, and this budget set is closed as well. Boundedness is equivalent to the budget set defined by $p'x \leq y$, and given if the smallest price is strictly greater than zero. Thus, the equality-problem budget set is compact by Heine-Borel's theorem.

An excellent formal discussion with examples of the Kuhn-Tucker method can be found on https://sites.google.com/site/nicolasschutz/teaching, download the PDF-file "Handout on Kuhn-Tucker" from the E701-course.[27]

For the problem $\mathcal{P}$ with *only inequality* constraints

$$\max_{x \in \text{dom}(f)} f(x) \qquad \text{subject to} \qquad \forall j \in \{1,\dots,k\} : h_j(x) \leq 0,$$

where f, h_j are all C^1 functions, the result may be summarized as follows:

Theorem 62. ***(Karush-Kuhn-Tucker Theorem)*** *For $\mu = (\mu_1,\dots,\mu_k)' \in \mathbb{R}^k$, consider the* ***optimality conditions***

- *(Feasibility)* $\forall j \in \{1,\dots,k\} : h_j(x) \leq 0$,
- *(FOC for x)* $\nabla f(x) = \sum_{j=1}^{k} \mu_j h_j(x)$,
- *(Complementary Slackness)* $\forall j \in \{1,\dots,k\} : \mu_j h_j(x) = 0$.

Then,

1. *(Necessity) If $x^* \in dom(f)$ is a local maximizer of $\mathcal{P}$ for which the set $\{\nabla h_j(x^*) : h_j(x^*) = 0\}$ is linearly independent, there exists $\mu^* \in \mathbb{R}^k$ such that (x^*, μ^*) satisfies the optimality conditions.*
2. *(Sufficiency) If f is concave, g is quasi-convex and there exists $x_0 \in dom(f)$ such that $g(x_0) < 0$, if the optimality conditions hold at $x^* \in dom(f)$, then x^* is a local maximizer of $\mathcal{P}$.*

In contrast to the Lagrange Theorem discussed earlier, what has changed in the Karush-Kuhn-Tucker (KKT) Theorem is that the rank condition has yet become more complex again. Intuitively, when varying candidates x^*, the set of "relevant" equality constraints changes because at different x^*, different $h_j(x^*) \leq 0$ will hold with equality. In consequence, we have a different rank condition at different x^*, expressed by the more unwieldy characterization above. Nonetheless, the concept is the same as before, and as there, we must *separately identify singularity candidates* from points that violate this condition, if there are any. The rest of the "necessity" part of the theorem follows the logic discussed earlier: points must be feasible, and either an inequality constraint holds with equality ($h_j(x) = 0$), or it does not constrain our ability to maximize f, i.e. the value of the objective function, locally ($\mu_j = 0$), which we summarize in the complementary slackness condition.

The "sufficiency" part of the theorem summarizes the key insights from the issue of "convex optimization". Deriving this result is beyond the scope of this class. However, note that when all conditions are met, that is, the linear independence rank condition (also called *constraint qualification*), the shape criteria for objective and constraint functions and $\exists x_0 \in dom(f) : g(x_0) < 0$ (also called *Slater's condition*) are all satisfied, then the optimality conditions are indeed **equivalent** to the local maximizer property, which greatly simplifies the search for local extremizers. Furthermore, note that for search of *local minimizers*, the conditions are exactly the same, with the exception that f must be convex rather than concave to obtain sufficiency.

As you may have noted, thus far, we have not addressed problems with both equality and inequality constraints, which are, however, very relevant to economics - for instance, you can think of any equality-constrained problem with added non-negativity constraints for some choice variables. So, what about problems $\mathcal{P}$ of the form

$$\max_{x \in \text{dom}(f)} f(x) \qquad \text{subject to} \qquad \forall i \in \{1,\dots,m\} : g_i(x) = 0 \qquad \wedge \qquad \forall j \in \{1,\dots,k\} : h_j(x) \leq 0?$$

Here, we also have a variant of KKT:

[27] Beware that the notation is slightly different as the one here, don't get confused!

Theorem 63. ***(Karush-Kuhn-Tucker with Equality Constraints)*** *For* $\Lambda = (\lambda_1, \ldots, \lambda_m)' \in \mathbb{R}^m$ *and* $\mu = (\mu_1, \ldots, \mu_k)' \in \mathbb{R}^k$*, consider the* ***optimality conditions***

- *(Feasibility)* $\forall j \in \{1, \ldots, k\} : h_j(x) \leq 0$ *and* $\forall i \in \{1, \ldots, m\} : g_i(x) = 0$,
- *(FOC for x)* $\nabla f(x) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j=1}^{k} \mu_j h_j(x)$,
- *(Complementary Slackness)* $\forall j \in \{1, \ldots, k\} : \mu_j h_j(x) = 0$.

Then, if $x^* \in dom(f)$ *is a local maximizer of the constrained problem for which the set* $\{\nabla h_j(x^*) : h_j(x^*) = 0\} \cup \{\nabla g_i(x^*) : i \in \{1, \ldots, m\}\}$ *is linearly independent, there exist* $\Lambda^* \in \mathbb{R}^m$ *and* $\mu^* \in \mathbb{R}^k$ *such that* (x^*, Λ^*, μ^*) *satisfies the optimality conditions.*

Thus, the "inequality-only" KKT transfers with two caveats: the first is technical and refers to the fact that constraint qualification has become more complex: because all equality constraints always "bind", i.e., hold with equality, they need to be considered in the linear independence test set. The second is more severe from a methodological point of view: sufficient conditions are no longer "readily" obtained, and for problems with both equality and inequality constraints, we usually only use necessary conditions.[28] While this may sound unfortunate, in many applications, it suffices to check for solution existence and then compare the values of candidates that remain after the FOC, which are usually not many.
That being said, it may still be quite tedious – or at least, time-consuming – to apply KKT, and it is also more error-prone relative to the simpler Lagrangian method. Fortunately, in economics, we can frequently avoid KKT by simplifying the problem to an equality-only or entirely unconstrained problem. This shall be the purpose of our remaining discussion.

4.4.1 Problem Simplifications

In the following, let us consider when and how we can re-write seemingly inequality-constrained problems as equality-constrained, which greatly facilitates the analytic solution. To start off, consider again the general constrained maximization problem,

$$\max_{x \in \text{dom}(f)} f(x) \quad \text{subject to} \quad \forall i \in \{1, \ldots, m\} : g_i(x) = 0 \quad \wedge \quad \forall j \in \{1, \ldots, k\} : h_j(x) \leq 0.$$

Instead of solving it using KKT, we may be able to re-write it as

$$\max_{x \in \text{dom}(f)} f(x) \quad \text{subject to} \quad \forall i \in \{1, \ldots, \tilde{m}\} : g_i(x) = 0,$$

which we can apply the more familiar and straightforward Lagrangian method to. Generally, there are two approaches to simplifying the problem: (i) imposing that an inequality constraint is binding, i.e. that it necessarily holds with equality (and thus is "not really an inequality constraint at all"), or (ii) dropping inequality constraints.

Let us study these methods in the context of the budget-constrained utility maximization problem we already introduced,

$$\max_{x \in \mathbb{R}^n} u(x) \quad \text{subject to} \quad p'x - y \leq 0, \ \forall i \in \{1, \ldots, n\} : x_i \geq 0$$

[28] A 'smart approach to represent equality constraints in the inequality-constrained problem is through two separate constraints, $-g(x) \leq 0$ and $g(x) \leq 0$. If g is quasi-linear, then both g and $-g$ are quasi-convex, such that the sufficient condition may apply! You should then, however, use the KKT with equality constraints to judge upon necessity, as $\nabla g = (-1) \cdot \nabla(-g)$, i.e. g and $-g$ are linearly dependent, and only the constraint qualification condition of the equality-constrained KKT may potentially hold.

with price vector p such that $p_i > 0$ for all i and $y > 0$. Utility functions that we typically consider are such that $u(x)$ is strictly increasing in all arguments, i.e. $\forall x \in \mathbb{R}^n \forall j \in \{1,\dots,n\} : \frac{\partial u}{\partial x_j}(x) > 0$, an assumption that we are usually ready to impose because strictly more is typically strictly better in the view of the consumer. Moreover, we assume that the marginal utility of the first bit is infinitely high, i.e. $\lim_{x_j \to 0} \frac{\partial u}{\partial x_j}(x) = \infty$. Common examples that satisfy these conditions are Cobb-Douglas or Constant Elasticity of Substitution utility functions.

The first simplification ("not really an inequality constraint") can be applied to the budget constraint: a point x with $p'x - y < 0$ can never be a solution, because the consumer can spend more to increase utility! Thus, **all solutions of the initial problem will be solutions of the re-written problem**

$$\max_{x \in \mathbb{R}^n} u(x) \qquad \text{subject to} \qquad p'x - y = 0,\ \forall i \in \{1,\dots,n\} : x_i \geq 0.$$

Whenever this condition in bold holds, the simplification is justified. Next, note that the marginal utility from spending money on good i is $\frac{\partial u}{\partial x_i}(x)/p_i$ (because 1 dollar gives $1/p_i$ units of the good). For any x such that $x_i = 0$ and $x_j > 0$, we can marginally decrease spending on x_j and increase spending on x_i - and, by our zero limit assumption for the marginal utility, generate a large leap in utility. Thus, the point x with $x_i = 0$ can not be an optimum. Thus, $x_i \geq 0$ will never bind, and if x^* solves our problem, it is also a local maximizer in the problem

$$\max_{x \in \mathbb{R}^n} u(x) \qquad \text{subject to} \qquad p'x - y = 0.$$

The Lagrangian approach of this problem gives $\nabla u(x^*) = \lambda p$, and we can ex-post impose that $x^* \in \mathbb{R}^n_+$, i.e. that all entries in x^* must be non-negative. In case of a bivariate Cobb-Douglas function $u(x) = x_1^\alpha x_2^{1-\alpha}$, $\alpha \in (0,1)$, it may be a good exercise to formally verify that the optimal input ratio x_1^*/x_2^* is given by $\alpha/(1-\alpha)$, pay special attention to the line of arguments that ensures that this is indeed the unique global maximum. Thus, we can restrict the role of inequality constraints to feasibility of points in the Lagrangian problem, making them a further criterion in the "cookbook recipe" outlined above according to which we may rule out candidates identified from the FOC.

Accordingly, we may be able to reduce the problem to either a standard Lagrangian problem if we are able to argue that the inequality constraints are actually equality constraints, or solve an equivalent Lagrangian problem with an additional feasibility condition that rules out some of the critical values. Indeed, most optimization problems you face during the Master's classes that feature inequality constraints can be reduced to such Lagrangian problems that you can solve using the methods discussed above.

As a final note on problem simplification, recall our motivation for deriving our one-equality-constraint procedure from the implicit function theorem: it may not always be possible to represent constraints using global, explicit functions. On the other hand, sometimes, this may indeed be the case. If so, you can *always* get rid of the respective constraint - solve for the explicit function representing a constraint and plug it into the problem! This usually makes the problem much easier to solve, as you reduce its dimensionality and the number of constraints. As an example, you can consult the elaborations on the explicit function in the section "Level Sets and Implicit Functions for Optimization" that we used to re-write the two-dimensional constrained problem as a one-dimensional, unconstrained one. When applying this trick, be sure to plug the explicit function not only in to the objective, but also to the other constraints, if any.

4.4.2 Exploiting Intuition: Lagrangian Multipliers as a Sufficient Condi-

TION

The intuition of Lagrangian multipliers, when combined with our re-writing approach presented above, yields an extremely simplistic alternative to the computation-intensive second order condition approach to determining the type of extreme values. For simplicity, we consider again the scenario of one constraint.

Suppose that we start from some problem (either the initial one, or a re-written inequality-constrained problem)

$$\max_{x\in\text{dom}(f)} f(x) \qquad \text{s.t.} \qquad g(x) = 0.$$

To use Lagrangian multipliers as a necessary condition, as stated above, we need a clear notion of how the constraint can be "relaxed". For this, a first requirement must hold: the problem must be *equivalent* to the inequality-constrained problem

$$\max_{x\in\text{dom}(f)} f(x) \qquad \text{s.t.} \qquad g(x) \leq 0.$$

For instance, this is the case for utility maximization subject to a budget constraint, but this equivalence also applies to a broad range of further economic optimization problems. Generally, as discussed in the previous section, this equivalence is more easily established when starting out from the inequality-constrained problem, and holds e.g. when both the objective and constraint are strictly monotonic in one variable. The equivalence to the inequality-constrained problem ensures that we have a well-defined idea of what "relaxing" the constraint means. For instance, in the example of the budget constraint, this notion would be to increase the budget y.
On this, make sure that your equality-constrained problem is equivalent to a problem of the form "$g(x) \leq 0$" rather than "$g(x) \geq 0$". The intuition of relaxing the constraint critically depends on this form, and if you assume a constraint "$g(x) \geq 0$" you are actually considering the problem with "$-g(x) \leq 0$", therefore inverting the sign of Lagrangian multipliers and confusing the intuition! Of course, in this case the method presented below continues to be applicable, but with constraint function $\tilde{g} = -g$ and the first order condition $\nabla f(x^*) = \lambda \nabla \tilde{g}(x^*)$.[29]

Now, let's turn to how to put this Lagrangian multiplier intuition to use in our context: if at a critical point, the *value cost of the constraint imposed on the objective function, equal to the Lagrangian multiplier*, is strictly larger zero at some candidate $x^* \in \text{dom}(f)$, then at x^*, what is constrained is our ability to *increase* the objective, and the candidate should be a local maximizer! Indeed, one can show formally that the following holds:

Theorem 64. ***(Lagrangian Multipliers and Type of Extremum)*** *Let $\mathcal{P}$ be an equality-constrained problem $\mathcal{P}$ with objective f and constraint function g (constraint: $g(x) = 0$), and suppose that $\mathcal{P}$ is equivalent to the inequality-constrained problem $\mathcal{P}^{ineq}$ with objective f and constraint function g (constraint: $g(x) \leq 0$). Then, if $x^* \in L_0(g)$ and $\lambda^* \in \mathbb{R}$ are such that $\nabla f(x^*) = \lambda^* \nabla g(x^*)$, $\nabla g(x^*) \neq \mathbf{0}$,*

- *if $\lambda^* > 0$, then x^* is a local maximizer of $\mathcal{P}$*
- *if $\lambda^* < 0$, then x^* is a local minimizer of $\mathcal{P}$*

Proof. Consider the directional derivative of f at x_0 with direction $z \neq \mathbf{0}$:

$$\left[\frac{d}{dt} f(x^* + tz)\right]\Big|_{t=0} = \nabla f(x^*)z \overset{\text{FOC}}{=} \lambda \nabla g(x^*)z = \lambda \left[\frac{d}{dt} g(x^* + tz)\right]\Big|_{t=0}.$$

[29] If you instead write the FOC as $\nabla f(x^*) = \lambda \nabla g(x^*)$, you get $\nabla f(x^*) = \lambda \cdot (-\nabla \tilde{g}(x^*)) = (-\lambda)\nabla \tilde{g}(x^*)$, introducing the sign error on the Lagrangian multiplier.

If z is a direction pointing to the interior of the constraint set in the inequality-constrained problem $C(\mathcal{P}^{ineq})$, $int(C(\mathcal{P}^{ineq})) = \{x \in \mathbb{R}^n : g(x) < 0\})$, then

$$\left[\frac{d}{dt}g(x^* + tz)\right]\Big|_{t=0} = \nabla g(x^*)z < 0,$$

so that

$$sgn\left(\left[\frac{d}{dt}f(x^* + tz)\right]\Big|_{t=0}\right) = sgn(\lambda) \cdot sgn(\nabla g(x^*)z) = (-1) \cdot sgn(\lambda).$$

Therefore, if $\lambda > 0$ ($\lambda < 0$), all x in a small neighborhood $N_\varepsilon(x^*) = B_\varepsilon(x^*) \cap int(C(\mathcal{P}^{ineq}))$, $\varepsilon > 0$, of x^* yield $f(x) < f(x^*)$ ($f(x) > f(x^*)$). Thus, for points on $L_0(g)$ around x^*, i.e. points $x \in B_\varepsilon(x^*) \cap L_0(g)$, it holds that $f(x) \leq f(x^*)$ ($f(x) \geq f(x^*)$).[30] Therefore, x^* is a constrained local maximizer (minimizer) of $\mathcal{P}$. □

For the multivariate case, an analogous variant of the theorem applies so long as you can re-write *all* inequality constraints as equality constraints ($g_i(x) \leq 0$ to $g_i(x) = 0$) and vice versa, in which case $\forall i \in \{1, \ldots, m\} : \lambda_i^* > 0$ is sufficient for a constrained local maximizer. As a note of caution, this condition classifies with certainty only critical points with $\lambda^* \neq 0$. Thus, points with $\lambda^* = 0$ are not classified, and must be kept as candidate solutions for comparison in the final step. Intuitively, $\lambda^* \neq 0$ tells us that the constraint does not matter, as relaxing it has no effect on the solution.[31] This scenario can occur in both minimization and maximization problems, so that $\lambda^* = 0$ gives you no information of the type of the extremizer.

This Lagrangian multiplier condition may be extremely helpful at times, as you can convince yourself of in the exercise problems and on the problem sets. The reason is that it offers an alternative, and in comparison to the Bordered Hessian criterion involving computation of multiple determinants for *each* candidate, fascinatingly simple way to determine the kind of local extremum a certain critical value of the Lagrangian corresponds to. The only "cost" this simplification comes at is that it applies only when there is equivalence in the equality- and inequality-constrained problem, which may be non-obvious to verify in some applications. Even if you can use it, you should be very careful to apply the trick correctly – one mess-up with the sign, and you're accidentally throwing away all candidates that you actually care about! To prevent this from happening, keeping in mind:

(1) Convince yourself thoroughly that the equality- and inequality-constrained problems are equivalent. Also make sure to have the correct form for the inequality-constrained problem, i.e. $g(x) \leq 0$. If you have $g(x) \geq 0$, multiply the constraint function by (-1) before proceeding.

(2) The proper way to write down the Lagrangian is with a **minus**: $\mathcal{L}(\lambda, x) = f(x) - \lambda g(x)$. Only like this, we get the FOC $\nabla f(x) = \lambda \nabla g(x)$, and you obtain the correct sign for λ.

For (2), if you use plus instead of minus, you're multiplying the constraint function by -1 and lose equivalence to the inequality-constrained problem! Thus, it may be advisable to simply proceed without the Lagrangian and start from the FOC $\nabla f(x^*) = \lambda \nabla g(x^*)$ directly. So, before using the Lagrange multiplier trick, ask yourself: *"Is the associated inequality-constrained problem equivalent? And am I using the correct FOC?"* If your answer is a definitive yes, you're good to go!

4.5 Conclusion

We have started by formally studying unconstrained optimization problems, which has allowed us to highlight these problems' key aspects and properties, to formally justify our first-and second order condition solution approach, and to make clear how we may generalize the respective insights to the case

[30] To see this, note that points $x \in B_\varepsilon(x^*) \cap L_0(g)$ can be written as the limit of a sequence $\{x_n^N\}_{n \in \mathbb{N}}$ over $N_\varepsilon(x^*)$, i.e. $x = \lim_{n \to \infty} x_n^N$. As weak inequalities are preserved under the limit, and since f is continuous, $f(x_n^N) < f(x^*)$ for all $n \in \mathbb{N}$ implies $f(x) \leq f(x^*)$.

[31] This statement applies only locally around x^*, it does not mean that the equality-constraint problem would have the same solution as an unconstrained problem with the same objective.

where optimization is subject to equality constraints. For notational simplicity, we have proceeded to a rigorous formal study of the problem with one equality constraint, where we outlined necessary and sufficient conditions for *interior* solutions. Thereafter, we generalized this approach to multiple constraints exploiting the structural similarity of the vector space $\mathbb{R}^n$ to the space of real numbers, and identified a three-step procedure to solving general equality-constrained problems using the Lagrangian method. Subsequently, we investigated problems that feature inequality constraints (such as budget constraints and non-negativity constraints) and outlined a step-by-step approach to reduce them to a standard Lagrangian problem. In case this reduction is not applicable, we may consult the Karush-Kuhn-Tucker conditions to find candidates for solutions, both from identifying interior points that satisfy optimality conditions and from a new singularity condition yielding "irregular" candidates in a similar spirit to before. It has emerged that thorough understanding of the Lagrangian method and especially the intuition for the Lagrangian multipliers suffices to understand the approach to most economic optimization problems.

4.6 Contents and Take-Aways

Chapter 4: Optimization discusses

- the formal representation of and basic concepts related to optimization problems (OPs)
- the step-by-step procedure to solve unconstrained OPs and its justification
- the step-by-step procedure to solve OPs with one equality constraint and its justification
- the previous methods' generalization to multiple equality constraints and inequality constraints
- approaches to re-writing and simplifying problems, and how to be computationally efficient

Someone with profound knowledge of the contents of this chapter should

- first and foremost: be able to solve unconstrained and constrained OPs
- be able to write down an OP formally and describe it using correct terminology (e.g. objective, choice variables, constraint types, etc.)
- know how $f|_S$, the restriction of a function f to a set S, is formally defined and how the concept is helpful in the optimization context
- be able to decscribe the objects $\max_{x \in C(\mathcal{P})} f(x)$ and $\arg\max_{x \in C(\mathcal{P})} f(x)$ and how they are related ($C(\mathcal{P})$ is the constraint set of the problem $\mathcal{P}$)
- be able to verbally describe how the first and second order necessary conditions for local extremizers come to be in the unconstrained problem
- know how to classify solution candidates using first and second order conditions
- be familiar with the univariate implicit function theorem and the intuition of how it helps in deriving first and second order conditions for problems with one equality constraint
- know where the Lagrangian function comes from, and how to write it down correctly
- be able to graphically illustrate economic level sets such as indifference "curves"
- know about the "value-cost" interpretation of the Lagrangian multipliers, and when and how it can be used as a simplistic sufficient condition for local extremizers
- be aware of the statement of the Karush-Kuhn-Tucker theorem

and be able to answer a number of related questions, including

- Why is the topic of mathematical optimization important for economists?
- What are maximizers and minimizers? How are they different from maxima and minima?
- What is the statement of the Weierstrass Extreme Value Theorem? What is its role in optimization?
- How does a "saddle point" of a bivariate function look like? Where does the label come from?
- How does convexity or concavity of the objective function help in unconstrained OPs?
- True or false: equality-constrained optimization can be viewed as optimization on the zero-level-set(s) of the constraint function(s).
- Which type of border solution candidate does equality-constrained optimization add relative to the unconstrained case?
- When and how can we re-write inequality-constrained problems as equality-constrained ones? Why would we be interested in doing so?
- How are "explicit function representations" of equality constraints helpful in facilitating an OP?
- What role does solution existence play in concrete applications, i.e. how does the issue relate to justifying our solution for the global extremizer of interest?

4.7 Recap Questions

Currently, no recap questions have been written for this chapter yet, later versions of the script may include some. In case you desire to solve some practice problems, please refer to the problem sets of earlier years.

5 Econometrics

The final chapter of this script gives a brief introduction into the realm of econometrics. Put simply, econometrics can be described as the study of methods that allow statistical analysis of economic issues. Such analyses are of central importance to the profession: no matter how sophisticated or convincing a theoretical model may be, the strongest proof of a theory is always its consistency with observations made in the real world. Therefore, economists need a set of tools allowing them to contrast their theoretical considerations against the real world. Conversely, econometric methods also allow for more exploratory analyses of empirical relationships, the results of which may themselves give rise to hypotheses that form the bases of interesting theoretical studies. Finally, independent of any economic theory, econometric methods can be used to assess the effectiveness of different political measures (*policy evaluation*), typically by treating the policy intervention as an exogenous shock to the economy, the causal effect of which is relatively easily identified through the use of appropriate methods.

The mathematics of econometrics are somewhat distinct from what we have seen thus far. It relies on new concepts, specifically random variables and their moments (expected values, variance, etc.), but at the same time, especially the concepts of differentiation, integration and optimization are also of central importance, which should facilitate our familiarization with econometrics. This chapter tries to serve as a refresher of/introduction to the very basics of econometrics, without going too deep into any specific direction. As such, it covers an introductory discussion of correlation and causation, the concept of random variables, and the linear regression model and its estimation.

5.1 Correlation does not mean or imply Causation

When engaging in an empirical analysis, our aim is usually to identify *causal* relationship, that is, we seek to arrive at conclusions such as *"X causes Y to increase"* or *"Productivity growth has slowed down because the rate of technology diffusion has decreased"*. In other sciences such as medicine, such statements are readily proved or disproved using controlled experiments. Put simply, to understand the effect of some treatment, you split a test population into two identical groups, apply the treatment to one group but not to the other, and observe how outcomes differ on average between the groups. While the outcomes will never be exactly identical across groups, we can use statistical methods that allow to quantify how likely it is to obtain the observed difference under the hypothesis that there is no treatment effect. If the likelihood is sufficiently small (usually: the difference of outcomes is sufficiently large), then we may claim to have identified a treatment effect.

In economics, however, things are rarely as simple. First, selecting two groups of identical economies/economic regions is already a vastly difficult task, as there are a multitude of economic characteristics along which regions may differ, including demographics, political orientation, trade integration, and many more. Especially if not all relevant characteristics are observable, defining two comparable groups may be difficult. Much more importantly, however, economic experiments would usually be *(i)* very costly, and *(ii)* potentially unethical. Considering the example of slower technology diffusion and lower productivity growth from before: if you artificially banned some firms from adopting technologies that could improve their performance, these firms may fall behind in competition, and you could cause, among others, worker layoffs, firm bankruptcies, weakened competition, and lower government income from revenue tax. Further, if your hypothesis is indeed true (and existing research seems to show that it is), you would weaken aggregate productivity growth, a central determinant of overall economic well-being.

Sometimes, researchers find themselves in the situation that nature takes over the role of the unethical interventionist, e.g. when natural disasters destroy a relevant share of economic infrastructure in some regions, but does not in other, neighboring and thus comparable regions. Here, *quasi-experimental*

methods can be applied to obtain causal conclusions. The more common scenario, however, is that economists use methods of correlation analysis, and need to argue that the identified correlational relationships are indeed causal. Verbally, correlation refers to the co-movement of certain quantities. A *perfect correlation* is when quantities co-move in a one-to-one fashion without any deviation; a positive perfect correlation would, for instance, observed for the sale of left-hand and right-hand gloves, if gloves are sold in pairs, and a perfect negative correlation can be assumed between the inventory stock and the quantity of sales, so long as there is no restocking. Less obvious real-world correlations are rarely perfect. Coming back to medical treatments, some patients may not respond as much to a treatment as others, but as long as the treatment is effective in some way, we should observe a positive correlation between the intensity of the treatment and the health outcome.

For economists that mainly analyze correlations, a critical conceptual circumstance is the following: *correlation does not imply causation*, or put differently, **correlation is not a sufficient condition for causation**. Two simple examples help illustrate this point.

1. Over the past decades, in the US, energy consumption has increased while the marriage rate (share of unmarried persons in legal marrying age getting married in a given year) has decreased.
2. Over a given year, in months with higher ice cream sales, the number of shark attacks is higher.

Example 1 describes a negative correlation between energy consumption and the marriage rate, while example 2 describes a positive correlation between ice cream sales and shark attacks over time. So, does energy consumption decrease people's willingness to get married, and do ice cream sales cause shark attacks? Clearly not. While the two quantities of example 1 are likely completely independent, in example 2, the quantities have a *common determinant*, but are not directly related. This common determinant is seasonality, or respectively the weather. On warm, sunny days, more people buy ice cream, and go to the beach swimming where they can be attacked by sharks. Therefore, the weather directly determines ice cream sales, and indirectly determines shark attacks.

A further limitation of correlations is that they only describe co-movement, but they also do not allow to infer on the direction of a potential causality between quantities. For instance, if employment and GDP are positively correlated, even if we are told there is a causal relationship between the two quantities, we don't know whether higher GDP causes employment to increase, or whether increasing employment causes higher GDP.[32] Such issues, unfortunately, can ex ante also be a concern in the more sophisticated models that we use. If we want to study how x affects y, but y may also affect x, we call this issue *reverse causality*.

The following example of a fictional economy describes how it may potentially be very harmful to base economic actions on insights from correlations. Consider an economy of a "saturated labor market" where strong population growth leads to only modest increases in the employment level:

period	total pop.	L	UR	Y	ΔY
0	11	10	1/11	100	-
1	14	12	1/7	140	1/6
2	20	15	1/4	200	1/5
3	30	20	1/3	300	1/4

For simplicity, production Y is assumed to follow $Y = 10L$, where L is employment. Suppose we were interested in understanding the effect of the unemployment rate UR on GDP growth ΔY. In this example, the higher the unemployment rate, the higher GDP growth. If a policy maker would infer causality from this correlation, they could conclude that forcing workers into unemployment could

[32] Crucially, as correlations may be in part driven by forces entirely unrelated to the causal relationship, it could even be the case that higher GDP *causes* lower employment, despite the quantities being positively correlated. That is, a positive correlation does not preclude a negative causal relationship, and vice versa, see also the example below.

boost GDP growth, possibly a desirable policy target. However, in the example, the unemployment rate and GDP growth are jointly determined by the total population (or population growth) and the induced effect on employment L, similar to the example of the seasonality origin of the correlation between ice cream sales and shark attacks. Indeed, if you increased the unemployment rate *at a given population level, i.e. holding this quantity constant*, you would *reduce* employment, and thus Y and ΔY, so that the unemployment rate has a negative effect on GDP growth, despite the positive correlation.

In this example, the true effect of the unemployment rate on the outcome of interest, GDP growth, could be uncovered from considering the relationship holding constant a third variable that drives the correlation but is unrelated to the causal relationship. This is precisely the idea of the methods we use in practice: they allow to "hold constant", or *control for* a set of observable quantities, and thereby enable us to zoom in on the *residual relationship of quantities of interest unexplained by controlled quantities*. By doing so, we are able to isolate the "direct" correlation between the two variables that abstracts from indirect dependence through third variables, and if we account for all such third variables, this direct correlation can be interpreted causally. The practical issue then is to know which third variables to account for, what to do if some of them are unobserved and/or unmeasurable, and how to identify the direction of the causality, that is not addressed by the mere correlation. To this end, a common tool, especially in macroeconometrics where we work with data that entail a time dimension, is to study dynamic relationships, i.e. the correlation between x_{t-1} and y_t, exploiting that causation works in only one direction through time.

To wrap up, you should take away that correlations are not always useful in arguing for causal relationships, as they can occur entirely at random (e.g. energy consumption and marriage rate), or due to joint dependence on a third variable (e.g. ice cream sales and shark attacks). Further, the sign of a correlation is not always informative about the sign of the causal relationship even if there is one (e.g. unemployment rate and GDP growth) if third variables play a role. Even if a correlation does describe a causal relationship (i.e., there are no third variables, or all relevant third variables are "held constant" by an appropriate model), it does not tell us the direction of the causal relationship, which needs to be argued for separately.

5.2 Probability Spaces and Random Variables

Before moving to our methods of empirical analysis, as per usual, a key step is to have all ingredients and our vocabulary straight. In the world of statistics, in which econometrics is a sub-discipline, the foundation of everything is probability theory, or stochastics. We will touch on this issue very briefly, focusing on the core concepts relevant to economists. The concepts we discuss here, especially random variables and their moments, are also key in many macroeconomic models that feature some uncertainty, and where agents have to form expectations about the future and adapt their behavior accordingly.

Usually, when engaging in empirical analysis, we try to learn something about the future, either directly or indirectly. A direct empirical question related to the future could be: "Will increasing university budgets augment GDP growth, and if so, by how much?" More indirectly, one may ask "Has increasing university budgets augmented GDP growth over a recent period, and if so, by how much?" The relevance of this question, however, also lies in the future: by looking at the (recent) past, we want to know whether one could expect similar relationships also in the future. Generally, we can learn about the future from the past if conditions remain similar enough. In this case, we exploit the circumstance that the events in the past used to be the future at some point, and outcomes were yet to materialize. In other words, as the outcomes were associated with some uncertainty, they were, to some degree, random (or: unpredictable), just as the future's outcomes are somewhat random from today's perspective. So, if conditions do not change much, we should be able to learn from the realizations of uncertainty in the past about what to *expect* for uncertainty in the future, and we can also quantify how accurate these

expectations are, based on how accurate they would have been for the period on which we already have data.

This describes the intuition of why probability theory is relevant to economics, and especially econometrics: it allows to model the future as an environment associated with uncertainty, and we are able to describe this uncertainty and form expectations based on the assumption that the past was characterized by the same model, with the only difference that the uncertainty has already materialized. Therefore, a crucial thing to keep in mind is that all predictions for the future obtained from econometric analysis, including the effectiveness of policy measures and the identification of structural economic relationships, inform about the future only to the degree that framework conditions do not change significantly relative to the period covered by analyzed data. This is why well-published empirical papers tend to use data that is as recent as possible, and it is also why economic crises such as the global financial crisis around 2008-2012, but also more recently the Covid pandemic and Russia's war on Ukraine and the ensuing economic repercussions, constitute considerable challenges for economists: these times of very unique economic conditions are unprecedented in history, and analyzing data from the past will only be partially able to guide effective policy in response to these crises.

5.2.1 Probability Space and Probability Measure

Definition 86. ***(Probability Space)*** *A probability space $\mathcal{P}$ is a triple $\mathcal{P} := (\Omega, \mathcal{A}, P)$, where*

- *Ω is the **sample space**, the set of all possible outcomes,*
- *$\mathcal{A}$ is the **event space**, the set of all possible events, and*
- *$P : \mathcal{A} \to [0,1]$ is the probability measure that assigns events $A \in \mathcal{A}$ a probability.*

Definition 86 gives the definition of the probability space. It is best illustrated at a simple example: consider the case of rolling a regular dice. Before the dice is rolled, the set of possible outcomes is $\Omega = \{1,2,3,4,5,6\}$. One example of an event that could occur is that a specific number is rolled, e.g. $A = \{2\}$ or $A = \{5\}$. However, it could also be that an even number is rolled, $A = \{2,4,6\}$, or that a 6 is *not* rolled, $A = \{1,2,3,4,5\}$. So, any combination of elements in Ω may constitute an event, and more generally, also beyond this simple example, we tend to consider $\mathcal{A} = \mathcal{P}(\Omega)$, i.e. the power set of the sample space, as the set of possible events.

It remains to define the probability measure $P : \mathcal{A} \to [0,1]$ that assigns events $A \in \mathcal{A}$ a probability between 0 and 1. Clearly, the measure should assign $A = \Omega$, the event that any number between 1 and 6 is rolled, a probability of 1. Further, for any specific number, i.e. $A = \{a\}$, $a \in \Omega$, the probability should be 1/6. More generally, as the dice is fair (i.e., all numbers have the same probability), we can define $P(A) = |A|/6$, where $|\cdot|$ is the cardinality of the set: the probability of an event is the amount of different numbers it allows, divided by the amount of all possible numbers.

Definition 87. ***(Event Space, σ-Algebra)*** *Consider a sample space Ω. Then, a set $\mathcal{A}$ is called an event space, or σ-algebra if*

- *$\varnothing \in A$ and $\Omega \in A$,*
- *for any $A \in \mathcal{A}$, for $A^c := \Omega \backslash A$, it holds that $A^c \in A$,*
- *for any $A, B \in A$, it holds that $A \cup B \in A$.*

Definition 88. ***(Probability Measure)*** *Consider a sample space Ω and an event space $\mathcal{A}$. Then, a function $P : \mathcal{A} \to [0,1]$ is called a probability measure on $(\Omega, \mathcal{A})$ if*

- *$P(\Omega) = 1$ and $P(\varnothing) = 0$ is the **sample space**, the set of all possible outcomes,*
- *if $A, B \in \mathcal{A}$ are disjoint, i.e. $A \cap B = \varnothing$, then $P(A \cup B) = P(A) + P(B)$.*

While Ω is always the "base set" of possible outcomes and requires little further characterization to allow generalization to richer stochastic environments, the other two components of the probability space do. These characterizations are given in Definitions 87 and 88. Verbally, the event space requires that we consider that "anything can happen" ($\Omega' in \mathcal{A}$), that for any event A, the opposite A^c can happen, and that for any two events A and B that can happen, the event that either one of them can happen, $A \cup B$, is also accounted for. The simplest way to achieve this is $\mathcal{A} = \mathcal{P}(\Omega)$, which is the default case for everything to follow.

The definition of the probability measure may seem surprisingly general. Next to the intuitive characteristic of assuming only numbers between 0 and 1, we only require that "anything will happen" with probability 1 ($P(\Omega) = 1$) while "nothing happens" with probability zero ($P(\varnothing) = 0$), and that further, the probability that either one of two distinct events occurs is the sum of their probabilities. To see the latter, imagine the dice roll with $A = \{1, 2\}$ and $B = \{3\}$ for which $P(A) = 2/6$, $P(B) = 1/6$ and $P(A \cup B) = P(\{1, 2, 3\}) = 3/6$. The focus on disjoint sets is crucial: if some elements $\omega \in \Omega$ would realize both A and B, for instance modifying $B = \{2, 3\}$, then additivity need not be assumed. Note that by the second property, it is usually sufficient to know the probability of events referring to individual elements of Ω, as any event A can be decomposed into a disjoint union of such events: $A = \bigcup_{i \in I} \{\omega_i\}$, $\omega_i \in \Omega$. For notational simplicity, we define $P(\omega) := P(\{\omega\})$ for $\omega \in \Omega$.

The reason that the definition of a probability measure is so broad is that we want to allow ourselves to adapt the concept flexibly to concrete contexts. While we are allowed to call many functions $\mathcal{P}(\{1, 2, 3, 4, 5, 6\}) \to [0, 1]$ a probability measure that describes the experiment of a dice roll in accordance with Definition 88, as we have seen above, only one of them suits the specific context of rolling a fair dice. Similar to the metric concept, the definition gives us a range of functions that we may attribute the label of a probability measure, and our exercise in concrete contexts is to find the most useful one. As an exercise, you can try to define the probability measure that characterizes rolling a rigged dice for which it is twice as likely to roll a 6 than it is to roll a 1, but all other numbers remain as likely as with a fair dice; the footnote gives the answer.[33]

5.2.2 Random Variable

In most practical scenarios, we care less about the realized events themselves, but rather about (numeric) outcomes that they imply. For instance, when you plan on going to the beach, you do not want to know all the meteorological conditions $\omega \in \Omega$ in the space of possible conditions, but you care about the functions $x(\omega) = \mathbb{1}[\omega \text{ leads to rain}]$ and about $c(\omega)$ that measures the temperature in degrees Celsius implied by the conditions ω. In other words, you care about two random variables: variables X on the real line that are determined by (possibly much richer) outcomes in the event set Ω of a probability space. As the beach example illustrates, the appeal of random variables is two-fold: on one hand, they can be defined to address directly what we are interested in, and they allow to reduce dimensionality and possibly reduce the prediction problem. To see the latter point, if we build a model to forecast exact meteorological conditions, this model will need a lot of data and may even then perform poorly, and if different conditions imply a similar temperature, we may be better off forecasting the temperature directly.

Definition 89. ***(Random Variable, Random Vector)*** *Consider a probability space $(\Omega, \mathcal{A}, P)$. Then, a function $x : \Omega \mapsto \mathbb{R}$ is called a random variable, and a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ where x_i, $i \in \{1, \ldots, n\}$ are random variables, is called a random vector.*

Definition 89 gives a broad definition of the random variable concept. Those with a stronger background in statistics or econometrics may excuse that it is very vague and simplistic; the concept is math-

[33] We know that $P(\omega) = 1/6$ for any $\omega \in \Omega \backslash \{1, 6\}$, so that $P(1) + P(6) = 2/6$. By $P(6) = 2P(1)$, we get $3P(1) = 2/6$, or $P(1) = 1/9$, and $P(6) = 2P(1) = 2/9$.

ematically not straightforward to define, and for our purposes, it suffices to focus on the characteristics stated in the definition.

Continuing the example of the fair dice roll, let's consider a game where I give you 100 Euros if you roll a 6, but you get nothing otherwise. Can you define the function $\pi(\omega)$ that describes your profit from this game, assuming for now I let you play for free? See the footnote for the answer.[34]

A key concept related to random variables X is their *distribution*, i.e. the description of probabilities characterizing the possible realizations of X. Our leading example of the dice roll, and the game where you can win 100 Euros, is an example of a discrete random variable, characterized by a discrete probability distribution. For such random variables, there are a finite number of realizations (in our case: two, either 0 Euros or 100 Euros), and every realization may occur with a strictly positive probability (5/6 vs. 1/6). The other case is a continuous probability distribution, with an infinite number of realizations that can not be attributed positive probability: $P(X = x) = 0$ for any $x \in \mathbb{R}$. One such example is a different game of dice: say I generate a random real number X between 0 and 25, and then I let you roll a dice with result D, giving you a profit $\pi = D \cdot X$, so that the result of the dice roll is multiplied by this random number. The first stage of this game, i.e. randomly choosing a number from $[0, 25]$, is characterized by the probability distribution $P(X \in [a, b]) = (a - b)/25$ for $a, b \in [0, 25]$ with $a > b$. More commonly, we set b as the minimal possible value (sometimes $b = -\infty$) and characterize the distribution as $P(X \leq a) = F(a)$, in our case $P(X \leq a) = a/25$. The **probability density function** is the derivative of this function: $f(a) = F'(a)$, in our case $f(a) = \mathbb{1}[x \in [0, 25]] \cdot 1/25$.

Definition 90. ***(Expected Value)*** *Consider a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $X : \Omega \mapsto \mathbb{R}$ on this space, with probability density function f_X. Then, the expected value $\mathbb{E}[X]$ of X is defined as the integral*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx,$$

and for a transformation $g(X)$ of X,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Definition 91. ***(Variance, Standard Deviation)*** *Consider a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $X : \Omega \mapsto \mathbb{R}$ on this space. Then, the variance of X is defined as $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$, and its square root is called the standard deviation of X, denoted $sd(X) = \sqrt{Var[X]}$. The variance can generally be computed as $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.*

Before moving to economic models, we need to introduce two more key concepts, given by Definitions 90 and 91. The expected value gives us the best possible prediction of X given its distribution. Crucially, you always use the probabilities referring to X, even if you consider a transformation $g(X)$ of X. For discrete distributions, when $\{x_i\}_{i \in I}$ is the set of values that X can take with positive probability, the expression simplifies to $\mathbb{E}[X] = \sum_{i \in I} P(X = x_i) \cdot x_i$. This shows the intuition of the expected value: you expect the value x_i with probability $p_i = P(X = x_i)$, and you expectation is then simply the sum of all these terms. While the expected value gives you a "mean" of what to expect for X, it is silent on how good this prediction is.[35] This is the purpose of the variance, measures the expected degree of deviation of X from the expected value. As the variance is based on a square term, we usually re-normalize it by taking the square root when talking about deviations from the expected value, which we call the *standard deviation*.

To close this section, you may test your understanding of the concepts thus far by answering two questions: First, if you were neutral to risk (i.e., you only care about your *expected* profit), would you

[34] The function is $\pi(\omega) = 0$ for $\omega \in \{1, 2, 3, 4, 5\}$ and $\pi(\omega) = 100$ for $\omega = 6$.

[35] Note that the expected value need not be a possible realization of X. In the case of the dice roll, you can convince yourself that $E[D] = 3.5$, which is not a number that the dice can show.

play the game where you can win 100 Euros if you roll a 6 if you needed to buy in for 15 Euros (also think about whether you would spontaneously agree to this game without doing the math)? And second, what game do you prefer in terms of expectations, the one where you get a 100 Euros for a 6, or the one where you get $D \cdot X$ Euros, and X is drawn randomly from $[0, 25]$? And which of these games has the more certain/less volatile outcome? For the latter game, you may use that the dice role is independent of the randomly drawn number X, so that $E[D^k X^k] = E[D^k]E[X^k]$ for any $k \in \mathbb{N}$. For the solution, you can go to the next page.

The expected value of game 1 is

$$\mathbb{E}[\pi] = \sum_{i \in I} P(X = x_i) \cdot x_i = \frac{5}{6} \cdot 0 + \frac{1}{6} \cdot 100 = \frac{100}{6} \approx 16.67.$$

So you should want to pay no more than 16.67 Euros to play this game. If I offer this to you for a buy-in of 15 Euros, if you are risk-neutral, you should do it.

On a side note, in practice, most people would probably not take this game for 15 Euros, as people are risk-averse (they don't assess games based on expected values) and they dislike losses more than they like gains. Behavioral economists argue that with these conditions, it is rational for people not to play such a game. Things were to change only if we played this game repeatedly, say a 1000 times, where you would expect to eventually be close to the expected profit on average.[36]

For game 2, the expected value is

$$\begin{aligned}
\mathbb{E}[D \cdot X] = \mathbb{E}[D]\mathbb{E}[X] &= 3.5 \cdot \int_{-\infty}^{\infty} x \mathbb{1}[x \in [0, 25]] \cdot \frac{1}{25} dx \\
&= 3.5 \cdot \frac{1}{25} \int_0^{25} x dx \\
&= 3.5 \cdot \frac{1}{25} [\frac{1}{2} x^2]_{x=0}^{x=25} \\
&= 3.5 \cdot \frac{625}{50} = 3.5 \cdot 12.5 = 43.75
\end{aligned}$$

Therefore, this game is much preferable to game 1 in terms of the expected profit. To address the volatility of outcomes, we need to compute the variance. Since $Var[X] = E[X^2] - E[X]^2$, it remains to compute the expectations of the squared terms. First, for game 1,

$$\mathbb{E}[\pi^2] = \sum_{i \in I} P(X = x_i) \cdot x_i^2 = \frac{5}{6} \cdot 0 + \frac{1}{6} \cdot 100^2 = \frac{10000}{6},$$

so that

$$Var[\pi] = \mathbb{E}[\pi^2] - \mathbb{E}[\pi]^2 = \frac{10000}{6} - \left(\frac{100}{6}\right)^2 = 10000\left(\frac{1}{6} - \frac{1}{36}\right) = \frac{10000 \cdot 5}{36} \approx 1388.89$$

with $sd(\pi) = \sqrt{1388.89} \approx 37.27$.

Next,

$$\mathbb{E}[D^2] = \sum_{d=1}^{6} P(D = d) \cdot d^2 = 1/6(1 + 4 + 9 + 16 + 25 + 36) = 91/6$$

and in analogy to above,

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \mathbb{1}[x \in [0, 25]] \cdot \frac{1}{25} dx = \frac{1}{25} [\frac{1}{3} x^3]_{x=0}^{x=25} = \frac{25^3}{75} = \frac{625}{3}$$

so that

$$Var(DX) = \mathbb{E}[D^2]\mathbb{E}[X^2] - \mathbb{E}[D \cdot X]^2 = \frac{91}{6}\frac{625}{3} - (43.75)^2 \approx 1245.66$$

with $sd(DX) = \sqrt{1245.66} \approx 35.29$.

Therefore, not only is game 2 better in terms of the expected value, but it is also slightly less un-

[36] This is indeed a central result of probability theory, the law of large numbers: $P(|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_i]| > \varepsilon) \to 0$ as $n \to \infty$ for any fixed $\varepsilon > 0$, if the realizations of X_i are independent and follow the same distribution (independent and identically distributed, i.i.d.). Verbally, by repeating a game/an experiment often enough, the average of realizations approaches the expected value with arbitrary precision.

certain/volatile. The more risk-averse an agent is, the more they value certainty, and the uncertainty comparison could therefore be an additional practical reason to opt for game 2 rather than game 1.

5.3 Linear Regression Model

Having laid the foundation of stochastic concepts, this final section proceeds to discuss basic econometric models used in empirical practice. Recalling the introduction to this chapter, these models aim to do the first step from correlations to causality by allowing to "control"/"hold constant" third confounding variables. We will thoroughly keep this intuition in mind when discussing their definition and estimation.

5.3.1 Specification

One final concept that we need to understand the relationship *between* random variables, as we intend to model in the following, is the one of a conditional expectation - the value we expect a random variable Y to take knowing the value of X. This quantity we call $\mathbb{E}[Y|X]$, the expected value of Y given X. To understand this concept in simple terms, recall the game 2 in dice roll experiment: in the first step, we drew a random number from $[0, 25]$ with a *uniform distribution*, i.e. all parts of equal length within $[0, 25]$ were equally likely. For the purpose of notational consistency, let's for now call this variable Y rather than X, as we did before. The expected value, as calculated earlier, was $\mathbb{E}[Y] = 12.5 = 25/2$. More generally, if we draw randomly with uniform distribution from $[0, a]$, then the expected value is $\mathbb{E}[Y] = a/2$. But what if we draw the upper bound at random in a first step? Then things become a lot more complicated, and depend on the distribution from which we draw the upper bound a. Still, if X is the random variable that gives us the upper bound a, we can say that $\mathbb{E}[Y|X = a] = a/2$. Note that $a \in \mathbb{R}$ is still a concrete realization of X. However, we do not need to condition on concrete realizations: no matter what X will be, given that we know X, we know that the expected value of Y will be $X/2$. Therefore, we write $\mathbb{E}[Y|X] = X/2$. This object, now, is a random variable that is a function of X: $\mathbb{E}[Y|X] = g(X)$ with $g(x) = x/2$.

This simple example illustrates an important point: when we condition on concrete values of X, the conditional expectation $\mathbb{E}[Y|X = x]$ will be a real number, as the "unconditional" expectation $\mathbb{E}[Y]$ is, too.[37] More generally, the conditional expectation of Y given X is a random variable itself, and functionally depends on X: $\mathbb{E}[Y|X] = g(X)$.

Now it is time to put these concepts to use. Recall that initially, we set out to come up with a model that allows to causally address how a variable X affects another variable Y. Since we want to speak to the future where the realizations of X and Y are not known, we treat these quantities as random variables, and exploit that we have *data* that records existing realizations of these random variables in the past. The simplest way to describe a relationship between Y and X using the conditional expectation is to define $e := Y - \mathbb{E}[Y|X]$, where e is the deviation of Y from what we would expect for it given X, or the *error* made from predicting Y using X. Re-arranging terms, one obtains

$$Y = \mathbb{E}[Y|X] + e$$

We know that generally, $\mathbb{E}[Y|X] = g(X)$ for an unknown function $g(\cdot)$. While we do not know $g(\cdot)$ itself, we do know Taylor's theorem, which tells us that a polynomial approximation to $g(\cdot)$ may be "good". The simple, linear model assumes that a first order approximation fares sufficiently well.[38] Under this assumption, $g(x) = \beta_0 + \beta x$, and the equation becomes

[37] Note that this statement refers to the type of object; in general, it will *not* be the case that $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$.

[38] There are a number of tests for linearity, which are however beyond the purpose of this script.

$$Y = \beta_0 + \beta X + e$$

In this model, if we observe that X is higher by $\Delta x > 0$ units, we expect Y to be higher by $\beta \cdot \Delta x$ units. Note that this does not assert anything about causality: a non-zero coefficient β simply tells us that X is useful in predicting Y, similar to the way ice cream sales are useful in predicting shark attacks. To move closer to a causal interpretation, let us consider a random vector of third variables Z of length k that may describe an indirect relationship between X and Y, such as seasonality in the ice cream/sharks example, or the population and employment level in the unemployment rate/GDP growth example. Similar to before, we can write

$$Y = \mathbb{E}[Y|X,Z] + e \overset{\text{Ass. of linearity}}{=} \beta_0 + \beta \cdot (X,Z) + e = \beta_0 + \beta^x X + \beta_1^z Z_1 + \ldots + \beta_k^z Z_k + e$$

Then, β^x measures the predictive potential of X for Y conditional on Z, i.e. holding constant Z. To see this more explicitly, consider the equation in concrete realizations,

$$Y = \mathbb{E}[Y|X = x, Z = z] + e = \beta_0 + \beta^x x + \beta_1^z z_1 + \ldots + \beta_k^z z_k + e$$

where the marginal effect of increasing the realization x of X on Y can be obtained as[39]

$$\frac{\partial Y}{\partial x} = \frac{\partial(\beta_0 + \beta^x x + \beta_1^z z_1 + \ldots + \beta_k^z z_k + e)}{\partial x} = \beta^x.$$

In this view, β^x is the effect that X has on Y, holding the *control variables* $Z_1, \ldots, Z_k$ constant at the levels $z_1, \ldots, z_k$. By the linear nature of the model, the levels at which these variables are held constant do not affect the marginal effect of X on Y. Therefore, coming back to the example of unemployment rate and GDP growth, an appropriate empirical model is

$$\text{GDP growth} = \beta_0 + \beta^x UR + \beta_1^z POP + \beta_2^z L + e.$$

In this model, β^x describes the relationship between UR and GDP growth, holding constant the levels of population and employment. Still, the sign and magnitude of β^x only tells you how useful UR is in predicting GDP growth at constant population and employment, and this relationship could, in practice, also arise due to further *omitted* variables that are not yet included in the model, or because GDP growth affects the unemployment rate. This aspect is discussed further in the last section of this chapter.

5.3.2 Estimation

Assume that you have arrived at a model that you think is useful in describing the relationship of interest, of the form

$$Y = \beta_0 + \beta^x X + \beta_1^z Z_1 + \ldots + \beta_k^z Z_k + e. \tag{23}$$

This model will tell you something interesting only if you manage to know/get a plausible estimate of the coefficient of interest, β^x. For what follows, we return to the simpler case of only one random variable on the right hand side of the model, but the methods easily extend (with some slight complexification of the algebra) to the multivariate case. So, for now, we return to

$$Y = \beta_0 + \beta^x X + e. \tag{24}$$

[39] Formally, one would have to argue that $\frac{\partial e}{\partial x} = 0$, but this detail is considered too complex for the purpose of this script.

Assume that you observe n units of data on Y, $\{x_i, y_i\}_{i=1}^n$. The intuition behind the approach to estimating the coefficients β_0 and β^x is relatively simple. We know that the expected value of Y given X, $\mathbb{E}[Y|X]$, which we assume to be linear, i.e. $\mathbb{E}[Y|X] = \beta_0 + \beta^x X$, is the best prediction that can be obtained for Y given X. Therefore, we when calculating the estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}^x)' \in \mathbb{R}^2$ of $\beta = (\beta_0, \beta^x) \in \mathbb{R}^2$ from the data, we proceed to choose $\hat{\beta}$ such that it gives the best linear prediction of the data points $\{y_i\}_{i\in I}$ given $\{x_i\}_{i\in I}$. If we slightly rephrase the issue as optimizing the quality of prediction, things begin to sound very familiar. Indeed, the only part that we need yet to specify is what we mean by "best" prediction. As per the economist's preference for the Euclidean space (recall that the Euclidean norm captures the direct distance), it seems natural to minimize the distance of the vectors $y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ of data points for Y and $\mathbf{X}\beta$, where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

is the *data matrix* for the equation's right hand side (RHS) that stacks the observations of RHS variables in its columns. To make sure that you understand the structure of this matrix, think about how it would look like if you had the more general model of equation (23) with $k = 1$, i.e. only one control variable $Z_1 = Z$. See the footnote for the answer.[40]

With this in mind, we can express the optimization problem as:

$$\min_{b\in\mathbb{R}^2} \|y - \mathbf{X}b\|_2 \qquad \text{or} \qquad \min_{b\in\mathbb{R}^2} \sqrt{\sum_{i=1}^n (y_i - b_0 - b^x x_i)^2}.$$

With the tools the previous chapters have equipped us with, solving this should be a relatively simple task: there are no constraints, and the objective function entails a polynomial of the solution. One source of complication could be the square root, but we can easily circumvent it by solving the equivalent problem

$$\min_{b\in\mathbb{R}^2} \sum_{i=1}^n (y_i - b_0 - b^x x_i)^2 \tag{25}$$

This formulation gives the most common estimators of the linear model their name: *least-squares estimators*. To see this name more intuitively, note that $y_i - b_0 - b^x x_i =: \hat{e}_i(b)$ is the error in predicting y_i using the coefficient vector $b = (b_0, b^x)'$ for prediction by x_i, which we also call the *residual*. Therefore, the solution $\hat{\beta} = \arg\min_{b\in\mathbb{R}^2} \sum_{i=1}^n (y_i - b_0 - b^x x_i)^2$, provided that it exists, minimizes the *residual sum of squares* $RSS(b) = \sum_{i=1}^n \hat{e}_i(b)^2$, i.e. the sum of squared prediction errors.

We are now set up to derive the (hopefully unique) solution for the estimator $\hat{\beta} = \arg\min_{b\in\mathbb{R}^2} \sum_{i=1}^n \hat{e}_i(b)^2$. To test your understanding of Chapter 4, you can try to derive it yourself before reading on. You may use without proof two helpful properties of differential calculus with matrices that are investigated on the problem sets of the course:

1. For $A \in \mathbb{R}^{n\times m}$ and $x \in \mathbb{R}^m$, $\frac{d}{dx} Ax = A$, and
2. For $A \in \mathbb{R}^{m\times m}$ and $x \in \mathbb{R}^m$, $\frac{d}{dx} x'Ax = x'(A + A')$

[40] The matrix with one control variable is

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \\ 1 & x_n & z_n \end{pmatrix}$$

where $\{z_i\}_{i\in I}$ are the observations for the random variable Z. The vector β would in this case be $\beta = (\beta_0, \beta^x, \beta^z)'$.

Further, it may be helpful to recall the discussion of positive definiteness and the rank; you may assume that $\mathbf{X}$ has full column rank. To help you get started, things remain more tractable if you stick with the vector notation, and start from the problem

$$\min_{b\in\mathbb{R}^2}(y-\mathbf{X}b)'(y-\mathbf{X}b).$$

If you have tried to solve it, or you want to read on directly, let's get to the solution of this problem. The first step is to multiply out the objective. Doing so, the problem becomes

$$\min_{b\in\mathbb{R}^2} y'y - y'\mathbf{X}b - b'\mathbf{X}'y + b'\mathbf{X}'\mathbf{X}b.$$

A first useful observation is that $b'\mathbf{X}'y$ is a scalar, so that $b'\mathbf{X}'y = (b'\mathbf{X}'y)' = y'\mathbf{X}b$, and the problem simplifies to

$$\min_{b\in\mathbb{R}^2} y'y - 2y'\mathbf{X}b + b'\mathbf{X}'\mathbf{X}b.$$

In obtaining the first order condition, the two properties of differential calculus with matrices mentioned above are very useful. Applying them, we know that

$$\frac{d}{db}y'\mathbf{X}b = y'\mathbf{X} \qquad \text{and} \qquad \frac{d}{db}b'\mathbf{X}'\mathbf{X}b = b'(\mathbf{X}'\mathbf{X} + (\mathbf{X}'\mathbf{X})') = 2b'\mathbf{X}'\mathbf{X}.$$

Putting these together, the first order condition is

$$\mathbf{0}' = \frac{d}{db}(y'y - 2y'\mathbf{X}b + b'\mathbf{X}'\mathbf{X}b) = -2y'\mathbf{X} + 2b'\mathbf{X}'\mathbf{X}$$

so that the solution $\hat{\beta}$ must satisfy

$$y'\mathbf{X} = \hat{\beta}'\mathbf{X}'\mathbf{X} \qquad \text{or equivalently} \qquad \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'y.$$

While it looks like we are almost there, there are two things that we need to address: *(i)* can we invert $\mathbf{X}'\mathbf{X}$ to obtain a unique solution? *(ii)* If we can, does this solution constitute a (strict) local minimum? (And is this local minimum global?)

We know that matrices of the form $A'A$ for $A \in \mathbb{R}^{n\times m}$ are always positive semi-definite: for $v \in \mathbb{R}^m$, $v'A'Av = (Av)'Av = w'w = \sum_{i=1}^m w_i^2 \geq 0$ for $w = Av \in \mathbb{R}^m$. Now, recalling the discussion of rank and definiteness after Corollary 5, this matrix is indeed positive definite if for any $v \neq \mathbf{0}$, $Av \neq \mathbf{0}$, which is the case if A has full column rank, i.e. no column of A can be written as a linear combination of the remaining columns. In our case, $A = \mathbf{X}$ has just two columns, and one of them is the vector $\mathbf{1} = (1,1,\ldots,1)'$ containing only ones. Therefore, the matrix does *not* have full column rank if there exists $\lambda \in \mathbb{R}$ such that $x_i = \lambda\cdot 1 = \lambda$ for all $i \in \{1,\ldots,n\}$, in which case the second column of $\mathbf{X}$, $x = (x_1, x_2, \ldots, x_n)'$, can be expressed as $x = \lambda\cdot 1$.

In less abstract terms, this occurs if the random variable X exhibits no variation across the observations in our data, i.e. for every observed unit i, we observe the same realization $x_i = \lambda$. In this case, we would say that X is *collinear to the constant*. However, in practice, the variables we usually consider do vary across observations, and the issue does not occur.[41] Therefore, a key assumption in least-squares estimation is that $\mathbf{X}$ has full column rank, or that the RHS random variables (including the "constant" that is always equal to one) are not linearly dependent. This assumption is usually met unless one adds

[41] In models with more RHS variables, the issue can be less direct to spot. If you have vectors x and z of observations for random variables X and Z, respectively, then collinearity occurs if $\mathbf{1} = \lambda x + \mu z$ for some $\lambda, \mu \in \mathbb{R}$. A simple example is $X = \mathbb{1}[i \text{ is older than 50 years old}]$ and $Z = \mathbb{1}[i \text{ is younger than 50 years old}]$ when units i refer to persons, another one is $X = age$ and $Z =$ years until 100th birthday, where $Z = 100 - age$ and $1 = 1/100X + 1/100Z$.

variables that essentially give the same information. In practice, statistical softwares will usually give you an error message if this condition is violated in the data you analyze.

Returning to the optimization problem, if the observations for X, $\{x_i\}_{i\in\{1,\dots,n\}}$ vary across individuals, then the unique solution is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

and it constitutes a strict local minimum by positive definiteness of $\mathbf{X}'\mathbf{X}'$, the second derivative of the objective function. This local solution is also global, as for any $b \in \mathbb{R}^2$, $\lim_{\lambda\to\infty}(y'y - 2y'\mathbf{X}(\lambda b) + (\lambda b)'\mathbf{X}'\mathbf{X}(\lambda b)) = \lim_{\lambda\to\infty}\lambda^2 b'\mathbf{X}'\mathbf{X}b = \infty$, so that the objective vanishes to $+\infty$ in any asymptotic direction.

This solution $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$, called the **ordinary least squares** (OLS) estimator, can be shown to have very strong theoretical properties: it is *unbiased*, i.e. $\mathbb{E}[\hat{\beta}] = \beta$, meaning that it is expected to estimate the true parameter β, and within the class of all estimators that are unbiased, it has the lowest variance, i.e. it is expected to provide the estimate closest to β for any given sample. This estimator *always* estimates the coefficients of the linear conditional expectations function $\mathbb{E}[Y|X] = \beta_0 + \beta^x X$, and as the excursion below discusses briefly, even if the conditional expectation function is not linear, the coefficients estimated by the OLS estimator recover the best linear prediction of Y given X. This is a crucial point: when you estimate OLS, you always estimate the best linear prediction, which corresponds to the conditional mean if it can be described by a linear function. Any discussion of "misspecification" or "estimation bias" does (or at least: should!) **not** argue that we are unable to recover the linear prediction coefficients, but that we are unable to interpret these coefficients in a causal/desired way. The last section addresses this issue to more detail.

Excursion. Non-linearity and best linear prediction. A crucial question concerns the quality of the OLS estimator when the true conditional expectation function is not linear, i.e. there exist no parameters $\beta_0, \beta^x \in \mathbb{R}$ such that $\mathbb{E}[Y|X] = \beta_0 + \beta^x X$. A first case that can generally be easily handled by the concepts introduced thus far is the one of *linearity in parameters*. One such example is $\mathbb{E}[Y|X] = \beta_0 + \beta_1^x X + \beta_2^x X^2$. Here, even though the conditional mean is not linear in X, the inclusion of $Z = X^2$ achieves a setup that is consistent with the linear model we introduced. Note that generally, $X^2 \neq \lambda X$ for any given $\lambda \in \mathbb{R}$, so that inclusion of the squared term does not create a collinearity issue. However, in this case the marginal effect of X on Y depends on the level of X: $\frac{\partial Y}{\partial X} = \beta_1^x + 2\beta_2^x X$. Another example is $\mathbb{E}[Y|X] = \beta_0 + \beta_1^x \log(X)$, the case of *log-linearity*. Also this setup is consistent with our model. If $Y = \log(\tilde{Y})$, then we have a log-log model where β_1^x can be interpreted as an *elasticity* capturing a relationship in percentages: if X increases by 1%, then $\tilde{Y}$ increases by β^x%.

However, linearity in parameters may often be a restrictive assumption. Without taking a stance on the functional form of $\mathbb{E}[Y|X]$, we can still interpret the linear model

$$Y = \beta_0 + \beta^x X + e = \tilde{X}'\beta + e; \qquad \beta = \begin{pmatrix}\beta_0 \\ \beta^x\end{pmatrix}, \tilde{X} = \begin{pmatrix}1 \\ X\end{pmatrix}.$$

and its multivariate extensions in a useful way. To see this, note that the OLS estimator recovers the best linear prediction of Y given X *in the data*, i.e. given the sample $\{(x_i, y_i)\}_{i\in\{1,\dots,n\}}$ of observations of pairs of X and Y. It relies only on minimizing the residual sum of squares, and not on linearity of the conditional mean for the underlying random variables. Thus, the question at hand is how to interpret this linear prediction if not through the conditional mean. For this, consider the linear forecast for Y using an arbitrary vector $b \in \mathbb{R}^2$. Then, the forecast $\tilde{X}'b$ for Y is good if

1. $\mathbb{E}[Y - \tilde{X}'b] = 0$, and
2. $\mathbb{E}[(Y - \tilde{X}'b)^2] = Var[Y - \tilde{X}'b]$ is as small as possible.

Equality in point 2 follows from $Var[Y - \tilde{X}'b] = \mathbb{E}[(Y - \tilde{X}'b)^2] - \mathbb{E}[Y - \tilde{X}'b]^2$ and $\mathbb{E}[Y - \tilde{X}'b] = 0$. Therefore, such forecast would be right "on average", and have the minimum amount of volatility. Let's see how

the optimum looks like:

$$\mathbb{E}[(Y - \tilde{X}'b)^2] = \mathbb{E}[Y^2 - 2Y\tilde{X}'b + (\tilde{X}'b)^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y\tilde{X}'b] + \mathbb{E}[(\tilde{X}'b)^2].$$

where the second equality uses linearity of the mean. By this property of linearity, we can also isolate non-random *vectors* at the beginning and end of expressions from the expected value, specifically the vector b. Writing $(\tilde{X}'b)^2 = b'\tilde{X}\tilde{X}'b$, the expression becomes

$$\mathbb{E}[(Y - \tilde{X}'b)^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[\tilde{X}Y]b + b'\mathbb{E}[\tilde{X}\tilde{X}']b.$$

Optimizing this expression with respect to b is analogous to the OLS problem, and provided that $\mathbb{E}[\tilde{X}\tilde{X}']$ is invertible, one obtains the unique solution

$$\beta = \mathbb{E}[\tilde{X}\tilde{X}']^{-1}\mathbb{E}[\tilde{X}Y].$$

This looks very familiar to the OLS solution. This is no accident: while OLS recovers the best linear prediction in the data, the estimated parameter is the best linear prediction in expectation, or *in the population*. To reiterate, **regardless of the functional form of the conditional expectation** $\mathbb{E}[Y|X]$, the vector β of parameters estimated by the OLS estimator $\hat{\beta}$ gives $\tilde{X}'\beta$ as the **best linear prediction** of Y given X, called the *linear projection of X onto Y*. If the conditional mean is indeed linear, it coincides with the linear projection of X onto Y. Still, the linear projection is the more general concept, as it always exists, regardless of the functional form of the conditional mean.

In terms of **interpretation**, without relying on the conditional mean and its intuition of marginal effects, the linear projection coefficients β tell you how much, on average, Y can be expected to be higher if X is observed to be higher by one unit. If you can argue that the model appropriately absorbs third variables, you can therefore claim that the linear projection model recovers average effects.

EXCURSION END

5.3.3 Inference

One thing we have only insufficiently addressed thus far is the quality of estimation. While the OLS estimator is not biased, and it is the "most efficient" estimator (lowest variance among the set of unbiased estimators), in practice, its variance (i.e., the expected squared error) can still be very high - just not as high as the one you would obtain with another estimator. Therefore, if you have a simple model $Y = \beta_0 + \beta^x X + e$ and I tell you that OLS has been applied to obtain $\hat{\beta}^x = 0.05$ based on 10 data points, would you be confident to preclude that the true value of β^x is 0? And would you rule out that it is 1?

To answer such questions, we need to address the uncertainty associated with our estimation. Only once we can quantify this uncertainty we will be able to perform meaningful inference on the estimates we obtained. To do so, we rely on the key concepts of *convergence in probability* and *convergence in distribution*.

Definition 92. *(Convergence in Probability) Let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, and let T be random variable. Then, we say that T_n converges to T in probability if for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|T_n - T| > \varepsilon) = 0 \qquad \textit{or equivalently} \qquad \lim_{n\to\infty} P(|T_n - T| \leq \varepsilon) = 1.$$

We write $T_n \xrightarrow{p} T$.

Definition 93. *(Convergence in Distribution) Let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, and let T be*

random variable. Then, we say that T_n converges to T in distribution if for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{T_n}(x) = \lim_{n\to\infty} P(T_n \leq x) = P(T \leq x) = F_T(x).$$

We write $T_n \xrightarrow{d} T$. If the distribution of T is known, e.g. T is normally distributed with mean μ and variance σ^2, i.e. $T \sim N(\mu, \sigma^2)$, we write $T_n \xrightarrow{d} N(\mu, \sigma^2)$.

To put it more simply, T_n converges to T in probability if for large n, T_n and T are as good as equal (deviation smaller than an arbitrarily small ε) with probability 1. The fairly abstract definition of convergence in distribution has a relatively simple interpretation: the sequence of distribution functions F_{T_n} associated with the sequence T_n of random variables approaches the distribution function F_T of T *pointwise*, i.e. at any given location x.[42] In simpler words, if n is large enough, then the distribution of T_n should be as good as identical to the one of T. This is good news especially if the distribution of T_n is difficult to characterize for finite n, but the one of T is easily obtained. Note that in distinction to convergence in probability, like you can draw two highly distinct numbers from the same normal distribution, specific realizations of T_n can still be very different from those of T, but their probability distributions will coincide. Therefore, this concept is weaker than convergence in probability. To link the two concepts, note that convergence in probability implies convergence of $T_n - T$ to the "random variable" L with expectation $\mathbb{E}[L] = 0$ and $Var[L] = 0$, i.e. the variable L that is equal to 0 with probability 1.[43] This gives the following convention: if $T_n \xrightarrow{p} C$, where C is a non-random object (e.g. real number or matrix), then this implies $T_n \xrightarrow{d} C$, where C is the "random variable" equal to C with probability 1.

In our context, the random variable T_n will be the OLS estimator. Treating the observations $\{(x_i, y_i)\}_{i\in\{1,\dots,n\}}$ as realizations of n independent and identically distributed random variables (X_i, Y_i), the OLS estimator without concrete realizations of data points is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'y) = \left(\sum_{i=1}^{n}\begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}\right)^{-1} \sum_{i=1}^{n}\begin{pmatrix} Y_i \\ X_i Y_i \end{pmatrix} = \left(\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 1 & X_i \\ X_i & X_i^2 \end{pmatrix}\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Y_i \\ X_i Y_i \end{pmatrix}$$

Adopting the convention $\tilde{X}_i = (1, X_i)'$, we can more compactly represent it as

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i Y_i$$

To address the error we make in estimation, it is useful to consider the quantity $\hat{\beta} - \beta$. To express it in a useful way, we plug in $Y_i = \tilde{X}_i'\beta + e_i$ with $e_i = Y_i - X_i\beta$ to the expression above, and we obtain

$$\begin{aligned}
\hat{\beta} - \beta &= \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i(\tilde{X}_i'\beta + e_i) - \beta \\
&= \left[\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right]\beta - \beta + \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i e_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i e_i.
\end{aligned}$$

This is the quantity that we need to study in order to characterize the error we make in estimating β

[42] This criterion is indeed weaker than the one of uniform convergence, where we would require convergence *across* locations x.

[43] Indeed, most probability limits you will come across are real numbers/matrices, rather than stochastic random variables.

using $\hat{\beta}$. The final ingredients are given in the Theorems below. For the sake of compactness, they are not proven here, but excellent resources exist online or in textbooks.[44]

Theorem 65. *(Law of large numbers) Consider a set $\{T_i\}_{i\in\{1,\dots,n\}}$ of n independently and identically distributed variables with $\mathbb{E}[T_i^2] < \infty$. Then,*

$$\bar{T}_n := \frac{1}{n}\sum_{i=1}^{n} T_i \xrightarrow{p} \mathbb{E}[T_i].$$

Theorem 66. *(Central limit theorem) Consider a set $\{T_i\}_{i\in\{1,\dots,n\}}$ of n independently and identically distributed variables with $\mathbb{E}[T_i^4] < \infty$. Then, for $\bar{T}_n := \frac{1}{n}\sum_{i=1}^{n} T_i$, it holds that*

$$\sqrt{n}(\bar{T}_n - \mathbb{E}[T_i]) \xrightarrow{d} N(0, Var[T_i]).$$

Theorem 67. *(Slutsky's theorem) Let $\{T_n\}_{n\in\mathbb{N}}$ and $\{W_n\}_{n\in\mathbb{N}}$ be sequences of random variables with $T_n \xrightarrow{d} N(0,\Sigma)$ and $W_n \xrightarrow{p} W$ and $\mathbb{E}[W^2] < \infty$. Then,*

$$W_n T_n \xrightarrow{d} N(0, \mathbb{E}[W]\Sigma\mathbb{E}[W])$$

and if W_n is invertible with probability 1, then also

$$W_n^{-1} T_n \xrightarrow{d} N(0, \mathbb{E}[W]^{-1}\Sigma\mathbb{E}[W]^{-1})$$

One last detail will help us put everything together: note that

$$\begin{aligned}
\mathbb{E}[\tilde{X}_i e_i] &= \mathbb{E}[\tilde{X}_i Y_i - \tilde{X}_i \tilde{X}_i'\beta] \\
&= \mathbb{E}[\tilde{X}_i Y_i] - \mathbb{E}[\tilde{X}_i \tilde{X}_i']\beta \\
&= \mathbb{E}[\tilde{X}_i Y_i] - \mathbb{E}[\tilde{X}_i \tilde{X}_i']\left(\mathbb{E}[\tilde{X}_i \tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i Y_i]\right) \\
&= \mathbf{0}.
\end{aligned}$$

For a derivation of $\beta = \mathbb{E}[\tilde{X}_i \tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i Y_i]$, see the excursion above. With this, we proceed as follows:

Step 1. Applying the central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{X}_i e_i = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i e_i - \underbrace{\mathbb{E}[\tilde{X}_i e_i]}_{=0}\right) \xrightarrow{d} N(0, Var[\tilde{X}_i e_i]).$$

The variance of the asymptotic distribution is

$$Var[\tilde{X}_i e_i] = \mathbb{E}[\tilde{X}_i \tilde{X}_i' e_i^2] - \underbrace{\mathbb{E}[\tilde{X}_i e_i]\mathbb{E}[\tilde{X}_i e_i]'}_{=\mathbf{0}} = \mathbb{E}[\tilde{X}_i \tilde{X}_i' e_i^2]$$

so that compactly,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{X}_i e_i \xrightarrow{d} N(0, \mathbb{E}[\tilde{X}_i \tilde{X}_i' e_i^2])$$

Step 2. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i' \xrightarrow{p} \mathbb{E}[\tilde{X}_i \tilde{X}_i'].$$

[44] These results are general enough that you will find comprehensive proofs on Wikipedia. However, for our purposes it is much more important to apply them properly, rather than to understand their theoretical justification.

Moreover, the law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i e_i \xrightarrow{p} \mathbb{E}[\tilde{X}_i e_i] = \mathbf{0}.$$

Step 3. We have two components that correspond to the setup of Slutsky's theorem. First, putting together the two results of step 2, we obtain the following from Slutsky's theorem (recalling that convergence in probability to a constant implies convergence in distribution to the same constant):

Corollary 10. *(OLS consistency) For the OLS estimator* $\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i Y_i$*, it holds that*

$$\hat{\beta} - \beta \xrightarrow{p} \mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i e_i] = \mathbf{0}$$

and thus also $\hat{\beta} \xrightarrow{p} \beta$*. Therefore, we say that* $\hat{\beta}$ *is a consistent estimator of* β*.*

This informs us that deviations of $\hat{\beta}$ from β will asymptotically be arbitrarily small with probability 1. However, it does not inform the distribution of $\hat{\beta}$ for concrete sample sizes n as we investigate in practice. For this, we consider instead the distribution of $\sqrt{n}(\hat{\beta} - \beta)$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'}_{\xrightarrow{p}\mathbb{E}[\tilde{X}_i\tilde{X}_i']}\right)^{-1} \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_i e_i}_{\xrightarrow{d}N(0,\mathbb{E}[\tilde{X}_i\tilde{X}_i' e_i^2])} \quad \xrightarrow{d} \quad N(0, \mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i\tilde{X}_i' e_i^2]\mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1})$$

We define $\Sigma_{\hat{\beta}} := \mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i\tilde{X}_i' e_i^2]\mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}$ as the asymptotic variance of $\hat{\beta}$. With this result, for sufficiently large samples, the distribution of $\sqrt{n}(\hat{\beta} - \beta)$ is approximately equal to the one of a random variable with distribution $N(0, \Sigma_{\hat{\beta}})$. This is useful because it also means that the distribution of $\hat{\beta} - \beta$ is approximately equal to $N(0, \frac{1}{n}\Sigma_{\hat{\beta}})$ (by $Var[aX] = a^2 Var[X]$), and finally that

$$\hat{\beta} \overset{a}{\sim} N(\beta, \frac{1}{n}\Sigma_{\hat{\beta}}).$$

Here, $\overset{a}{\sim}$ is used to indicate that asymptotically, the distribution of $\hat{\beta}$ can be described as $N(\beta, \frac{1}{n}\Sigma_{\hat{\beta}})$. With some oversimplification, for large enough n, the distribution of $\hat{\beta}$ is as good as indistinguishable from a normal distribution with mean β and variance $\frac{1}{n}\Sigma_{\hat{\beta}}$.

The important bit of information obtained from this is that we now have an approximation to the finite sample variance of $\hat{\beta}$ that is "good" if the sample is "large". In practice, a few hundred observations are usually considered to suffice for simple models and a few thousand for more complex models, i.e. longer vectors β. The matrix $\Sigma_{\hat{\beta}} = \mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}\mathbb{E}[\tilde{X}_i\tilde{X}_i' e_i^2]\mathbb{E}[\tilde{X}_i\tilde{X}_i']^{-1}$ is straightforward to estimate using averages instead of expectations, and replacing the unobserved e_i with its sample counterpart $\hat{e}_i = Y_i - \tilde{X}_i'\hat{\beta}$:

$$\hat{\Sigma}_{\hat{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\hat{e}_i^2\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right)^{-1}.$$

In the matrix $\frac{1}{n}\hat{\Sigma}_{\hat{\beta}}$, the (1,1) element gives the estimated variance of $\hat{\beta}_0$, and the $(2,2)$ element the one of $\hat{\beta}^x$, denoted as $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta^x}$, respectively. One can show that

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\beta_0}} \overset{a}{\sim} N(0,1) \quad \text{and} \quad \frac{\hat{\beta}^x - \beta^x}{\hat{\sigma}_{\beta^x}} \overset{a}{\sim} N(0,1).$$

This circumstance can be used to test hypotheses such as "$\beta^x = 0$" or "$\beta^x = 1$" as mentioned in the introduction of this subsection. The interested reader is referred to test theory, but in a nutshell, the procedure is as follows: you assume that your hypothesized value, let's say $\beta^x = 0$, is indeed the true value of β^x. Under this hypothesis, $\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}$ should be normally distributed with zero mean and unit variance. You then investigate what is the probability of observing a deviation of at least $\left|\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}\right|$ from zero under this hypothesis, i.e. in a scenario where $\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}$ is indeed distributed $N(0,1)$: if $Z \sim N(0,1)$, then for $c > 0$,

$$P(|Z| > c) = P(Z < -c) + P(Z > c).$$

By symmetry of the normal distribution, $P(Z < -c) = P(Z > c)$, and

$$P(|Z| > c) = 2P(Z > c) = 2(1 - P(Z \leq c)) = 2(1 - \Phi(c))$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution $N(0,1)$. Therefore, the probability of observing a deviation from zero of at least $\left|\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}\right|$ is

$$p = 2(1 - \Phi\left(\left|\frac{\hat{\beta}_0}{\hat{\sigma}_{\beta_0}}\right|\right)$$

This quantity is called the *p-value*, and is typically reported in regression output tables. If this probability is sufficiently large, you can not reject the hypothesis that $\beta^x = 0$. If it is sufficiently low, however, you reject the hypothesis. If your model is specified in a way that allows for a marginal effect interpretation, this rejection is required to claim that the model identifies a non-zero effect of X on Y.

To conclude the discussion of the linear model, we have introduced the concept of the conditional expectation, and the linear conditional expectation regression model. This model is also useful if the true conditional expectation is not linear, among others due to a Taylor approximation justification. In this model, coefficients can be interpreted as marginal effects of one variable on the outcome Y, holding the other variables constant. Therefore, it allows to describe the partial relationship between X and Y that is unexplained by third variables Z. This model can be estimated using the OLS estimator, which could be derived from a simple unconstrained optimization problem. This estimator has powerful theoretical properties, and always estimates the best linear prediction of Y given X. The practical exercise of the econometrician is usually not to allow for "consistent" estimation of the linear prediction (which is guaranteed), but to write down a linear prediction that is useful for studying a given practical issue at hand. Beyond estimation, we have covered inference, i.e. methods of quantifying estimation uncertainty, as well as their justification through large sample asymptotics. These methods allow to test for, and especially to reject certain values for the estimated parameters in the model.

5.4 Correlation and Causality Revisited

The mathematics of econometric models is one side of the coin the empirical economist has to understand, their practical application is the other. While this script intends to give an introduction to the mathematical foundation of the economic profession, the latter aspect is still important enough to deserve a brief discussion also here. As mentioned in the first sections of this chapter, our goal is usually to arrive at a *causal* conclusion ("X causes Y") while our methods measure only a *correlational* relationship ("X predicts Y"). While the linear model goes one step from correlation to causality allowing to hold fixed some third variables, its interpretation is still inherently correlational: the most we can say is that "holding constant Z, X predicts Y", or equivalently "X predicts Y given Z". So, how do we argue for causality, and what needs to be kept in mind in these arguments?

The typical approach is very simple: you assume that there exists a true, causal model of the form

$$Y = \gamma_0 + \gamma^x X + e$$

i.e. that Y has some baseline level γ_0, and is linearly driven by X ($\gamma^x X$) and other factors e. This is an assumption that you can not test in practice, just like you can not test any assumption you impose on a theoretical model. Your research questions are typically:

1. Does X affect Y, i.e. does the hypothesis $\gamma^x \neq 0$ hold?
2. If $\gamma^x \neq 0$, is the causal effect economically relevant?[45]

There are two possible reasons that prevent you from estimating the causal γ-coefficients through the predictive model:

1. *Omitted variables:* there are third variables Z that cause Y and are correlated with X.
2. *Reverse causality:* Y causes X.

While reverse causality is not a problem in the causal model, it is an issue in the predictive model: suppose you are a firm that sells backpacks at 50$ a piece. An employee working in marketing wants to convince you that you should spend more on marketing, as he thinks business expansion could greatly boost the number of backpacks you sell. He shows you some estimates from a linear model he estimated:

$$x_t = 0.02 \cdot r_t + e_t$$

where x_t is the number of backpacks sold in month t and r_t is your firm's revenue in this month. He has also learned about inference and shows you that his estimates are highly statistically significant. Would you, then, believe that if you managed to grow as a firm in terms of revenue, this would attract more new customers?

Well, if you re-write his estimates, you obtain

$$r_t = 50 \cdot x_t - 50 \cdot e_t$$

Given that your revenue comes from sales of backpacks which you sell at 50$ a piece, this equation does not look very surprising. This issue captures the essence of reverse causality: just because X is a good predictor of Y, even if X and Y are causally related, it does not mean that X causes Y, but the predictive quality could also be because Y causes X. Indeed, his estimates from the predictive equation are perfectly consistent with causal equations where X does not cause Y, i.e. $\gamma^x = 0$, but Y causes X with coefficient 50.

So, how do we manage to estimate the causal coefficients using the predictive model? For omitted variables, since they cause Y, they should be contained in the remaining factors. If they enter Y linearly, we can write $e = \gamma^z Z + \varepsilon$, so that the causal equation becomes

$$Y = \gamma_0 + \gamma^x X + \gamma^z Z + \varepsilon.$$

Intuitively, this equation explicitly accounts for all variables Z that would induce X to predict Y for non-causal reasons. Indeed, we can show that estimation of this model allows to recover the causal

[45] Not any non-zero effect concerns/is considered relevant the economist. If you find that the introduction of 1000 robots to an industry reduces employment in this industry by 0.0001%, and there are no industries in your sample that introduce more than 1000 robots in a given year, then you would not claim that robot adoption reduces employment. The threshold to relevance, however, is subjective and depends also on the context.

coefficient γ^x, the formalities of this investigation are however beyond the purpose of this course.[46]

Control variables can also be used to address issues of reverse causality: usually, if we are in the comfortable situation to have data for several periods, the first step to circumvent reverse causality is to adjust the timing, exploiting that causation can only work forward in time, not backwards. Continuing the example of the backpack firm, as this month's sales of backpacks can not be caused by last month's revenues, this step would suggest to specify the model

$$x_t = \beta_0 + \beta^r r_{t-1} + e_t.$$

However, a second step may still be necessary if backpack sales are *serially dependent*, i.e. if x_t depends on x_{t-1}. For instance, if customers are satisfied with your product and recommend it to others, then if sales numbers are high in one month, they may also be higher in following months because more people recommend your product. In your model, this means that x_{t-1} causes x_t (not directly, but indirectly through increased mouth-to-mouth advertising). This is an issue because x_{t-1} is clearly correlated with your explanatory variable of interest, r_{t-1}. This discussion suggests that you should include x_{t-1} as a control variable. Your model becomes[47]

$$x_t = \beta_0 + \beta^r r_{t-1} + \beta^x x_{t-1} + e_t.$$

To summarize, we have briefly reviewed the two most common issues in empirical analysis, omitted variables and reverse causality, and we have seen how they can be addressed using control variables and the appropriate use of variable timing. In practice, especially omitted variables are tricky for two reasons: *(i)* you do not know ex-ante which third variables may be relevant for the relationship of interest,[48] and *(ii)* not all of the variables you think are relevant are measured in the data available for your analysis. To address the former point, we usually estimate relatively rich models with many variables, as the cost of missing one important variable are much larger than including a few irrelevant ones, especially with large datasets. Next to specific variables, we also include *fixed effects* aimed at absorbing unmeasured variables that vary at higher levels; the interested reader can look up the standard fixed effects and the two-way fixed effects model. To address the latter point, several techniques have been developed to estimate causal parameters in environments where (we can) not (be sure that) all relevant third variables are observed. To this end, the interested reader is referred to instrumental variables estimation, difference-in-difference estimation, regression discontinuity design and propensity score matching. All these techniques are frequently used in the economic literature, where the former two have been most prominent.

[46] Note that we design the model to estimate the causal coefficient for X, γ^x. The causal coefficient(s) for Z may not be estimated consistently if there are third variables that determine Y and are correlated with Z. Thus, always limit your interpretations as much as possible to the coefficients corresponding to variable(s) for which you construct the model, and abstain from speculating on the coefficients of control variables.

[47] Whether it is sufficient to include just one lag for the dependent (i.e., left-hand side) variable is not a trivial question, but an issue of central interest to the field of time series analysis. Intuitively, a second lag would be needed if it explains today's values even conditional on the first lag, which may occur for more complex dynamic processes. In practice, a simple but efficient approach is to estimate the model once with and once without the second lag, and compare the estimates obtained for the coefficient of interest. If it is different across models, the second lag should be included, and an analogous test should be performed on whether or not to include a third lag, and the process should be iterated until the coefficient of interest converges.

[48] This issue is especially relevant to the economist, who typically studies environments where dependent variables (think of a firm's profits, or an industry's level of total employment) have a multitude of causes, and it is not clear which of them would be correlated with an analytical variable you are interested in.

6 Solutions to the Recap Questions

Chapter 0

For Questions 1 and 2, please consult the text.

3. (a) true, (b) false, (c) false; the set describes the uneven natural numbers, (d) true, since the biggest uneven number in $[0,10]$ is 9 and thus strictly smaller than 10.

4. (a) $\exists x \in A : x \notin B$, (b) the statement is identical to (a); in set notation one could write $A \not\subseteq B$, (c) $\exists x \in X : (\forall y \in X : y \leq x)$, (d) the statement is equivalent to $A \cap B^c \neq \emptyset$. Thus, the negation is $A \cap B^c = \emptyset$. Any equivalent characterization (e.g. $\forall x \in X : (x \notin A \vee x \notin B^c)$) is fine as well.

5. This question admittedly is an exception to the "not that difficult" rule, as it is easy to get confused here. The solution is

$$\mathcal{P}(\mathcal{P}(A)) = \Big\{\emptyset, \{\emptyset\}, \{\{1\}\}, \{\{\pi\}\}, \{\{1,\pi\}\}, \{\emptyset,\{1\}\}, \{\emptyset,\{\pi\}\}, \{\emptyset,\{1,\pi\}\}, \{\{1\},\{\pi\}\}, \{\{1\},\{1,\pi\}\},$$
$$\{\{\pi\},\{1,\pi\}\}, \{\emptyset,\{1\},\{\pi\}\}, \{\emptyset,\{1\},\{1,\pi\}\}, \{\emptyset,\{\pi\},\{1,\pi\}\}, \{\{1\},\{\pi\},\{1,\pi\}\}, \{\emptyset,\{1\},\{\pi\},\{1,\pi\}\}\Big\}.$$

   The universal superset for $\mathcal{P}(\mathcal{P}(A))$ is $\mathcal{P}(\mathcal{P}(\mathbb{R}))$.

6. (a) Since $\ln(10) > 0$, $|\ln(10)| = \ln(10)$, and $f(\ln(10)) = \exp(\ln(10)) = 10$.
   (b) $\mathrm{im}(f) = \{y \in \mathbb{R}_+ : (\exists x \in \mathbb{R} : y = \exp(|x|)\}$. Note that f is discontinuous at $x = 0$ and continuous and monotonically increasing on $(0,\infty)$, and that $\lim_{x\to\infty} \exp(|x|) = \infty$. Therefore, $\mathrm{im}(f) \supseteq (\exp(0),\infty) = (1,\infty)$. Since for $x \in (-\infty, 0)$, $f(x) = f(-x)$, this part of the domain does not add new values. Thus,
$$\mathrm{im}(f) = (1,\infty) \cup f[\{0\}] = (1,\infty) \cup \{1\} = [1,\infty).$$
   (c) For $x \in \mathbb{R}$, $x \in f^{-1}[\{2\}] \Leftrightarrow 2 = f(x) = \exp(|x|) \Leftrightarrow \ln(2) = |x|$. Thus, $f^{-1}[\{2\}] = \{-\ln(2), \ln(2)\}$.
   (d) For the inverse function of f to exist at $y = 2$, it must have exactly one image at this point. From (c), we know that this is not the case. Thus, $f^{-1}(2)$ does not exist.

7. (a) $f'(x) = f(x) = 3\exp(x)$.
   (b) Using the Quotient Rule, $f'(x) = (x^4 \cdot \cos(x) - 4x^3 \cdot \sin(x))/x^8 = \cos(x)/x^4 - 4\sin(x)/x^5$. (c) One may either use the Chain Rule, or exploit the rules for the logartihm.[1] Doing the latter, we can re-write $f(x) = x \cdot \ln(3)$, which gives $f'(x) = \ln(3)$.
   (d) Now, we have to use the chain rule: Let $f_1(x) = \sin(x)$ and $f_2(x) = \cos(x^2)$. To obtain $f_2'(x)$, we have to apply chain rule as well: Let $f_{2,1}(x) = \cos(x)$ and $f_{2,2}(x) = x^2$. Then, $f_{2,1}'(x) = -\sin(x)$, $f_{2,2}'(x) = 2x$, and by chain rule $f_2'(x) = f_{2,2}'(x) f_{2,1}'(f_{2,2}(x)) = -2x \cdot \sin(x^2)$. With $f_1'(x) = \cos(x)$, one obtains $f'(x) = f_2'(x) f_1'(f_2(x)) = -2x \cdot \sin(x^2) \cdot \cos(\cos(x^2))$.

8. (a) $\lim_{x\to 0} \sin(x) = \lim_{x\to 0} x = 0$. Because additionally, numerator and denominator are differentiable, we can use L'Hôpital's rule, which gives $\lim_{x\to 0} \frac{\sin(x)}{x} = \frac{\cos(x)}{1} = \cos(x)$.
   (b) Using the product rule with $f(x) = \cos(x)$ and $g(x) = 1/(1+x)$, by continuity of $\cos(x)$, $\lim_{x\to\pi} \frac{\cos(x)}{1+x} = \frac{\cos(\pi)}{1+\pi} = -\frac{1}{1+\pi}$.
   (c) $\lim_{x\to 0} \exp(x) - 1 - x = \lim_{x\to 0} x^2 = 0$. Again, numerator and denominator are differentiable, and we can use L'Hôpital's rule. Let $f(x) = \exp(x) - 1 - x$ and $g(x) = x^2$. Then, we have from L'Hôpital's rule that
$$\lim_{x\to 0} \frac{\exp(x) - 1 - x}{x^2} = \lim_{x\to 0} \frac{f(x)}{g(x)} = \lim_{x\to 0} \frac{f'(x)}{g'(x)}.$$

[1] The logarithm rules are $\ln(a \cdot b) = \ln(a) + \ln(b)$ and $\ln(a^b) = b \cdot \ln(a)$.

Here, $f'(x) = \exp(x) - 1$ and $g'(x) = 2x$, and both expressions still go to 0 as $x \to 0$, so that we can not use product rule directly. Thus, we consult the second derivative: $f''(x) = \exp(x)$ and $g''(x) = 2$. Thus, the product rule yields $\frac{f''(x)}{g''(x)} = \frac{\exp(x)}{2}$, and L'Hôpital's rule gives

$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = \lim_{x\to 0} \frac{f''(x)}{g''(x)} = \frac{\exp(x)}{2} \quad \Rightarrow \quad \lim_{x\to 0} \frac{\exp(x) - 1 - x}{x^2} = \frac{\exp(x)}{2}.$$

CHAPTER 1

1. (a) yes; vector addition, (b) yes; scalar/real number addition, (c) no, (d) yes; scalar multiplication, (e) yes; scalar product or inner/vector/dot product.

2. Let $\tilde{B} \subseteq B$, and denote $\tilde{B} = \{b_{i(1)}, \dots, \tilde{b}_{i(k)}\}$ where $i(j)$ is the index of the j-th element in $\tilde{B}$ in the set B, and $k \leq m$. We want to check that $\tilde{B}$ satisfies the implication in Theorem 8 to show that the set is linearly independent. Denote by $E = \{1, \dots, m\} \backslash \{i(1), \dots, i(k)\}$ the set of indices excluded from $\tilde{B}$. Consequently, if $\sum_{j=1}^{k} \lambda_j b_{i(j)} = \mathbf{0}$, then

$$\mathbf{0} = \sum_{j=1}^{k} \lambda_j b_{i(j)} + \sum_{i \in E} 0 \cdot b_i = \sum_{l=1}^{m} \lambda_l b_l \quad with \quad \lambda_l = \begin{cases} 0 & \text{if } l \in E \\ \lambda_j & \text{if } l = i(j) \end{cases}$$

   Since this is a linear combination of B, and B is linearly independent, it follows that $\lambda_l = 0 \forall l \in \{1, \dots, m\}$, and thus $\lambda_j = 0 \forall j \in \{1, \dots, k\}$. Thus, by Theorem 8, it follows that $\tilde{B}$ is linearly independent.

3. There is no sample solution, you can ask after class or per mail if you are unsure whether your solution makes sense.

4. We have to check properties *(i)-(iii)* in Def. 15 for the norm-induced metric $d_N(x, y) = \|x - y\|$. For *(i)*, because $\|\cdot\|$ is a norm, clearly, $\forall x, y \in X : d_N(x, y) \geq 0$, and $d_N(x, y) = 0 \Leftrightarrow \|x - y\| = 0 \Leftrightarrow x - y = \mathbf{0} \Leftrightarrow x = y$, where the second equivalence follows from non-negativity of the norm. This establishes *(i)*. Next, let $x, y \in X$. Then, the norm absolute homogeneity gives

$$d_N(x, y) = \|x - y\| = \|(-1)(y - x)\| = |-1| \|y - x\| = d_N(y, x),$$

   proving *(ii)*, i.e. symmetry of d_N. Finally, let $x, y, z \in X$. Then, from the norm triangle inequality, it follows that

$$d_N(x, z) = \|x - z\| = \|x - y + y - z\| \leq \|x - y\| + \|y - z\| = d_N(x, y) + d_N(y, z),$$

   which establishes *(iii)*, the triangle inequality of the norm-induced metric.

CHAPTER 2

1. The square matrix has as many rows as columns, but the concept does not restrict the entries otherwise. A diagonal matrix on the other hand is a *necessarily square* matrix that additionally allows only entries on the diagonal to be non-zero. The identity matrix is both square and diagonal.

2. (i) A is invertible with inverse A^{-1}. Note that $\mathbf{I_n}' = \mathbf{I_n}$. Thus,

$$AA^{-1} = \mathbf{I_n} \quad \Leftrightarrow \quad \mathbf{I_n} = (AA^{-1})' = (A^{-1})'A'.$$

Therefore, $(A^{-1})'$ is the inverse matrix of A'.
(ii) By associativity of the matrix product,

$$[B^{-1}A^{-1}]\cdot AB = B^{-1}\underbrace{(A^{-1}A)}_{=\mathbf{I_n}}B = B^{-1}B = \mathbf{I_n}.$$

(iii) $\lambda \in \mathbb{R}$, $\lambda \neq 0$ is a 1×1 matrix, and we know that it is invertible with $\lambda^{-1} = 1/\lambda$. Thus, (iii) is a direct implication of (ii).

CHAPTER 3

1. I suggest before doing the calculations, you draw the functions, then you should have a clear expectation about your results for this question. Here, I confine to computing the derivatives and applying point 3. We have

$$f_1'(x) = 2x, \qquad f_2'(x) = \begin{cases} 1 & x < 5 \\ 0 & x > 5 \\ \text{undefined} & x = 5 \end{cases}, \quad \textit{and} \quad f_3'(x) = \cos(x).$$

Thus, f_1 is strictly increasing on $(0,\infty)$ and thus also increasing on this interval and constant on no interval, f_2 is strictly increasing on $(-\infty,5)$, constant on $(5,\infty)$ and weakly increasing on both these intervals – and indeed also on $\mathbb{R}$, but this is beyond the rule of point 3 because $f'(x)$ is not defined in $x = 5$! Finally, f_3 is strictly increasing on all intervals $((4z-1)\pi/2,(4z+1)\pi/2)$, $z \in \mathbb{Z}$, and thus also weakly increasing on these, and constant on no interval.

2. For formal definitions, see the text. The gradient is the first and the Hessian the second derivative of a multivariate real-valued function, i.e. $f : X \mapsto \mathbb{R}$ with $X \subseteq \mathbb{R}^n$ but $\mathbb{R}^n$ as the codomain. The Jacobian is the first derivative of a general multivariate function with potentially vector-valued output, i.e. codomain $\mathbb{R}^m$, $m \in \mathbb{N}$. If $m = 1$, the Jacobian and the gradient coincide.

3. $\frac{\partial}{\partial x_j}$ is an operator mapping functions f that are partially differentiable with respect to the j-th argument everywhere onto the partial derivative with respect to x_j: $\frac{\partial}{\partial x_j} : D_j^1(X,\mathbb{R}) \mapsto F_X$. $\frac{\partial f}{\partial x_j}$ is a function with domain X and codomain $\mathbb{R}$ and represents the partial derivative of f with respect to x_j; it corresponds to the value of $\frac{\partial}{\partial x_j}$ when evaluated at a specific function $f \in D_j^1(X)$. Thus, we would typically write $\frac{\partial}{\partial x_j}(f)$, but we commonly use the established notational simplification for this expression, $\frac{\partial f}{\partial x_j}$. Finally, $\frac{\partial f}{\partial x_j}(x_0)$ is a real number, equal to the specific value one obtains when evaluating the partial derivative at $x_0 \in X$. Finally, $\frac{\partial f(x_0)}{\partial x_j}$ is bad notation that you should not use!

4. The total derivative of $f(x_1,x_2,x_3)$ is

$$df = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \frac{\partial f}{\partial x_3}dx_3.$$

It is a function for which, evaluated at a specific point x_0, we have

$$df(x_0) = \frac{\partial f}{\partial x_1}(x_0)dx_1 + \frac{\partial f}{\partial x_2}(x_0)dx_2 + \frac{\partial f}{\partial x_3}(x_0)dx_3.$$

It is similar in interpretation to a directional derivative, and tells us how f changes when marginally varying the argument x from any given point x_0, the evaluation point of the partial derivatives,

in a specific way, or "direction" $(dx_1, \ldots, dx_n)$. It is derived from Taylor's theorem, using that the approximation error vanishes as deviations from x_0 becomes arbitrarily small.

5. f is multiplicatively separable, such that

$$\begin{aligned}\int_{[0,\pi/2]\times[0,\ln(2)]} f(x)dx &= \int_0^{\pi/2} \cos(x)dx \int_0^{\ln(2)} e^{2x}dx = [\sin(x)]_{x=0}^{x=\pi/2}[1/2e^{2x}]_{x=0}^{x=\ln(4)} \\ &= (\sin(\pi/2) - \sin(0)) \cdot (e^{\ln(4)} - e^0) = (1-0)(4-1) = 3\end{aligned}$$

where I used that $2\ln(2) = \ln(2^2) = \ln(4)$, $\sin(\pi/2) = 1$ and $\sin(0) = 0$.