Study guide

I will update this study guide after each lecture. It specifies, per lecture, the material I discussed from my notes, as well as related sections from three recommended texts:

1. Carter, M., 2001, Foundations of Mathematical Economics, MIT Press. Consult the book's website `http://michaelcarteronline.com/FOME/` for solutions and additional resources.

2. Sydsæter, K., Hammond, P., Seierstad, A., and Strøm, A., 2008, Further Mathematics for Economic Analysis, Prentice Hall, 2nd edition. A solutions manual can be found at

   `http://media.pearsoncmg.com/intl/ema/ema_uk_he_sydsaeter_fmea_2/smfmea2.pdf`

3. Sorger, G., 2015, Dynamic Economic Analysis; Deterministic Models in Discrete Time, Cambridge University Press.

In fact, the Sorger text only concerns a very small part of the course (dynamic optimization); it won't be mentioned until then. The lecture notes contain the compulsory reading; the recommended parts in these three other books are optional and can serve as a source of additional (economic) examples and applications. I assume that you know the material specified in the file of prerequisites and will use it frequently.

Lecture 1

- ☒ Notes: Preface, section 1.
- ☒ Carter: pp. 66–76.

Please read the preface: it describes most of the administrative matters that are important for the problem sets and exam and provides some motivation for the contents. I expect you to read each update of this study guide. I also recommend the file about how to read math texts.

The French mathematician Poincaré wrote that 'mathematics is the art of giving the same name to different things'. We give the name ***vector space*** to any set with two operations, addition and scalar multiplication (multiplication with a real/complex number), that satisfy standard arithmetic properties. That is the intuition I want you to keep in mind. Consult Definition 1.1 for the precise formulation. There is really no point in learning that definition by heart; the exam will be open-book!

We're not doing this out of a morbid fascination with horripilative abstraction, but because it gives us extra mileage: if we can prove a result, like the simple claims in Theorem 1.1, for a *general* vector space, then it automatically holds in each *particular* example of such a space. This is what we will be doing in much of the sequel: our theorems will state things under pretty general circumstances and in many examples and exercises, we will invoke these theorems to draw conclusions about special cases.

It is cumbersome to prove that a set V is a vector space by checking that it is closed under addition (it contains the sum of any pair of its elements) and scalar multiplication (it contains any scalar multiple of any of its elements) and arithmetic properties (V1) to (V8). Fortunately, there are at least two convenient ways of making new vector spaces from old ones:

1. larger spaces from smaller ones: if, for each index i in some index set I, V_i is a vector space, then Example 1.8 tells us that the product space $V = \times_{i \in I} V_i$ of all functions x on I such that $x(i) \in V_i$ for all $i \in I$ is a vector space as well. In class, I wrote v_i instead of $x(i)$ to obtain vectors $(v_i)_{i \in I}$, but that's just a different notation for the same thing. For instance, $\mathbb{R}$ is a vector space, so $\mathbb{R}^3$, consisting of three 'copies' of $\mathbb{R}$ is a vector space!

2. smaller spaces from larger ones: Theorem 1.2 provides an easy test to see when a subset W of a vector space V is a vector space itself. The reason that it suffices to check way fewer conditions is this: many of the properties (V1) to (V8) make statements about *all* elements of V. So they apply in particular to the elements of the smaller set $W \subseteq V$! I strongly recommend checking out Example 1.9, which illustrates this theorem.

Make sure you read the examples and get used to the notation, because we will be using this throughout the course. The proof of Theorem 1.1 is a bit pedantic. Its theoretical *raison d'être* is to illustrate that there is no need to add other familiar arithmetic rules to the definition of a vector space, since these properties are implied by the definition; the practical one is that from now on, you can use these properties to draw helpful conclusions about vector spaces. It suffices to browse through its proof.

In addition to the notational conventions on page 6, the following might be good to keep in mind: I try to be careful to distinguish between the *number* zero, denoted 0, and the *zero vector* of a vector space, denoted $\mathbf{0}$ or, when I write by hand, $\underline{0}$. By definition (property (V3)), the zero vector is a special element of the vector space under consideration. So depending on the vector space, the zero vector can be a lot of things. For instance:

☒ in $\mathbb{R}^3$: $\mathbf{0} = (0,0,0)$,

☒ in $\mathbb{R}^{2\times 3}$: $\mathbf{0} = \left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right]$,

☒ in $C[a,b]$: $\mathbf{0}$ is the function $\mathbf{0} : [a,b] \to \mathbb{R}$ with $\mathbf{0}(x) = 0$ for all $x \in [a,b]$, the function that is constant at zero.

Lecture 2

☒ Notes: Section 2 and the first part of section 5. Read the latter until (including) Example 5.2. Skip section 3. I will briefly talk about section 4 next class.

☒ Carter: 77–83 (until 1.4.3), 263–273 (until 3.1.1), pp. 114–118 ('Normed Linear Spaces' until Exercise 1.200).

☒ Sydsæter et al.: Section 1,2 pp. 86–87 (Subsection 'Linear Transformations'), 6 ('Vectors'), 465–466 ('Point Set Topology in $\mathbb{R}^n$').

Bases are the building blocks of vector spaces. A basis for a vector space V is a maximal linearly independent subset. This means two things:

1. The basis is so large that every vector in V can be written as a linear combination of basis vectors. Formally, it must span the vector space.

2. The basis is so small that you cannot omit basis vectors and still span the entire space V. Formally, it must be linearly independent.

In section 2, read carefully through the definitions, examples, and theorems. The proof of Theorem 2.1 is skipped entirely; you don't need to read the proof of Theorem 2.2 either, but that of Theorem 2.3 is just a clever application of an earlier result.

If a basis has some desirable property, the fact that every vector can be expressed in terms of basis vectors might allow you to infer the same property for *all* vectors. This is why bases are so important in linear algebra.

But how do you know that a vector space actually *has* a basis? Sometimes explicit bases are easy to find: Examples 2.6 and 2.7 provide standard bases for $\mathbb{R}^n$ and spaces of polynomials, respectively. The notation e_i for the i-th standard basis vector of $\mathbb{R}^n$ is important: we will be using it a lot. In general, however, the existence of bases is not so obvious. How would you go about finding one?

The argument I sketched was as follows: start with a some small linearly independent subset — let's call it I — of vector space V. If I spans V, it is a basis! If it doesn't, find some vector that is not in its span and add it to I. The new set remains linearly independent and now has a strictly larger span. Now repeat the process, generating larger and larger linearly independent sets that consequently span larger and larger subspaces of V.

With the help of a very abstract result, Zorn's lemma, from axiomatic set theory, this indeed assures that there is a sufficiently large set that has the desired properties of being both linearly independent and spanning V. The proof is skipped (section 3), but have a look if you want to.

Next, I discussed the formalization of 'length', starting with the familiar setting of $\mathbb{R}^2$ and using some standard properties there that went into the definition of a norm on a general vector space. I will continue the discussion, with more examples, next lecture.

A proof I skipped in class

In response to a question in class, I promised to include the following:

Remember that I said that *the span of a set W of vectors in some vector space V is a (linear) subspace of V*. In the lecture notes, this is stated below Definition 2.1. It is said to be a consequence of the subspace theorem (Thm. 1.2). Let me do the argument in extreme detail (sorry).

If we want to show that span(W) is a subspace of vector space V, Theorem 1.2 tells us that there are three things to check:

1. The sum of two vectors in span(W) is again a vector in span(W).

   Since span(W) is the set of all linear combinations of vectors in W, this states the hopefully obvious fact that if you add two linear combinations of vectors in W you again get a linear combination of vectors in W.

   In precise mathematical terms: let x and y lie in span(W): they are linear combinations of vectors in W. For instance, we may write

   $$x = \alpha_1 w_1 + \cdots + \alpha_m w_m \qquad \text{and} \qquad y = \beta_1 \bar{w}_1 + \cdots + \beta_n \bar{w}_n \tag{1}$$

   for vectors $w_1, \dots, w_m, \bar{w}_1, \dots, w_n$ in W and scalars $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_n$. This notation is meant to stress two things: we only know – by definition of span(W) — that x and y can be written as linear combinations of vectors in W, but which vectors (w_i's for x and $\bar{w}_i$'s for y) and how many of them (m for x and n for y) may well be different.

   Then
   $$x + y = \alpha_1 w_1 + \cdots + \alpha_m w_m + \beta_1 \bar{w}_1 + \cdots + \beta_n \bar{w}_n$$
   is a linear combination of the vectors $w_1, \dots, w_m, \bar{w}_1, \dots, \bar{w}_n$ in W: it lies in span(W) as desired.

2. If a vector lies in span(W), then so does every scalar multiple of this vector.

   The argument is pretty similar: let x lie in span(W) and let α be a scalar. To see that αx lies in span(W), write x as a linear combination of vectors in W as in (1). Then

   $$\alpha x = \alpha(\alpha_1 w_1 + \cdots + \alpha_m w_m) = (\alpha\alpha_1) w_1 + \cdots + (\alpha\alpha_m) w_m$$

   shows that αx is again a linear combination of vectors $w_1, \dots, w_m$ in W: it lies in span(W) as desired.

3. Finally, the zero vector $\mathbf{0}$ of V must lie in span(W).

   Well, take any vector w in W. We just showed that span(W) contains all scalar multiples of this vector. Picking scalar $\alpha = 0$, vector $0w$ lies in span(W). But this is precisely the zero vector, since $0w = \mathbf{0}$ (see Thm. 1.1).

Having verified all properties in Thm. 1.2, we conclude that span(W) is indeed a linear subspace of V.

Lecture 3

- ☒ Notes: sections 5, 6. Skip footnote 1 on page 20: we won't talk about complex numbers in this course.
- ☒ Carter: pp. 114–118 ('Normed Linear Spaces' until Exercise 1.200), 290–291
- ☒ Sydsæter et al.: pp. 6 ('Vectors')

For this lecture, the big picture to keep in mind is that in $\mathbb{R}^2$ we have a standard notion of how to measure the length of a vector (see p. 17) by means of Pythagoras' theorem and this notion of length easily extends to vectors in $\mathbb{R}^n$. We related that to the properties of inner products and proved that (1) the way we usually define the inner product in $\mathbb{R}^n$ has a number of nice properties; these properties (I1) to (I4) define an inner product space; (2) the properties of an inner product imply some nice properties of the length/norm of a vector; these properties (N1) to (N4) define a normed vector space.

Proving those properties wasn't entirely straightforward; I do slightly different proofs in my lecture notes. It is my task to prove such things, essentially so that you don't have to...

Since I did not manage to do so in class, I wrote some more extensive remarks about a few examples from the notes; see the next two pages. In it, I refer to cities with a rectangular grid of roads. Here is an example from Athens:

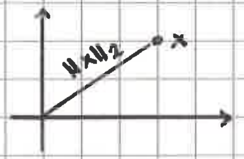
Next class I will start talking about how to measure distance; in mathematics distance functions are called metrics. I will also include in the study guide an overview of how three concepts (inner products, norms, metrics) relate to each other after next class.

I didn't have time to talk about other norms than the Euclidean norm. Here are a few others:

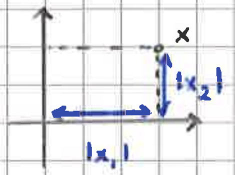
Taxicab norm

Inspired by some cities having roads in two directions only: north-south and east-west. I include a picture from Athens below.

The Euclidean norm involves drawing a straight line from the origin to x and computing its length:



But you cannot travel along such a straight line if the map of your city has a rectangular grid that allows you to move only horizontally or vertically. In that case, to move from the origin to $x = (x_1, x_2)$ you need to travel x_1 steps horizontally (or $-x_1$ if x_1 happens to be negative, i.e., $|x_1|$ steps, the absolute value of x_1) and $|x_2|$ steps vertically for a total distance of $|x_1| + |x_2|$ steps:



the length of a trip from the origin to $x \in \mathbb{R}^2$ is

$$|x_1| + |x_2|$$

In n dimensions this generalizes to

$$|x_1| + |x_2| + \cdots + |x_n|$$

This norm is often denoted by $\|x\|_1$; see Ex. 5.3 on p.18 ◁

Supremum norm

If, as above, you can only travel horizontally or vertically, what is the worst-case/maximal distance you need to travel in any direction to reach $x = (x_1, x_2)$?

Well, you travel $|x_1|$ units horizontally, $|x_2|$ units vertically, so the maximal number of steps in any direction is

$$\max\{|x_1|, |x_2|\}$$

steps. In n dimensions this generalizes to

$$\max\{|x_1|, \ldots, |x_n|\}$$

This norm is often denoted by $\|x\|_\infty$; see Ex. 5.4 on p.18 ◁

Numerical examples

Compute $\|x\|_2$, $\|x\|_1$ and $\|x\|_\infty$ if $x = (-3, 6, 4)$

Answer: $\|x\|_2 = \|(-3, 6, 4)\|_2 = \sqrt{(-3)^2 + 6^2 + 4^2} = \sqrt{9 + 36 + 16} = \sqrt{61}$

$\|x\|_1 = \|(-3, 6, 4)\|_1 = |-3| + |6| + |4| = 3 + 6 + 4 = 13$

$\|x\|_\infty = \|(-3, 6, 4)\|_\infty = \max\{|-3|, |6|, |4|\} = \max\{3, 6, 4\} = 6$

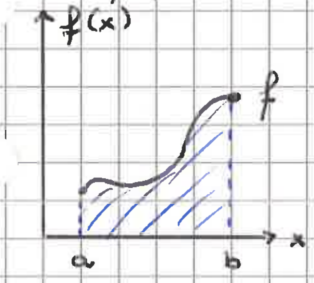
So far we looked at lengths in $\mathbb{R}^n$: how far do you need to travel from the zero vector to reach some $x \in \mathbb{R}^n$.

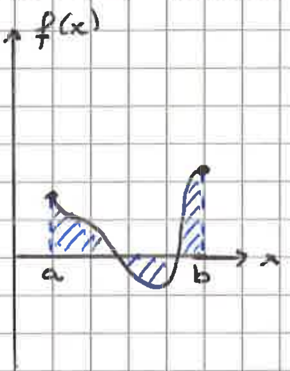
The next examples are about functions. Remember: also functions can be treated as special cases of vectors (Ex. 1.6 on p.3).

Again, the central question, now given some continuous function $f : [a,b] \to \mathbb{R}$, is how to define some intuitive notion of the "length" between the zero vector and f.

Norms of functions

In $C[a,b]$, the set of continuous functions $f : [a,b] \to \mathbb{R}$, the zero vector is just the function that is constant and equal to 0 for each $x \in [a,b]$: its graph lies on the horizontal axis. One way to measure how far f is away from the horizontal axis is to measure the area between them: if f lies close to the horizontal axis, that area is small. That area in mathematical terms is just the integral of $|f|$:





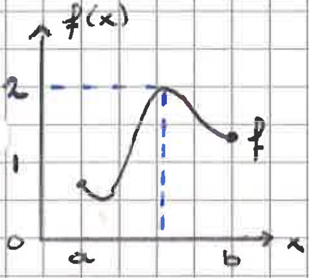
$$\int_a^b |f(x)| \, dx, \qquad (*)$$

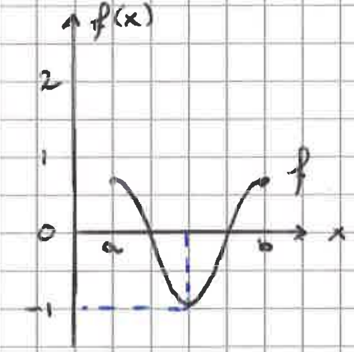
where we need the absolute value simply to make sure that the area under the horizontal axis adds (instead of subtracts) to the area between the horizontal axis and the graph of f. This norm $(*)$ is often denoted $\|f\|_1$; see Ex. 5.8 on p. 19

Another way would be to measure how far from 0 the function values are at most:

$$\max \{ |f(x)| : x \in [a,b] \} \qquad (**)$$

In the figures below, that happens to be 2 and 1 units, respectively.





Again, if that number is small, f intuitively lies close to the horizontal axis. The norm $(**)$ is often denoted $\|f\|_\infty$ and analogous to an earlier example in $\mathbb{R}^n$ referred to as the supremum norm; see Ex 5.7 on p. 19

Lecture 4

- ☒ Notes: section 6, 7, 8.1 and 8.2. Skip section 8.3.
- ☒ Carter: pp. 45–50.
- ☒ Sydsæter et al.: pp. 465–467 (‘Point Set Topology in $\mathbb{R}^n$’, the part about open and closed sets), 513.

I finished the discussion about lengths and distances. Then used balls to define some different properties of sets (there are a few more on pages 29 and 30) and proceeded to continuity of functions. I separately (see below) give an overview of the big picture from inner product spaces to normed vector spaces to metric spaces together with some practical motivation. It might be nice to have all that in one place. Also, I promised during yesterday’s Q and A to answer a question about polynomials.

This is fairly common calculus material. The definition of a *closed* set in mathematics is peculiar and you will have to take it for granted: a set U in metric space (X,d) is called *closed* if its complement $U^c = X \setminus U = \{x \in X : x \notin U\}$ is open. Recall that the complement of the set U simply consists of whatever elements of X that do *not* belong to U. This is good to keep in mind, since it is one of the most common mistakes in problem sets and exams: as opposed to doors, which are either open or closed, in metric spaces (X,d) there may be sets which are neither open nor closed and sets (like $\emptyset$ and X) which are both open and closed.

Figure 1 gives some examples of subsets of $\mathbb{R}$ with its usual distance. For instance, $[0,1)$ is not open: 0 belongs to the set, but is not an interior point. Likewise, $[0,1)$ is not closed, since its complement is $(-\infty,0) \cup [1,\infty)$, which is not open: 1 belongs to this set, but is not an interior point.

set	is
$(0,1)$	open, not closed
$[0,1]$	not open, but closed
$[0,1)$	neither open nor closed
$\mathbb{R}$	both open and closed

Figure 1: Exampes involving open and closed sets in $\mathbb{R}$.

I agree that checking continuity using for instance the (ε,δ)-definition (expression (22) is tedious. I told you not to spend too much time on it: there is a relatively easy exercise on this in the next problem set, but it won’t be on the exam. Why? This is an MSc level course: roughly speaking, we don’t want to know if one particular function happens to be continuous, we want to prove and understand general theorems that hold for *all* continuous functions. Examples 8.3 to 8.6 do some important cases like linear or affine functions, sums, products, quotients. Together with the result (Thm. 8.3) that compositions of continuous functions are continuous, this allows you to build up complicated continuous functions from simpler ones using addition, multiplication, division (whenever this is not by zero), etc: that is the main message of Theorem 8.4. For instance, since $f(x) = x^2$ and $g(x) = x^4 + 1$ are continuous, so are the sum $(f+g)(x) = x^2 + x^4 + 1$, the fraction $(f/g)(x) = x^2/(x^4+1)$, the composition $(f \circ g)(x) = f(g(x)) = (x^4+1)^2$, etc.

So in much of the literature, you will see sentences like ‘As a composition of continuous functions, ... is continuous.’ This is a useful skill to have: it allows you, in practice, to avoid doing complicated continuity proofs with messy ε’s and δ’s.

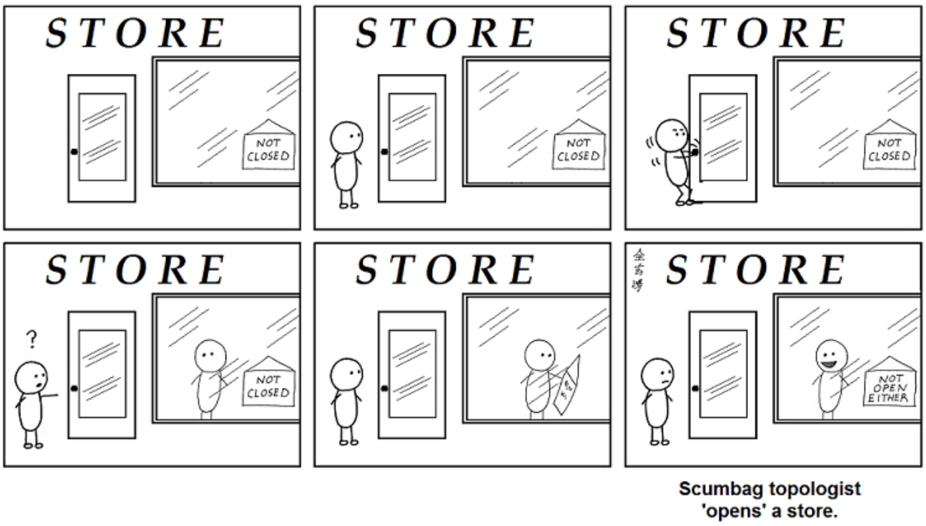
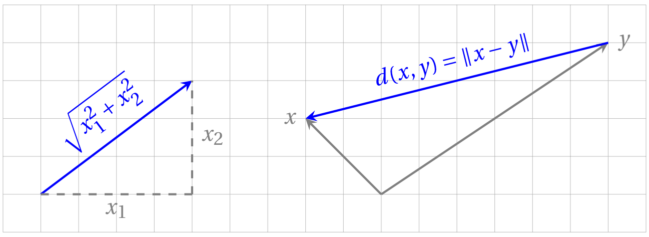


Figure 2: How to irritate people with topology.

Overview: inner products, norms, metrics

Here's the big picture you might want to keep in mind. Using Pythagoras' Law, the length of a vector $x = (x_1, x_2) \in \mathbb{R}^2$ is defined as $\|x\| = \sqrt{x_1^2 + x_2^2}$ and the distance between two vectors x and y is the length of their difference: $d(x, y) = \|x - y\|$.



This is straightforwardly extended to vectors in $\mathbb{R}^n$: we can define the length of $x = (x_1, \ldots, x_n)$ as

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2} \tag{2}$$

and the distance between x and y as the length of their difference:

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{3}$$

This particular socalled ***Euclidean norm*** and ***distance*** are often denoted with a subscript 2 — $\|x\|_2$ and $d_2(x, y)$ — to remind you about the squares in these expressions. Expression (2) uses the inner product of two vectors x and y in $\mathbb{R}^n$, defined as

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i. \tag{4}$$

In particular,

$$\|x\| = \sqrt{\langle x, x\rangle}.$$

For instance, if $x = (2, -4, 3)$ and $y = (1, 2, -6)$, their inner product is

$$\langle x, y\rangle = 2 \cdot 1 + (-4) \cdot 2 + 3 \cdot (-6) = 2 - 8 - 18 = -24,$$

the length of x and y is

$$\begin{aligned} \|x\| &= \sqrt{2^2 + (-4)^2 + 3^2} = \sqrt{4 + 16 + 9} = \sqrt{29}, \\ \|y\| &= \sqrt{1^2 + 2^2 + (-6)^2} = \sqrt{1 + 4 + 36} = \sqrt{41}, \end{aligned}$$

and the distance from x to y is

$$d(x, y) = \|x - y\| = \|(2 - 1, (-4) - 2, 3 - (-6))\| = \|(1, -6, 9)\| = \sqrt{1^2 + (-6)^2 + 9^2} = \sqrt{1 + 36 + 81} = \sqrt{118}$$

Here we have three notions:

1. an ***inner product***,
2. a notion of length or ***norm***,
3. a distance function or ***metric***,

and the relations between them are summarized below:

$$\begin{tikzcd} \text{inner product space} \arrow[r, "\|x\| = \sqrt{\langle x, x\rangle}"] & \text{normed vector space} \arrow[r, "d(x, y) = \|x - y\|"] & \text{metric space} \end{tikzcd}$$

The basic idea is to generalize the notions of inner product, norm, and metric as follows:

1. We realize that the inner product as defined in (4) has a bunch of nice properties referred to as (I1) to (I4); we now give the name ***inner product*** to *any* function satisfying these properties; see Definition 5.3.
2. Similarly, the length/norm defined in (2) has nice properties referred to as (N1) to (N4); we now give the name ***norm*** to *any* function satisfying these properties; see Definition 5.1.
3. Similarly, the distance/metric defined in (3) has nice properties referred to as (D1) to (D4); we now give the name ***metric*** to *any* function satisfying these properties; see Definition 6.1.

In the notes you will find roughly a dozen particular examples. For detailed proofs that some important examples are norms, you can check the solutions manual to exercise 5.3.

Why would we at all be interested in measuring things like length/distance? There are evident physical reasons (how far is it from your home to the Stockholm School of Economics?), but also more technical ones: in mathematics, notions like continuity, differentiability, compactness, boundedness, etc. are often defined in terms of length and distance. And for a very concrete economic application, you will encounter this in ordinary least squares estimation, where the goal is to find a linear combination of vectors of exogenous/independent/explanatory variables that is *as close as possible* to a vector of endogenous/dependent/... variables. And the 'squares' in least squares estimation means that you're using the distance function defined above in (3)!

I also recommend that you take a quick look at Exercise 6.10, which shows how metric spaces are used to provide online customer recommendations.

In real life, you often use the triangle inequality to find estimates of how far away things are. Suppose you are on a trip, look out of your car window and see the sign in Figure 3. How far is it between Köln and Dortmund? Well, we're not sure, but driving from Köln to your current location is 106 km, and from your current location to Dortmund takes 24 km. So taking into account that this may very well be a detour, the distance can't be more than $106 + 24 = 130$ km.

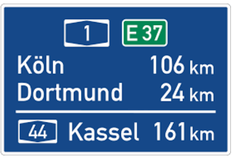


Figure 3: A traffic sign

For those of you who try Exercise 5.2, you might like the comic below. Oh, believe me, I'm aware that math humor is at best an acquired taste, at worst a *contradictio in terminis*. Try to see such interludes for what they are: my way of trying to convince you that there is an actual human being behind this course. If you're the artistic type and have comics, caricatures, collages, photo impressions, aquarelles of study group members banging their heads against nearby walls in frustration over the problem sets, or other things pertaining to this course that might embellish the study guide or lecture notes and make them a little less arid, please send them to me!

Figure 4: The struggle to reverse triangle inequality

Which are the most important norms/metrics? On $\mathbb{R}^n$ it is the Euclidean norm from Example 5.2; for functions, it is the supremum norm from Example 5.7. These are so common that not everybody even bothers to mention that they are using them; see the convention at the bottom of page 24.

Exercises 6.1 to 6.4 are really good practice. They also help to convey that balls can have different shapes (including a square) depending on what metric you use to measure how far points are away from each other.

A question about polynomials

During the first Q-and-A session I promised to answer a question about polynomials in the study guide; so here we go:
QUESTION: Why is the set of polynomials $\{1, x, x^2, x^3, x^4, \dots\}$ a basis of the vector space of polynomials?
ANSWER: For background, this is claimed in Example 2.7; the vector space of polynomials itself is in Example 1.7. If needed, read those first to refresh your memory!

According to the definition of a basis (Def. 2.3) there are two things to show:

1. That the set of polynomials $\{1, x, x^2, x^3, x^4, \dots\}$ spans the vector space of all polynomials, i.e., that each polynomial is a linear combination of $1, x, x^2, x^3, x^4, \dots$. Well, by definition a polynomial is of the form
$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$
for some nonnegative integer n and coefficients $a_0, a_1, \dots, a_n$ in $\mathbb{R}$. This is precisely a linear combination of $1, x, \dots, x^n$ with scalars $a_0, a_1, \dots, a_n$, respectively.

2. That the set of polynomials $\{1, x, x^2, x^3, x^4, \dots\}$ is linearly independent. To verify this, we must show (Def. 2.2) that if we have an arbitrary linear combination
$$a_0 + a_1 x + \cdots + a_n x^n$$
of finitely many such polynomials that equals the zero vector which here is the special polynomial
$$\mathbf{0}(x) = 0 + 0x + 0x^2 + 0x^3 + \cdots = 0,$$
then all scalars a_i must be zero. Well, by definition (Ex. 1.7), if two polynomials are equal, then equal powers must have equal coefficients. The coefficients of the first expression are $a_0, a_1, \dots$, and those of the zero polynomial are $0, 0, \dots$, so it follows that $a_0 = a_1 = \cdots = 0$, as we needed to show.

Lecture 5

- ☒ Notes: Sections 9.1 (skip 9.2) and 10 (last thing I covered was Thm 10.1). You should definitely skip — unless you are specifically interested — the proof of Theorem 10.3.
- ☒ Carter: pp. 56–65 (the part of section 1.3.2 on sequences).
- ☒ Sydsæter et al.: sections 13.2 (‘Topology and Convergence’) and A.3 (‘Sequences of Real Numbers’).

Establishing that sequences converge somewhere is not always easy. Have a look at exercise 9.1 and the solution in the back, where I try to give more hands-on advice on how to prove such things in elementary examples. Remember, I sometimes give intricate proofs so that you don’t have to: I try to convey the big picture. The important thing is to keep track of the results and use them to your advantage in solving exercises.

Theorem 9.3 is often useful in computing more complicated limits. It implies, for instance, that if sequence $(x_k)_{k\in\mathbb{N}}$ converges to x and function f is continuous at x, then $f(x_k)$ converges to $f(x)$. Now, the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^5$ is continuous, so if x_k converges to x, then $f(x_k) = x_k^5$ converges to x^5!

We argued that each convergent sequence is a Cauchy sequence, and that the converse is true in $\mathbb{R}^n$ with its usual Euclidean distance. Spaces where this converse is true (each Cauchy sequence has a limit) are called ***complete***. In the next lecture we will use (Cauchy) sequences and completeness to prove the

Banach contraction theorem, a fixed-point theorem that is important in mathematics and economics for many reasons. Our main application towards the end of the course will concern the existence of optimal policies in dynamic optimization, one of the crucial results in dynamic macroeconomics. This also explains the need for talking about function spaces like $(B(X,Y), d_\infty)$ and $(C(X,Y), d_\infty)$ in Theorem 10.3, of which I only discussed the special case mentioned in Example 10.3, because economic policies are functions: they translate the current state of the economy into an action.

Lecture 6

- ☒ Notes: Sections 11 and 13 (skip section 12, skip part (b) of Theorem 13.6).
- ☒ Carter: 238–241, 218–220, 61–65 (compactness).
- ☒ Sydsæter et al.: section 14.3 ('Fixed Points for Contraction Mappings'), 106–107 (Extreme value theorem), 474, 477 (compactness).

The first part of the lecture discussed the Banach contraction theorem, a result about fixed points. Many problems in economics formulate equilibria or optima as fixed points. It is used in Theorem 23.2 and the value iteration algorithm to characterize and find solutions to dynamic optimization problems. Another economic application is the best-reply dynamic in Cournot oligopoly, where firms repeatedly choose optimal quantities in response to the ones in the previous period. Under common assumptions, this process is a contraction and the quantities converge over time to those in the Nash equilibrium. It is also very useful for approximating solutions to difficult equations; Example 11.5 illustrates this. In Section 11.1 I show how it is used to rank the importance of webpages. That section is just a cute application, but will not be on the exam.

In the second part of the lecture we discussed the notion of compactness. I am aware that compactness in terms of coverings is a difficult thing to grasp, but I hope the umbrella simile conveys some of its intuition. And as usual, the important thing is to be able to use the theorems rather than being able to prove long and tricky ones.

Analysts like to pull things apart into infinitely many pieces and see if a finite subcollection can represent the original thing.
I tried that with a bike once...
now he has a scooter.
INTRODUCTION TO MATH ANALYSIS.
2006
© COURTNEY GIBBONS

The Heine-Borel theorem, that tells you that a set in $\mathbb{R}^n$ with its usual distance is compact if and only if it is closed and bounded, is very useful. Read through the theorems in section 13; forget about

Theorem 13.6(b) entirely. Theorem 13.4 is good to know, but I won't harass you with the proof. The proofs of Theorem 13.1 are short and practice with simple applications of the definition of compactness, so read through at least some of them to get a grasp of how this type of thing is established.

In specific economic models, you can often solve for optima/equilibria explicitly (see Paul's course). In more general, abstract models, you cannot do this and a crucial first step is to prove that the problem you study actually *has* an optimal solution. The Extreme Value Theorem provides sufficient conditions for this: if the goal function is continuous and the set of feasible alternatives is nonempty and compact, a maximum and a minimum of the goal function exist. Typically you will be interested in only one of these: cost minimization, profit maximization, etc.

In exercises, it is good to keep the Heine-Borel theorem in mind, which says that in $\mathbb{R}^n$ *with its usual distance*, a set is compact if and only if it is closed and bounded. This is not true if you give $\mathbb{R}^n$ a different metric: Example 13.2 (applied to $X = \mathbb{R}^n$) illustrates this for the discrete metric.

A left-over question from the Q-and-A session

I promised to include in the study guide a proof that the set $U = \{x \in \mathbb{R}^2 : x_1 < 0\}$ is open.

METHOD 1: I will prove explicitly that each element $x \in U$ is an interior point. So let $x \in U$. Since $x_1 < 0$, it follows that $-x_1 > 0$. Choose radius $\varepsilon = -x_1$; I will prove that the entire ball around x with radius ε lies in U, i.e., that $B(x, \varepsilon) \subseteq U$.

To see this, let $y \in B(x, \varepsilon)$. We need to establish that $y \in U$, i.e., that $y_1 < 0$. Suppose, to the contrary, that $y_1 \geq 0$; we will derive a contradiction. If $y_1 \geq 0$, then $y_1 - x_1 \geq 0 + \varepsilon = \varepsilon$. Hence also

$$d_2(y, x) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2} \geq \sqrt{\varepsilon^2 + 0} = \varepsilon.$$

This contradicts our choice of $y \in B(x, \varepsilon)$: by definition, $d_2(y, x) < \varepsilon$.

METHOD 2: Function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x) = f(x_1, x_2) = x_1$ is linear, hence continuous (Ex. 8.3). The interval $(-\infty, 0)$ of real numbers smaller than zero is open. Therefore (Thm. 8.2) its preimage

$$f^{-1}((-\infty, 0)) = \{x \in \mathbb{R}^2 : f(x) \in (-\infty, 0)\} = \{x \in \mathbb{R}^2 : x_1 < 0\} = U$$

is open.

Lecture 7

- ☒ Notes: sections 14, 15.1 (you may skip 15.2), 16.
- ☒ Carter: 88–94, 98, 104–108, 125, 306–317.
- ☒ Sydsæter et al.: sections 2.2 and 13.5 (both with title 'Convex Sets').

Convex sets

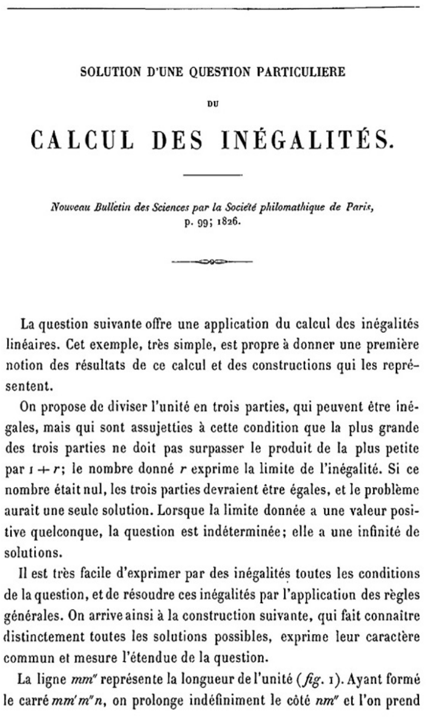
This lecture was about convex sets, sets that contain the line segment between each pair of its points. We mentioned many examples:

1. polytopes;
2. polyhedra;
3. finitely generated cones;

4. hyperplanes;
5. halfspaces;
6. and a way of extending a nonconvex set to a convex one by adding all convex combinations of its elements: the convex hull of a set.

There are a few more in the notes and it is good to familiarize yourself with these examples: you can then refer to them to argue that certain specific sets are convex.

Fourier-Motzkin elimination



SOLUTION D'UNE QUESTION PARTICULIERE

DU

CALCUL DES INÉGALITÉS.

Nouveau Bulletin des Sciences par la Société philomathique de Paris,
p. 99; 1826.

La question suivante offre une application du calcul des inégalités linéaires. Cet exemple, très simple, est propre à donner une première notion des résultats de ce calcul et des constructions qui les représentent.

On propose de diviser l'unité en trois parties, qui peuvent être inégales, mais qui sont assujetties à cette condition que la plus grande des trois parties ne doit pas surpasser le produit de la plus petite par $1 + r$; le nombre donné r exprime la limite de l'inégalité. Si ce nombre était nul, les trois parties devraient être égales, et le problème aurait une seule solution. Lorsque la limite donnée a une valeur positive quelconque, la question est indéterminée; elle a une infinité de solutions.

Il est très facile d'exprimer par des inégalités toutes les conditions de la question, et de résoudre ces inégalités par l'application des règles générales. On arrive ainsi à la construction suivante, qui fait connaître distinctement toutes les solutions possibles, exprime leur caractère commun et mesure l'étendue de la question.

La ligne mm'' représente la longueur de l'unité (*fig.* 1). Ayant formé le carré $mm'm''n$, on prolonge indéfiniment le côté nm'' et l'on prend

Figure 5: The first page of Fourier's note from 1826 on elimination for linear inequalities

A polyhedron is the set of solutions to a finite system of linear inequalities. We saw that Fourier-Motzkin elimination is a simple method of solving such systems by eliminating variables; it is the sibling of Gaussian elimination for systems of linear equations. Fourier-Motzkin elimination gives you a new polyhedron, i.e. a new system of linear inequalities, but now with fewer variables, that has a solution if and only if the original system has a solution.

For the historically curious I include the first page of Fourier's note from 1826 on what is now called Fourier-Motzkin elimination (Motzkin rediscovered the method in his 1936 dissertation). He formulates a specific problem, namely "to divide the unit into three parts, which may be distinct, but which are subject to the condition that the largest of the three parts does not exceed the product of the smallest with $1 + r$".

So, given a parameter $r \geq 0$ and denoting the 'three parts' as x_1, x_2, x_3, he considers the (in)equalities

$$x_1 + x_2 + x_3 = 1 \qquad \text{and} \qquad \max\{x_1, x_2, x_3\} \leq (1+r)\min\{x_1, x_2, x_3\}.$$

Using that $\max\{a_1, \ldots, a_m\} \leq \min\{b_1, \ldots, b_n\}$ if and only if $a_i \leq b_j$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$, this can be written as a system of linear (in)equalities $x_1 + x_2 + x_3 = 1$ and

$$\begin{array}{lll} x_1 \leq (1+r)x_1 & x_2 \leq (1+r)x_1 & x_3 \leq (1+r)x_1 \\ x_1 \leq (1+r)x_2 & x_2 \leq (1+r)x_2 & x_3 \leq (1+r)x_2 \\ x_1 \leq (1+r)x_3 & x_2 \leq (1+r)x_3 & x_3 \leq (1+r)x_3 \end{array}$$

It's terribly messy to solve.

Farkas' lemma and separating hyperplanes

I also discussed Farkas' lemma (section 15) and separating hyperplanes (section 16). I recommend Exercise 16.2 — in addition to my banana-fruitfly example — which illustrates via a number of examples why convexity plays a crucial role in separation. I give one particular application below, to the existence of stationary distributions for certain stochastic processes. This is just an illustration and not obligatory reading.

The practical motivation for treating these results is that whole branches of economics arise as specific applications of such separating hyperplane theorems or variants of Farkas' Lemma. These include:

1. Zero-sum game theory. See, for instance, chapter 2 of González-Díaz, J., García-Jurado, I., and Fiestras-Janeiro, G., 2010, An Introductory Course on Mathematical Game Theory, American Mathematical Society, and page 2 of my book review at

   `http://dx.doi.org/10.1016/j.geb.2010.12.006`

   on why this class of games is important.

2. Mechanism design. See, for instance, Vohra, R.V., 2011, Mechanism Design: A Linear Programming Approach, Cambridge University Press.

3. Input-output analysis. See, for instance, Ten Raa, T., 2005, The Economics of Input-Output Analysis, Cambridge University Press.

4. Discrete arbitrage theory. See, for instance, Kallio, M., Ziemba, W.T., 2007, Using Tucker's theorem of the alternative to simplify, review and expand discrete arbitrage theory, *Journal of Banking & Finance* 31, 2281–2302; `http://dx.doi.org/10.1016/j.jbankfin.2007.02.004`.

5. Also the 'second fundamental welfare theorem' of Walrasian equilibrium theory is often proved using a separating hyperplane theorem. See, for instance, Debreu, G., 1959, The Theory of Value, Yale University Press.

In general, I want you to know the theorems in these sections. I will not harass you with their proofs. Have a quick look at Gordan's theorem as well: it will be our main tool when we find optimality conditions for static optimization problems.

Farkas application: stationary distributions of Markov chains

Let $n \in \mathbb{N}$. The ***unit simplex***

$$\Delta_n = \{x \in \mathbb{R}^n : x \geq \mathbf{0}, x_1 + \cdots + x_n = 1\}$$

consists of all probability vectors in $\mathbb{R}^n$: vectors whose coordinates are nonnegative and add up to one. A square matrix $A \in \mathbb{R}^{n \times n}$ is a ***stochastic matrix*** if each column (or each row, depending on an arbitrary choice of direction) is a probability vector. In the theory on Markov chains, stochastic matrices are used to describe transition probabilities: a_{ij} is the probability (notice the order) of moving from state j in the current period ('today') to state i in the next period ('tomorrow'). For instance, suppose there are two states, state 1 being good weather and state 2 being bad, and the transition probability matrix is

$$\begin{bmatrix} \frac{4}{5} & \frac{1}{3} \\ \frac{1}{5} & \frac{2}{3} \end{bmatrix}.$$

The first column says that if the weather is good today, the probability of the weather being good tomorrow is 4/5 and the probability of the weather being bad is 1/5. The second column is interpreted likewise.

If a probability vector $x \in \Delta_n$ specifies the probability of being in any of the n states today, then Ax is the probability distribution over the states tomorrow. Observe that Ax indeed lies in Δ_n: it is a convex combination of the columns of A. Since each column of A lies in the convex set Δ_n, so does their convex combination.

We call $x \in \Delta_n$ a ***stationary distribution*** if the distribution over states remains unchanged over time: $Ax = x$. We now prove that *each stochastic matrix has a stationary distribution.*

A stationary distribution is a nonnegative solution x to $Ax = x$ whose coordinates sum to one (probabilities!), i.e., a nonnegative solution to

$$\begin{bmatrix} a_{11}-1 & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn}-1 \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{5}$$

We show that such a solution exists via Farkas' Lemma, arguing that the second alternative in that lemma cannot hold. This second alternative says that there is a vector $y = (y_1, \ldots, y_n, y_{n+1})$ whose inner product with each column of the matrix in (5) is nonnegative:

$$\begin{aligned} (a_{11}-1)y_1 + a_{21}y_2 + \cdots + a_{n1}y_n + 1y_{n+1} &\geq 0, \\ a_{12}y_1 + (a_{22}-1)y_2 + \cdots + a_{n2}y_n + 1y_{n+1} &\geq 0, \\ &\vdots \\ a_{1n}y_1 + a_{2n}y_2 + \cdots + (a_{nn}-1)y_n + 1y_{n+1} &\geq 0, \end{aligned} \tag{6}$$

but whose inner product with the vector on the right side of (5) is less than zero:

$$0y_1 + \cdots + 0y_n + 1y_{n+1} = y_{n+1} < 0.$$

Let y_k be the largest number among $y_1, \ldots, y_n$. Since the numbers in the k-th column of A are probabilities (i.e., nonnegative with sum one), we find that

$$y_k = a_{1k}y_k + \cdots + a_{nk}y_k \geq a_{1k}y_1 + \cdots + a_{nk}y_n.$$

But rewriting the k-th inequality in (6) gives the opposite:

$$y_k \le a_{1k}y_1 + \cdots + a_{nk}y_n + 1\underbrace{y_{n+1}}_{<0} < a_{1k}y_1 + \cdots + a_{nk}y_n.$$

So the second alternative in Farkas' Lemma has no solution: a stationary distribution exists.

Notational reminder

We have seen a variety of different names for special kinds of linear combinations of vectors. This might be a good opportunity to remind you that an expression of the form

$$\lambda_1 v_1 + \cdots + \lambda_m v_m$$

is called:

- ☒ a ***linear combination*** of $v_1, \ldots, v_m$ if the scalars $\lambda_1, \ldots, \lambda_m$ are arbitrary real numbers,
- ☒ a ***convex combination*** of $v_1, \ldots, v_m$ if the scalars $\lambda_1, \ldots, \lambda_m$ are nonnegative and add up to one:

$$\lambda_1, \ldots, \lambda_m \ge 0, \qquad \lambda_1 + \cdots + \lambda_m = 1,$$

- ☒ a ***nonnegative combination*** of $v_1, \ldots, v_m$ if the scalars $\lambda_1, \ldots, \lambda_m$ are nonnegative:

$$\lambda_1, \ldots, \lambda_m \ge 0.$$

Reminder: differentiability prerequisites

The file of prerequisites that the MSc program director to send to you at the time of admission and that is also available under the downloads on the courseweb, contains among other things the following list of required knowledge:

- ☒ continuous and differentiable functions of one real variable
- ☒ elementary rules of differentiation for functions like $\sin x$, e^x, x^n, $\ln x$, ...
- ☒ differentiation of the sum, difference, product, and quotient of differentiable functions, the chain rule
- ☒ partial derivatives

Since we will be doing a lot of differentiation in the next couple of lectures, this would be the right time to review that material once more. In our recommended books, short reviews can be found in:

- ☒ Carter: sections 4.1 – 4.3.1
- ☒ Sydsæter et al.: sections 2.1 ('Gradients and Directional Derivatives') and 2.9 ('Differentiability')

Lecture 8

- ☒ Notes: section 17. As a general rule, you can always skip sections of postponed proofs. Here I recommend reading selectively. What you need most is spelled out in the introductory remarks of section 17; you can skip the rest. And the only things you really need about differentiability are what I discussed in class. Section 18 gives a lot more detail, but you do not need to read it. To follow the remainder of the course, you only need what I discussed during the second half of today's lecture.
- ☒ Carter: 323–343 (convex functions and variants)
- ☒ Sydsæter et al.: sections 2.3 and 2.4 ('Concave and Convex Functions')

After class someone sent me a meme:

Lecture 9

☒ Notes: section 19 (always skip the postponed proofs). Do exercise 19.1!

☒ Carter: chapter 5 (static optimization).

☒ Sydsæter et al.: chapter 3 (static optimization).

I went in detail through the development of optimality conditions for maximization problems with inequality constraints. The big picture is this:

1. If x^* is a maximum, you cannot find a direction in which to move from x^* that leads to feasible points with a higher function value.

2. Rewriting that in terms of linear (in)equalities and using Gordan's theorem, this means that there must be a solution to another system of linear (in)equalities, the so-called Fritz John (FJ) conditions.

3. Splitting the Fritz John conditions into two cases, one where the multiplier of the goal function f is 0 and one where the multiplier of the goal function is 1, leads to other necessary conditions: the gradients of the binding constrains are linearly dependent or the Karush-Kuhn-Tucker conditions must hold.

This is the theory behind the necessary conditions for maxima, i.e., my job. And of course we need to treat that first before you can start applying it to concrete problems.

Having treated the case of maximization subject to inequality constraints in substantial detail, I will not go through the case of equality constraints and mixed (allowing both inequality and equality) constraints, the topic of Section 19.4. Especially problems with only equality constraints (Remark 19.1 on page 103) are particularly easy because the tricky complementary slackness conditions disappear. So you should read that section (quickly) yourself and then move on to the worked examples in sections 19.5 and 19.6 to get some hands-on experience.

My goal for today was to explain the theory; in terms of practical skills for this section, you should be able to:

☒ find candidate optima in the interior of the feasible set by setting the partial derivatives of the goal function equal to zero, the standard first-order conditions (Theorem 19.1);

☒ find candidate optima in problems with inequality constraints using the Fritz John or Karush-Kuhn-Tucker conditions;

☒ and likewise for problems with equality and inequality constraints (not in class; read yourself);

☒ argue whether the candidates you found indeed are maximum locations using, for instance, the Extreme Value Theorem or Theorems 19.5 and 19.9.

My notes, as well as the recommended literature and the old problem sets and exams contain numerous worked examples.

A few years ago I found that someone with an unhealthy preoccupation with Fritz John (FJ) conditions had scratched this on the wall of a culture center near my home:

Lecture 10

☒ Notes: section 20 and pages 136 and 137 of section 23. You may skip section 22 (I will do 21 later).

☒ Carter: none.

☒ Sydsæter et al.: section 12.1 ('Dynamic Programming')

I introduced the standard form of a dynamic optimization problem. One way to solve a finite-horizon problem is the dynamic programming algorithm (DPA): you start in the final period and work backward, at each stage using that you already know what is optimal from the next period onward. The important intuition is on pages 119:

> Suppose you have come to the final period, $t = T$, and state $x(T) = x$. You don't have to worry about future consequences: the only part of the goal function you can still affect is the term $f(T, x, u)$ by choosing a control $u \in U(T, x)$ that makes it as large as possible. So $J_T(x)$ should satisfy[1]
>
> $$J_T(x) = \sup_{u \in U(T,x)} f(T, x, u).$$
>
> Now let $s \in \{0, \dots, T-1\}$ and suppose you already figured out what is optimal in tail problems starting at time $s+1$. This helps you to decide what is optimal at time s. Suppose you find yourself in state $x(s) = x$ at time s. Your feasible controls are those in $U(s, x)$. So choosing $u \in U(s, x)$ leads to instantaneous payoff $f(s, x, u)$ and a next state $x(s+1) = g(s, x, u)$. But by assumption, you know that the optimal value from the next state $x(s+1)$ onward is $J_{s+1}(x(s+1)) = J_{s+1}(g(s, x, u))$. So the best thing you can do is to optimize the sum of these two expressions: the value functions at time s and $s+1$ are related via the equation
>
> $$J_s(x) = \sup_{u \in U(s,x)} \big(f(s, x, u) + J_{s+1}(g(s, x, u))\big).$$

You must be able to use this algorithm to solve specific dynamic optimization problems. I will do an exercise in class next time.

I also introduced the infinite-horizon model and in the end argued that the optimal value function needs to satisfy a certain equality, the Bellman equation. Also this will be elaborated upon next time.

[1] Remember: $J_T(x)$ is the highest payoff you can get if at time T you find yourself in state x.

Lecture 11

- ☒ Notes: section 23 on the Bellman equation; skip page 141, although you may come across it in the macro courses.
- ☒ Carter: none.
- ☒ Sydsæter et al.: section 12.3.
- ☒ Sorger: section 5.5.

Intuition behind the Bellman equation

The ***Bellman equation*** for the infinite-horizon problem in section 23 (see Theorem 23.1) states that the optimal value function J satisfies

$$\text{for each } x \in X: \qquad J(x) = \sup_{u \in U(x)} \left\{ f(x,u) + \beta J(g(x,u)) \right\}. \tag{7}$$

The rough intuition is this:

- ☒ If you are in state x at time 0, the maximal payoff you can get, by definition, is $J(x)$, the left-hand side of the expression.
- ☒ If you are in state x at time 0 and you choose control u, you

    1. receive payoff $\beta^0 f(x,u) = f(x,u)$ at time $t = 0$;
    2. move to state $g(x,u)$ at time $t = 1$. Choosing optimally from that time onward will give you a maximal payoff of $J(g(x,u))$, but delayed by one period and consequently discounted by a factor $\beta^1 = \beta$. Thus, the sum of your future payoffs is $\beta J(g(x,u))$.

  Consequently, the best you can do at time $t = 0$ is to choose a control that maximizes the sum of your immediate payoff $f(x,u)$ and the discounted optimal payoff $\beta J(g(x,u))$. That is the right-hand side of the expression.

How to use this in practice

Assuming that instantaneous payoffs f are bounded, the right-hand side of the Bellman equation assigns to each function V in $B(X,\mathbb{R})$, the space of bounded real-valued functions on the state space, a new function in $B(X,\mathbb{R})$. Call this function $T(V)$:

$$T(V)(x) = \sup_{u \in U(x)} f(x,u) + \beta V(g(x,u)).$$

The Bellman equation then says that the optimal value function is a ***fixed point*** of this transformation:

$$J = T(J).$$

The mapping T turns out to be a contraction and $B(X,\mathbb{R})$ is complete, so the Banach contraction theorem assures that the value function J is the *unique* fixed point in $B(X,\mathbb{R})$. This gives us two tools for figuring out the solution of stationary discounted problems:

1. FINDING A CANDIDATE VALUE FUNCTION: we know from Banach's contraction theorem that function iteration converges to the fixed point. So if we start with an arbitrary bounded function $V_0 : X \to \mathbb{R}$ and successively compute $V_{k+1} = T(V_k)$, these approximations V_k of the value function will converge to the value function J. This is exactly the value iteration algorithm described in section 23.3.

In stylized examples and applications, a few iterations typically suffice to see some pattern in the shape of the functions V_k. So it seems an educated guess that the optimal value function is going to be of this shape as well!

2. VERIFYING YOUR CANDIDATE: Once you have an idea what the value function J ought to look like:

   ☒ substitute it on both sides of the Bellman equation;
   ☒ solve the optimization problem on the right-hand side;
   ☒ equate the coefficients to make sure that you have a fixed point.

   I illustrated this using a life-time consumption problem.

Where many things come together

This stuff about the Bellman equation ties together a lot of the material that we looked at earlier: we're talking about a (vector) space $B(X,\mathbb{R})$ of bounded functions on the state space X. We give $B(X,\mathbb{R})$ its usual metric d_∞ and show that the right-hand side of the Bellman equation defines a contraction. By Theorem 10.3, metric space $(B(X,\mathbb{R}), d_\infty)$ is complete. So we know from Banach's contraction theorem that we can use function iteration to generate a Cauchy sequence that converges to the fixed point we're interested in: the value function of the stationary infinite-horizon problem.

Movie time

You will doubtlessly be interested to know that Richard Bellman's grandson Gabriel Bellman, who kindly allowed me to use the photograph below for teaching purposes, made a documentary about the Bellman equation; you can find it on YouTube.

(the) BELLMAN equation
a documentary film by Gabriel Leif Bellman Screening in 2011
the bellman equation
THE EQUATION THAT CHANGED HIS WORLD.

Figure 6: Richard Bellman (left), the movie poster (middle) and DVD cover (right).

Lecture 12

☒ Notes: section 21.1 and 21.2; you can skip subsection 21.3.
☒ Carter: none.
☒ Sydsæter et al.: sections 12.4 ('The Maximum Principle') and 12.5 ('More Variables')

My aim was to drive home three key points:

1. Some dynamic optimization problems can be solved using static optimization tools.

2. Optimality conditions using the Lagrangian can also be found using the Hamiltonian, which is an easier-looking function.

3. Since the first and last period are special in any dynamic optimization problem, there will be separate optimality conditions for those.

I did so using a simplified case of our general finite-horizon problem, just to make the main points as easily as possible.

How should you read this section? For practical purposes, as little as possible. . . The main idea was presented in class, so there is no need to struggle through the proofs. A minimalist reading would be:

☒ Read the text on page 123 before subsection 21.1.

☒ On pages 125–127, start with Theorem 21.1 and read until (and including) Example 21.2, jumping over proofs if necessary.

☒ Section 21.2 treats more complicated cases which you can skip if you understand just one thing: in a maximum of problem (134) – (138), the Fritz John conditions must hold. And many of these conditions can be found using the Hamiltonian as well.

I would like to rewrite this section in the future, although I am not entirely sure how: no matter how you do it, it is going to involve a lot of messy notation because there are many time periods and many restrictions. So in the meantime, it is more than enough to just practice with it in practical examples.

The practical side

You can interpret a finite-horizon dynamic optimization problem as a special static optimization problem with a typically huge domain: in a problem with horizon T there are $T+1$ controls $u(0),\dots,u(T)$ and $T+1$ states $x(0),\dots,x(T)$, and if these lie in — say — the set of real numbers, we're maximizing over a $2(T+1)$-dimensional vector $(u(0),\dots,u(T),x(0),\dots,x(T))$. And if all the functions in our model are nicely differentiable, we can bring tools like the Lagrangian to bear upon such a problem. That is the main conceptual insight behind section 21. Moreover, it turns out that the derivatives of the Lagrangian — with typical exceptions for the initial and final period — can be found by differentiating the Hamiltonian instead.

In the notes I provide two rather general versions of corresponding optimality conditions that were motivated predominantly by macroeconomic texts, like the books by Sargent and Ljungqvist, by Stokey and Lucas, or by Acemoglu. These theorems formally state that you can forget about the Lagrangian and use the Hamiltonian instead; plus it provides the additional conditions that take care of the initial and final period. In the lecture, I sketched the connection between the Lagrangian and the Hamiltonian, both by rewriting the Lagrangian of a fairly general problem in terms of the Hamiltonian, and by illustrating it for a specific dynamic optimization problem. And I am sure you will agree with me that the Hamiltonian looks more friendly than the Lagrangian. You should be able

☒ to recognize when a simple dynamic optimization problem is a special case of a static optimization problem that can be solved by the methods discussed in the section on nonlinear programming;

☒ to do so in simple examples like the one in class and the examples and exercises in the lecture notes and problem sets. If you feel more comfortable using the Lagrangian than the Hamiltonian, you should feel perfectly free to use the former: they just give different ways to arrive at necessary optimality conditions.

On this part of the course, the recommended book by Sydsæter et al. is pretty good. They also give many worked examples and exercises.

Some mathematical history

Important parts of dynamic optimization were developed more or less in parallel in the early 1950s by an American 'school' around Bellman and a Russian one around Pontryagin. The American school concentrated to a large extent on the dynamic programming algorithm for finite-dimensional problems (section 25) and the Bellman equation for infinite-horizon problems (section 28). The Russian school was stronger in applying calculus-related tools and derived several versions of the maximum principles discussed in section 26.

Here is some historical background from Richard Bellman's autobiography *Eye of the Hurricane*, which I took from Dreyfus, S. (2002): "Richard Bellman on the birth of dynamic programming", Operations Research 50(1), 48–51:

> "I spent the Fall quarter (of 1950) at RAND. My first task was to find a name for multistage decision processes.
>
> An interesting question is, 'Where did the name, dynamic programming, come from?' The 1950s were not good years for mathematical research. We had a very interesting gentleman in Washington named Wilson. He was Secretary of Defense, and he actually had a pathological fear and hatred of the word, research. I'm not using the term lightly; I'm using it precisely. His face would suffuse, he would turn red, and he would get violent if people used the term, research, in his presence. You can imagine how he felt, then, about the term, mathematical. The RAND Corporation was employed by the Air Force, and the Air Force had Wilson as its boss, essentially. Hence, I felt I had to do something to shield Wilson and the Air Force from the fact that I was really doing mathematics inside the RAND Corporation. What title, what name, could I choose? In the first place I was interested in planning, in decision making, in thinking. But planning, is not a good word for various reasons. I decided therefore to use the word, 'programming.' I wanted to get across the idea that this was dynamic, this was multistage, this was time-varying — I thought, let's kill two birds with one stone. Let's take a word that has an absolutely precise meaning, namely dynamic, in the classical physical sense. It also has a very interesting property as an adjective, and that is it's impossible to use the word, dynamic, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. It's impossible. Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities."

A question about time indices

Sometimes the lecture notes use time indices like $x(t)$ and $u(t)$ for the state and control at time t, sometimes not. Does that matter?
ANSWER: I do this when formulas involve controls and states at multiple different times. If calculations involve only a single state and control, it is common in the literature on dynamic optimization to simplify notation as much as possible and just call them x and u instead: expression (123) on page 119, for instance, is of the form

$$J_T(x) = \sup_{u \in U(T,x)} f(T, x, u),$$

since I only try to find a single clever control u at time T, depending on the state x I happen to find myself in. If you prefer to continue denoting the state and control at time T by $x(T)$ and $u(T)$, that is perfectly fine. It just makes the notation a bit more cluttered: expression (123) becomes

$$J_T(x(T)) = \sup_{u(T) \in U(T,x(T))} f(T, x(T), u(T))$$

and similar changes apply to the equation for J_s at earlier times $s = 0, \ldots, T-1$.

No matter if you write x or $x(T)$, u or $u(T)$ in two formulas above, they are just variable names. The number on the righthand side of these formulas is the same.

Lecture 13

- ☒ Notes: section 24.
- ☒ Carter: none.
- ☒ Sydsæter et al.: section 12.2 ('The Euler Equation').
- ☒ Sorger: section 5.2 ('Euler equation and transversality condition')

I illustrated the Euler equations and transversality condition for a simple one-dimensional case and used them to give an alternative solution to the infinite-horizon problem I solved using the Bellman equation last lecture.

For the economic interpretation in special cases, the following article is nice:

> T. Kamihigashi (2008) "Transversality conditions and dynamic economic behaviour", in *The New Palgrave Dictionary of Economics*, 2nd edition, Eds. Steven N. Durlauf and Lawrence E. Blume, Palgrave Macmillan; doi:10.1111/b.9780631218234.2003.X
> Online: `http://www.dictionaryofeconomics.com/article?id=pde2008_T000217` or `https://www.rieb.kobe-u.ac.jp/academic/ra/dp/English/dp180.pdf`

Sydsæter et al. discuss the finite-horizon version only. A little trick: you can often get statements about problems with finite horizon $T \in \mathbb{N}$ from infinite-horizon problems by setting $f(t, x, u) = 0$ for $t > T$!

The important thing is that you can compute these conditions; it is, in practice, very hard to describe exactly when they give an optimum. This is still an active but really difficult area of mathematical research some new results along those lines were proved as late as last year. But not in a form that is legible at the MSc level, so in the next couple of years I will try to see if I can simplify those arguments for inclusion in later installments of the notes. I give one set of conditions on page 143, but not all of them apply in the example I gave in the lecture (which can be covered by a more difficult theorem from section 4.2 in the Kamihigashi paper referred to on page 143).

Lecture 14

No new material. I discussed common mistakes and how to avoid them and did a couple of requests. I promised to include the following in the Study Guide:

A brief reminder about the use of preimages of continuous functions

The useful Theorem 8.2 says that a function is continuous exactly if pre-images of open sets are open sets (pre-images of closed sets are closed sets). This is an easy short-cut to show that certain sets are open. For instance, if $f: X \to \mathbb{R}$ is a continuous, real-valued function on some metric space (X, d), then for each $r \in \mathbb{R}$, the following sets are open:

$$\{x \in X : f(x) < r\}, \qquad \{x \in X : f(x) > r\}, \qquad \{x \in X : f(x) \neq r\}.$$

The first set can be rewritten as

$$\{x \in X : f(x) < r\} = \{x \in X : f(x) \in (-\infty, r)\} = f^{-1}((-\infty, r)), \quad \text{i.e., the pre-image of open set } (-\infty, r).$$

And since f is continuous, this pre-image is open. The second set is the pre-image of open set (r, ∞) and the third set is the union of the former two and the union of two open sets is an open set by Theorem 7.1. Likewise, the following sets are closed:

$$\{x \in X : f(x) \leq r\}, \qquad \{x \in X : f(x) \geq r\}, \qquad \{x \in X : f(x) = r\}$$

For a specific example, the function $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1, x_2) = 3x_1 + 4x_2$ is continuous (Example 8.3), so taking $r = 4$, the set $\{x \in \mathbb{R}^2 : 3x_1 + 4x_2 < 4\}$ is open and the set $\{x \in \mathbb{R}^2 : 3x_1 + 4x_2 \leq 4\}$ is closed.

A slightly more difficult one: the set

$$\{x \in \mathbb{R}^2 : 3x_1 + 4x_2 < 4, x_1 x_2 > 5\} = \{x \in \mathbb{R}^2 : 3x_1 + 4x_2 < 4\} \cap \{x \in \mathbb{R}^2 : x_1 x_2 > 5\}$$

is an open set: it is the intersection of two sets, each of which is open by the argument above. And the intersection of two open sets is an open set.

For exercises where this is used, see for instance pages 65 and B29 (solution to exc. 13.1(e) and (h)), as well as the first exercise in the exams of 2020, 2021, and 2022.

When can you be sure that an optimization problem has a solution?

You may find the following list of results helpful.

- ☒ The main tool was the Extreme Value Theorem, Theorem 13.3.
- ☒ Theorem 19.5 tells when solutions to the KKT conditions are automatically maxima (and not just maximum candidates). Theorem 19.9 does something similar for the Fritz John conditions.
- ☒ You can sometimes apply the Extreme Value Theorem by imposing an artificial additional restriction — in particular in cases where the feasible set is not bounded. You can find an example of this in the first step of the proof of Theorem 16.1. I discussed this more generally in class and Manuel will do a special case on Friday.
- ☒ Theorem 17.13: if a concave function has a local maximum, it is automatically a global maximum. And if its partial derivatives are zero at some interior point of the feasible set, it is automatically a global maximum.
- ☒ The second part of Theorem 20.2 "Moreover..." tells that controls and states that you find when optimizing in the steps of the dynamic programming algorithm are automatically optimal in the whole problem.
- ☒ Theorem 21.2 tells when solutions to the Maximum Principle are optimal.
- ☒ Theorem 23.4 is the infinite-horizon counterpart of Theorem 20.2: once you know the optimal value function, it tells that controls and states that you find as maxima in (the corresponding fixed point of) the Bellman equation are automatically optimal in the whole problem.

Finally...

I am sorry that we were so brusquely interrupted during our Q-and-A: when we booked the room, it was available for longer, so I anticipated that we could have it for at least an hour longer. Thank you all for attending the course; I very much enjoyed teaching it and you did a great job so far. Good luck with next week's exams! Until the weekend, I will be abroad and won't have much email access, so please contact Yifan with questions. I will be back in action on Monday and more than happy to answer questions then.