Mathematics 2

Lecture notes, fall 1997

(Preliminary and incomplete)

Paul Klein

November 10, 1997

Contents

Chapter 1

Introduction

1.1 About the course

This course is mainly about dynamic systems and infinite-dimensional ('dynamic') optimization, but contains a number of supporting topics as well. It is primarily designed for macroeconomics (it is a prerequisite for Macroeconomics 1), but the material has plenty of applications in micro, finance and econometrics as well. It builds on Mathematics 1 in the sense that it draws upon many of the definitions and results established there.

The main emphasis is on conceptual understanding and practical calculation rather than on the details of the proofs. Often proofs are omitted (with a reference to one usually given instead), and sometimes a heuristic argument ('kind-of proof') will replace or complement a real proof to give some intuition for why a result holds and also a sense of why in the world anyone might dream up the result of a theorem.

Similarly, when concepts are defined I have tried as far as possible to say something about the intuitive notion that the formal definition is an attempt to

capture. Notice that we can always ask of a definition whether it does a good job in capturing an intuitive notion and in generating useful theorems. In spite of the conventional wisdom, definitions are not at all arbitrary, and should be studied critically.

However, although the proofs are sometimes omitted, the definitions are kept precise and the results usually stated exactly.

Even though Jörgen's compendium and this one are rather thick, it mustn't be thought that Math 1 and 2 deal with all of the mathematics that an economist needs in order (1) to understand the literature and (2) to do research (although we do give at least an introduction to most of what you'll need). Examples of important topics excluded from both Math I and these lecture notes include (1) partial differential equations, stochastic differential equations, Itô calculus and continuous time stochastic dynamic optimization and (2) line integrals on $\mathbb{C}$ and the theory of transforms (Fourier, Laplace). If you are into finance, you will eventually want to understand (1)[1], and if you are into probability and statistics and especially the spectral analysis of time series, you will eventually want to understand (2).

Moreover, most econometricians will want to learn even more matrix differential calculus and linear algebra than I have included in chapters 4 and 8. And if you are into numerical methods (which certainly any empirically oriented macroeconomist or econometrician should be) you will want to know much more than is covered in chapter 11. See [37] or [36] for useful surveys of the methods available. If you are really keen, a treasure trove for numerical methods is Ellen McGrattan's ftp site at [48].

Other omissions reflect my occasionally idiosyncratic views about what is im-

[1] Good references on these topics are [24] and [44].

portant. One example is the matrix Riccati differential equation and the algebraic matrix Riccati equation, which arise in linear-quadratic dynamic optimization problems. As you will see, one can do perfectly well without these equations (we will use similarity transforms instead) so I will simply ignore them. If you are interested in the matrix Riccati differential equation, I don't even know where to look, although apparently entire journals are devoted to them. On the other hand, a book that is full of algebraic matrix Riccati equations is [5].

The omissions notwithstanding; if you work hard, you should be able to acquire enough sophistication during Math 1 and 2 in order to be able to read up on the omitted topics yourself if and when you need to, or at least the ability to follow courses (e.g. the Finance and Econometrics sequences) which contain some of these omitted topics. To the extent that we focus on at least some of the proofs rather than just the results, this is the reason: it should help you lose respect for mathematics in general and achieve at least some confidence in approaching initially scary concepts and expressions.

The only required reading for this course is (selected parts of) these lecture notes. However, if you want to learn more, I strongly recommend looking at some of the books, handouts and internet sites in the bibliography, especially those on the Reading List (Lang and Chow).

1.2 Notation

An ordered n-tuple or sequence will be denoted by $\langle x_1, ..., x_n \rangle$ or sometimes just $\langle x_k \rangle$. Sequences $\langle x_k \rangle$ will often be written as $\mathbf{x}$, and so will functions $x : A \to B$.

A set is written using curly brackets, e.g. $\{1, 2, ...\}$ or $\{x_1, ..., x_n\}$.

An equality put forward as a definition will be written $f(x) \triangleq \sin x$.

An equality which is also an identity will be written $\cos^2 x + \sin^2 x \equiv 1$.

The notation y_k will usually mean the kth element of the vector y, but sometimes the kth vector y. What is intended is hopefully clear from the context.

Chapter 2

The Riemann integral

2.1 Definition

The purpose of this section is to make sense of expressions like

$$\int_a^b f(x)\,dx. \tag{2.1}$$

The idea will be to capture in a precise way the notion that an integral is the area under a curve. The strategy will be to approximate the area under a curve by the union of rectangles whose area we can calculate in an obvious way, and then try to make the approximation arbitrarily good. Note that our approximation will be based on a partition of the x-axis.

Let $f : [a,b] \to \mathbb{R}$ be a bounded function. Now divide the interval $[a,b]$ into subintervals, choosing $a = x_0 < x_1 < x_2 \ldots < x_{n-1} < x_n = b$. Since f is bounded, we can find numbers h_k and H_k such that, for each $k = 1, 2, ..., n$, we have

$$h_k \leq f(x) \leq H_k \quad \text{for all } x \in [x_{k-1}, x_k]. \tag{2.2}$$

Now introduce lower sums of the form

$$s = \sum_{k=1}^{n} h_k (x_k - x_{k-1}) \tag{2.3}$$

and upper sums of the form

$$S = \sum_{k=1}^{n} H_k \left(x_k - x_{k-1} \right) \tag{2.4}$$

Note that for each partition of the interval $[a, b]$ and choice of numbers h_k and H_k we can define an upper and lower sum. We now consider the set of all such upper and lower sums. Define the sets

$$A \triangleq \{s : s \text{ is a lower sum}\} \tag{2.5}$$

and

$$B \triangleq \{S : S \text{ is an upper sum}\} \tag{2.6}$$

Apparently A has an upper bound and B has a lower bound. Hence we can define

$$\underline{R} \triangleq \sup A \tag{2.7}$$

and

$$\overline{R} \triangleq \inf B \tag{2.8}$$

Definition 2.1.1 *Let f be a bounded function on the interval $[a, b]$. Let $\underline{R}$ and $\overline{R}$ be defined as above. If $\underline{R} = \overline{R} \triangleq R$, then f is said to be* Riemann integrable *on $[a, b]$ and we define*

$$\int_a^b f(x)\, dx \triangleq R \tag{2.9}$$

We now want to know which functions are integrable, but a full answer to that question would take us further than we need to go right now. For our present purposes, it suffices to note the following theorems.

Theorem 2.1.1 *Let f be bounded and continuous, except possibly at a countable number of points, on the compact interval $[a, b]$. Then f is Riemann integrable on $[a, b]$.*

Proof *See [17].*

Theorem 2.1.2 *Let f and g be Riemann integrable on the compact interval $[a,b]$. Then fg is Riemann integrable on $[a,b]$.*

Proof. See [17]. ■

Remark 2.1.1 *Unfortunately, there are many bounded functions which are not Riemann integrable. Indeed, it may even happen that $\langle f_k \rangle_{k=1}^{\infty}$ is a uniformly bounded sequence (a sequence with a common upper bound) of Riemann integrable functions defined on a compact interval, yet the pointwise limit f (the function defined by setting, for each x, $f(x) \triangleq \lim_{k\to\infty} f_k(x)$) is not Riemann integrable. A famous example is the following. Let $\langle q_j \rangle_{j=1}^{\infty}$ be an enumeration of the rational numbers, and let $f_k : [0,1] \to \mathbb{R}$; $k = 1, 2, ...$ be defined via*

$$f_k(x) \triangleq \begin{cases} 1 \text{ if } x = q_j \text{ for some } j \leq k \\ 0 \text{ otherwise} \end{cases} \tag{2.10}$$

Then the pointwise limit f is

$$f(x) \triangleq \begin{cases} 1 \text{ if } x \text{ is rational} \\ 0 \text{ if } x \text{ is irrational} \end{cases} \tag{2.11}$$

and all the f_k are Riemann integrable on $[0,1]$, yet f is not. This pathological property of the Riemann integral provides good motivation for (although historically it has little to do with) the development of the Lebesgue integral (see chapter 6).[1]

We now note some important properties of the Riemann integral. Note that they are intimately connected with the fact that the Riemann integral is the limit of a sum.

[1] In chapter 6, we will define an integral concept such that $\lim_{k\to\infty} \int_a^b f_k(x)\,dx = \lim_{k\to\infty} \int_a^b \lim_{k\to\infty} f_k(x)\,dx$ so long as the convergence of $\langle f_k \rangle$ is monotone. However, even with the Riemann integral, the result holds if the convergence of $\langle f_k \rangle$ is uniform. See [17].

Theorem 2.1.3 *1. Let f be Riemann integrable on $[a,c]$ and let $a<b<c$. Then f is Riemann integrable on $[a,b]$ and $[b,c]$ and*

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx \tag{2.12}$$

2. Let f,g be Riemann integrable on $[a,b]$ and let c be a scalar. Then $f+g$ and cf are Riemann integrable, and we have

(a) $\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$ *and*

(b) $\int_a^b [cf(x)]\,dx = c\int_a^b f(x)\,dx.$

Proof. See [17]. ■

2.1.1 The improper Riemann integral

We would also like to know how to deal with unbounded functions and integration over unbounded intervals rather than just compact ones. Within the Riemann theory, this is done by introducing the so-called *improper* Riemann integral. For example, suppose f is unbounded on $[a,b)$ but bounded on each closed subset of $[a,b)$. Then we can try the definition

$$\int_a^b f(x)\,dx \triangleq \lim_{\beta\to b}\int_a^\beta f(x)\,dx. \tag{2.13}$$

If this limit exists, then the improper Riemann integral is said to exist and be equal to the right hand side. Note that this definition works for $[a,\infty)$ as well. The extension to integration over $(a,b]$, $(-\infty,b]$ is obvious. The extension to $(-\infty,\infty)$ is more tricky, however. One option is

$$\int_{-\infty}^{\infty} f(x)\,dx \triangleq \lim_{K\to\infty}\int_{-K}^{K} f(x)\,dx. \tag{2.14}$$

This definition is called the Cauchy principal value. Another option is

$$\int_{-\infty}^{\infty} f(x)\,dx \triangleq \lim_{K\to\infty}\int_{-K}^{0} f(x)\,dx + \lim_{K\to\infty}\int_{0}^{K} f(x)\,dx. \tag{2.15}$$

By looking at the function $f(x) = x$ you may want to convince yourself that these two definitions are not equivalent. The point is that $\lim_{K\to\infty}\int_{-K}^{K} f(x)\,dx$ may exist even when $\lim_{K\to\infty}\int_{0}^{K} f(x)\,dx$ does not.

2.2 The fundamental theorem of calculus

There is a very strong sense in which integration is the inverse of differentiation, as revealed by the following theorems.

Theorem 2.2.1 (the fundamental theorem of calculus, part 1) *Let f be Riemann integrable on $[a,b]$ and define*

$$F(x) \triangleq \int_{a}^{x} f(t)\,dt. \tag{2.16}$$

Then F is continuous on $[a,b]$, and at any point x_0 where f is continuous, F is differentiable with derivative $F'(x_0) = f(x_0)$.

Proof. See [17]. ■

Kind-of proof. Define the 'discrete' integral

$$y_n \triangleq \sum_{i=1}^{n} f_i\,(x_i - x_{i-1}) \tag{2.17}$$

Let $\Delta y_n \triangleq y_n - y_{n-1}$ and $\Delta x_i \triangleq x_i - x_{i-1}$. Then

$$\Delta y_n = f_n \Delta x_n \tag{2.18}$$

Hence

$$\frac{\Delta y_n}{\Delta x_n} = f_n \tag{2.19}$$

Our theorem is just a continuous version of this rather banal conclusion.

Theorem 2.2.2 (the fundamental theorem of calculus, part 2) *Let f be Riemann integrable on $[a,b]$ and suppose there is a function F on $[a,b]$ such that $F' = f$. Then $\int_a^b f(x)\,dx = F(b) - F(a)$.*

Proof. See [17]. ■

A useful consequence of the linearity of the integral is that it enables one to 'differentiate under the integral sign'. Combining this fact with a version of Theorem (2.2.1), we get the following theorem.

Theorem 2.2.3 (Leibniz' formula) *Let $f(x,t)$ and $\frac{\partial f(x,t)}{\partial x}$ be continuous on a closed rectangle $[a,b] \times [t_0,t_1]$, let $u(x)$ and $v(x)$ be continuously differentiable on $[a,b]$, and suppose that $u([a,b]) \subset [t_0,t_1]$ and $v([a,b]) \subset [t_0,t_1]$. Then, for all $x \in [a,b]$, we have*

$$\frac{d}{dx}\left[\int_{u(x)}^{v(x)} f(x,t)\,dt\right] = f(x,v(x))\,v'(x) - f(x,u(x))\,u'(x) + \int_{u(x)}^{v(x)} \frac{\partial f(x,t)}{\partial x} dt. \tag{2.20}$$

Proof. See [7]. ■[7] also contains conditions under which one can differentiate under the integral sign in the case where the integration limits are infinite.

2.3 Calculating the Riemann integral

Having acquainted ourselves with the Riemann integral, we now want to calculate it in concrete instances. In some cases, this is easy; the fact that integration is

the inverse of differentiation is enough. For example, for $k \neq -1$,

$$\int_a^b x^k dx = \frac{1}{k+1}\left(b^{k+1} - a^{k+1}\right) \tag{2.21}$$

and for $k = -1$, we have, so long as $0 \notin [a, b]$,

$$\int_a^b \frac{dx}{x} = \ln|b| - \ln|a| \tag{2.22}$$

In a similar fashion, it is sometimes possible to integrate functions by just guessing a function whose derivative is that function. Usually, however, that doesn't work, and we need more sophisticated tricks. We will present some of the most basic tricks below. These tricks are useful, and there are other, even more ingenious and powerful tricks used to calculate integrals. (The most ingenious trick is perhaps the calculus of residues. See [39].) But it is often hard to figure out what trick to use, and sometimes there is no trick; the integral just can't be rewritten in terms of elementary functions. (An example is the integral of the density function of a normal random variable.) We then have to resort to numerical methods; see [37] or [36]. But these methods can be rather slow, so before going that far, you should at least be familiar with the following basic pencil-and-paper techniques.

2.3.1 Change of variables

Theorem 2.3.1 *Let f be continuous on $[g(a), g(b)]$, let g be continuously differentiable on $[a, b]$ and let $f \circ g$ be continuous on $[a, b]$. Then*

$$\int_a^b f(g(t))\, g'(t)\, dt = \int_{g(a)}^{g(b)} f(x)\, dx \tag{2.23}$$

Proof. See [17]. ■

Kind-of proof.

$$\begin{aligned} \int_a^b f(g(t))\, g'(t)\, dt &= \int_a^b f(g(t))\, \frac{dg}{dt} dt = \\ &= \int_{g(a)}^{g(b)} f(g)\, dg \end{aligned} \tag{2.24}$$

2.3.1.1 The f'/f method

Proposition 2.3.1 *Let f and g be such that $g(x) = \dfrac{f'(x)}{f(x)}$ on $[a,b]$ and suppose $f(x) \neq 0$ for all $x \in [a,b]$. Then*

$$\int_a^b g(x)\, dx = \ln|f(b)| - \ln|f(a)| \tag{2.25}$$

Proof. Change of variables. ■

Example 2.3.1 *Consider*

$$\int_1^t \frac{1}{s(s+1)} ds \tag{2.26}$$

where $t \geq 1$. Now note that

$$\frac{1}{s(s+1)} = \frac{s^{-2}}{1+s^{-1}} = -\frac{-s^{-2}}{1+s^{-1}}. \tag{2.27}$$

Hence

$$\begin{aligned} \int_1^t \frac{1}{s(s+1)} ds &= -\int_1^t \frac{-s^{-2}}{1+s^{-1}} ds = \\ = -\ln(1+t^{-1}) + \ln(2) &= \ln t - \ln(t+1) + \ln 2 \end{aligned} \tag{2.28}$$

2.3.2 Integration by parts and the Neatest Trick

From the product rule of differentiation, we can derive the following theorem.

Theorem 2.3.2 *(integration by parts). Suppose F and G are differentiable functions on $[a,b]$, and that $F'=f$ and $G'=g$ are Riemann integrable on $[a,b]$. Then*

$$\int_a^b F(x)\,g(x)\,dx = F(b)\,G(b) - F(a)\,G(a) - \int_a^b f(x)\,G(x)\,dx. \tag{2.29}$$

Proof. Define $H(x) \triangleq F(x)\,G(x)$. By the product rule of differentiation, $H'(x) = f(x)\,G(x) + F(x)\,g(x)$. Hence

$$\int_a^b F(x)\,g(x)\,dx = \int_a^b H'(x)\,dx - \int_a^b f(x)\,G(x)\,dx \tag{2.30}$$

and the theorem follows by the fundamental theorem of calculus. ■

Usually, the way to take advantage of partial integration is to let F be a function that becomes simpler when differentiated, and let g be a function that does not become much more complicated by being integrated. This is seen clearly in the following example.

Example 2.3.2 *When calculating*

$$\int_0^1 xe^x dx \tag{2.31}$$

it makes sense to define $F(x) \triangleq x$ and $g(x) \triangleq e^x$. Consistent with this choice, set $f(x) = F'(x) = 1$ and $G(x) = e^x$. Using this, we find that

$$\int_0^1 xe^x dx = 1\cdot e^1 - 0\cdot e^0 - \int_0^1 1\cdot e^x dx = e - e + e^0 = 1. \tag{2.32}$$

Example 2.3.3 *The Γ function. Define, for $0<x<\infty$,*

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt \tag{2.33}$$

We now use integration by parts to show that, for $0<x<\infty$, we have

$$\Gamma(x+1) = x\Gamma(x) \tag{2.34}$$

Let $F(t) = t^x$, $g(t) = e^{-t}$ *and* $G(t) = -e^{-t}$. *Then* $G'(t) = g(t)$ *and* $F'(t) = f(t) = xt^{x-1}$. *Using integration by parts, we find that*

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt = -\lim_{b\to\infty} b^x e^{-b} + 0^x e^{-0} + \int_0^\infty xt^{x-1} e^{-t} dt =$$

$$= x \int_0^\infty t^{x-1} e^{-t} dt = x\Gamma(x) \qquad (2.35)$$

where we used that the exponential function eventually grows faster than any polynomial and, also, the linearity of the integral.

Example 2.3.4 *The Neatest Trick. Suppose we want to calculate*

$$R = \int_0^\pi e^x \sin x dx \qquad (2.36)$$

where we know that $\frac{d}{dx}\sin x = \cos x$. *(See chapter 3.) Neither the exponential function nor the sine function becomes more or less complicated by integration or differentiation, so we seem to be in trouble. Nevertheless, upon repeated use of partial integration, we find that*

$$R = \int_0^\pi e^x \sin x dx = e^\pi \sin\pi - e^0 \sin 0 - \int_0^\pi e^x \cos x dx =$$

$$= -\int_0^\pi e^x \cos x dx = -e^\pi \cos\pi + e^0 \cos 0 - \int_0^\pi e^x \sin x dx =$$

$$= e^\pi + 1 - R. \qquad (2.37)$$

Hence

$$R = \int_0^\pi e^x \sin x dx = \frac{e^\pi + 1}{2} \qquad (2.38)$$

Example 2.3.5 *A tricky one. Suppose we want to calculate*

$$\int_1^y \ln x dx \qquad (2.39)$$

The way forward is to note that $\ln x = \ln x \cdot 1$. *Now* 1 *does not become much more complicated by integration, so set* $F(x) = \ln x$ *and* $g(x) = 1$. *It follows that* $F'(x) = f(x) = \dfrac{1}{x}$ *and we may choose* $G(x) = x$. *We get*

$$\int_1^y \ln x dx = \ln y \cdot y - \ln 1 \cdot 1 - \int_1^y \frac{1}{x} x dx = \\ = y(\ln y - 1) + 1 \tag{2.40}$$

2.4 The Riemann-Stieltjes integral

The purpose of this section is to make sense of expressions like

$$\int_a^b f(x)\, dF(x). \tag{2.41}$$

where F is a non-decreasing function. These expressions appear, for example, when we want to calculate the expected value of a random variable that doesn't have a density function. (A random variable fails to have a density function whenever its distribution function is discontinuous, and sometimes even when it is continuous).

Definition 2.4.1 *Let* $f : [a,b] \to \mathbb{R}$ *be a bounded function and let* $F : [a,b] \to \mathbb{R}$ *be a non-decreasing function. Now partition* $[a,b]$ *and define numbers* h_k *and* H_k *as in Definition 2.1.1. This time, the lower and upper sums take the form*

$$s = \sum_{k=1}^{n} h_k (F(x_k) - F(x_{k-1})) \tag{2.42}$$

and

$$S = \sum_{k=1}^{n} H_k (F(x_k) - F(x_{k-1})) \tag{2.43}$$

and the definition proceeds exactly as Definition 2.1.1.

Proposition 2.4.1 *Suppose F is differentiable with derivative f. Then*

$$\int_a^b g(x)\, dF(x) = \int_a^b g(x)\, f(x)\, dx \tag{2.44}$$

Proof. Exercise. ■

For more properties of the Riemann-Stieltjes integral, see [17].

Chapter 3

Trigonometry and complex numbers

3.1 Trigonometry

3.1.1 Geometric definition of the trigonometric functions

Consider the unit circle in $\mathbb{R}^2$.

We first establish the unit of measurement for angles.

Definition 3.1.1 *Consider two rays from the origin. The angle between them is the length of the arc on the unit circle between the points of intersection between each of the rays and the unit circle. This unit of measurement is called* radians. *By the definition of the number π, the angle of an entire revolution is 2π radians.*

Having established a unit of measurement, we can now define the trigonometric functions.

Definition 3.1.2 *Let $0 \leq \theta \leq 2\pi$ be a real number. Then $\cos\theta$ ($\sin\theta$) is the abscissa (ordinate) of the point of intersection between the unit circle and a ray*

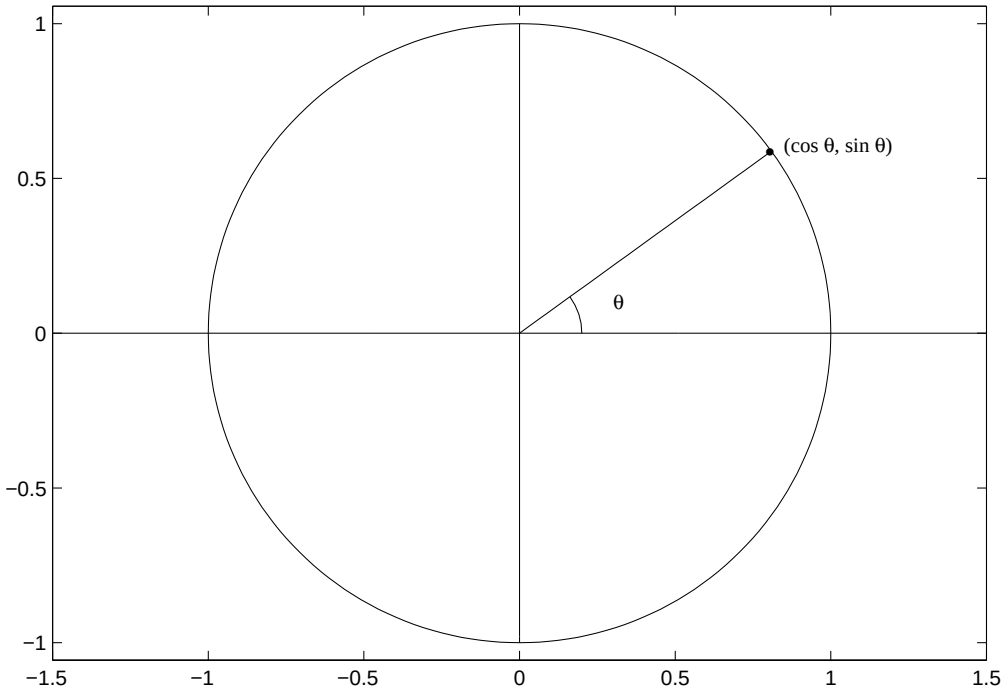


Figure 3.1:

from the origin which is at an angle θ with a ray from the origin through $(1,0)$. For θ outside this range, we extend the definition via

$$\cos(\theta + 2n\pi) = \cos\theta \tag{3.1}$$

$$\sin(\theta + 2n\pi) = \sin\theta \tag{3.2}$$

for all $n \in \mathbb{Z}$.

Definition 3.1.3 *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be periodic with period p and amplitude a if*

1. *For all $x \in \mathbb{R}$, $f(x+p) = f(x)$ and*
2. $\sup_{x \in \mathbb{R}} |f(x)| = a$.

Proposition 3.1.1 *Let f be defined via $f(x) \triangleq A\sin(\omega x)$. Then f is a periodic function with period $\dfrac{2\pi}{\omega}$ and amplitude $|A|$.*

Proof. Obvious. ■

Remark 3.1.1 *The number ω is sometimes called the* frequency *of f.*

When calculating the value of sin and cos in simple cases, the following table, derived from basic geometry and Pythagoras' theorem, is useful. Notice the pattern.

Table 1.

θ	$\sin\theta$	$\cos\theta$
0	$\frac{1}{2}\sqrt{0}$	$\frac{1}{2}\sqrt{4}$
$\pi/6$	$\frac{1}{2}\sqrt{1}$	$\frac{1}{2}\sqrt{3}$
$\pi/4$	$\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{2}$
$\pi/3$	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{1}$
$\pi/2$	$\frac{1}{2}\sqrt{4}$	$\frac{1}{2}\sqrt{0}$

We now define some of the most commonly used trigonometric functions.

Definition 3.1.4 *Let θ be a real number such that $\cos\theta \neq 0$. Then $\tan\theta \triangleq \dfrac{\sin\theta}{\cos\theta}$.*

Definition 3.1.5 *Let θ be a real number such that $\sin\theta \neq 0$. Then $\cot\theta \triangleq \dfrac{\cos\theta}{\sin\theta}$.*

Definition 3.1.6 *Let θ be a real number such that $\cos\theta \neq 0$. Then $\sec\theta \triangleq \dfrac{1}{\cos\theta}$.*

Definition 3.1.7 *Let θ be a real number such that $\sin\theta \neq 0$. Then $\operatorname{cosec}\theta \triangleq \dfrac{1}{\sin\theta}$.*

Definition 3.1.8 *Let $x \in [-1,1]$. Then $\arcsin x$ is that angle $\theta \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$ such that $\sin\theta = x$.*

Remark 3.1.2 *It follows from the definition that $\sin(\arcsin x) = x$ on the entire domain $[-1,1]$. But $\arcsin(\sin\theta) = \theta$ only holds for $\theta \in \left[-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right]$.*

Definition 3.1.9 *Let $x \in [-1,1]$. Then $\arccos x$ is that angle $\theta \in [0,\pi]$ such that $\cos\theta = x$.*

Definition 3.1.10 *Let $x \in \mathbb{R}$. Then $\arctan x$ is that angle $\theta \in \left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$ such that*

$$\tan\theta = x. \tag{3.3}$$

We now list some properties of the trigonometric functions.

3.1.2 Trigonometric identities

Below, we will write $\sin^2\theta$ for $(\sin\theta)^2$ etc.

Theorem 3.1.1 *For all real numbers θ and φ such that the expressions below are defined, we have*

1. $\cos^2\theta + \sin^2\theta \equiv 1$

2. $\sec^2\theta \equiv 1 + \tan^2\theta$

3. $\cos\theta \equiv \sin\left(\frac{\pi}{2} - \theta\right)$

4. $\sin(\theta + \pi) \equiv -\sin\theta$

5. $\cos(\theta + \pi) \equiv -\cos\theta$

6. $\sin(-\theta) \equiv -\sin\theta$

7. $\cos(-\theta) \equiv \cos\theta$

8. $\sin(\theta + \varphi) \equiv \sin\theta\cos\varphi + \cos\theta\sin\varphi$

9. $\cos(\theta + \varphi) \equiv \cos\theta\cos\varphi - \sin\theta\sin\varphi$

10. $\sin\theta + \sin\varphi \equiv 2\sin\left(\frac{\theta+\varphi}{2}\right)\cos\left(\frac{\theta-\varphi}{2}\right)$

11. $\cos^2\theta \equiv \frac{1+\cos 2\theta}{2}$

12. $\sin^2\theta \equiv \frac{1-\cos 2\theta}{2}$.

Sketch of proof. *(1) and (2) are consequences of Pythagoras' theorem, (3-7) are geometrically obvious, and (8-9) are a little bit tricky to prove geometrically (see [6]), but follow from Euler's formula (see below). (It is essential that they can be proved geometrically, though, since otherwise we can't prove Euler's formula itself.) (10) follows from the others.*

There are many more trigonometric identities, but the above are more than enough for our purposes.

3.1.3 Derivatives of trigonometric functions

Lemma 3.1.1

$$\lim_{h\to 0}\frac{\sin h}{h}=1. \tag{3.4}$$

Proof. See [38]. ■

Theorem 3.1.2 *1. Let* $f(x)\triangleq\sin x$. *Then* $f'(x)=\cos x$.

2. Let $f(x)\triangleq\cos x$. *Then* $f'(x)=-\sin x$.

3. Let $f(x)\triangleq\tan x$. *Then* $f'(x)=\sec^2 x$.

Proof *It is tempting to use Euler's formula (see below) here (and the quotient rule for the derivative of* tan*). But since our sketch of the proof of Euler's formula invokes the derivatives of the trigonometric functions, that would be circular. So here follows a direct proof of (1). We use the definition of the derivative, and find that*

$$\begin{aligned}\lim_{h\to 0}\frac{\sin(x+h)-\sin x}{h} &= \lim_{h\to 0}\frac{2\cos(x+h/2)\sin(h/2)}{h} = \\ &= \lim_{h\to 0}\cos(x+h/2)\frac{\sin(h/2)}{h/2} = \\ &= \cos(x)\end{aligned} \tag{3.5}$$

where we used that cos *is a continuous function.*

Corollary 3.1.1 *1. Let* $f(x)\triangleq\arcsin x$. *Then* $f'(x)=\dfrac{1}{\sqrt{1-x^2}}$.

2. Let $f(x)\triangleq\arccos x$. *Then* $f'(x)=-\dfrac{1}{\sqrt{1-x^2}}$.

Example 3.1.1 *Let's calculate* $f'(x)$ *when* $f(x) = \arctan x$. *Note that, by definition,* $f(x) \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ *for each* $x \in \mathbb{R}$. *By the definition of the* arctan *function, it follows that*

$$\tan f(x) = x. \tag{3.6}$$

Differentiating both sides (using the chain rule on the left hand side), we get

$$\sec^2 f(x) f'(x) = 1. \tag{3.7}$$

Hence

$$f'(x) = \frac{1}{\sec^2 f(x)} = \frac{1}{1 + \tan^2 f(x)} = \tag{3.8}$$
$$= \frac{1}{1 + x^2}$$

where the final equality holds since, by definition, $f(x) \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. *The somewhat surprising consequence is that*

$$\int_{-\infty}^{\infty} \frac{1}{1 + x^2} dx = \lim_{x \to \infty} [\arctan x - \arctan(-x)] = \pi. \tag{3.9}$$

3.1.4 Integrating trigonometric functions

Generally speaking, trigonometric functions are integrated by making good use of trigonometric identities, integration by parts and change of variables. Plenty of practice is needed. Good exercises are found in [38].

Here we will just take some examples. An example where a trigonometric identity comes in handy is the following.

$$\int_0^x \cos^2 \theta d\theta = \int_0^x \frac{1 + \cos 2\theta}{2} d\theta = \frac{1}{2}x + \frac{1}{4}\sin 2x \tag{3.10}$$

An example of where the fact that $\cos\theta$ is the derivative of $\sin\theta$ is useful is the following.

$$\int_0^x \sin\theta \cos\theta d\theta = \int_0^{\sin x} t dt = \frac{1}{2}\sin^2 x \tag{3.11}$$

3.2 Complex numbers

3.2.1 Motivation

The set $\mathbb{R}$ of real numbers is rich, but not quite rich enough for some purposes. For example, it does not contain solutions to all polynomial equations (and hence not all eigenvalue problems). An example is $x^2 + 1 = 0$. We will define a set of complex numbers $\mathbb{C}$ that contains solutions to all polynomial equations with real coefficients. Some mathematicians in the 19th century worried that, if we allowed complex coefficients, we would have to invent an even richer set of numbers to house the solutions to the resulting polynomial equations, and so on ad infinitum. They were wrong. The fundamental theorem of algebra (see below) shows that the extension from $\mathbb{R}$ to $\mathbb{C}$ is enough to solve polynomial equations with complex coefficients too.

3.2.2 Definition

Definition 3.2.1 *The set $\mathbb{C}$ of complex numbers is the set $\mathbb{R}^2$ together with the usual (elementwise) addition and (real) scalar multiplication operations in $\mathbb{R}^2$ and the following multiplication operation. Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$ be two members of $\mathbb{C}$. Then*

$$xy \triangleq (x_1 y_1 - x_2 y_2, x_1 y_2 + x_2 y_1) \tag{3.12}$$

So as to distinguish $\mathbb{R}^2$ from $\mathbb{C}$ (so that it is clear which rules of arithmetic apply) we adopt the following notation.

Definition 3.2.2 *Let $x = (x_1, x_2)$ be a complex number. Then we write $x = x_1 + ix_2$. The complex number $(0, 1) = i$ is called the imaginary unit.*

Theorem 3.2.1

$$i^2 = -1. \tag{3.13}$$

Proof. Exercise. ■

Note that, with our definition of multiplication and the notation $x = x_1 + ix_2$, multiplication is just like calculating $(a + b)(c + d)$ in the usual way, keeping in mind that $i^2 = -1$. Also, note that all the usual laws of arithmetic such as the distributive and associative laws hold for complex numbers as well as real ones.

Often one writes, for $z \in \mathbb{C}$, $z = x + iy$ where $x = z_1$ and $y = z_2$. Also, one sometimes defines the real part of z via $Re(z) = x$ and the imaginary part via $Im(z) = y$. (Note that the imaginary part of a complex number is a real number.) When $Im(z) = 0$ we will regard z as a real number, and in this sense we will say that $\mathbb{R} \subset \mathbb{C}$.

Actually, the rule of multiplication mentioned above isn't the only difference between $\mathbb{C}$ and $\mathbb{R}^2$. Another (and this is the final difference between the two sets) is that the word 'scalar' changes meaning when one passes from $\mathbb{R}^2$ to $\mathbb{C}$. When talking about $\mathbb{R}^2$, a scalar is any member of $\mathbb{R}$. When talking about $\mathbb{C}$, a scalar is any member of $\mathbb{C}$. This is of crucial importance in defining the set of linear operators on $T : \mathbb{C} \to \mathbb{C}$ and hence in defining the notion of differentiability for complex-valued functions defined on $\mathbb{C}$ (see below).

Definition 3.2.3 *A linear operator $T : \mathbb{C} \to \mathbb{C}$ is a function such that, for all scalars $\alpha \in \mathbb{C}$ and all complex numbers $x, y \in \mathbb{C}$, we have*

$$T(\alpha(x + y)) = \alpha T(x) + \alpha T(y). \tag{3.14}$$

Proposition 3.2.1 *The linear functions $T : \mathbb{C} \to \mathbb{C}$ are precisely those representable as $T(z) = \alpha z$ for some $\alpha \in \mathbb{C}$. As stressed below, this is a much smaller class of linear operators than the set of linear operators $T : \mathbb{R}^2 \to \mathbb{R}^2$.*

Proof. Copy the corresponding proof for $\mathbb{R}$. ■

Definition 3.2.4 *The complex conjugate $\overline{z}$ of a complex number $z = x + iy$ is defined via*

$$\overline{z} \triangleq x - iy \tag{3.15}$$

Definition 3.2.5 *The modulus $|z|$ of a complex number is defined via*

$$|z| \triangleq \sqrt{z\overline{z}}. \tag{3.16}$$

If $z = x + iy$, then we have the practical formula

$$|z| \triangleq \sqrt{x^2 + y^2}. \tag{3.17}$$

Remark 3.2.1 *For this to make sense, $z\overline{z}$ must be a non-negative real number. As the 'practical formula' shows, it always is.*

Remark 3.2.2 *This modulus serves as a norm (see chapter 5) and hence defines the open sets in $\mathbb{C}$.*

Remark 3.2.3 *The modulus of a complex number z is the length of the corresponding vector in $\mathbb{R}^2$.*

Proposition 3.2.2 *Let x, y be complex numbers. Then*

$$\overline{x \cdot y} = \overline{x} \cdot \overline{y}. \tag{3.18}$$

Proof. Exercise. ■

3.2.3 The fundamental theorem of algebra

Theorem 3.2.2 (the fundamental theorem of algebra) *Let $n \geq 1$ be a natural number and let $a_0,\ a_1, ..., a_n$ be complex numbers with $a_n \neq 0$. Define the function $f : \mathbb{C} \to \mathbb{C}$ via*

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n. \tag{3.19}$$

Then there is a $z \in \mathbb{C}$ such that $f(z) = 0$.

Proof. See [39]. ■

3.2.4 The exponential function on $\mathbb{C}$ and Euler's formula

Definition 3.2.6 *For all $z \in \mathbb{C}$, define*

$$e^z \triangleq \exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} \tag{3.20}$$

where we adopt the convention that $0^0 = 1$.

For this definition to make sense, we must prove that the sum converges for all $z \in \mathbb{C}$. It does. See [17].

In many cases we want to differentiate a function $f : \mathbb{C} \to \mathbb{C}$. For example, we want it to be true that $\exp'(z) = \exp(z)$ and for that we must at the very least know what this would mean. A fuller treatment of this (very exciting) topic is given in [39]. Here we just give the definition, which is almost the same as in $\mathbb{R}^2$. As usual, the idea is to capture the notion that the local behavior of f is well approximated by a linear function plus a constant, keeping in mind what the class of linear functions $T : \mathbb{C} \to \mathbb{C}$ is.

Definition 3.2.7 *Let $f : \Omega \to \mathbb{C}$ be a function, where $\Omega \subset \mathbb{C}$ is an open set. If there for each $z_0 \in \Omega$ is a complex number $f'(z_0)$ such that for each $\varepsilon > 0$ there*

is a $\delta_\varepsilon > 0$ *such that*

$$\left| \frac{f(z) - f(z_0) - f'(z_0) \cdot (z - z_0)}{z - z_0} \right| < \varepsilon \tag{3.21}$$

for all $z \in \{z \in \Omega : 0 < |z - z_0| < \delta_\varepsilon\}$ *then* $f(z)$ *is said to be* differentiable *(holomorphic, analytic) on* Ω *with derivative* $f'(z)$.

Remark 3.2.4 *Because of the smaller class of linear operators on* $\mathbb{C}$ *than* $\mathbb{R}^2$*, the requirement that a function* $f : \mathbb{C} \to \mathbb{C}$ *is differentiable is much stronger than the requirement that* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *is (in the latter case, all that is needed is that the partial derivatives exist). Essentially the reason for the difference is that a linear operator* $T : \mathbb{R}^2 \to \mathbb{R}^2$ *is represented by a* 2×2 *matrix, but a linear operator* $T : \mathbb{C} \to \mathbb{C}$ *is represented by a single complex number, i.e. essentially a* 2×1 *vector. For the local behavior of a function to be captured by just two numbers rather than four requires consistency between the partial derivatives in a certain sense, as expressed by the so-called Cauchy-Riemann equations. For details, see [39].*

Theorem 3.2.3 *The function* exp *is holomorphic in* $\mathbb{C}$ *and* $\exp'(z) = \exp(z)$.

Proof. Just use the definition of exp and differentiate term by term. To show that term-by-term differentiation is justified, we need to confirm that the sum in the definition of exp converges uniformly. For details, see [17]. ■

Theorem 3.2.4 (Euler's formula) *For all real numbers* θ,

$$e^{i\theta} \equiv \cos\theta + i \sin\theta. \tag{3.22}$$

Sketch of proof. *Show that both sides have the same Taylor series expansion. Or, more suggestively, show that both sides solve the same differential equation, which has a unique solution. (See section 9.1.) Hint: the differential*

equation is $f'(t) = if(t)$, $f(0) = 1$. *Alternatively, we can* define *the* cos *and* sin *functions on* $\mathbb{C}$ *via Euler's formula. Then we would have to show that these definitions agree with the geometric ones for real arguments.*

3.2.5 The Cartesian and polar representation of a complex number

Euler's formula enables us to represent complex numbers in polar coordinates (modulus-argument form). We write $z = re^{i\theta}$ where the nonnegative real number r is the modulus of z (check that this is consistent with the definition of modulus given above!) and the real number $\theta \in (-\pi, \pi]$ is the argument.

Suppose we have a complex number z in so-called Cartesian form, i.e. we know the numbers x and y in the representation $z = x + iy$. Our project now is to translate that into the modulus-argument form $z = re^{i\theta}$. Finding the modulus is easy - for that there is a simple formula. And finding the argument is not too hard once we see the geometry involved.

Take a new look at Figure 3.1.1. Let the point $(\cos\theta, \sin\theta)$ represent the point $\frac{1}{r}z$. Knowing x and y evidently implies knowing $\cos\theta$ and $\sin\theta$! Indeed, it seems that all we need to find θ is the ratio $\dfrac{\cos\theta}{\sin\theta} = \dfrac{x}{y}$, and it is tempting to say that $\theta = \arctan\left(\dfrac{y}{x}\right)$. But that unfortunately isn't always right, because the range of arctan is confined to the first and fourth quadrants, whereas the argument of z could easily be in the second or third quadrant (x could be negative). If x is negative, one right answer is evidently $\theta = \arctan\left(\dfrac{y}{x}\right) + \pi$. Another is $\theta = \arctan\left(\dfrac{y}{x}\right) - \pi$. A common convention is to choose the value of θ that lies in $[0, 2\pi)$. Another is to choose the value that lies in $(-\pi, \pi]$ The case $x = 0$ may seem to be problematic here. But again the geometry makes things obvious. If

$y = 0$ the choice of θ is arbitrary. If $y < 0$, then $\theta = \frac{3\pi}{2}$ (or $-\frac{\pi}{2}$) and if $y > 0$, then $\theta = \frac{\pi}{2}$.

3.3 The space $\mathbb{C}^n$

The space $\mathbb{C}^n$ is just like the set $\mathbb{R}^n$ except that the conventional inner ('dot') product is defined via

$$(x, y) \triangleq \sum_{i=1}^{n} x_i \overline{y_i} \tag{3.23}$$

This makes $\mathbb{C}^n$ into a Hilbert space (see chapter 5) with complex scalars.

Chapter 4

Calculus with vectors and matrices

4.1 Matrix differential calculus

4.1.1 The gradient and the Hessian

The purpose of this section is to make sense of expressions like

$$\frac{\partial f(x)}{\partial x^T} = \nabla_x f(x) = \nabla f(x) = f'(x) = f_x(x) \tag{4.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^m$. Of course, we already know what a partial derivative is and how to calculate it. What this section will tell us is how to arrange the partial derivatives into a matrix (gradient), and the rules of arithmetic that follow from adopting our particular arrangement convention.

Definition 4.1.1 *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ have partial derivatives at x. Then*

$$\underbrace{\frac{\partial f(x)}{\partial x^T}}_{m \times n} \triangleq \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix} \tag{4.2}$$

and

$$\underbrace{\frac{\partial f(x)}{\partial x}}_{n\times m} \triangleq \left(\frac{\partial f(x)}{\partial x^T}\right)^T \tag{4.3}$$

where A^T is the transpose of A.

Definition 4.1.2 *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ have partial derivatives at x. Then the (scalar-valued)* Jacobian *of f at x is defined via*

$$J_f(x) \triangleq \det \frac{\partial f(x)}{\partial x^T}. \tag{4.4}$$

Remark 4.1.1 *Sometimes the gradient $\dfrac{\partial f(x)}{\partial x^T}$ itself is called the Jacobian. Here the Jacobian is defined as the* determinant *of the gradient.*

The following properties of the gradient follow straightforwardly from the definition.

Proposition 4.1.1 *1. Let x be an $n \times 1$ vector and A an $m \times n$ matrix. Then*

$$\frac{\partial}{\partial x^T}[Ax] = A. \tag{4.5}$$

2. Let x be an $n \times 1$ vector and A an $n \times m$ matrix. Then

$$\frac{\partial}{\partial x^T}\left[x^T A\right] = A^T. \tag{4.6}$$

3. Let x be an $n \times 1$ vector and A an $n \times n$ matrix. Then

$$\frac{\partial}{\partial x^T}\left[x^T Ax\right] = x^T\left(A + A^T\right). \tag{4.7}$$

4. Let x be an $n \times 1$ vector and A an $n \times n$ symmetric matrix. Then

$$\frac{\partial}{\partial x^T}\left[x^T Ax\right] = 2x^T A \tag{4.8}$$

If f is *scalar*-valued, it is straightforward to define the second derivative (Hessian) as follows.

Definition 4.1.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ have continuous first and second partial derivatives at x (so as to satisfy the requirements of Young's theorem). Then*

$$\underbrace{\frac{\partial^2 f(x)}{\partial x \partial x^T}}_{n \times n} \triangleq \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} \triangleq f''(x). \tag{4.9}$$

Note that, by Young's theorem, the Hessian of a scalar-valued function is symmetric.

Proposition 4.1.2 *Let $f(x) \triangleq x^T A x$ where A is symmetric. Then*

$$\frac{\partial^2 f(x)}{\partial x \partial x^T} = 2A \tag{4.10}$$

Occasionally we run into matrix-valued functions, and the way forward then is to vectorize and then differentiate.

Definition 4.1.4 *Let $A = \begin{bmatrix} \underbrace{\mathbf{a}_1}_{m \times 1} & \underbrace{\mathbf{a}_2}_{m \times 1} & \cdots & \underbrace{\mathbf{a}_n}_{m \times 1} \end{bmatrix}$ be an $m \times n$ matrix. Then*

$$\underbrace{\operatorname{vec}(A)}_{mn \times 1} \triangleq \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} \tag{4.11}$$

Definition 4.1.5 *Let $f : \mathbb{R}^k \to \mathbb{R}^{n \times m}$ have partial derivatives at x. Then*

$$\underbrace{\frac{\partial f(x)}{\partial x^T}}_{nm \times k} \triangleq \frac{\partial \operatorname{vec} f(x)}{\partial x^T} \tag{4.12}$$

Having defined the vec operator, we quickly run into cases where we need the Kronecker product, defined as follows.

Definition 4.1.6 *Let* $\underset{m\times n}{A}$ *and* $\underset{k\times l}{B}$ *be matrices. Denote the element in the i:th row and j:th column of A by* a_{ij}*. Then*

$$\underbrace{A\otimes B}_{mk\times ln}\triangleq\begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}. \tag{4.13}$$

Proposition 4.1.3 *Let* $\underset{k\times l}{A}$, $\underset{m\times n}{B}$ *and* $\underset{p\times q}{C}$ *be matrices. Then*

$$\operatorname{vec}(ABC)=\left(C^T\otimes A\right)\operatorname{vec}(B) \tag{4.14}$$

Proof. Exercise. ■

Occasionally we find ourselves wanting to differentiate a vector-valued function with respect to a matrix. Again the way forward is to vectorize.

Definition 4.1.7 *Let* $f:\mathbb{R}^{n\times m}\to\mathbb{R}^k$ *have partial derivatives at x. Then*

$$\underbrace{\frac{\partial f(x)}{\partial A^T}}_{nm\times k}\triangleq\frac{\partial f(x)}{\partial(\operatorname{vec}A)^T} \tag{4.15}$$

Example 4.1.1 *Let* $f:\mathbb{R}^{n\times m}\to\mathbb{R}^n$ *be defined via* $f(\Phi)\triangleq\Phi k$ *where* $k\in\mathbb{R}^m$ *is a constant vector. Then* $f(\Phi)=\left(k^T\otimes I_n\right)\operatorname{vec}\Phi$ *and hence*

$$\frac{\partial f(x)}{\partial\Phi^T}=\left(k^T\otimes I_n\right). \tag{4.16}$$

We are now in a position to state rather general versions of the product and chain rule for matrices.

4.1.2 The product rule

Proposition 4.1.4 (the product rule) *Let $A : \mathbb{R}^l \to \mathbb{R}^{n\times m}$ and $B : \mathbb{R}^l \to \mathbb{R}^{m\times k}$ have partial derivatives at $x \in \mathbb{R}^l$. Then*

$$\frac{\partial}{\partial x^T}\left[A\left(x\right)B\left(x\right)\right] = \left(B\left(x\right)^T \otimes I_n\right)\frac{\partial \operatorname{vec} A\left(x\right)}{\partial x^T} + \left(I_k \otimes A\left(x\right)\right)\frac{\partial \operatorname{vec} B\left(x\right)}{\partial x^T}. \quad (4.17)$$

Kind-of proof. *Suppose $A\left(x\right) \equiv A$. Then, by Proposition 4.1.3,*

$$\operatorname{vec}\left(AB\left(x\right)\right) = \left(I_k \otimes A\right)\operatorname{vec} B\left(x\right). \quad (4.18)$$

Since differentiation is a linear operator, it follows that

$$\frac{\partial \operatorname{vec}\left(AB\left(x\right)\right)}{\partial x^T} = \left(I_k \otimes A\right)\frac{\partial \operatorname{vec} B\left(x\right)}{\partial x^T} \quad (4.19)$$

Conversely, assume that $B\left(x\right) \equiv B$. Then

$$\frac{\partial \operatorname{vec}\left(A\left(x\right)B\right)}{\partial x^T} = \left(B^T \otimes I_n\right)\frac{\partial \operatorname{vec} A\left(x\right)}{\partial x^T} \quad (4.20)$$

Combining the two results yields the product rule.

Corollary 4.1.1 *When we have vector- rather than matrix-valued functions, the formula is drastically simplified. Let $f : \mathbb{R}^l \to \mathbb{R}^m$ and $g : \mathbb{R}^l \to \mathbb{R}^m$ have partial derivatives at $x \in \mathbb{R}^l$. Then*

$$\frac{\partial}{\partial x^T}\left[f\left(x\right)^T g\left(x\right)\right] = g\left(x\right)^T\frac{\partial f\left(x\right)}{\partial x^T} + f\left(x\right)^T\frac{\partial g\left(x\right)}{\partial x^T} \quad (4.21)$$

4.1.3 The chain rule

Proposition 4.1.5 (the chain rule) *Let f and g have partial derivatives at x, and let $h\left(x\right) = \left(f \circ g\right)\left(x\right) = f\left(g\left(x\right)\right)$. Define $y = g\left(x\right)$. Then h has partial derivatives at x and*

$$\frac{\partial h\left(x\right)}{\partial x^T} = \frac{\partial f\left(y\right)}{\partial y^T}\frac{\partial g\left(x\right)}{\partial x^T}. \quad (4.22)$$

With an alternative piece of notation, we have

$$\frac{\partial f\left(g\left(x\right)\right)}{\partial x^{T}}=\frac{\partial f\left(g\left(x\right)\right)}{\partial g^{T}}\frac{\partial g\left(x\right)}{\partial x^{T}}. \tag{4.23}$$

Proof. The scalar chain rule and the definition of matrix multiplication. ■

4.1.4 Taylor's formula in n dimensions

Proposition 4.1.6 *Let $f:\mathbb{S}\to\mathbb{R}^{m}$ be differentiable on the open set $\mathbb{S}\subset\mathbb{R}^{n}$. Let $x_{0}\in\mathbb{S}$. Then there is a function $r:\mathbb{R}^{n}\to\mathbb{R}^{m}$ (which typically depends on x_{0}) such that*

1. *For all $x\in\mathbb{S}$,*

$$f\left(x\right)=f\left(x_{0}\right)+\frac{\partial f\left(x_{0}\right)}{\partial x^{T}}\left(x-x_{0}\right)+r\left(x-x_{0}\right). \tag{4.24}$$

2.

$$\lim_{h\to0}\frac{r\left(h\right)}{\left\|h\right\|}=0 \tag{4.25}$$

Proof. See [26]. ■

Proposition 4.1.7 *Let $f:\mathbb{S}\to\mathbb{R}$ be twice differentiable on the open set $\mathbb{S}\subset\mathbb{R}^{n}$. Let $x_{0}\in\mathbb{S}$. Then there is a function $r:\mathbb{R}^{n}\to\mathbb{R}$ such that*

1. *For all $x\in\mathbb{S}$,*

$$f\left(x\right)=f\left(x_{0}\right)+\frac{\partial f\left(x_{0}\right)}{\partial x^{T}}\left(x-x_{0}\right)+\frac{1}{2}\left(x-x_{0}\right)^{T}\frac{\partial f^{2}\left(x_{0}\right)}{\partial x\partial x^{T}}\left(x-x_{0}\right)+r\left(x-x_{0}\right). \tag{4.26}$$

2.

$$\lim_{h\to0}\frac{r\left(h\right)}{\left\|h\right\|^{2}}=0 \tag{4.27}$$

Proof. See [26]. ■

4.2 Integrating over $\mathbb{R}^n$

4.2.1 Definition

Within the Riemann theory, integrating over rectangles (and the $n-$dimensional counterparts) is just a matter of iterating the process of integration. More precisely, suppose $E \subset \mathbb{R}^2$ is a closed rectangle, i.e. $E = [a_1, b_1] \times [a_2, b_2]$ where we require $a_1 \leq b_1$ and $a_2 \leq b_2$ so that the orientation of our set E is not an issue. We then have the following definition.

Definition 4.2.1 *Let $E \subset \mathbb{R}^2$ be a closed rectangle and let $f : E \to \mathbb{R}$ be a continuous function. Let $\mathbf{x} = \langle x, y \rangle$. Define*

$$\varphi(y) = \left[\int_{a_2}^{b_2} f(x, y)\, dx \right] \tag{4.28}$$

Then

$$\int_E f(\mathbf{x})\, \mathbf{dx} = \int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(x, y)\, dx \right] = \int_{a_1}^{b_1} \varphi(y)\, dy \tag{4.29}$$

Happily, the order of integration does not matter under our assumptions. We have the following proposition.

Proposition 4.2.1 *Let E be a closed rectangle and let $f : E \to \mathbb{R}$ be a continuous function. Then*

$$\int_{a_1}^{b_1} \left[\int_{a_2}^{b_2} f(x, y)\, dx \right] dy = \int_E f(\mathbf{x})\, \mathbf{dx} = \int_{a2}^{b_2} \left[\int_{a_1}^{b_1} f(x, y)\, dy \right] dx. \tag{4.30}$$

Proof. See [17]. ■

Remark 4.2.1 *If you think this is a surprising result, recall that integrals are just sums, and sums (avoiding pathologies where infinity is involved) are the same independent of the order of the terms.*

We can of course generalize this definition and proposition to integration over closed rectangles $E \subset \mathbb{R}^n$, i.e. sets of the form $E = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$. Just keep on iterating the process of integration!

4.2.2 Change of variables

Definition 4.2.2 *Let f be an arbitrary function on $\mathbb{R}^n$ into $\mathbb{R}$. Then the set*

$$S_f = \{x \in X : f(x) \neq 0\} \tag{4.31}$$

is called the support *of f. If S_f is a compact set, then f is said to have compact support.*

Theorem 4.2.1 *Let T be a 1-1 (injective) continuously differentiable function from an open set $E \subset \mathbb{R}^n$ into $\mathbb{R}^n$ such that the Jacobian $J_T(x) \neq 0$ for all $x \in E$. Let f be a continuous function from $\mathbb{R}^n$ into $\mathbb{R}$ such whose support is compact and lies in $T(E)$. Then*

$$\int_{\mathbb{R}^n} f(y)\, dy = \int_{\mathbb{R}^n} f(T(x)) \left|J_T(x)\right| dx. \tag{4.32}$$

Proof. See [17]. ■

Remark 4.2.2 *The reason for having $|J_T(x)|$ instead of $J_T(x)$ is that, with the definition of the integral used in this section, we integrate over subsets of $\mathbb{R}^n$ without regard for their orientation. For example, in the scalar case, we consider $\int_a^b f(x)\, dx$ and $\int_b^a f(x)\, dx$ to be the same. Given that these are defined to be the*

same, we must take steps to assure that, say, the change of variables $T(x) = -x$ *makes no difference, and that is guaranteed by taking the absolute value of the Jacobian.*

Example 4.2.1 (from Econometrics II; calculating the volume of a cylinder) *Let* $c, k \geq 0$. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be defined via*

$$f(x,y) = \begin{cases} c & \text{if } x^2 + y^2 \leq k^2 \\ 0 & \text{otherwise} \end{cases} \tag{4.33}$$

(Draw a picture of this!) We now want to calculate

$$\int_{\mathbb{R}^2} f(x,y)\, dx \tag{4.34}$$

and it turns out to be convenient to use the change of variables approach, noting with satisfaction that f *has compact support. Looking at the picture, it seems that a switch to polar coordinates makes sense. So define*

$$E = \left\{ \begin{bmatrix} r \\ \theta \end{bmatrix} \in \mathbb{R}^2 : 0 < r < k \text{ and } 0 < \theta < 2\pi \right\} \tag{4.35}$$

and T *on* E *via*

$$T(r,\theta) = \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} \tag{4.36}$$

Apparently the Jacobian is

$$J_T(r,\theta) = \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r \tag{4.37}$$

and

$$f(T(r,\theta)) = \begin{cases} c & \text{if } 0 \leq r \leq k \text{ and } 0 \leq \theta \leq 2\pi \\ 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

Hence

$$\int_{\mathbb{R}^2} f(x,y) = \int_0^{2\pi} \left[\int_0^k cr\, dr \right] d\theta = ck^2\pi. \tag{4.39}$$

Chapter 5

Abstract spaces

5.1 Metric spaces

5.1.1 Introduction

The idea of a metric space is to generalize from a Euclidean space to any space for which it makes sense to talk about the *distance* between two points. This distance is called a *metric*, as defined below. Notice how the axioms are tailor-made to fit our intuitions about distances in physical space.

5.1.2 Definitions

Definition 5.1.1 *A* metric space *is a non-empty set X associated with a function (called a* metric*) $\mu : X \times X \to \mathbb{R}$ such that, for all $x, y, z \in X$,*

1. $\mu(x, y) \geq 0$ *with equality iff* $x = y$,
2. $\mu(x, y) = \mu(y, x)$, *and*
3. $\mu(x, z) \leq \mu(x, y) + \mu(y, z)$ *(the triangle inequality).*

Remark 5.1.1 *Subsets $Y \subset X$ inherit the same metric as X is associated with and are themselves metric spaces.*

Definition 5.1.2 *A subset O of a metric space (X, μ) is said to be* open *if for each $x \in O$ there is a real number $\varepsilon_x > 0$ such that*

$$\{y \in X : \mu(x, y) < \varepsilon_x\} \subset O. \tag{5.1}$$

Remark 5.1.2 *By defining the open sets, we have also defined the continuous mappings on X (into some other space where the open sets are defined). Just declare any mapping $f : X \to Y$ to be continuous whenever $f^{-1}(O)$ is open for every open set $O \subset Y$. Similarly, we have defined the convergent sequences. Just declare any sequence $\langle x_n \rangle$ to converge to the element $x \in X$ if for every neighborhood O of x (recall that a neighbor hood of x is any open set O such that $x \in O$) there is a natural number N with the property that $x_n \in O$ for all $n > N$.*

Definition 5.1.3 *A subset F of a metric space (X, μ) is said to be* closed *if its complement $F^c = \{x \in X : x \notin F\}$ is open.*

Definition 5.1.4 *A* Cauchy *sequence is a sequence $\langle x_n \rangle_{n=0}^{\infty}$ such that for each $\varepsilon > 0$ there exists an N_ε such that $\mu(x_n, x_m) < \varepsilon$ for all $n, m > N_\varepsilon$.*

Definition 5.1.5 *A sequence $\langle x_n \rangle_{n=0}^{\infty}$ is said to* converge *to an element $x \in X$ if for each $\varepsilon > 0$ there exists an N_ε such that $\mu(x_n, x) < \varepsilon$ for all $n > N_\varepsilon$.*

Definition 5.1.6 *A metric space (X, μ) is said to be* complete *if every Cauchy sequence from X converges to some element $x \in X$.*

The property of completeness is closely linked to the property of closedness. Indeed, we have the following propositions.

Proposition 5.1.1 *Let X be a complete metric space and let $Y \subset X$ be closed. Then Y is complete.*

Proof. Let $\langle x_k \rangle_{k=1}^{\infty}$ be a Cauchy sequence from Y. By the completeness of X, $\langle x_k \rangle_{k=1}^{\infty}$ converges to some element $x \in X$. But since Y is closed, it contains all its limit points. Hence $x \in Y$. ■

Proposition 5.1.2 *Let X be a complete metric space and let $Y \subset X$ be complete. Then Y is closed.*

Proof. It suffices to show that Y contains all its limit points. But this follows from the fact that every convergent sequence is a Cauchy sequence. ■

5.1.3 The contraction mapping theorem

We now come to the highlight of this section, which is the contraction mapping theorem, also known as the contraction principle, or Banach's fixed point theorem.

Definition 5.1.7 *Let (X, μ) be a metric space. A function $\varphi : X \to X$ is called a* contraction *if there is a $0 \leq \beta < 1$ such that, for all $x, y \in X$,*

$$\mu(\varphi(x), \varphi(y)) \leq \beta \mu(x, y) \tag{5.2}$$

Lemma 5.1.1 *Let φ be a contraction on (X, μ). Then φ is uniformly continuous.*

Proof. We need to show that, for each $\varepsilon > 0$ there corresponds a $\delta_\varepsilon > 0$ such that $\mu(\varphi(x), \varphi(y)) \leq \varepsilon$ for all $x, y \in$ satisfying $\mu(x, y) \leq \delta_\varepsilon$. Put $\delta_\varepsilon = \dfrac{\varepsilon}{\beta}$ and we are done. ■

Theorem 5.1.1 (Banach's fixed point theorem) *Let (X, μ) be a complete metric space and let φ be a contraction. Then there is a unique $x \in X$ such*

that $\varphi(x) = x$. (φ has a unique fixed point.) Moreover, the proof of this theorem is constructive*, in the sense that it gives us an algorithm which delivers an arbitrarily good approximation to the fixed point x.*

Proof. Uniqueness is obvious, since if $\varphi(y) = y$ and $\varphi(x) = x$ then $\mu(\varphi(x), \varphi(y)) = \mu(x, y) \leq \beta\mu(x, y)$ for some $\beta < 1$ which can only happen for $\mu(x, y) = 0$ and hence $x = y$. To show existence, pick x_0 arbitrarily and define $\langle x_n \rangle$ recursively by setting

$$x_{n+1} = \varphi(x_n) \qquad n = 0, 1, 2, ... \tag{5.3}$$

We now show that $\langle x_n \rangle$ is a Cauchy sequence. Choose $\beta < 1$ so that (5.2) holds for all $x, y \in X$. For $n \geq 1$ we have

$$\mu(x_{n+1}, x_n) = \mu(\varphi(x_n), \varphi(x_{n-1})) \leq \beta\mu(x_n, x_{n-1}). \tag{5.4}$$

By induction,

$$\mu(x_{n+1}, x_n) \leq \beta^n \mu(x_1, x_0) \tag{5.5}$$

Putting (without loss of generality since $\mu(x_n, x_m) = \mu(x_m, x_n)$) $n < m$, it follows, by the triangle inequality, that

$$\mu(x_n, x_m) \leq \sum_{i=n+1}^{m} \mu(x_i, x_{i-1}). \tag{5.6}$$

By (5.5),

$$\sum_{i=n+1}^{m} \mu(x_i, x_{i-1}) \leq \left(\beta^n + \beta^{n+1} + \cdots + \beta^{m-1}\right) \mu(x_1, x_0). \tag{5.7}$$

But

$$\begin{aligned}
\left(\beta^n + \beta^{n+1} + \cdots + \beta^{m-1}\right)\mu\left(x_1, x_0\right) &= \\
= \beta^n\left(1 + \beta + \cdots + \beta^{m-n-1}\right)\mu\left(x_1, x_0\right) &\leq \\
\leq \beta^n\left[\sum_{k=0}^{\infty}\beta^k\right]\mu\left(x_1, x_0\right) &= \\
= \frac{\beta^n}{1-\beta}\mu\left(x_1, x_0\right)
\end{aligned} \tag{5.8}$$

Hence $\mu\left(x_n, x_m\right)$ is arbitrarily small when n and m are sufficiently large. It follows that $\langle x_n \rangle$ is a Cauchy sequence. Since X is complete, $x_n \to x$ for some $x \in X$. Since φ is continuous,

$$\varphi\left(x\right) = \varphi\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty}\varphi\left(x_n\right) = \lim_{n\to\infty} x_{n+1} = x \tag{5.9}$$

so x is a fixed point of φ. ■

5.2 Banach spaces

In the following, a 'scalar' will refer to a real number. With minor modifications of some of the axioms and proofs, we could have used complex scalars instead.

Definition 5.2.1 *A vector space is a non-empty set $\mathcal{S}$ associated with an addition operation $+ : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$ and a scalar multiplication operation $\cdot : \mathbb{R} \times \mathcal{S} \to \mathcal{S}$ such that for all $x, y \in \mathcal{S}$ and all scalars α, $\alpha\left(x + y\right) \in \mathcal{S}$. For $+$ and $\cdot$ to qualify as addition and scalar multiplication operators, we require the following axioms.*

1. $x + y = y + x$ *for all* $x, y \in \mathcal{S}$
2. $x + \left(y + z\right) = \left(x + y\right) + z$ *for all* $x, y, z \in \mathcal{S}$

3. *There exists an element* $\theta \in \mathcal{S}$ *such that* $x + \theta = x$ *for all* $x \in \mathcal{S}$

4. $\alpha(x + y) = \alpha x + \alpha y$ *for all* $\alpha \in \mathbb{R}$ *and all* $x, y \in \mathcal{S}$

5. $(\alpha + \beta) x = \alpha x + \beta x$ *for all* $\alpha, \beta \in \mathbb{R}$ *and all* $x \in \mathcal{S}$

6. $\alpha(\beta x) = (\alpha\beta) x$ *for all* $\alpha, \beta \in \mathbb{R}$ *and all* $x \in \mathcal{S}$

7. $0x = \theta$ *for all* $x \in \mathcal{S}$

8. $1x = x$ *for all* $x \in \mathcal{S}$

Remark 5.2.1 *It is not hard to see that the zero element* θ *is unique. Also, subtraction is defined via* $x - y = x + (-1) y$*. It follows that* $x - x = (1 + (-1)) x = 0x = \theta$ *for all* $x \in \mathcal{S}$*.*

Definition 5.2.2 *A* norm *on a vector space* $\mathcal{S}$ *is a function* $\|\cdot\| : \mathcal{S} \to \mathbb{R}$ *such that*

1. *For each* $x \in \mathcal{S}$, $\|x\| \geq 0$ *with equality for and only for the zero element* $\theta \in \mathcal{S}$*.*

2. *For each* $x \in \mathcal{S}$ *and every scalar* α, $\|\alpha x\| = |\alpha| \|x\|$.

3. *(The triangle inequality.) For all* $x, y \in \mathcal{S}$, $\|x + y\| \leq \|x\| + \|y\|$.

Proposition 5.2.1 *Let* $(\mathcal{S}, \|\cdot\|)$ *be a vector space with an associated norm. Define the function* $\mu : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ *via*

$$\mu(x, y) = \|x - y\| \tag{5.10}$$

Then $(\mathcal{S}, \mu)$ *is a metric space, and* μ *is called the metric generated by* $\|\cdot\|$*.*

Proof *Exercise.*

Definition 5.2.3 *A Banach space is an ordered pair* $(\mathcal{S}, \|\cdot\|)$ *such that* $\mathcal{S}$ *is a vector space and* $\mathcal{S}$ *is complete in the metric generated by the norm* $\|\cdot\|$*. (Once the norm has been defined, it is usually suppressed in the notation for convenience.)*

The property of completeness is closely linked to the property of closedness. Indeed, we have the following propositions.

Proposition 5.2.2 *Let* $\mathcal{S}$ *be a Banach space and let* $\mathcal{T} \subset \mathcal{S}$ *be closed. Then* $\mathcal{T}$ *is complete.*

Proof. Let $\langle x_k \rangle_{k=1}^{\infty}$ be a Cauchy sequence from $\mathcal{T}$. By the completeness of $\mathcal{S}$, $\langle x_k \rangle_{k=1}^{\infty}$ converges to some element $x \in \mathcal{S}$. Since the sequence is taken from $\mathcal{T}$, x must be a point of closure of $\mathcal{T}$. But since $\mathcal{T}$ is closed, it contains all its points of closure. Hence $x \in \mathcal{T}$. □

Proposition 5.2.3 *Let* $\mathcal{S}$ *be a Banach space and let* $\mathcal{T} \subset \mathcal{S}$ *be complete. Then* $\mathcal{S}$ *is closed.*

Proof. It suffices to show that $\mathcal{T}$ contains all its limit points. But this follows from the fact that every convergent sequence is a Cauchy sequence. ■

5.3 Hilbert spaces

5.3.1 Definitions and basic properties

Definition 5.3.1 *An* inner product *on a vector space* $\mathcal{H}$ *is a function* $(\cdot,\cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ *such that*

1. *For all* $x, y \in \mathcal{H}$, $(x, y) = (y, x)$.[1]

[1] When the scalar field is $\mathbb{C}$ rather than $\mathbb{R}$, this axiom becomes $(x, y) = \overline{(y, x)}$ where $\overline{z}$ is the complex conjugate of z. You may want to check that this holds for the Euclidean inner product in $\mathbb{C}^n$, i.e. $(x, y) = \sum_{i=1}^{n} x_i \overline{y_i}$.

2. *For all* $x, y \in \mathcal{H}$ *and all scalars* α, β, $(\alpha x + \beta y, z) = \alpha (x, z) + \beta (y, z)$.

3. *For all* $x \in \mathcal{H}$, $\langle x, x \rangle \geq 0$ *with equality iff* $x = \theta$.

Proposition 5.3.1 *The function* $\|\cdot\|$ *defined via* $\|x\| = \sqrt{(x, x)}$ *is a norm. This norm is called the norm* generated *by* $\langle \cdot, \cdot \rangle$.

Proof *To show the triangle inequality, square both sides and use the Cauchy-Schwarz inequality (see below).*

Definition 5.3.2 *A Hilbert space is an ordered pair* $(\mathcal{H}, (\cdot, \cdot))$ *such that*

1. $\mathcal{H}$ *is a vector space.*

2. $(\cdot, \cdot)$ *is an inner product.*

3. *The normed space* $(\mathcal{H}, \|\cdot\|)$ *is complete, where* $\|\cdot\|$ *is the norm generated by* $(\cdot, \cdot)$.

Henceforth whenever the symbol $\mathcal{H}$ appears, it will denote a Hilbert space (with some associated inner product).

Intuitively, a Hilbert space is a generalization of $\mathbb{R}^n$ with the usual inner product $(x, y) = \sum_{i=1}^{n} x_i y_i$ (and hence the Euclidean norm), preserving those properties which have to do with *geometry* so that we can exploit our ability to visualize (our intuitive picture of) physical space in order to deal with problems that have nothing whatever to do with physical space. In particular, the ideas of *distance, length,* and *orthogonality* are preserved. As expected, we have

Definition 5.3.3 *The distance between two elements* $x, y \in \mathcal{H}$ *is defined as* $\|x - y\|$, *where* $\|\cdot\|$ *is the norm generated by the inner product associated with* $\mathcal{H}$.

Definition 5.3.4 *Two elements* $x, y \in \mathcal{H}$ *are said to be orthogonal if* $(x, y) = 0$.

Some further definitions and facts are needed before we can proceed.

Definition 5.3.5 *A Hilbert subspace $\mathcal{G} \subset \mathcal{H}$ is a subset $\mathcal{G}$ of $\mathcal{H}$ such that $\mathcal{G}$, too, is a Hilbert space.*

The following proposition is very useful, since it guarantees the well-definedness of the inner product between two elements of finite norm.

Proposition 5.3.2 *(The Cauchy-Schwarz inequality). Let $\mathcal{H}$ be a Hilbert space. Then, for any $x, y \in \mathcal{H}$, we have*

$$|(x, y)| \leq \|x\|\|y\| \tag{5.11}$$

Proof. *If $x = \theta$ or $y = \theta$ the inequality is trivial. So suppose $\|x\|, \|y\| > 0$ and let $\lambda > 0$ be a real number. We get*

$$\begin{aligned} 0 \leq &\| x - \lambda y\|^2 = (x - \lambda y, x - \lambda y) = \\ &= \|x\|^2 + \lambda^2 \|y\|^2 - 2\lambda \langle x, y \rangle \end{aligned} \tag{5.12}$$

Dividing by λ, it follows that

$$2 \langle x, y \rangle \leq \frac{1}{\lambda} \|x\|^2 + \lambda \|y\|^2 \tag{5.13}$$

Clearly this is true for all $\lambda > 0$. In particular it is true for $\lambda = \dfrac{\|x\|}{\|y\|}$ which is strictly positive by assumption. It follows that $(x, y) \leq \|x\|\|y\|$. To show that $-(x, y) \leq \|x\|\|y\|$, note that $-\langle x, y \rangle = (-x, y)$, and since $(-x) \in \mathcal{H}$ we have just shown that $(-x, y) \leq \|x\|\|y\|$.[2] $\square$

Proposition 5.3.3 (The parallelogram identity) *Let $\mathcal{H}$ be a Hilbert space. Then, for any $x, y \in \mathcal{H}$, we have*

$$\|x + y\|^2 + \|x - y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right) \tag{5.14}$$

[2] In the case where the scalar field is $\mathbb{C}$ rather than $\mathbb{R}$, the proof is slightly different, but just as simple. See [18].

Proof. Exercise. ■

The following proposition states that the inner product is (uniformly) continuous.

Proposition 5.3.4 *Let $\mathcal{H}$ be a Hilbert space, and let $y \in \mathcal{H}$ be fixed. Then for each $\varepsilon > 0$ there exists a $\delta > 0$ such that $|(x_1, y) - (x_2, y)| \leq \varepsilon$ for all $x_1, x_2 \in \mathcal{H}$ such that $\|x_1 - x_2\| \leq \delta$.*

Proof.

$$\begin{aligned} |(x_1, y) - (x_2, y)| &= |(x_1 - x_2, y)| \leq \{\text{Cauchy-Schwarz}\} \\ &\leq \|x_1 - x_2\|\|y\| \end{aligned} \tag{5.15}$$

Set $\delta = \dfrac{\varepsilon}{\|y\|}$, and we are done. ■

Definition 5.3.6 *Let $\mathbb{G} \subset \mathcal{H}$ be an arbitrary subset of $\mathcal{H}$. Then the orthogonal complement $\mathbb{G}^\perp$ of $\mathbb{G}$ is defined via $\mathbb{G}^\perp \triangleq \{y \in \mathcal{H} : (x, y) = 0 \text{ for all } x \in \mathbb{G}\}$.*

Proposition 5.3.5 *The orthogonal complement $\mathbb{G}^\perp$ of a subset $\mathbb{G}$ of a Hilbert space $\mathcal{H}$ is a Hilbert subspace.*

Proof. It is easy to see that $\mathbb{G}^\perp$ is a vector space. By the fact that any closed subset of a Hilbert space is complete, all that remains to be shown is that $\mathbb{G}^\perp$ is closed. To show that, we use the continuity of the inner product. Consider first the set

$$x^\perp = \{y \in \mathcal{H} : (x, y) = 0\} \tag{5.16}$$

But this is easily recognized as the inverse image of the closed set $\{0\}$ under a continuous mapping. Hence $x^\perp$ is closed. Now notice that

$$\mathbb{G}^\perp = \bigcap_{x \in \mathbb{G}} x^\perp \tag{5.17}$$

Hence $\mathbb{G}^\perp$ is the intersection of a family of closed sets. It follows that $\mathbb{G}^\perp$ is itself closed. ■

A useful corollary is that if $\mathbf{G}$ is dense in $\mathbb{G}$, then $\mathbf{G}^\perp$=$\mathbb{G}^\perp$. More explicitly, we have the following definition and proposition.

Definition 5.3.7 *A set* $\mathbf{G} \subset \mathbb{G} \subset \mathcal{H}$ *(where* $\mathcal{H}$ *is a Hilbert space) is said to be dense in* $\mathbb{G}$ *if the closure of* $\mathbf{G}$ *(relative to* $\mathbb{G}$*) is equal to* $\mathbb{G}$*. Equivalently,* $\mathbf{G}$ *is said to be dense in* $\mathbb{G}$ *if* $\mathbf{G} \subset \mathbb{G}$ *and if for every* $z \in \mathbb{G}$ *there exists a sequence* $\{z_k\}_{k=1}^{\infty}$ *such that* $z_k \in \mathbf{G}$ *for each* k *and* $z_k \to z$ *as* $k \to \infty$*.*

Proposition 5.3.6 *Let* $\mathbb{G}$ *be an arbitrary subset of a Hilbert space* $\mathcal{H}$*, and let* $\mathbf{G}$ *be a dense subset of* $\mathbb{G}$*. Then* $\mathbb{G}^\perp = \mathbf{G}^\perp$*.*

Proof. Exercise. ■

We now state the most important theorem in Hilbert space theory.

5.3.2 The projection theorem

Theorem 5.3.1 (the projection theorem) *Let* $\mathcal{G} \subset \mathcal{H}$ *be a Hilbert subspace and let* $x \in \mathcal{H}$*. Then*

1. *There exists a unique element* $\hat{x} \in \mathcal{G}$ *(called the* projection *of* x *onto* $\mathcal{G}$*) such that*

$$\|x - \hat{x}\| = \inf_{y \in \mathcal{G}} \|x - y\| \tag{5.18}$$

   where $\|\cdot\|$ *is the norm generated by the inner product associated with* $\mathcal{H}$*.*

2. $\hat{x}$ *is (uniquely) characterized by*

$$(x - \hat{x}) \in \mathcal{G}^\perp \tag{5.19}$$

Remark 5.3.1 *The word 'characterize' is used here in the strong sense, i.e. it is claimed that (5.18) and (5.19) are equivalent.*

Proof. In order to prove part 1 we begin by noting that $\mathcal{G}$, since it is a Hilbert subspace, is both complete and convex. Note that any vector space is convex, but that the converse does not hold. Now fix $x \in \mathcal{H}$ and define

$$d = \inf_{y \in \mathcal{G}} \|x - y\|^2 \tag{5.20}$$

Clearly d exists since the set of squared norms $\|x - y\|^2$ is a set of real numbers bounded below by 0. Now since d is the greatest lower bound of $\|x - y\|^2$ there exists a sequence $\langle y_k \rangle_{k=1}^{\infty}$ from $\mathcal{G}$ such that, for each $\varepsilon > 0$, there exists an N_ε such that

$$\|x - y_k\|^2 \leq d + \varepsilon \tag{5.21}$$

for all $k \geq N_\varepsilon$. We now want to show that any such sequence $\langle y_k \rangle$ is a Cauchy sequence. For that purpose, define

$$u = x - y_m \tag{5.22}$$

$$v = x - y_n \tag{5.23}$$

Now applying the parallelogram identity to u and v, we get

$$\|2x - y_m - y_n\|^2 + \|y_n - y_m\|^2 = 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) \tag{5.24}$$

which may be manipulated to become

$$4\|x - \frac{1}{2}(y_m + y_n)\|^2 + \|y_m - y_n\|^2 = 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) \tag{5.25}$$

Now since $\mathcal{G}$ is convex, $\frac{1}{2}(y_m + y_n) \in \mathcal{G}$ and consequently $\|x - \frac{1}{2}(y_m + y_n)\|^2 \geq d$. It follows that

$$\|y_m - y_n\|^2 \leq 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) - 4d \tag{5.26}$$

Now consider any $\varepsilon > 0$, choose a corresponding N_ε such that $\|x - y_k\|^2 \leq d + \varepsilon/4$ for all $k \geq N_\varepsilon$ (such an N_ε exists as we have seen). Then, for all $n, m \geq N_\varepsilon$, we have

$$\|y_m - y_n\|^2 \leq 2\left(\|x - y_m\|^2 + \|x - y_n\|^2\right) - 4d \leq \varepsilon \tag{5.27}$$

Hence $\langle y_k \rangle$ is a Cauchy sequence. By the completeness of $\mathcal{G}$, it converges to some element $\widehat{x} \in \mathcal{G}$. By the continuity of the inner product, $\|x - \widehat{x}\|^2 = d$. Hence $\widehat{x}$ is the projection we seek. To show that $\widehat{x}$ is unique, consider another projection $y \in \mathcal{G}$ and the sequence $\langle \widehat{x}, y, \widehat{x}, y, \widehat{x}, y... \rangle$. By the argument above, this is a Cauchy sequence. But then $\widehat{x} = y$. Hence (1) is proved. The proof of part (2) comes in two parts. First we show that any $\widehat{x}$ that satisfies (5.18) also satisfies (5.19). Suppose, then, that $\widehat{x}$ satisfies (5.18). Define $\varepsilon = x - \widehat{x}$ and consider an element $y = \widehat{x} + \alpha z$ where $z \in \mathcal{G}$ and $\alpha \in \mathbb{R}$. Since $\mathcal{G}$ is a vector space, it follows that $y \in \mathcal{G}$. Now since $\widehat{x}$ satisfies (5.18), y is no closer to x than $\widehat{x}$ is. Hence

$$\begin{aligned} \|\varepsilon\|^2 &\leq \|\varepsilon - \alpha z\|^2 = (\varepsilon - \alpha z, \varepsilon - \alpha z) = \\ &= \|\varepsilon\|^2 + \alpha^2 \|z\|^2 - 2\alpha\,(\varepsilon, z) \end{aligned} \tag{5.28}$$

Simplifying, we get

$$0 \leq \alpha^2 \|z\|^2 - 2\alpha\,(\varepsilon, z) \tag{5.29}$$

This is true for all scalars α. In particular, set $\alpha = (\varepsilon, z)$. We get

$$0 \leq (\varepsilon, z)^2 \left(\|z\|^2 - 2\right) \tag{5.30}$$

For this to be true for all $z \in \mathcal{G}$ we must have $(\varepsilon, z) = 0$ for all $z \in \mathcal{G}$ such that $\|z\|^2 < 2$. But then (why?) we must have $(\varepsilon, z) = 0$ for all $z \in \mathcal{G}$. Hence $\varepsilon \in \mathcal{G}^\perp$. Now we want to prove the converse, i.e. that if $\widehat{x}$ satisfies (5.19), then it also satisfies (5.18). Thus consider an element $\widehat{x} \in \mathcal{G}$ which satisfies (5.19) and let

$y \in \mathcal{G}$. Mechanical calculations reveal that

$$\begin{aligned} \|x-y\|^2 &= (x-\widehat{x}+\widehat{x}-y, x-\widehat{x}+\widehat{x}-y) = \\ &= \|x-\widehat{x}\|^2 + \|\widehat{x}-y\|^2 + 2\,(x-\widehat{x}, \widehat{x}-y) \end{aligned} \tag{5.31}$$

Now since $(x-\widehat{x}) \in \mathcal{G}^{\perp}$ and $(\widehat{x}-y) \in \mathcal{G}$ (recall that $\mathcal{G}$ is a vector space), the last term disappears, and our minimization problem becomes (disregarding the constant term $\|x-\widehat{x}\|^2$)

$$\min_{y \in \mathcal{G}} \|\widehat{x}-y\| \tag{5.32}$$

Clearly $\widehat{x}$ solves this problem. (Note that it doesn't matter for the solution whether we minimize a norm or its square.) Indeed, since $\|\widehat{x}-y\| = 0$ implies $\widehat{x} = y$ we may conclude that if some $\widehat{x}$ satisfies (5.19), then it is the *unique* solution. ■

Remark 5.3.2 *Given our definition of distance, $\hat{x}$ is by definition that element in $\mathcal{G}$ which is closest to x.*

Remark 5.3.3 *A good way to understand intuitively why the projection theorem is true is to visualize the projection of a point in $\mathbb{R}^3$ onto a 2-dimensional plane through the origin.*

Remark 5.3.4 *Note that our characterization in terms of orthogonality of the projection onto $\mathcal{G}$ is closely linked to the fact that $\mathcal{G}$ is assumed to be a vector space. Meanwhile, our proof that the projection problem*

$$\min_{y \in \mathbf{G}} \|x-y\| \tag{5.33}$$

has a unique solution works for any closed and convex subset $\mathbf{G} \subset \mathcal{H}$. For such general $\mathbf{G}$, then, we know that there is a unique solution, but we have not shown that the projection error must be orthogonal to $\mathbf{G}$. Indeed it need not be. Consider

for example the projection of an arbitrary point $x \in \mathbb{R}^n$ *onto the* k*-dimensional plane* $\mathbf{G} = \left\{z \in \mathbb{R}^n : z = z^1 + A\beta \text{ for some } \beta \in \mathbb{R}^k\right\}$ *where* $z^1 \in \mathbb{R}^n$ *and* A *is an* $n \times k$ *matrix. (Note that* $\mathbf{G}$ *is not a vector space unless* $z^1 = \theta$*). Then it is (at least intuitively) clear that the projection of* x *onto* $\mathbf{G}$ *must be orthogonal to every vector* along *the plane rather than to every vector* in *the plane. More precisely, the projection* $\widehat{x}_{\mathbf{G}}$ *satisfies*

$$(z - y, x - \widehat{x}_{\mathbf{G}}) = 0 \tag{5.34}$$

for all $z, y \in \mathbf{G}$*.*

Corollary 5.3.1 (the repeated projection theorem) *Let* $\mathcal{G} \subset \mathcal{F} \subset \mathcal{H}$ *be Hilbert spaces. Let* $\hat{x}_G$ *be the projection of* $x \in \mathcal{H}$ *onto* $\mathcal{G}$ *and let* $\hat{x}_F$ *be the projection of* x *onto* $\mathcal{F}$*. Then the projection of* $\hat{x}_F$ *onto* $\mathcal{G}$ *is simply* $\hat{x}_G$*.*

Proof. By the characterization of the projection, it suffices to prove that $(\hat{x}_F - \hat{x}_G) \in \mathcal{G}^\perp$. To do this, we rewrite $(\hat{x}_F - \hat{x}_G) = (x - \hat{x}_G) - (x - \hat{x}_F)$. Since the orthogonal complement of a Hilbert subspace is a vector space and hence closed under addition and scalar multiplication, it suffices to show that $(x - \hat{x}_G) \in \mathcal{G}^\perp$ and that $(x - \hat{x}_F) \in \mathcal{G}^\perp$. The truth of the first statement follows immediately from the characterization of the projection. The truth of the second one can be seen by noting that $\mathcal{F}^\perp \subset \mathcal{G}^\perp$. □

5.3.3 Fourier analysis

The previous section asserted that the projection existed, was unique and had a certain orthogonality property. But how is the projection calculated in concrete cases? That question was left unanswered, but will be addressed in this section.

You will recall from linear algebra in $\mathbb{R}^n$ that a useful way of representing a vector $x \in \mathbb{R}^n$ is as a linear combination of some orthogonal basis $\mathbb{B} =$

$\{x_k;\ k = 1, 2, ..., n\}$. You will remember that if $\mathbb{B}$ spans $\mathbb{R}^n$, then for each $x \in \mathbb{R}^n$ there exist scalars $\{\varphi_k;\ k = 1, 2, ..., n\}$ such that

$$x = \sum_{k=1}^{n} \varphi_k x_k \tag{5.35}$$

where the x_k are the elements of $\mathbb{B}$. A popular choice of orthogonal basis vectors is of course the unit vectors, but there are many other orthogonal bases of $\mathbb{R}^n$.

These ideas can easily be generalized to any Hilbert space, since the essential part of the whole program is orthogonality, and orthogonality is, as we have seen, what Hilbert spaces are all about.

There are, however, two issues to be dealt with before taking the leap from $\mathbb{R}^n$ into Hilbert space. One is that Hilbert spaces don't always have a finite basis (i.e. they may be infinite-dimensional). Actually, there are many interesting finite-dimensional Hilbert spaces (e.g. the set of polynomials of a fixed degree defined on a compact interval), but we don't want to exclude the infinite-dimensional ones. Nevertheless, in some cases we would like our Hilbert spaces to have a *countable* basis. After all this loose talk it is time for some definitions and propositions.

Definition 5.3.8 *Let $\langle \mathcal{H}, (\cdot, \cdot) \rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k : k = 1, ...\}$ be a countable subset of $\mathcal{H}$. Then $\mathbb{B}$ is called* orthogonal *if, for all $j \neq k$ we have*

$$(x_j, x_k) = 0. \tag{5.36}$$

If, in addition, $(x_j, x_j) = 1$ for all $j = 1, 2, ...$, then $\mathbb{B}$ is called orthonormal.

Definition 5.3.9 *Let $\langle \mathcal{H}, (\cdot, \cdot) \rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then the* closed span $\overline{\text{sp}}\,(\mathbb{B})$ *of $\mathbb{B}$ is defined as follows. Let* $\text{sp}\,(\mathbb{B})$ *be the set of elements $x \in \mathcal{H}$ such that there exist scalars $\{\varphi_k;\ k = 1, 2, ...\}$ with the property that*

$$x = \sum_{k=1}^{n} \varphi_k x_k \tag{5.37}$$

where the x_k are elements of $\mathbb{B}$. We then define $\overline{\text{sp}}(\mathbb{B})$ as the closure of $\text{sp}(\mathbb{B})$. Intuitively, taking the closure means that we include all the infinite sums as well as the finite ones.

Definition 5.3.10 *Let $\langle \mathcal{H}, (\cdot,\cdot)\rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then $\mathbb{B}$ is said to* span $\mathcal{H}$ *if $\overline{\text{sp}}(\mathbb{B}) = \mathcal{H}$.*

Proposition 5.3.7 *Let $\langle \mathcal{H}, (\cdot,\cdot)\rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Suppose $\mathbb{B}$ spans $\mathcal{H}$. Then, for each $x \in \mathcal{H}$, there are scalars $\{\varphi_k;\ k = 1, 2, ...\}$ such that*

$$\lim_{n\to\infty} \left\| x - \sum_{k=1}^{n} \varphi_k x_k \right\| = 0 \tag{5.38}$$

and we sometimes write

$$x = \sum_{k=1}^{\infty} \varphi_k x_k. \tag{5.39}$$

Proof *Obvious.*

Proposition 5.3.8 *Let $\langle \mathcal{H}, (\cdot,\cdot)\rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable subset of $\mathcal{H}$. Then $\overline{\text{sp}}(\mathbb{B})$ is a Hilbert subspace.*

Proof *Obvious.*

Definition 5.3.11 *Let $\langle \mathcal{H}, (\cdot,\cdot)\rangle$ be a Hilbert space and suppose $\mathbb{B}$ is orthogonal (orthonormal) and spans $\mathcal{H}$. Then $\mathbb{B}$ is called an orthogonal (orthonormal) basis for $\mathcal{H}$.*

Definition 5.3.12 *A Hilbert space $\langle \mathcal{H}, (\cdot,\cdot)\rangle$ is called separable if it has a countable dense subset.*

Proposition 5.3.9 *A Hilbert space has a countable orthogonal basis iff it is separable.*

Usually it is hard to figure out whether a particular Hilbert space is separable or not. Indeed, even if we are given a countable subset $\mathbb{B}$, it is not always easy (for an economist, that is) to check whether this set spans $\mathcal{H}$. Happily, we can often draw upon standard results, such as the Stone-Weierstrass theorem which asserts that the set of polynomials of arbitrary degree spans the set of continuous functions on $\mathbb{R}$. We also have the remarkable result that the set of trigonometric polynomials (linear combinations of $\sin nx$ and $\cos nx$ where n is an arbitrary integer) span the set of square integrable functions defined on a compact interval.

Usually, though, orthogonality is much more important than spanning, and orthogonality is usually not too hard to confirm in concrete cases. To see why orthogonality is more important than spanning, consider the following example. Suppose we want to approximate a function by a polynomial of fixed degree. Then (we will see why!) it makes a lot of sense to project our function onto a set of orthogonal polynomials. In this context, we are not too worried about the fact that the polynomials of a fixed degree may fail to span the set of functions that we are trying to approximate. After all, what we sought was approximation, not perfect representation.

In any case, suppose $\langle \mathcal{H}, (\cdot, \cdot) \rangle$ is a Hilbert space with the (countable!) orthonormal subset $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$. Now consider an arbitrary element $x \in \mathcal{H}$. Our project now is to find the projection $\hat{x}$ onto $\overline{\text{sp}}(\mathbb{B})$. By the definition of the closed span, there are scalars $\{\varphi_k;\ k = 1, 2, ...\}$ such that

$$\hat{x} = \sum_{k=1}^{\infty} \varphi_k x_k. \tag{5.40}$$

The scalars $\{\varphi_k;\ k = 1, 2, ...\}$ are called the Fourier coefficients of x (with respect to $\mathbb{B}$). But what values do they have? To find out, recall the characterization of the projection. The idea is to choose the φ_k so that the projection error is orthogonal to every vector $y \in \overline{\text{sp}}(\mathbb{B})$. Actually, it suffices to set the projection

error orthogonal to every vector $x_k \in \mathbb{B}$. Then, for every $j = 1, 2, ...$ we have

$$\left(x - \sum_{k=1}^{\infty} \varphi_k x_k, x_j \right) = 0. \tag{5.41}$$

But

$$\begin{aligned} \left(x - \sum_{k=1}^{\infty} \varphi_k x_k, x_j \right) &= (x, x_j) - \sum_{k=1}^{\infty} \varphi_k (x_j, x_k) = \\ = \{\text{orthogonality!}\} &= (x, x_j) - \varphi_j (x_j, x_j) = \{\text{orthonormality!}\} = \\ = (x, x_j) - \varphi_j \end{aligned} \tag{5.42}$$

Hence $\varphi_j = (x, x_j)$ for each $j = 1, 2, ...$, and we have the following proposition.

Proposition 5.3.10 *Let $\langle \mathcal{H}, (\cdot, \cdot) \rangle$ be a Hilbert space and let $\mathbb{B} = \{x_k;\ k = 1, 2, ...\}$ be a countable orthonormal subset of $\mathcal{H}$. Let $x \in \mathcal{H}$. Then the projection $\hat{x}$ of x onto $\bar{s}(\mathbb{B})$ is*

$$\hat{x} = \sum_{k=1}^{\infty} (x, x_k) x_k. \tag{5.43}$$

Corollary 5.3.2 (Bessel's inequality) *Since (why?) $\|x\| = \|\hat{x}\| + \|x - \hat{x}\|$, we have $\|\hat{x}\| \leq \|x\|$ and consequently*

$$\sum_{k=1}^{\infty} |(x, x_k)|^2 \leq \|x\|^2.$$

Corollary 5.3.3 (Parseval's identity) *If $\mathbb{B}$ spans $\mathcal{H}$ then $x = \hat{x}$ and hence*

$$\sum_{k=1}^{\infty} |(x, x_k)|^2 = \|x\|^2.$$

Remark 5.3.5 *This is a generalization of Pythagoras' theorem.*

Chapter 6

The Lebesgue integral

The purpose of this chapter is to make sense of expressions like

$$\int_A f d\mu = \int_A f(x)\, d\mu(x) = \int_A f(x)\, \mu(dx). \tag{6.1}$$

6.1 Motivation

From our point of view, there are at least two reasons for introducing the Lebesgue integral. The first is that we want certain limit theorems of the following form

$$\lim_{k\to\infty} \int_A f_k(x)\, dx = \int_A \lim_{k\to\infty} f_k(x)\, dx \tag{6.2}$$

to hold under weaker conditions than we need for the Riemann integral (weaker, at any rate, than uniform convergence of $\langle f_k \rangle$). As we have seen, the Riemann integral does not have this property even when very strong restrictions (short of uniform convergence) are put on $\langle f_k \rangle$. This means that if we try to construct an abstract space (say a Banach space) of Riemann integrable functions, we will find that it is incomplete (and hence not a Banach space), and this renders the Riemann integral almost useless in functional analysis.

The second reason is that we often want to integrate over a more or less arbitrary space X, not just $\mathbb{R}$ or $\mathbb{R}^n$. This is particularly important in probability theory, since we want to define the expected value of a random variable as its integral over the sample space.

Note that, in this chapter, we will be integrating real- (scalar-)valued functions nearly all of the time, but that it is very easy to extend our theory to integrating functions with values in $\mathbb{R}^n$, $\mathbb{C}$ or $\mathbb{C}^n$. The idea is just to integrate component by component. We could have been even more abstract and tried to integrate functions with values in more or less arbitrary spaces, but that would take us much further than we usually need to go as economists.

Many of the definitions and results in this chapter nevertheless *are* somewhat abstract, and to avoid getting dizzy, it is worth always thinking about the case when X is a finite set. Then all the results are trivial, and it is easy to believe the more general results even without the proof. When you find a concept difficult to grasp, always take a simple example with a finite X, or if you have a talent for spatial visualization, think of X as $\mathbb{R}^2$ or $\mathbb{R}^3$.

6.2 Definition

Recall that the basic idea behind the Riemann integral is to partition the x-axis and see what happens as this partition becomes arbitrarily fine. Lebesgue chose the opposite strategy and decided to partition the y-axis instead. Roughly, the program ahead of us looks like this. Let $f : X \to \mathbb{R}$ be a real-valued function on an arbitrary space X. Suppose the range of f is compact, and divide this range into subintervals. For each sub-interval $[y_{k-1}, y_k]$, define an upper sum as follows.

$$S = \sum_{k=1}^{n} y_k \mu \left(f^{-1} \left([y_{k-1}, y_k] \right) \right) \tag{6.3}$$

where $\mu\left(f^{-1}\left([y_{k-1}, y_k]\right)\right)$ is the (generalized) length of the set of points $x \in X$ with their image in $[y_{k-1}, y_k]$.

Now if $X = \mathbb{R}$ and every set $f^{-1}\left([y_{k-1}, y_k]\right)$ is an interval, we have no problems. We know how to calculate the length of an interval. But even if $X = \mathbb{R}$ and f is continuous, there is no guarantee whatsoever that the inverse image of $[y_{k-1}, y_k]$ is an interval. So we need to generalize our notion of length to apply to a wider class of sets than just intervals. This is the subject of measure theory.

6.2.1 Measure theory

Suppose we have a set X and want to be able to define a measure μ on the subsets of X, i.e. a function which assigns a non-negative real number to some (or perhaps all) of the subsets of X. It turns out that we don't always want to include all subsets of X in the domain $\mathcal{F}$ of μ. One reason is that some subsets may be so bizarre that it is impossible to assign a measure to them in a reasonable way (this happens when $X = \mathbb{R}$; see below). Another is that a restriction of the domain of our measure is a very elegant way to model information in probability theory (see chapter 7).

In any case, we want the domain $\mathcal{F}$ of our measure μ to be well-behaved in certain ways. In particular, we want it to be 'closed' under certain set operations, so that we don't suddenly leave $\mathcal{F}$ when we perform the usual set operations on members of $\mathcal{F}$. The most convenient approach is to require $\mathcal{F}$ to be a σ-algebra (σ-field). Below we will denote the complement of a set A by A^c and the family of all subsets of X (the power set of X) by 2^X.

Definition 6.2.1 *Let X be an arbitrary set. Let $\mathcal{F} \subset 2^X$ be a family of subsets of X. Then $\mathcal{F}$ is called a* σ-algebra *if*

1. $X \in \mathcal{F}$,

2. *for each* $A \in \mathcal{F}$, $A^c \in \mathcal{F}$, *and*

3. *for every countable family of sets* $\{A_k\}_{k=1}^{\infty}$ *with* $A_k \in \mathcal{F}$ *for* $k = 1, 2, ...,$ *we have*

$$\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}. \tag{6.4}$$

By definition, then, we stay within $\mathcal{F}$ when we perform complementation and countable unions. By de Morgan's laws, it follows that we also stay within $\mathcal{F}$ when we perform countable intersections (why?). Note that it is essential that these operations are countable. We do not require $\mathcal{F}$ to be closed under arbitrary unions and intersections.

In case you should have forgotten de Morgan's laws, I state them here.

Proposition 6.2.1 (de Morgan's laws) *Let* $\mathbb{A}$ *be an arbitrary family of sets. Then*

$$\left[\bigcup_{A\in\mathbb{A}} A\right]^c = \bigcap_{A\in\mathbb{A}} A^c \tag{6.5}$$

and

$$\left[\bigcap_{A\in\mathbb{A}} A\right]^c = \bigcup_{A\in\mathbb{A}} A^c \tag{6.6}$$

Proof. Exercise. ■

We now note some useful and simple-to-prove properties of σ-algebras. (We will use these properties later.)

Proposition 6.2.2 *Let* $\{\mathcal{F}_\alpha : \alpha \in I\}$ *be an arbitrary (not necessarily countable) family of* σ*-algebras. Then the intersection* $\mathcal{F} = \bigcap_{\alpha\in I} \mathcal{F}_\alpha$ *is a* σ*-algebra.*

Proof. To show that $\mathcal{F}$ is closed under complementation, suppose $A \in \mathcal{F}$. Then $A \in \mathcal{F}_\alpha$ for all $\alpha \in I$. Since all the $\mathcal{F}_\alpha$ are $\sigma-$algebras, we have $A^c \in \mathcal{F}_\alpha$ for all $\alpha \in I$ as well. Hence $A^c \in \mathcal{F}$. The other parts of the proof have the same form. ■

Remark 6.2.1 *The σ in the term $\sigma-$algebra indicates that it is closed under countable unions and not just finite unions. Recall that an infinite series (sum) has countably many terms and begins with the letter s. There has to be a connection here.*

Proposition 6.2.3 *Let $\mathbb{A}$ be an arbitrary (not necessarily countable) family of subsets of some basic set X. Then there is a unique smallest extension of this family to a σ-algebra, i.e. a unique σ-algebra $\mathcal{F}$ such that*

1. *$\mathbb{A} \subset \mathcal{F}$, and*

2. *For any $\sigma-$algebra $\mathcal{G}$ such that $\mathbb{A} \subset \mathcal{G}$, we have $\mathcal{F} \subset \mathcal{G}$.*

We write $\mathcal{F} = \sigma(\mathbb{A})$ and call $\mathcal{F}$ the $\sigma-$algebra generated by $\mathbb{A}$.

Proof. Let $\{\mathcal{F}_\alpha\}$ be the family of σ-algebras such that $\mathbb{A} \subset \mathcal{F}_\alpha$ for each $\mathcal{F}_\alpha$. This family is non-empty, since 2^X is a σ-algebra with this property. Now define

$$\mathcal{F} = \bigcap_\alpha \mathcal{F}_\alpha \tag{6.7}$$

and it is clear (why?) that this defines a σ-algebra with the desired properties. ■

We sometimes want to take the union of a family of σ-algebras, but since that union is not necessarily itself a σ-algebra, we usually want to extend it to a σ-algebra in a minimal way. The preceding proposition shows that this can be done in a unique way, and we have the following definition.

Definition 6.2.2 *Let $\{\mathcal{F}_\alpha : \alpha \in I\}$ be a family of σ-algebras. Then the smallest σ-algebra $\mathcal{F}$ such that $A \in \mathcal{F}$ whenever there is an $\alpha \in I$ such that $A \in \mathcal{F}_\alpha$ is denoted by $\mathcal{F} = \bigvee_{\alpha \in I} \mathcal{F}_\alpha$.*

Definition 6.2.3 *Let X be an arbitrary non-empty set and let $\mathcal{F}$ be a σ-algebra. Then $(X, \mathcal{F})$ is called a* measurable space. *The sets $A \in \mathcal{F}$ are called the $\mathcal{F}$-*measurable sets. *When the σ-algebra $\mathcal{F}$ is taken as understood, we will sometimes suppress it and talk simply about the* measurable sets.

We now want to define a measure on a measurable space $(X, \mathcal{F})$, and we want to impose some reasonable requirements on a function $\mu : \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ for it to qualify as a measure. (The number $+\infty$ is defined to be a positive number greater than any real number; we want to allow μ to take this value since X might be unbounded in some sense.) When studying the definition, keep in mind that we are trying to generalize the notion of length (or area, or volume, or mass, or something like that).

Definition 6.2.4 *Let $(X, \mathcal{F})$ be a measurable space and let $\mu : \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ be a function. Then μ is called a (positive) measure if*

1. $\mu(\varnothing) = 0$,
2. *for each $A \in \mathcal{F}$, $\mu(A) \geq 0$, and*
3. *(countable additivity) for each family $\{A_k\}_{k=1}^{\infty}$ with $A_k \in \mathcal{F}$ and $A_j \cap A_k = \varnothing$ for $j \neq k$ (for every countable family of disjoint measurable sets) we have*

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k). \tag{6.8}$$

An important subfamily of $\mathcal{F}$ are the sets of measure zero. If a certain property holds at all points of X except on a set of points $A \in \mathcal{F}$ such that $\mu(A) = 0$, then we say that it holds *almost everywhere* (μ), and we abbreviate this by a.e. (μ). Sometimes when the measure μ is taken as understood, we suppress it and write just a.e.

Definition 6.2.5 *Let $(X, \mathcal{F})$ be a measurable space and let $\mu : \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ be a measure. Then $(X, \mathcal{F}, \mu)$ is called a (positive)* measure space*, and if $\mu(X) < \infty$ it is called a* finite *measure space.*

6.2.2 The Lebesgue integral

We now go on to define the Lebesgue integral, and the idea will be to approximate the integrand by functions which are simple in a certain sense, integrate the simple functions in an obvious way, and then try to make the approximation arbitrarily good. We also characterize the functions which can be arbitrarily well approximated by simple functions. Those functions turn out to be the *measurable* ones, defined as follows.

Definition 6.2.6 *Let $(X, \mathcal{F})$ be a measurable space, and let $f : X \to \mathbb{R}$ be a real-valued function on X. Then f is said to be* measurable with respect to $\mathcal{F}$ *(or $\mathcal{F}-$measurable) if, for every (open or closed) interval $I \subset \mathbb{R}$,*

$$f^{-1}(I) \in \mathcal{F}. \tag{6.9}$$

Actually, once the inverse images of all intervals are measurable, there is a much larger class of sets whose inverse image is measurable. In fact, we have the following proposition.

Proposition 6.2.4 *Let $(X, \mathcal{F})$ be a measurable space and let $f : X \to Y$ be a function. Then the family of subsets $A \subset Y$ such that $f^{-1}(A) \in \mathcal{F}$ is a σ-algebra.*

Proof. The proof uses that $\mathcal{F}$ is a σ-algebra, and that the inverse image preserves the relevant set operations, noting that

1. $f^{-1}(Y) = X \in \mathcal{F}$.

2. Let A be a set such that $f^{-1}(A) \in \mathcal{F}$. Then $f^{-1}(A^c) = [f^{-1}(A)]^c \in \mathcal{F}$.

3. Let $\{A_k\}_{k=1}^{\infty}$ be a family of sets such that $f^{-1}(A_k) \in \mathcal{F}$. Then $f^{-1}\left(\bigcup_{k=1}^{\infty} A_k\right) = \bigcup_{k=1}^{\infty} f^{-1}(A_k) \in \mathcal{F}$. ■

So when f is a measurable function, the family of all sets $A \subset \mathbb{R}$ such that $f^{-1}(A)$ is measurable is a σ-algebra. Clearly it contains all the intervals. At the very least, then, it contains all the sets in the so-called *Borel* σ-algebra on $\mathbb{R}$ which is the *smallest* σ-algebra on $\mathbb{R}$ that contains all the intervals. This example is so important that we restate it in a formal definition.[1]

Definition 6.2.7 *Let $\mathbb{I} \subset 2^{\mathbb{R}}$ be the set of all intervals. Then $\mathcal{B}(\mathbb{R}) = \sigma(\mathbb{I})$ is called the Borel σ-algebra on $\mathbb{R}$. Note that $\mathcal{B}(\mathbb{R}) = \sigma(\text{open subsets of } \mathbb{R}) = \sigma(\text{closed subsets of } \mathbb{R})$. A set $A \in \mathcal{B}(\mathbb{R})$ is called a Borel (measurable) set. A function $f : \mathbb{R} \to \mathbb{R}$ which is $\mathcal{B}(\mathbb{R})$-measurable is called a Borel (measurable) function.*

We can now characterize measurability in an interesting way, which is often used to *define* measurability.

Proposition 6.2.5 *Let $(X, \mathcal{F})$ be a measurable space, and let $f : X \to \mathbb{R}$ be a real-valued function on X. Then f is $\mathcal{F}-$measurable iff $f^{-1}(\mathcal{B}(\mathbb{R})) \subset \mathcal{F}$.*

Remark 6.2.2 *$f^{-1}(\mathbb{A})$ where $\mathbb{A}$ is a family of sets is of course the family of inverse images $f^{-1}(A)$ with $A \in \mathbb{A}$.*

Having defined measurability of a function, it is interesting to note that we can always construct a σ-algebra so that a given family of functions is measurable.

[1] In 1924, Émile Borel became active in the French government serving in the French Chamber of Deputies (1924-36) and as Minister of the Navy (1925-40). After his arrest and brief imprisonment under the Vichy regime he worked for the Resistance. He was awarded The Resistance Medal (1945), and the Grand Croix Légion d'Honneur (1950).

Just take the union of all inverse images of Borel sets under the functions in the family. Then extend to a σ-algebra in a minimal way as in Proposition (6.2.3). This leads us to the following definition.

Definition 6.2.8 *Let $(X, \mathcal{F})$ be a measurable space, and let $\{f_\alpha : \alpha \in I\}$ be a family of real-valued functions on X. Then the minimal $\sigma-$algebra $\mathcal{F}$ such that all the f_α are $\mathcal{F}-$measurable is called the σ-algebra* generated *by $\{f_\alpha : \alpha \in I\}$, and we write $\mathcal{F} = \sigma(\{f_\alpha : \alpha \in I\})$.*

We also have the following almost trivial but very important result.

Proposition 6.2.6 *Let $(X, \mathcal{F})$ and $(X, \mathcal{G})$ be two measurable spaces such that $\mathcal{G} \subset \mathcal{F}$, and let f be an $\mathcal{F}$-measurable function on X. Then f is also $\mathcal{G}$-measurable.*

Proof *Exercise.*

Happily, measurability is preserved under most common operations. In particular, we have the following theorem.

Theorem 6.2.1 *Let $(X, \mathcal{F})$ be a measurable space. Let f, g be measurable functions on X and let α be a scalar. Then*

1. *$f + g$, αf, and fg are measurable,*
2. *if $g \neq 0$, $\dfrac{f}{g}$ is measurable,*
3. *$h = \max[f, g]$ is measurable,*
4. *if $\langle f_n \rangle_{n=1}^{\infty}$ is a (countable!) sequence of measurable functions, then the following functions are also measurable:*
   (a) $\sup_{n \in \mathbb{N}} f_n$,

(b) $\inf_{n\in\mathbb{N}} f_n$,

(c) $\limsup_{n\to\infty} f_n$, *and*

(d) $\liminf_{n\to\infty} f_n$.

Proof. See [18]. ■

Just in case you should have forgotten, the definition of lim sup (limit superior) and lim inf (limit inferior) are as follows.

Definition 6.2.9 *Let* $\langle x_n\rangle$ *be a sequence of real numbers. Then*

$$\limsup_{n\to\infty} x_n = \inf_{n\in\mathbb{N}} \left\{ \sup_{k\geq n} x_k \right\} \tag{6.10}$$

and

$$\liminf_{n\to\infty} x_n = \sup_{n\in\mathbb{N}} \left\{ \inf_{k\geq n} x_k \right\} \tag{6.11}$$

In words, if $\limsup x_n = M$ *then* x_n *eventually stops ever going above* M*, and* M *is the smallest number with this property. Note that, by the least upper bound property of* $\mathbb{R}$*, every bounded sequence of real numbers has a* lim sup *and a* lim inf*, and when they are equal, the sequence is convergent with a limit equal to the common value.*

Also, we have the following very useful fact.

Theorem 6.2.2 *Let* $(X, \mathcal{F})$ *be a measurable space. Let* g *be an* $\mathcal{F}-$*measurable function on* X*, and let* $f : \mathbb{R} \to \mathbb{R}$ *be a Borel function. Then the composite mapping* $h = f \circ g$ *is* $\mathcal{F}-$*measurable.*

Proof *Exercise.*

Measurability is a rather abstract concept, but there is a way of making it concrete in an illustrative class of cases.

Definition 6.2.10 *Let X be a set and let $\mathbb{P} = \{P_k\}_{k=1}^n$ be a finite family of subsets of X. Then $\mathbb{P}$ is called a* finite partition *of X if*

1. $P_j \cap P_k = \varnothing$ *whenever* $j \neq k$ *and*
2. $\bigcup_{k=1}^{n} P_k = X$.

Theorem 6.2.3 *Let X be a set and let $\mathbb{P} = \{P_k\}_{k=1}^n$ be a finite partition of X. Let $\mathcal{F} = \sigma(\mathbb{P})$ be the σ-algebra generated by $\mathbb{P}$. Let f be a real-valued function on X. Then f is $\mathcal{F}$-measurable iff it is constant on each P_k, i.e. if there are numbers f_k; $k = 1, 2, ..., n$ such that $f(x) = f_k$ for each $x \in A_k$; $k = 1, 2, ..., n$.*

Proof *Exercise.*

Although, 'usually', σ-algebras are not generated by partitions, it is very helpful to *think* of them as having been so generated, since this makes both the notion of a σ-algebra and that of measurability easier to grasp intuitively.

Definition 6.2.11 (indicator function) *Let $(X, \mathcal{F})$ be a measurable space and let $A \in \mathcal{F}$. Then the function I_A is defined via*

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{6.12}$$

Remark 6.2.3 *Sometimes an indicator function I_A is called a* characteristic *function and is denoted by χ_A or K_A.*

Definition 6.2.12 *Let $(X, \mathcal{F})$ be a measurable space, and let $\varphi : X \to \mathbb{R}$ be a real-valued function on X. Then φ is called* $\mathcal{F}$-simple *if it is $\mathcal{F}$-measurable and its range is a finite set. It follows that an $\mathcal{F}$-simple function can be represented as*

$$\varphi(x) = \sum_{k=1}^{n} \varphi_k I_{A_k}(x) \tag{6.13}$$

where φ_k, $k = 1, 2, ..., n$ *are real numbers and the sets* A_k *are* $\mathcal{F}$*-measurable. Note that the sum on the right hand side has finitely many terms. Sometimes the* $\sigma-$*algebra* $\mathcal{F}$ *will be taken as understood, and we will talk (simply!) about* simple *functions.*

Definition 6.2.13 *Let* $(X, \mathcal{F}, \mu)$ *be a measure space and let* φ *be a simple function with the representation*

$$\varphi(x) = \sum_{k=1}^{n} \varphi_k I_{A_k}(x) \tag{6.14}$$

Then, rather naturally,

$$\int_X \varphi(x)\, d\mu(x) = \int_X \varphi d\mu = \sum_{k=1}^{n} \varphi_k \mu(A_k) \tag{6.15}$$

and it is not too hard to establish that this definition is independent of the precise representation of φ *(which unfortunately isn't unique). Recall that the* A_k *are measurable by definition, and that this is essential for the definition of the integral of a simple function to make sense.*

Having defined the integral of a simple function, we now generalize in three steps. First, in order to state a theorem relating integrability and measurability, we define the integral for a bounded function on a finite measure space. Then we define the integral of a measurable non-negative function. Finally, we define the integral of an arbitrary integrable function.

Definition 6.2.14 *Let* $(X, \mathcal{F}, \mu)$ *be a finite measure space and let* f *be a bounded real-valued function. Define*

$$\overline{L} = \inf_{\substack{\varphi \text{ simple} \\ \varphi \geq f}} \int_X \varphi d\mu \tag{6.16}$$

and

$$\underline{L} = \sup_{\substack{\varphi \text{ simple} \\ \varphi \leq f}} \int_X \varphi d\mu \tag{6.17}$$

If $\overline{L} = \underline{L} = L$ *then* f *is said to be Lebesgue integrable on* $(X, \mathcal{F}, \mu)$ *and*

$$\int_X f d\mu = L. \tag{6.18}$$

We now want to know what functions are Lebesgue integrable and the reader may already suspect what the answer is.

Theorem 6.2.4 *Let* $(X, \mathcal{F}, \mu)$ *be a finite measure space and let* f *be a bounded real-valued function. Then* f *is Lebesgue integrable iff it is* $\mathcal{F}$*-measurable.*

Proof *See [40].*

The idea of the proof is that $\mathcal{F}$-measurable functions, and only $\mathcal{F}$-measurable functions, can be arbitrarily well approximated by $\mathcal{F}$-simple functions. This idea will be reiterated in a slightly different context below.

Example 6.2.1 (a non-measurable function) *Let* $X = [0, 1]$*,let* $\mathcal{F} = \left\{[0, 1], \varnothing, \left[0, \frac{1}{2}\right], \left(\frac{1}{2}, 1\right]\right\}$*, and let* $f(x) = x$*. Then* $\overline{L} = \frac{3}{4}$ *and* $\underline{L} = \frac{1}{4}$*.*

We now define the integral of a non-negative measurable function.

Definition 6.2.15 *Let* $(X, \mathcal{F}, \mu)$ *be a measure space and let* f *be a non-negative measurable real-valued function. Then we define*

$$\int_X f d\mu = \sup_{\substack{\varphi \text{ simple} \\ \varphi \leq f}} \int_X \varphi d\mu. \tag{6.19}$$

Remark 6.2.4 *If the right hand side has no upper bound, we write* $\int_X f d\mu = +\infty$*.*

Remark 6.2.5 *The reason for confining our attention initially to non-negative functions is to avoid getting into trouble with functions such as $f(x) = x$ on $\mathbb{R}$ into $\mathbb{R}$ where the improper Riemann integral over $\mathbb{R}$ of f is finite but that of $|f|$ is not.*

We now state a very important theorem which we have already hinted at and which we'll find very useful later on. It motivates (again) why we confined our attention to *measurable* non-negative functions in the preceding theorem.

Theorem 6.2.5 *Let $(X, \mathcal{F})$ be a measurable space and let $f : X \to \mathbb{R}$ be non-negative and $\mathcal{F}$-measurable. Then, and only then, do there exist $\mathcal{F}$-simple functions φ_n; $n = 1, 2, ...$ such that*

1. $0 \leq \varphi_1 \leq \varphi_2 \leq \cdots \leq f$
2. $\varphi_n(x) \to f(x)$ *for each* $x \in X$.

Proof 1. *($\Rightarrow$) To each positive integer n and each real number y corresponds a unique integer $k = k_n(y)$ such that $k \cdot 2^{-n} \leq y < (k+1) \cdot 2^{-n}$. Define*

$$s_n(y) = \begin{cases} k_n(y) \cdot 2^{-n} & \text{if } 0 \leq y < n \\ n & \text{otherwise} \end{cases} \tag{6.20}$$

Note that the s_n are Borel functions. Now define for each $n = 1, 2, ...$ the functions

$$\varphi_n = s_n \circ f \tag{6.21}$$

which are $\mathcal{F}$-simple by Theorem (6.2.2).

2. *($\Leftarrow$) The converse is left as an exercise.*

Remark 6.2.6 *Note that it is possible for the convergence to be monotone. This will be essential for the usefulness of the Monotone Convergence Theorem (see below).*

It is now time to define the integral of signed functions. That is done as follows. The idea is to integrate the positive and negative parts of a function separately. The point of this, as we pointed out above, is to avoid the problematic phenomenon of infinities cancelling out as in the case of $f(x) = x$ on $\mathbb{R}$.

Definition 6.2.16 *Let $f : X \to \mathbb{R}$ be an arbitrary real-valued function on X. Then the positive part of f is defined via*

$$f^+(x) = \max[f(x), 0] \tag{6.22}$$

and the negative part of f is defined via

$$f^-(x) = \max[-f(x), 0] \tag{6.23}$$

Note that the negative part of f is a nonnegative function and that

$$f = f^+ - f^-. \tag{6.24}$$

Definition 6.2.17 *Let $(X, \mathcal{F}, \mu)$ be a measure space and let f be a measurable function such that*

$$\int_X |f| \, d\mu < \infty \tag{6.25}$$

(Note that $|f|$ is measurable since $|f| = f^+ + f^-$ and also nonnegative so the Lebesgue integral is defined by Definition 6.2.15.) Then f is said to be Lebesgue integrable on $(X, \mathcal{F}, \mu)$. We write $f \in \mathcal{L}^1(X, \mathcal{F}, \mu)$ and define

$$\int_X f d\mu = \int_X f^+ d\mu - \int_X f^- d\mu \tag{6.26}$$

Finally, we want to define the integral over subsets of X. That is done in the following way.

Definition 6.2.18 *Let $(X, \mathcal{F}, \mu)$ be a measure space and let f be an integrable function. Let $A \in \mathcal{F}$. Then*

$$\int_A f d\mu = \int_X (I_A \cdot f)\, d\mu \tag{6.27}$$

We now list some basic properties of the Lebesgue integral. As with the Riemann integral, they are intimately related to the fact that the integral is a limit of a sum.

Theorem 6.2.6 *Let $f, g \in \mathcal{L}^1(X, \mathcal{F}, \mu)$ and let α, β be scalars. Then*

1. $\int_X (\alpha f + \beta g)\, d\mu = \alpha \int_X f d\mu + \beta \int_X g d\mu$,
2. $f \leq g \;\Rightarrow\; \int_X f d\mu \leq \int_X g d\mu$, *and*
3. $\left| \int_X f d\mu \right| \leq \int_X |f|\, d\mu$.

Proof *See [17].*

6.3 The Monotone Convergence Theorem

6.3.1 Motivation

Many theorems are easy to prove if the functions involved are indicator functions, and by the linearity of the integral, they also hold for simple functions. The next step is to show that the theorem holds for arbitrary non-negative measurable functions, and this is done by invoking the Monotone Convergence Theorem (MCT). The final step, which proves the theorem for arbitrary integrable functions, proves it for f^+ and f^- separately by invoking the MCT and then adds up. So the only hard part is taken care of by the MCT; the rest is trivial. Having motivated it,

we now present the Monotone Convergence Theorem, which is the centerpiece of Lebesgue's achievement.

6.3.2 The theorem

Theorem 6.3.1 *Let $(X, \mathcal{F}, \mu)$ be a measurable space and let $\langle f_n \rangle$ be a sequence of non-negative measurable functions such that $0 \leq f_1 \leq f_2 \leq \cdots f_n \leq \cdots$ and $f_n(x) \to f(x)$ for all $x \in X$. Then*

$$\int_X f d\mu = \lim_{n\to\infty} \int_X f_n d\mu \tag{6.28}$$

Proof *See [17].*

This is by far the most important convergence theorem. Others include the dominated convergence theorem and Fatou's lemma. See, for example, [24].

6.4 Integrating on $\mathbb{R}$

Although we have stressed that X can be virtually any set, the case $X = \mathbb{R}$ is an important special case. Preferably we would like to show that the Lebesgue integral of f on $[a, b] \subset \mathbb{R}$ is the same as the (proper) Riemann integral on $[a, b]$ whenever f is Riemann integrable, and that turns out to be possible.

First, however, we need a measure m on $\mathbb{R}$ which is a generalization of the length of an interval. Ideally, we would like it to have the following properties.

1. m is defined for every $A \in 2^{\mathbb{R}}$,
2. m is translation invariant, i.e. if a is a real number, $A \subset \mathbb{R}$ and $B = \{x + a \in \mathbb{R} : x \in A\}$, then $m(A) = m(B)$,
3. m is countably additive, and

4. when I is an interval, $m(I)$ is its length.

(Translation invariance looks abstract, but what it means is that the measure of a set should not change if the set is merely shifted by a.)

Unfortunately, there is no such measure (if we accept the axiom of choice, see [18].) So we will drop requirement (1), and define a measure m on as many subsets of $\mathbb{R}$ as we can. This can be done in several ways, but the following seems to me to be the most intuitive one (if not perhaps technically the most elegant).

6.4.1 Lebesgue measure

To make things simple, let's try to measure the subsets of the bounded interval $[0,1]$. We begin with the open intervals. They are easy to measure; their measure is just their length. Then take the open sets. Fortunately, (see [18]) every open set is the countable union of disjoint open intervals. So we use countable additivity to measure all the open sets. For the closed sets $F \subset [0,1]$, we set $m(F) = 1 - m(F^c)$ where F^c is open by definition. For an arbitrary set $A \subset [0,1]$, put

$$\overline{m}(A) = \inf_{\substack{O \supset A \\ O \text{ open}}} m(O), \tag{6.29}$$

and

$$\underline{m}(A) = \sup_{\substack{A \supset F \\ F \text{ closed}}} m(F). \tag{6.30}$$

Whenever $\overline{m}(A) = \underline{m}(A) = m(A)$ we declare A to be Lebesgue measurable, and define its measure to be $m(A)$.

We now want to know whether the family of Lebesgue measurable sets is suitable as the domain of a measure (is it a σ-algebra?) and what relation, if any, it has to the family of Borel sets (are they the same?). We also want to know whether the function we have defined is a measure (is it countably additive?). We have the following results.

Theorem 6.4.1 *The family $\mathcal{L}([0,1])$ of Lebesgue measurable subsets of $[0,1]$ is a $\sigma-$algebra. Moreover, every Borel set is Lebesgue measurable, i.e. $\mathcal{B}([0,1]) \subset \mathcal{L}([0,1])$. Also, for any Lebesgue measurable set A that is not Borel measurable, we have $m(A) = 0$. Finally, the function m is a (positive) measure on both $\mathcal{B}([0,1])$ and $\mathcal{L}([0,1])$.*

Proof *See [30].*

It is possible to extend m from $\mathcal{B}([0,1])$ to $\mathcal{B}(\mathbb{R})$ in a unique reasonable way (see [30]) and for our purposes that will suffice as the definition of the Lebesgue measure m.

Example 6.4.1 *Consider the set of rational numbers $\mathbb{Q}$. Since $\mathbb{Q}$ is countable, it is the countable union of intervals of the form $[q_k, q_k]$ and hence of length zero. By the countable additivity of m, we have $m(\mathbb{Q}) = \sum_{k=1}^{\infty} m([q_k, q_k]) = \sum_{k=1}^{\infty} 0 = 0$. Consequently $\mathbb{Q}$ and hence $I_{\mathbb{Q}}$ is Borel measurable, and $\int_{\mathbb{R}} I_{\mathbb{Q}} dm = 0$.*

We now want to reassure ourselves that the object we've just constructed is the same as the Riemann integral whenever the latter exists. To our lasting relief, we have the following result.

Theorem 6.4.2 *Let a, b be real numbers and let f be Riemann integrable (and hence bounded) on $[a,b]$. Then $f \in \mathcal{L}^1([a,b], \mathcal{B}([a,b]), m)$ and*

$$\int_{[a,b]} f dm = \int_a^b f(x)\, dx \tag{6.31}$$

where the left hand side is a Lebesgue integral and the right hand side is a Riemann integral.

Proof. See [40]. ■

Idea of proof. The upper and lower Riemann sums approximate f by step functions. But every step function is also a simple function, so $\underline{R} \leq \underline{L} \leq \overline{L} \leq \overline{R}$. Hence when $\underline{R} = \overline{R}$ we also have $\underline{L} = \overline{L}$.

Remark 6.4.1 *It may happen that the* improper *Riemann integral exists but the Lebesgue integral does not. An example is $f(x) = x$ on $\mathbb{R}$. However, so long as $|f|$ is both Lebesgue and improperly Riemann integrable on $\mathbb{R}$, the two integrals are the same.*

Remark 6.4.2 *Henceforth, the notation $\int_a^b f(x)\,dx$ will be used to mean the Lebesgue integral of f over $[a,b]$ with respect to m.*

Equipped with the Lebesgue apparatus, it is possible to prove (although we won't) that the Riemann integrable functions are precisely those which are such that the set of points of discontinuity has Lebesgue measure zero. So one (sloppy) way to think about our results so far is that the Riemann integrable functions are the nearly continuous ones, and the Lebesgue integrable functions are the slightly less nearly continuous ones, where 'slightly less nearly continuous' means measurable. To see how measurability is really a generalization of continuity, recall that the inverse image of an open set under a continuous function is again an open set. Similarly, the inverse image of a Borel set under a Borel function is again a Borel set.

Mathematicians often stress that the step from Riemann to Lebesgue integration is a process of closure, i.e. of 'filling in the holes' in a certain sense. The following theorem substantiates that claim.

Theorem 6.4.3 *Let $R([a,b])$ be the space of Riemann integrable functions f on*

$[a,b]$ with the norm

$$\|f\| = \int_a^b |f(x)|\, dx \tag{6.32}$$

Then $\mathcal{L}^1([a,b], \mathcal{B}([a,b]), m)$ is the closure of $R([a,b])$.

Proof *See [41].*

We now know that the Riemann integrable functions are also Lebesgue integrable. This implies, for example, that the bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$ are Lebesgue integrable on $([a,b], \mathcal{B}([a,b]), m)$. More generally, Theorem 6.2.4 tells us that the bounded Borel functions are Lebesgue integrable on this space.

Exercise 6.4.1 *Show that every continuous function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable. Hint: Use that the inverse image of every open set under a continuous mapping is open.*

6.5 The $\mathcal{L}^p$ spaces

Here the idea is to define interesting metric, Banach and Hilbert spaces of functions, taking advantage of the fact that (given suitable norms) spaces of Lebesgue integrable functions are complete.

Definition 6.5.1 *Let $(X, \mathcal{F}, \mu)$ be a measure space and let p be a positive integer or the number $+\infty$. For $p < \infty$, $\mathcal{L}^p(X, \mathcal{F}, \mu)$ is the space of measurable functions f such that*

$$\int_X |f|^p \, d\mu < \infty \tag{6.33}$$

with the norm

$$\|f\|_p = \left[\int_X |f|^p \, d\mu \right]^{1/p} \tag{6.34}$$

For $p = \infty$, we require for $f \in \mathcal{L}^{\infty}(X, \mathcal{F}, \mu)$ that there is a number M such that $|f| \leq M$ almost everywhere (μ), and define $\|f\|_{\infty}$ to be the smallest such number. Formally, we write

$$\|f\|_{\infty} = \inf_{M \in \mathbb{R}} \{|f| \leq M \quad a.e. \ (\mu)\} \tag{6.35}$$

and call $\|f\|_{\infty}$ the essential supremum of f.

Remark 6.5.1 *For this to make sense, we need to confirm that the functions $\|\cdot\|_p$ really are norms, and that is actually hard work. But it can be done. See [18].*

We now have the following theorems, which establish the usefulness of the Lebesgue integral for functional analysis.

Theorem 6.5.1 *Let $(X, \mathcal{F}, \mu)$ be a measure space and let p be a positive integer or the number $+\infty$. Suppose also that we regard two functions f and g as the same if $f = g$ a.e. Now associate the space $\mathcal{L}^p(X, \mathcal{F}, \mu)$ with the natural addition and scalar multiplication operations, i.e. define $(f+g)(x) = f(x) + g(x)$ and $(\alpha f)(x) = \alpha f(x)$. Then $\mathcal{L}^p(X, \mathcal{F}, \mu)$ is a Banach (and hence also a complete metric) space.*

Proof. See [18]. ■

Remark 6.5.2 *The bit about identifying functions which are equal a.e. is to guarantee that the metric $\lambda(f, g) = \|f - g\|$ satisfies the axiom $\lambda(f, g) = 0$ iff $f = g$. Note that whenever applying a metric, Banach or Hilbert space theorem which guarantees a unique element which satisfies some set of conditions to the $\mathcal{L}^p$ spaces, it will mean that if f, g are two elements which both satisfy the conditions, then $f = g$ a.e.*[2]

[2] A formal way of dealing with this problem is to say that the elements of $\mathcal{L}^p$ are not functions but equivalence classes of functions under the equivalence relation a.e.-equality. See

Chapter 7

Probability

7.1 Introduction

This chapter provides the basic definitions in probability theory without containing much in the way of applications or concrete results. In particular, we won't talk about concrete probability distributions such as the normal or Poisson distribution; nor do I prove or even state any laws of large numbers or central limit theorems. For all of that, see the books in the bibliography (and many of the other courses on the graduate program).

Nevertheless, the material in this chapter does provide a very powerful toolbox which, once grasped, can be applied with a very small effort/payoff ratio. See, for example, sections 9.2.8 and 10.3.9. In any case, it gives the necessary conceptual foundations for any further study of probability theory.

7.2 Probability spaces and random variables

Probability theory begins with a measure space $(\Omega, \mathcal{F}, P)$ where Ω is a set of points called *outcomes* (denoted generically by ω), $\mathcal{F}$ is a $\sigma-$algebra of subsets

of Ω called *events*, and P is a (positive) measure which assigns probabilities to the events in $\mathcal{F}$. We will say that an event $A \in \mathcal{F}$ *occurs* if $\omega \in A$.

Definition 7.2.1 *A* probability space *is a measure space* $(\Omega, \mathcal{F}, P)$ *with* $P(\Omega) = 1$. *A measure* P *with this property is called a* probability measure. *If something is true a.e.* (P)*, we often say that it holds a.s. (almost surely).*

Scientists are fond of thinking of a probability space as describing an experimental situation, but we can think of it more broadly as capturing a situation of uncertainty.

Definition 7.2.2 *A* stochastic variable *(random variable) on a probability space* $(\Omega, \mathcal{F}, P)$ *is an* $\mathcal{F}$*-measurable mapping* $X : \Omega \to \mathbb{R}$.

We may interpret this (informally) as the (unknown) outcome ω giving rise to the (observed) value $X(\omega)$.

Definition 7.2.3 *The (unconditional) expectation of a random variable* X *is defined via*

$$E[X] = \int_{\Omega} X(\omega)\, dP(\omega). \tag{7.1}$$

Definition 7.2.4 *A random variable is said to be* integrable *if* $E[|X|] < \infty$ *and* square integrable *if* $E\left[|X|^2\right] < \infty$. *We write* $X \in L^1(\Omega, \mathcal{F}, P)$ *and* $X \in L^2(\Omega, \mathcal{F}, P)$, *respectively.*

Definition 7.2.5 *Let* X *be square integrable. Then its variance is defined via*

$$V[X] = E\left[(X - E[X])^2\right]. \tag{7.2}$$

Remark 7.2.1 *If* X *is complex-valued, the formula is*

$$V[X] = E\left[|X - E[X]|^2\right]. \tag{7.3}$$

Definition 7.2.6 *Let X and Y be square integrable. Then their covariance is defined via*

$$Cov[X,Y] = E\left[(XY - E[XY])^2\right]. \tag{7.4}$$

Definition 7.2.7 *Let $\mathbf{X}$ be a vector of stochastic variables. Then its covariance matrix is defined via*

$$V[\mathbf{X}] = E\left[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T\right]. \tag{7.5}$$

Remark 7.2.2 *If the elements of $\mathbf{X}$ are complex valued, the ordinary transpose T is replaced by the Hermitian transpose H.*

Definition 7.2.8 *The* distribution measure *of the random variable X is the measure μ_X defined on $\mathcal{B}(\mathbb{R})$ via $\mu_X(B) = P(X^{-1}(B))$.*

Remark 7.2.3 *Instead of $X^{-1}(B)$ we often write $\{X \in B\}$. In both cases we mean the set $\{\omega \in \Omega : X(\omega) \in B\}$.*

Proposition 7.2.1 *The function $F_X : \mathbb{R} \to \mathbb{R}_+$ defined via*

$$F_X(x) = \mu_X((-\infty, x]) \tag{7.6}$$

is a distribution function and is called the distribution function of the random variable X.

Proof. See [31]. ■

Remark 7.2.4 *While there is a one-one relation between distribution measures and distribution functions, there is a many-one relation between random variables and distribution functions.*

Happily, what you thought you knew about the unconditional expectation (and how to calculate it in concrete instances) remains correct. More precisely, we have

$$E[X] = \int_{\mathbb{R}} x d\mu_X = \int_{\mathbb{R}} x dF_X(x) \tag{7.7}$$

where the second equality is simply the definition of the Lebesgue-Stieltjes integral.

If $\mu_X \ll m$ where m is the Lebesgue measure, then the Radon-Nikodym theorem guarantees the existence of an a.e. (m) unique density function $f_X : \mathbb{R} \to \mathbb{R}_+$ such that

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx. \tag{7.8}$$

A necessary and sufficient condition for $\mu_X \ll m$ is that F_X is absolutely continuous (for a definition of absolute continuity for functions, see [18]). Then $f_X = F_X'$ (which is defined almost everywhere since F_X is monotone). A sufficient condition is that F_X is differentiable everywhere.

A nice example, taken from [24] which shows that this abstract apparatus is really capable of delivering concrete results, is the following.

Proposition 7.2.2 *Let X be a non-negative stochastic variable. Define $\{X \geq t\} = \{\omega \in \Omega : X(\omega) \geq t\}$. Then*

$$E[X] = \int_0^{\infty} P(\{X \geq t\})\, dt. \tag{7.9}$$

Proof. By using Fubini's theorem, we get

$$E[X] = \int_{\Omega} X(\omega)\, dP(\omega) = \int_{\Omega} \left[\int_{0}^{X(\omega)} 1 \cdot dt \right] dP(\omega) =$$

$$= \int_{\Omega} \left[\int_{0}^{\infty} I_{\{X(w) \geq t\}} dt \right] dP(\omega) = \int_{0}^{\infty} \left[\int_{\Omega} I_{\{X(w) \geq t\}} dP(\omega) \right] dt = \tag{7.10}$$

$$= \{\text{Fubini!}\} = \int_{0}^{\infty} P(\{X \geq t\})\, dt.$$

■

Another example is a general version of Chebyshev's inequality, which has an extremely simple proof in this abstract setting.

Proposition 7.2.3 *Let X be a non-negative stochastic variable and let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a non-decreasing with function (with $\varphi(x) > 0$ whenever $x > 0$) such that $\varphi(X)$ is integrable. Then, for each $\varepsilon > 0$,*

$$P(\{X(\omega) \geq \varepsilon\}) \leq \frac{1}{\varphi(\varepsilon)} E[\varphi(X)]. \tag{7.11}$$

Proof.

$$E[\varphi(X)] = \int_{\Omega} \varphi(X(\omega))\, dP \geq \int_{\{X(\omega) \geq \varepsilon\}} \varphi(X(\omega))\, dP \geq \tag{7.12}$$

$$\geq \int_{\{X(\omega) \geq \varepsilon\}} \varphi(\varepsilon)\, dP = \varphi(\varepsilon)\, P(\{X(\omega) \geq \varepsilon\}).$$

■

Corollary 7.2.1 *Let X be a square integrable random variable with expected value μ. Then*

$$P(\{|X(\omega) - \mu| \geq \varepsilon\}) \leq \frac{1}{\varepsilon^2} E\left[|X - \mu|^2\right] = \frac{1}{\varepsilon^2} V[X]. \tag{7.13}$$

7.3 Information and σ-algebras

When considering σ-algebras $\mathcal{G} \subset \mathcal{F}$ one may interpret $\mathcal{G}$ as the amount of available information. Intuitively, if our information is given by $\mathcal{G}$, we can distinguish between the events in $\mathcal{G}$ in the sense that for any event $G \in \mathcal{G}$ we know with perfect certainty whether or not it has occurred. Given this, it makes sense to say that if $\mathcal{G} \subset \mathcal{H}$, then $\mathcal{H}$ contains no less information than $\mathcal{G}$. Also, it is tempting to say that $\mathcal{G} = \sigma\{\text{singletons}\}$ corresponds to *full* information since it should enable us to tell exactly what ω has been drawn. But this turns out to be an awkward way of defining full information in general (see Exercise 7.5.6) although admittedly it makes perfect sense when Ω is a finite set. Instead, we will define full information as $\mathcal{G} = \mathcal{F}$, since then our information enables us to forecast perfectly the realized value of every random variable. Finally, we will say that $\mathcal{G} = \{\Omega, \varnothing\}$ (the trivial σ-algebra) corresponds to *no* information.

Alternatively, we might follow [24] in telling the following story. Suppose our σ-algebra $\mathcal{G}$ is generated by a finite partition $\mathbb{P}$.

(i) Someone (Tyche, the norns, Gustaf Lindencrona, or whoever it is) chooses an outcome $\omega \in \Omega$ without telling us which. (On a computer, you can 'play norn' by setting the seed for the random number generator.)

(ii) While we don't know which $\omega \in \Omega$ has been chosen, we *are,* however, told (by an oracle, Hugin & Munin, or Doktorandmeddelandet or whatever) in which component $P_k \in \mathbb{P}$ ω lies. In practice, this could be arranged by allowing us to observe a stochastic variable defined via

$$X(\omega) = \sum_{k=1}^{n} k I_{P_k}(\omega). \tag{7.14}$$

To flesh this out a little bit more, you may want to think that getting 'more information' in this context would correspond to having a 'finer' partition, where

a partition $\mathbb{Q}$ finer than $\mathbb{P}$ arises from chopping up the components of $\mathbb{P}$. It follows, of course, that $\sigma(\mathbb{P}) \subset \sigma(\mathbb{Q})$, which was our original (and more general) definition of 'more information'.

In any case, notice that the axioms that characterize a $\sigma-$algebra accord well with our intuitions about information. Obviously, we should know whether Ω, since it always occurs by definition. Also, if we know whether A, we should know whether not-A too. Similarly, if we know whether A and whether B, we should know whether $A \cup B$. Countable unions are perhaps a little trickier to motivate intuitively; they are there essentially for technical reasons. In particular, they allow us to prove various limit theorems which are part of the point of the Lebesgue theory.

In economic modelling, it is plausible to allow decisions to depend only upon the available information. Mathematically, this means that if the agent's information is given by $\mathcal{G}$, then her decision must be a $\mathcal{G}$-measurable random variable. The interpretation of this is that the information in $\mathcal{G}$ suffices to give us perfect knowledge of the decision. Thus when it is time for the agent to act, she knows precisely what to do.

At this stage it is worth thinking about what it means for a stochastic variable X to be $\mathcal{G}$-measurable. Intuitively, it means that the information in $\mathcal{G}$ suffices in order to know the value $X(\omega)$. To make this more concrete, suppose that $\mathcal{G}$ is generated by a partition $\mathbb{P}$. Then for X to be $\mathcal{G}$-measurable, X has to be constant on each element $P_k \in \mathbb{P}$. It follows that knowing which element P_k has occurred is enough to be able to tell what the value of $X(\omega)$ must be.

As a further illustration of the fact that σ-algebras do a good job in modelling information, we have the following result.

Definition 7.3.1 *Let $\{X_\alpha, \alpha \in I\}$ be a family of random variables. Then the*

σ-algebra generated by $\{X_\alpha, \alpha \in I\}$, denoted by $\sigma\{X_\alpha, \alpha \in I\}$ is the smallest σ-algebra $\mathcal{G}$ such that all the random variables in $\{X_\alpha, \alpha \in I\}$ are $\mathcal{G}$-measurable.

Remark 7.3.1 *Such a σ-algebra exists by Proposition 6.2.3. (Recall the proof: consider the intersection of all σ-algebras on Ω such that $\{X_\alpha, \alpha \in I\}$ are measurable.)*

Proposition 7.3.1 *Let $\mathbb{X} = \{X_1, X_2, ..., X_n\}$ be a finite set of random variables. Let Z be a random variable. Then Z is $\sigma\{\mathbb{X}\}$-measurable iff there exists a Borel measurable function $f : \mathbb{R}^n \to \mathbb{R}$ such that, for all $\omega \in \Omega$,*

$$Z(\omega) = f(X_1(\omega), X_2(\omega), ..., X_n(\omega)). \tag{7.15}$$

Proof. The case when $\sigma\{\mathbb{X}\}$ is generated by a finite partition (i.e. when the mapping $T : \Omega \to \mathbb{R}^n$ defined via $T(\omega) = \langle X_1, X_2, ..., X_n \rangle$ is $\mathcal{F}$-simple) is not too hard and is left as an exercise. For the general case, see [31]. ■

7.3.1 Filtrations

It is often useful to let the available information change as time goes by. This is done formally by introducing a *filtration*.

Definition 7.3.2 *A filtration in discrete time is a nondecreasing sequence of σ-algebras, i.e. a sequence of σ-algebras $\langle \mathcal{F}_t \rangle$ such that $s \leq t \Rightarrow \mathcal{F}_s \subset \mathcal{F}_t$. Note that this rules out forgetting by definition.*

Definition 7.3.3 *Let $\mathcal{F}$ be a σ-algebra and let $\langle \mathcal{F}_t \rangle$ be a filtration such that $\mathcal{F}_t \subset \mathcal{F}$ for each t. Then $\langle \mathcal{F}_t \rangle$ is called a filtration of $\mathcal{F}$.*

Definition 7.3.4 *The σ-algebra $\mathcal{F}_\infty = \bigvee_{t=0}^{\infty} \mathcal{F}_t$ is called the tail σ-algebra of $\langle \mathcal{F}_t \rangle$.*

7.4 The conditional expectation

We now want to define the conditional expectation. The idea will be to capture formally the intuitive idea that the conditional expectation of the random variable X is our best guess of the realized value $X(\omega)$ given the information we have available.

Definition 7.4.1 *Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, P)$. Then the conditional expectation $Y = E[X|\mathcal{G}]$ is the projection of X onto $L^2(\Omega, \mathcal{G}, P)$.*

Remark 7.4.1 *By the Hilbert space projection theorem, the conditional expectation solves*

$$Y = \min_{Z \in \mathcal{L}^2(\Omega, \mathcal{G}, P)} E\left[(X - Z)^2\right]. \tag{7.16}$$

Remark 7.4.2 *The conditional expectation is itself a random variable. Its value is uncertain because it depends on precisely which events $G \in \mathcal{G}$ actually occur. In other words, it is a (contingent) forecasting* rule *whose output (the forecast) depends on the content of the information revealed. For example, suppose our information set is such that we know whether the president has been shot. Then our actions may depend on whether he is or is not shot.*[1]

The projection-based definition is intuitively the most appealing one, but unfortunately it only applies to *square* integrable stochastic variables. One way to extend the definition to merely *integrable* stochastic variables is to note that $\mathcal{L}^2$ is dense in $\mathcal{L}^1$ and define $E[X|\mathcal{G}]$ as the limit of the sequence $\langle E[X_n|\mathcal{G}]\rangle$ where $X_n \in \mathcal{L}^2$ and $X_n \to X$ (in $\mathcal{L}^1$). Another way is the following.

[1] A real-life example of a forecasting rule is 'A red morning is a sailor's warning - a red night is a sailors delight'. Note that the output of this rule, the forecast, is random since it depends on whether the morning is red or not and whether the night is red or not. Nevertheless, this example is problematic since if both the morning and the evening are red, then we have both a warning and a delight, so that the conditional expectation is strictly speaking not a function on the set of colors of the sky at various times into the set {warning, delight}.

Proposition 7.4.1 *Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$. Then there is an a.s. (P) unique integrable random variable Z such that*

1. *Z is $\mathcal{G}$-measurable and*

2. $\int_G X dP = \int_G Z dP$ *for each $G \in \mathcal{G}$.*

   Using this result, we define $E[X|\mathcal{G}] = Z$.

Proof. The Radon-Nikodym theorem. ■

Remark 7.4.3 *Since the conditional expectation is only a.s. (P) unique, most of the equations below strictly speaking need a qualifying 'a.s. (P)' appended to them to be true. But since this is a bit tedious, we adopt instead the convention that the statement $X = Y$ means $P(\{\omega \in \Omega : X(\omega) = Y(\omega)\}) = 1$. If two random variables W and Z both qualify as the conditional expectation $E[X|\mathcal{G}]$, then we will sometimes call them* versions *of $E[X|\mathcal{G}]$.*

This $\mathcal{L}^1$-based definition can be intuitively motivated independently of the projection-based definition in the following way. On events such that we know whether they have occurred, our best guess of X should track X perfectly.

In any case, it had better be true that our two definitions of the conditional expectation coincide when they both apply, i.e. on $\mathcal{L}^1 \cap \mathcal{L}^2 = \mathcal{L}^2$. They do. You will be asked in an exercise to prove this.

Having defined the conditional expectation with respect to a σ-algebra, we now define the conditional expectation with respect to a family of stochastic variables.

Definition 7.4.2 *Let $Y \in L^1(\Omega, \mathcal{F}, P)$ and let $\{X_\alpha, \alpha \in I\}$ be a family of random variables. Then the conditional expectation $E[Y|\{X_\alpha, \alpha \in I\}]$ is defined as $E[Y|\sigma\{X_\alpha, \alpha \in I\}]$*

Since $E[Y|X]$ is a $\sigma(X)-$measurable random variable, there is a Borel function f such that $E[Y|X] = f(X)$. Sometimes we use the notation $f(x) = E[Y|X = x]$ where the expression on the right hand side is defined by the left hand side.

Definition 7.4.3 *Let $Y \in L^1(\Omega, \mathcal{F}, P)$ and let X be a stochastic variable. Then the function $E[Y|X = x]$ is defined as any Borel function $f : \mathbb{R} \to \mathbb{R}$ with the property that $f(X)$ is a version of $E[Y|X]$. Note that $E[Y|X = x]$ is not always uniquely defined, but that this does not matter in practice.*

Having defined the conditional expectation, we now note some of its properties. Let the given probability space be $(\Omega, \mathcal{F}, P)$.

Proposition 7.4.2 *Let $\mathcal{G} = \{\Omega, \varnothing\}$. Then $E[X|\mathcal{G}] = E[X]$.*

Proof. Exercise. ■

Proposition 7.4.3 *Let X and Y be integrable random variables, let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let α, β be scalars. Then*

$$E[\alpha X + \beta Y|\mathcal{G}] = \alpha E[X|\mathcal{G}] + \beta E[Y|\mathcal{G}] \tag{7.17}$$

Proof. Exercise. ■

Proposition 7.4.4 (Law of iterated expectations) *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$ be σ-algebras. Then*

$$E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}] \tag{7.18}$$

Proof *We check that the left hand side satisfies the conditions for being the conditional expectation of X with respect to $\mathcal{G}$. Clearly it is $\mathcal{G}$-measurable. Now let $G \in \mathcal{G}$ and we have, since $\mathcal{G} \subset \mathcal{H}$ and consequently $G \in \mathcal{H}$,*

$$\int_G E[E[X|\mathcal{H}]|\mathcal{G}]\, dP = \int_G E[X|\mathcal{H}]\, dP = \int_G X dP. \tag{7.19}$$

□

Corollary 7.4.1 *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then $E[E[X|\mathcal{G}]] = E[X]$.*

Proposition 7.4.5 *Let X and Y be random variables such that XY is integrable. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and suppose X is $\mathcal{G}$-measurable. Then*

1. *$E[X|\mathcal{G}] = X$ and*
2. *$E[XY|\mathcal{G}] = XE[Y|\mathcal{G}]$.*

Proof. (1) is trivial. To prove (2), note first that, by Theorem 6.2.1, the right hand side is $\mathcal{G}$-measurable. To show that the right hand side integrates to the right thing, suppose $X = I_G$ where $G \in \mathcal{G}$. Let $F \in \mathcal{G}$. Then

$$\int_F XE[Y|\mathcal{G}]\,dP = \int_F I_G E[Y|\mathcal{G}]\,dP = \int_{G\cap F} E[Y|\mathcal{G}]\,dP =$$

$$= \{(G \cap F) \in \mathcal{G}!\} = \int_{G\cap F} Y\,dP = \int_F I_G Y\,dP = \tag{7.20}$$

$$= \int_F XY\,dP$$

To show the more general case, show it for simple functions and then use the MCT. ■

We end the discussion of the conditional expectation by defining the conditional probability of an event. We then note with satisfaction that our formal definition substantiates our claim above that if our information is given by $\mathcal{G}$, then we know, for all events $G \in \mathcal{G}$ whether or not they have occurred.

Definition 7.4.4 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G} \subset \mathcal{F}$ be a $\sigma-$algebra. Let $A \in \mathcal{F}$. Then the conditional $P(A|\mathcal{G})$ probability of A given $\mathcal{G}$ is defined via*

$$P(A|\mathcal{G}) = E[I_A|\mathcal{G}] \tag{7.21}$$

It follows from this definition (why?) that if $G \in \mathcal{G}$ then $P(A|\mathcal{G}) = 1$ when G occurs and $P(A|\mathcal{G}) = 0$ when it does not.

7.5 Stochastic independence

Definition 7.5.1 *Two events $A, B \in \mathcal{F}$ are said to be* independent *if*

$$P(A \cap B) = P(A) P(B) \tag{7.22}$$

Remark 7.5.1 *Note that the concept of independence is strongly tied to the particular probability measure that we are considering. Strictly speaking, we should say P-independent etc. rather than just independent.*

The intuitive idea behind this definition of independence is that A is as likely to happen when we know that B happens as when we don't. This may be formalized as (draw a picture of this!)

$$\frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(\Omega)} \tag{7.23}$$

of which (7.22) is a generalization to cover the possibility that $P(B) = 0$.

Definition 7.5.2 *Two σ-algebras $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ are said to be independent if*

$$P(G \cap H) = P(G) P(H) \tag{7.24}$$

for all $G \in \mathcal{G}$ and $H \in \mathcal{H}$.

Remark 7.5.2 *Note that the property of independence is not transitive. Hence one must take care when defining independence for families of sets or $\sigma-$algebras with more than two members. See [30] for the conventional definition.*

Definition 7.5.3 *Two random variables X and Y are said to be independent if $\sigma\{X\}$ and $\sigma\{Y\}$ are independent.*

Proposition 7.5.1 *Let X and Y be independent square integrable random variables. Then*

$$E[XY] = E[X]\,E[Y]. \tag{7.25}$$

Proof. The first part of the proof is to show the result for indicator functions. So let $A \in \sigma\{X\}$ and $B = \sigma\{Y\}$. Then

$$\begin{aligned} E[I_A I_B] &= E[I_{A\cap B}] = \int_\Omega I_{A\cap B}\,dP = P(A\cap B) = \\ &= P(A)\,P(B) = E[I_A]\,E[I_B] \end{aligned} \tag{7.26}$$

The next step is, as usual, to use the linearity of the integral to show the result for simple functions. The proof is then finished by invoking the MCT to show the general result. ■

We end this chapter with a result that states that if a σ-algebra $\mathcal{G}$ is independent of the σ-algebra generated by a random variable X, then $\mathcal{G}$ is useless in forecasting X.

Proposition 7.5.2 *Let X be an integrable random variable and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra which is independent of $\sigma(X)$. Then*

$$E[X|\mathcal{G}] = E[X]. \tag{7.27}$$

Proof. Exercise. ■

Just as for the expected value, there is a Jensen's inequality.

Proposition 7.5.3 *Suppose $f : \mathbb{R} \to \mathbb{R}$ is convex, and suppose $X, f(X) \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then, for $P-$almost all ω,*

$$f(E[X|\mathcal{G}](\omega)) \leq E[f(X)|\mathcal{G}](\omega). \tag{7.28}$$

Proof. See [34]. ■

Exercise 7.5.1 *What class of random variables are measurable with respect to the trivial σ-algebra $\{\Omega, \varnothing\}$?*

Exercise 7.5.2 *Show that the Radon-Nikodym theorem guarantees the existence of the conditional expectation of an integrable random variable.*

Exercise 7.5.3 *Show that the two definitions of the conditional expectation agree on $\mathcal{L}^2$. Hints: There are (at least) two ways of doing this. The first way shows that the expectation error arising from the $\mathcal{L}^1$-based definition is orthogonal to what it's supposed to be orthogonal to. The second way shows that the $\mathcal{L}^1$-based conditional expectation really solves the projection problem. This is done by 'completing the square'.*

Exercise 7.5.4 *Show that, on a finite measure space, $\mathcal{L}^2 \subset \mathcal{L}^1$. Hint: On finite measure spaces, the constants are integrable.*

Exercise 7.5.5 *Show that, on a finite measure space, $\mathcal{L}^2$ is dense in $\mathcal{L}^1$, i.e. for each $X \in \mathcal{L}^1$ there is a sequence of stochastic variables $\langle X_n \rangle$ such that $X_n \in \mathcal{L}^2$ and $\|X_n - X\|_1 \to 0$. Hints: Note that bounded functions on finite measure spaces are square integrable. Then use the MCT.*

Exercise 7.5.6 *Consider the probability space* $([0,1], \mathcal{B}([0,1]), m)$. *Let* X *be an arbitrary integrable random variable. Show that*

$$E[X|\sigma\{singletons\}] = E[X]. \tag{7.29}$$

so that, far from giving full information, $\mathcal{G} = \sigma\{singletons\}$ *in a sense is as bad as no information at all.*

Exercise 7.5.7 *(This exercise confirms that our definition of the conditional expectation agrees with the one given in elementary treatments whenever the latter applies.) Let* $(\Omega, \mathcal{F}, P)$ *be a probability space and let* $\mathcal{G} \subset \mathcal{F}$ *be generated by the finite partition* $\mathbb{S} = \{S_1, S_2, ..., S_n\}$. *Suppose* $P(S_k) > 0$ *for all* $k = 1, 2, ..., n$. *Let* $B \in \mathcal{F}$. *Show that*

$$Z(\omega) = \sum_{k=1}^{n} I_{S_k} \frac{P(S_k \cap B)}{P(S_k)} \tag{7.30}$$

qualifies as the conditional probability $E[I_B|\mathcal{G}]$.

Exercise 7.5.8 *Prove all the propositions in this section where no reference to a proof is given.*

7.6 Stochastic processes in discrete time

A stochastic process in discrete time is a sequence of random variables. More formally, we have the following definition.

Definition 7.6.1 *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. Let* $\mathbb{Z}_+$ *be the set of non-negative integers. A* stochastic process in discrete time *is a mapping* $X : \mathbb{Z}_+ \times \Omega \to \mathbb{R}$ *such that for each* $t \in \mathbb{Z}_+$, *the mapping* $X(t, \cdot) \to \mathbb{R}$ *is a random variable (and hence* $\mathcal{F}-$*measurable). Instead of* $X(t, \omega)$ *we sometimes write* $X_t(\omega)$ *or just* X_t.

Definition 7.6.2 *Let X be a stochastic process. Holding $\omega \in \Omega$ fixed, the mapping $X(\cdot, \omega) : \mathbb{Z}_+ \to \mathbb{R}$ is called a* trajectory *or* sample path *of X.*

It is of course easy to extend this definition to the vector case; a vector stochastic process is just a vector of scalar stochastic processes.

7.6.1 Adapted processes

Definition 7.6.3 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\underline{\mathcal{F}} = \langle \mathcal{F}_t \rangle_{t=0}^{\infty}$ be a filtration of $\mathcal{F}$. Let X be a stochastic process on $(\Omega, \mathcal{F}, P)$. Then X is said to be adapted to $\underline{\mathcal{F}}$ (or $\underline{\mathcal{F}}$-adapted) if, for each t, X_t is $\mathcal{F}_t-$measurable.*

Since adapted processes are so useful, it is nice to know that we can always construct a filtration such that any given stochastic process X is adapted to it. The idea is to let the filtration contain all the information revealed by the stochastic process so far.

Proposition 7.6.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a stochastic process. Define a sequence of σ-algebras via*

$$\mathcal{F}_t = \sigma\left(\{X_s : s \leq t\}\right). \tag{7.31}$$

Then $\langle \mathcal{F}_t \rangle$ is a filtration of $\mathcal{F}$ and X is adapted to it.

Proof. Obvious except to show that $\mathcal{F}_\infty \subset \mathcal{F}$. [Look into this in more detail.]

The extension to the vector case is trivial: just require each element to be adapted.

7.6.2 Markov processes

The idea of a Markov process is to capture the idea of a short-memory stochastic process: once its current state is known, past history is irrelevant from the point

of view of predicting its future.

Definition 7.6.4 *Let $(\Omega, \mathcal{F})$ be a measurable space and let $(P, \underline{\mathcal{F}})$ be, respectively, a probability measure on and a filtration of this space. Let X be a stochastic process in discrete time on $(\Omega, \mathcal{F})$. Then X is called a* $(P, \underline{\mathcal{F}})$-Markov process *if*

1. *X is $\underline{\mathcal{F}}-$adapted, and*

2. *For each $t \in \mathbb{Z}_+$ and each Borel set $B \subset \mathcal{B}(\mathbb{R})$*

$$P(X_{t+1} \in B | \mathcal{F}_t) = P(X_{t+1} \in B | \sigma(X_t)). \tag{7.32}$$

Sometimes when the probability measure and filtration are understood, we will talk simply of a Markov process.

Remark 7.6.1 *Often (see, for example, [30]) the filtration $\underline{\mathcal{F}}$ is taken to be that generated by the process X itself.*

Proposition 7.6.2 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let X be a $(P, \underline{\mathcal{F}})$-Markov process. Let $t, k \in \mathbb{Z}_+$. Let $f : \mathbb{R} \to \mathbb{R}$ be a Borel function such that $f(X_{t+k})$ is integrable. Then,*

$$E[f(X_{t+k}) | \mathcal{F}_t] = E[f(X_{t+k}) | \sigma(X_t)] \tag{7.33}$$

and hence there is a Borel function $g : \mathbb{Z}_+ \times \mathbb{Z}_+ \times \mathbb{R} \to \mathbb{R}$ such that, for each t,

$$E[f(X_{t+k}) | \mathcal{F}_t] = g(t, k, X_t). \tag{7.34}$$

Proof. To show it for $k = 1$, use that $f(X_{t+1})$ is a $\sigma(X_{t+1})-$measurable random variable and hence is the limit of a monotone increasing sequence of $\sigma(X_{t+1})-$simple random variables. But such random variables are linear combinations of indicator functions of sets $X_{t+1}^{-1}(B)$ with B a Borel set. This completes

the proof for $k = 1$. To prove it for arbitrary positive k, use induction. To prove it for $k+1$ assuming it true for k, use the law of iterated expectations. ■

The vector case is a simple extension of the scalar case. However, it is important that the definition of a vector Markov process is *not* that each component is Markov. Instead, we require that all the relevant (for the future of X) bits of information in $\mathcal{F}_t$ are in the $\sigma-$algebra generated by *all* the stochastic variables in X_t, i.e. $\sigma(X_t)$ is defined as the *single* $\sigma-$algebra $\sigma(\{X_{1,t}, X_{2,t}, ..., X_{n,t}\})$. This means that, for each Borel function $f : \mathbb{R}^n \to \mathbb{R}^m$,

$$E[f(X_{t+k})|\mathcal{F}_t] = E[f(X_{t+k})|\sigma(X_t)] \tag{7.35}$$

and hence that there is a Borel function $g : \mathbb{Z}_+ \times \mathbb{R}^n \to \mathbb{R}^m$ such that, for each t,

$$E[f(X_{t+1})|\mathcal{F}_t] = g(t, X_t). \tag{7.36}$$

(The case $k = 1$ is so important that we stress it here by ignoring greater values of k.)

7.6.2.1 Transition probability function and time homogeneity

Definition 7.6.5 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let X be a $(P, \underline{\mathcal{F}})$-Markov process. Then, for each $t = 0, 1, 2...$ its transition probability function $Q_t : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ is defined via*

$$Q_t(x, B) = P(X_t \in B | X_{t-1} = x). \tag{7.37}$$

Note that any Markov process has a sequence of probability transition functions. Note also that for each fixed t and x, $Q_{t+1}(x, \cdot)$ is a probability measure on $\mathcal{B}(\mathbb{R})$. Meanwhile, if we fix B, $Q_{t+1}(X_t(\cdot), B)$ is a random variable. Indeed, it is the conditional probability of $X_{t+1} \in B$ given X_t, i.e. $Q_{t+1}(X_t, B) = E\left[I_{X_{t+1}^{-1}(B)} \middle| \sigma(X_t)\right]$. Moreover, the conditional expectation of any $\sigma(X_{t+1})$-measurable

random variable (given X_t) is an integral with respect to the measure Q_{t+1} in the following sense.

Proposition 7.6.3 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let X be a $(P, \underline{\mathcal{F}})$-Markov process. Let $\langle Q_t \rangle$ be its transition probability functions and let $Z \in \mathcal{L}^1(\Omega, \sigma(X_{t+1}), P)$. Then, for each $t = 0, 1, ...$*

$$E[Z|X_t] = \int_{\mathbb{R}} f(y) Q_{t+1}(X_t, dy) \tag{7.38}$$

or, put differently, we have for each $t = 0, 1, ...$ and each x,

$$E[Z|X_t = x] = \int_{\mathbb{R}} f(y) Q_{t+1}(x, dy). \tag{7.39}$$

Proof. We will show it first for an indicator variable $Z = I_{X_{t+1}^{-1}(A)}$ where $A \in \mathcal{B}(\mathbb{R})$. Then $f(y) = I_A(y)$. We now need to show that the random variable $\int_{\mathbb{R}} f(y) Q_{t+1}(X_t, dy)$ qualifies as the conditional expectation $E[Z|X_t]$. Clearly it is $\sigma(X_t)-$measurable. But does it integrate to the right thing? Well, let $G \in \sigma(X_t)$ and recall that, by definition, $Q_{t+1}(X_t, A) = E\left(I_{X_{t+1}^{-1}(A)}|\sigma(X_t)\right)$. Hence

$$\int_G \int_{\mathbb{R}} f(y) Q_{t+1}(X_t, dy) P(d\omega) = \int_G \int_{\mathbb{R}} I_A(y) Q_{t+1}(X_t, dy) P(d\omega) =$$

$$= \int_G Q_{t+1}(X_t, A) P(d\omega) = \int_G E\left(I_{X_{t+1}^{-1}(A)}|\sigma(X_t)\right) P(d\omega) = \int_G I_{X_{t+1}^{-1}(A)} P(d\omega). \tag{7.40}$$

Meanwhile, since $Z = I_{X_{t+1}^{-1}(A)}$ we obviously have

$$\int_G ZP(d\omega) = \int_G I_{X_{t+1}^{-1}(A)} P(d\omega). \tag{7.41}$$

To show the theorem for an arbitrary $Z \in \mathcal{L}^1(\Omega, \sigma(X_{t+1}), P)$, use the MCT. ■

We now use the transition probability function to define a *time homogeneous* Markov process.

Definition 7.6.6 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let X be a $(P, \underline{\mathcal{F}})$-Markov process. Let $\langle Q_t \rangle_{t=1}^{\infty}$ be its transition probability functions. If there is a Q such that $Q_t = Q$ for all $t = 1, 2, ...$ then X is called a* time homogeneous *Markov process.*

Proposition 7.6.4 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let X be a time homogeneous $(P, \underline{\mathcal{F}})$-Markov process. For any nonnegative integers k, t, let $Y_{t+k} \in \mathcal{L}^1(\Omega, \sigma(X_{t+k}), P)$. Then for each $k = 0, 1, ...$ there is a Borel function $g_k : \mathbb{R} \to \mathbb{R}$ such that, for each $t = 0, 1, ...$*

$$E[Y_{t+k}|\mathcal{F}_t] = g_k(X_t). \tag{7.42}$$

In particular, there is a Borel function h such that, for each $t = 0, 1, ...$

$$E[Y_{t+1}|\mathcal{F}_t] = h(X_t). \tag{7.43}$$

7.6.2.2 Markov chains

7.6.3 Processes bounded in $\mathcal{L}^2$

Sometimes we want to consider abstract spaces of stochastic processes, and then it is useful to have a norm for them. Let's begin with a piece of notation. Let $\mathbf{x} : \mathbb{Z}_+ \times \Omega \to \mathbb{R}^n$ be a vector-valued stochastic process. We write $\mathbf{x}(t, \omega) = x_t$.

One attractive norm for a vector-valued stochastic process $\mathbf{x}$ is

$$\|\mathbf{x}\| = \limsup_{t \to \infty} E[\|x_t\|] \tag{7.44}$$

and the corresponding metric is of course

$$\mu(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|. \tag{7.45}$$

For this norm and metric to be well-defined, two conditions suffice. First, all the components of x_t are square integrable random variables for each t. (If this

condition alone is satisfied, we call $\mathbf{x}$ a *square integrable process.*) Second, there is an M such that $E[\|x_t\|] \leq M$ for all $t = 0, 1, ...$ If both these conditions are satisfied, $\limsup_{t\to\infty} E[\|x_t\|] < \infty$ and we call $\mathbf{x}$ a process *bounded in* $\mathcal{L}^2$.

Theorem 7.6.1 *The space of processes bounded in* $\mathcal{L}^2$ *is a Banach space.*

7.6.4 Martingales

Definition 7.6.7 *Let* $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ *be a filtered probability space and let* X *be a stochastic process. Then* X *is called a* $(P, \underline{\mathcal{F}})$-martingale *if*

1. X *is adapted to* $\underline{\mathcal{F}}$,
2. $E[|X_t|] < \infty$ *for each* $t = 0, 1, ...$
3. $E[X_{t+1}|\mathcal{F}_t] = X_t$ *for each* $t = 0, 1, ...$

When the probability measure P *and filtration* $\underline{\mathcal{F}}$ *are understood, we will talk simply of a* martingale.

Remark 7.6.2 *You should think carefully about how the notion of a martingale is dependent on the particular choice of probability measure and filtration. In particular, note that the expectations operator* E *represents integration with respect to the probability measure* P.

7.6.4.1 Martingale differences

Definition 7.6.8 *Let* $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ *be a filtered probability space and let* X *be a martingale. Define the stochastic process* $\boldsymbol{\varepsilon}$ *via*

$$\varepsilon_t = X_t - X_{t-1}. \tag{7.46}$$

Then $\boldsymbol{\varepsilon}$ *is called a* $(P, \underline{\mathcal{F}})$-martingale difference *or* white noise process.

Proposition 7.6.5 *Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space and let $\boldsymbol{\varepsilon}$ be a $(P, \underline{\mathcal{F}})$-martingale difference. Then, for all $t = 0, 1...$*

$$E\left[\varepsilon_{t+1}|\mathcal{F}_t\right] = 0. \tag{7.47}$$

Corollary 2 *For all $Z \in \mathcal{L}^1\left(\Omega, \mathcal{F}_t, P\right)$, we have*

$$E\left[Z\varepsilon_{t+1}\right] = 0. \tag{7.48}$$

7.6.5 Stochastic integration in discrete time

Definition 7.6.9 *A process is said to be predictable if*

Predictable processes. Semimartingales. Super- and submartingales.

7.6.5.1 The martingale convergence theorem

Chapter 8

Some linear algebra

8.1 Introduction

The purpose of this chapter is to present those parts of linear algebra that are essential for the analysis of dynamic systems. It will take many basic concepts in linear algebra for granted. If you paid attention during Mathematics 1, you have little to worry about, but to refresh your memory you may occasionally want to refer to an introductory text on linear algebra. An excellent choice is [20]. (Actually, if you remember every word of [20], including the Schur form of a matrix, you can skip to section 8.5 right away.)

Since this chapter is full of eigenvalues which may easily be complex, all matrices here will be matrices of complex numbers (unless I explicitly state otherwise). Moreover, all matrices in this chapter will be $n \times n$ and consequently square. Sometimes we will write $A \in \mathbb{C}^{n\times n}$ and the meaning of that is obvious.

8.2 Four important theorems

Proposition 8.2.1 *Let $A \in \mathbb{C}^{n\times n}$ be a square matrix. Then its determinant is*

the product of the eigenvalues and its trace (the sum of the diagonal elements) is the sum of the eigenvalues.

Remark 8.2.1 *It is not claimed that the diagonal elements actually are the eigenvalues.*

Proof. By definition, the eigenvalues are the zeros of a polynomial. By the fundamental theorem of algebra, this polynomial can be factorized. By the definition of the set of eigenvalues $\{\lambda_k : k = 1, 2, ..., n\}$ as the zeros of this polynomial, we have, for all complex λ,

$$\det(A - \lambda I) = \prod_{k=1}^{n} (\lambda - \lambda_k).$$

In particular, it is true for $\lambda = 0$ and the first part of the proposition follows. For the second part, see [20]. ■

Corollary 8.2.1 *A square matrix A is invertible iff all of its eigenvalues are non-zero.*

Proposition 8.2.2 *Let A be a square matrix with real entries. Let λ be an eigenvalue of A. Then so is its complex conjugate $\overline{\lambda}$. Moreover, the corresponding eigenvectors are also each others' complex conjugates. Hence the complex eigenvalues and eigenvectors (those with a non-zero imaginary part) appear in complex conjugate pairs.*

Proof. By definition,

$$A\lambda = \lambda x. \tag{8.1}$$

Now take the complex conjugate of both sides. Since A is real, it is unchanged by this operation, so

$$A\overline{\lambda} = \overline{\lambda}\overline{x}.$$

Hence $\overline{\lambda}$ is an eigenvector and $\overline{x}$ an associated eigenvector. ■

Proposition 8.2.3 *A square symmetric matrix A with real entries is positive (negative) definite iff all its eigenvalues are positive (negative).*

Remark 8.2.2 *The eigenvalues are guaranteed to be real since A is symmetric.*

Proof. ($\Rightarrow$) Suppose A is positive definite, let λ be an eigenvalue and let x be the corresponding eigenvector. Then $x^T Ax = \lambda x^T x$ so λ is the ratio between two positive numbers. ($\Leftarrow$) See [20]. ■

Corollary 8.2.2 *The eigenvalues of a real positive definite symmetric matrix are all real and strictly positive.*

Proposition 8.2.4 *Let A be a square invertible matrix and let λ be an eigenvalue of A. Then $1/\lambda$ is an eigenvalue of A^{-1}.*

Proof. Exercise. ■

8.3 Similarity transforms

8.3.1 Motivation

When analyzing dynamic systems (systems of differential or difference equations), we often want to uncouple the equations so that we can solve them row by row (scalar by scalar) rather than the whole system at once. You will see in chapter 9 exactly how this is done. In this chapter, we offer the toolbox needed to perform this uncoupling. Algebraically, what this is all about is the factorization ('decomposition') of matrices.

8.3.2 Definitions and basic results

We have already said that we are interested in factorizing matrices. One important class of factorizations arises from the concept of similarity.

Definition 8.3.1 *Let A and B be two $n \times n$ matrices of complex numbers. If there exists an invertible matrix C such that*

$$A = CBC^{-1}, \tag{8.2}$$

then A and B are said to be similar.

Remark 8.3.1 *Note that, as we promised in the previous section, CBC^{-1} is a factorization of A.*

Remark 8.3.2 *The (invertible!) function that takes us from A to B is sometimes called a* similarity transform.

Proposition 8.3.1 *If A and B are similar, then they have the same eigenvalues.*

Proof. Let λ be an eigenvalue of A. Then, for some $x \in \mathbb{C}^n$ such that $x \neq \theta$, we have $Ax = \lambda x$. Hence $CBC^{-1}x = \lambda x$ and consequently $BC^{-1}x = \lambda C^{-1}x$. Hence $y = C^{-1}x$ is such that $By = \lambda y$, and $y \neq \theta$ since C^{-1} is non-singular and $x \neq \theta$. ■

Definition 8.3.2 *A matrix is called lower (upper) triangular if all the elements above (below) the main diagonal are equal to zero.*

Proposition 8.3.2 *The eigenvalues of a lower or upper triangular matrix are its diagonal elements.*

Proof. The formula for the determinant. See [25]. ■

Definition 8.3.3 *Let A be a matrix of complex numbers. Then the Hermitian (or conjugate) transpose of A, denoted by A^H, is the (elementwise) complex conjugate of the transpose A^T. In other words, to find A^H, first take the complex conjugate of each element and then transpose (or vice versa).*

Definition 8.3.4 *A square matrix A is called* unitary *if $A^H A = AA^H = I$.*

Remark 8.3.3 *A unitary matrix is a matrix with orthogonal columns of norm one.*

Remark 8.3.4 *The inverse of a unitary matrix A is just the Hermitian transpose A^H.*

Remark 8.3.5 *A unitary matrix is the polar opposite of a singular matrix. While attempts to invert singular and close-to-singular matrices wreak havoc with numerical calculations, unitary matrices can be inverted quickly and precisely on a computer.*

To show how much fun we can have with the Hermitian transpose, we now throw in a definition and a result which is otherwise quite unimportant for our purposes (but hugely important in other contexts).

Definition 8.3.5 *A square matrix A such that $A = A^H$ is called a* Hermitian *matrix.*

Proposition 8.3.3 *The eigenvalues of a Hermitian matrix are real.*

Proof. Let A be a Hermitian matrix, let $\lambda \in \mathbb{C}$ be one of its eigenvalues, and let $x \in \mathbb{C}^n$ an associated eigenvector. Then

$$Ax = \lambda x \tag{8.3}$$

and hence

$$x^H Ax = \lambda x^H x \tag{8.4}$$

Now the left hand side is a *scalar*, and since it is equal to its conjugate transpose (why?), it is equal to its complex conjugate; hence it is real. Meanwhile, $x^H x$ is real and strictly positive since $x \neq \theta$. Hence λ is the ratio of a real number and a positive real number. ■

Getting back on track, we now define unitary similarity, which *is* important for our purposes.

Definition 8.3.6 *Two matrices A and B are said to be* unitarily similar *to each other if there is a unitary matrix Q such that*

$$A = QBQ^H. \tag{8.5}$$

We now come to the most important concrete examples of similarity transforms.

8.3.3 The eigenvalue/eigenvector decomposition

Definition 8.3.7 *A matrix A is said to be* diagonalizable *if it is similar to a diagonal matrix.*

Proposition 8.3.4 *An $n \times n$ matrix is diagonalizable iff it has a set of n linearly independent eigenvectors.*

Remark 8.3.6 *Note that the proof is constructive, so you had better read it or you'll miss the main point of this subsection.*

Remark 8.3.7 *Eigenvectors of distinct eigenvalues are linearly independent, so A having distinct eigenvalues is sufficient to ensure that A is diagonalizable. It*

is not however necessary; consider for example $A = I$. But when eigenvalues are repeated, care must be taken to select linearly independent eigenvalues, and unfortunately that it is not always possible.

Proof. Let Λ be the diagonal matrix that results from putting the eigenvalues of A on the main diagonal and zeros elsewhere. Now create a matrix Ω by letting its columns be a set of eigenvectors of A, ordered by the associated eigenvalues in the order that they appear in Λ. By hypothesis, these eigenvectors can be chosen to be linearly independent, and hence Ω is invertible. Now consider the equation

$$A\Omega = \Omega\Lambda. \tag{8.6}$$

Column by column, it says

$$Ax_i = \lambda_i x_i; \quad i = 1, 2, ..., n \tag{8.7}$$

where λ_i is an eigenvalue, and x_i an associated eigenvector. Hence the equation must be true, by definition! Inverting Ω, we find that

$$A = \Omega\Lambda\Omega^{-1}. \tag{8.8}$$

So A is similar to the diagonal matrix Λ, and the similarity transform is given by $\Omega\,(\cdot)\,\Omega^{-1}$. ■

Corollary 8.3.1 *Λ and Ω can be constructed so that the eigenvalues appear in any order along the diagonal of Λ.*

Example 8.3.1 *Sadly, however, not all matrices are diagonalizable. An example is the following.*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{8.9}$$

8.3.4 Schur form

So the bad news is that there exist non-diagonalizable ('defective') matrices. But the good news, and this is what really matters when we want to solve dynamic systems, is that every square matrix is (unitarily!) similar to an upper triangular matrix. This result is known as Schur's lemma, and the resulting factorization is called the *Schur form*.

Theorem 8.3.1 (Schur's lemma) *Every square matrix A is unitarily similar to an upper triangular matrix T, i.e. there exists an upper triangular matrix T and a unitary matrix Q such that*

$$A = QTQ^H. \tag{8.10}$$

Moreover, the Q and T matrices can be constructed so that the eigenvalues of A appear in any order along the diagonal of T.

Proof. See [20] or [23]. ■

Theorem 8.3.2 *Every square matrix is unitarily similar to a* lower *triangular matrix.*

Remark 8.3.8 *The fact that the similarity guaranteed here is unitary is important, since the inversion of a unitary matrix is a fast and precise numerical calculation.*

8.4 Symplectic matrices

Definition 8.4.1 *A $2n \times 2n$ real matrix M is said to be* symplectic *if*

$$M^T J M = J \tag{8.11}$$

where

$$J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}. \tag{8.12}$$

Proposition 8.4.1 *If M is symplectic, then M^T is similar to M^{-1} and hence the eigenvalues of M appear in reciprocal pairs, i.e. if λ is an eigenvalue of M, then so is $\frac{1}{\lambda}$.*

Proof. Exercise. ■

Proposition 8.4.2 *Let α be an invertible $n \times n$ real matrix and let β and γ be $n \times n$ real symmetric matrices. Then the $2n \times 2n$ matrix M defined via*

$$M = \begin{bmatrix} \alpha & \alpha\beta \\ \gamma\alpha & \alpha^{-T} + \gamma\alpha\beta \end{bmatrix} \tag{8.13}$$

is symplectic. Note that α^{-T} is the inverse of α^T.

Proof. Exercise. ■

8.5 Matrix pencils

8.5.1 Motivation

Singular difference equations. Consider

$$Ax_{t+1} = Bx_t + z_t \tag{8.14}$$

where A is possibly singular. For details, consider [29].

8.5.2 Basic definitions

Definition 9 *Let A and B be $n \times n$ matrices of complex numbers. Then the function $P(z) = B - zA$ is called a* matrix pencil.

Definition 10 *Let $B - zA$ be a matrix pencil. This pencil is called* regular *if there is a $z \in \mathbb{C}$ such that $|B - zA| \neq 0$.*

Definition 11 *Let $B - zA$ be a matrix pencil. Then the set of generalized eigenvalues $\lambda(B, A)$ is defined via*

$$\lambda(B, A) = \{z \in \mathbb{C} : |B - zA| = 0\} \tag{8.15}$$

8.5.3 Generalized Schur form

Theorem 12 (the complex generalized Schur form) *Let $B - zA$ be a regular matrix pencil. Then there exist unitary $n \times n$ matrices of complex numbers Q and Z such that*

1. *$QAZ = S$ is lower triangular,*
2. *$QBZ = T$ is lower triangular,*
3. *For each i, s_{ii} and t_{ii} are not both zero,*
4. *$\lambda(B, A) = \left\{ \frac{t_{ii}}{s_{ii}} : s_{ii} \neq 0 \right\}$,*
5. *The pairs (s_{ii}, t_{ii}), $i = 1, ..., n$ can be arranged in any order.*

Proof. See Golub & van Loan (1996). ■

Chapter 9

Dynamic systems

9.1 Ordinary differential equations (ODEs)

9.1.1 The problem and existence of a solution

9.1.1.1 General case

Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ be a function and consider the first–order system of ordinary differential equations

$$\begin{cases} \dot{x}(t) = f(t, x(t)) \\ x(t_0) = a \end{cases} \tag{9.1}$$

where an overdot denotes a derivative with respect to t. This system is said to be of *first-order*, since there are first derivatives but no higher-order derivatives. It is a system of *ordinary* (as opposed to partial) differential equations since only derivatives with respect to one variable (t) appear.

Equivalently, we might have written

$$x(t) = a + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{9.2}$$

Proposition 9.1.1 *The function $x : \mathbb{R} \to \mathbb{R}^n$ satisfies (9.1) iff it satisfies (9.2).*

Proof. Fundamental theorem of calculus. ■

You may want to ponder a while about the sort of equation we are considering. It is a *functional* equation in the sense that a solution is an entire function rather than just a number or even a vector.

We now ask ourselves when our system might have a unique solution. Very loosely speaking, there seems to be hope: we have an initial position and we have the direction we're heading in. So if nothing strange is going on, we should be able to figure out where we'll be.

To show this formally, we might draw upon what we learned in chapter 5 and define a suitable space of functions in which to search for a solution, then define an operator on that space into itself under which the solution is a fixed point. The final step would be to take an arbitrary initial function, and apply the operator repeatedly to construct a sequence of functions and then show that (given certain assumptions) this sequence converges to the unique solution. We won't go through this program here (see, for example, [27] if you are interested). Instead we will just give the final result.

Definition 9.1.1 *A function $f : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is said to be* Lipschitz continuous *if there is a function $M : \mathbb{R} \to \mathbb{R}_+$ such that, for all $x, y \in \mathbb{R}^n$,*

$$\|f(t,x) - f(t,y)\| \leq M(t)\|x - y\| \tag{9.3}$$

Remark 9.1.1 *If $n = 1$ and $\dfrac{\partial f(t,x)}{\partial x}$ is bounded, then f is Lipschitz.*

Theorem 9.1.1 (Cauchy-Picard) *Suppose*

1. *f is Lipschitz and that*
2. *The M function in the Lipschitz condition can be chosen so that $\int_I M(u)\,du < \infty$ where I is a closed interval such that $t_0 \in I$.*

Then (9.1) has a unique solution on I.

Remark 9.1.2 *A natural case to consider is $I = [t_0, t_1]$. But in fact the point t_0 is allowed to be in the interior of I or even to be the upper limit.*

Corollary 9.1.1 *If $\int_I M(u)\,du < \infty$ for every bounded interval I, then (9.1) has a unique solution defined on the whole of $\mathbb{R}$.*

Corollary 9.1.2 *If I is a closed interval and f is Lipschitz on $I \times \mathbb{R}^n$ with $M(t) = M$ (constant), then (9.1) has a unique solution on I.*

Remark 9.1.3 *The integrability condition is there to prevent the solution from exploding off into infinity in finite time.*

Example 9.1.1 *Consider the scalar ODE*

$$\begin{cases} \dot{x}(t) = x^2(t) \\ x(t_0) = a \end{cases} \tag{9.4}$$

A solution, if it exists, is either identically zero (in which case we must have $a = 0$) or has the form

$$x(t) = \frac{1}{-t + t_0 + a^{-1}}. \tag{9.5}$$

so that any solution is not defined on I if $t_0 + a^{-1} \in I$. However, all hope is not lost for this equation. If $t_0 = 1$, $a = 1$ then the solution $x(t) = \dfrac{1}{2-t}$ is defined for all $t \in (-\infty, 2)$.

Example 9.1.2 *(taken from [27]). Consider the scalar ODE*

$$\begin{cases} \dot{x}(t) = 3x^{2/3}(t) \\ x(0) = 0. \end{cases} \tag{9.6}$$

where, to avoid complex numbers, we define $x^{2/3}$ *as* $\sqrt[3]{x^2}$*. Then, for any* $\alpha \geq 0$,

$$x_\alpha(t) = \begin{cases} (t+\alpha)^3 & \text{when } t < -\alpha \\ 0 & \text{when } -\alpha \leq t < \alpha \\ (t-\alpha)^3 & \text{when } t \geq \alpha \end{cases} \tag{9.7}$$

is a solution. Hence our ODE has infinitely many solutions.

9.1.1.2 Linear systems

Consider the system

$$\begin{cases} \dot{x}(t) = A(t)\, x(t) \\ x(t_0) = a. \end{cases} \tag{9.8}$$

To apply Cauchy-Picard to this, we need another definition.

Definition 9.1.2 *Let* A *be square matrix. Then its norm is defined via*

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \tag{9.9}$$

Remark 9.1.4 *It may of course be confirmed that this really is a norm. See [17].*

We can now derive the following result from Cauchy-Picard's theorem.

Proposition 9.1.2 *If* $\int_I \|A(t)\|\, dt < \infty$ *for each bounded interval* I*, then (9.8) has a unique solution defined on the whole of* $\mathbb{R}$*.*

Proof. By the definition of the matrix norm, we have, for all $x, y \in \mathbb{R}^n$,

$$\|A(t)\, x - A(t)\, y\| \leq \|A(t)\|\, \|x - y\| \tag{9.10}$$

Hence the Lipschitz condition is satisfied, and the other (integrability) condition is satisfied by assumption. ■

Remark 9.1.5 *If $A(t) \equiv A$, then (9.8) has a unique solution defined on the whole of $\mathbb{R}$.*

Notice that the uniqueness of the solution requires that we fix the value $x(t_0) = a$. One useful perspective on this is that, for a linear system, the set of solutions (ignoring the condition $x(t_0) = a$) is an $n-$dimensional vector space. The condition $x(t_0) = a$ then imposes n independent linear restrictions on the space of solutions, and hence picks out a unique solution. There are of course other ways of imposing n linear restrictions than putting $x(t_0) = a$; we will see other examples below.

Exercise 9.1.1 *Show that the set of solutions $x(t)$ to the system $\dot{x}(t) = A(t)x(t)$ is a vector space, i.e. that if α is a scalar and $x(t)$ and $y(t)$ are solutions, then $\alpha[x(t) + y(t)]$ is also a solution.*

9.1.2 Solving scalar equations in special cases

Very few differential equations have explicit solutions (whatever that means exactly), so very often one has to resort to numerical methods (see [37] or [36]) or settle for a qualitative characterization of the solution (see section 9.1.3). But mere qualitative characterization is often not good enough (although it is occasionally interesting), and numerical techniques may be slow and/or tedious to implement. So before you decide to use one of them, you should be aware of at least the following basic pencil-and-paper techniques.

9.1.2.1 Separable

The simplest ODEs are the separable ones. The general format is the following.

$$\begin{cases} \dot{x}(t) = f(t)\, g(x(t)) \\ x(t_0) = a. \end{cases} \tag{9.11}$$

We will motivate the solution using a rather wild argument, and then clean up our act by confirming that our conclusion is in fact correct in a special case. Using Leibniz's notation, the ODE can be written as

$$\frac{dx}{dt} = f(t)\, g(x(t)). \tag{9.12}$$

Now suppose for the sake of argument that dx and dt are numbers (which of course they are not). Ignoring the points $x = a_j$ at which $g(x) = 0$, we write

$$\frac{dx}{g(x(t))} = f(t)\, dt. \tag{9.13}$$

Now integrate both sides from t_0 to t. We get

$$\int_{x(t_0)}^{x(t)} \frac{dx}{g(x)} = \int_{t_0}^{t} f(s)\, ds. \tag{9.14}$$

Calculating these integrals, one can sometimes solve for $x(t)$, or at least find an equation that implicitly defines $x(t)$. In addition, since we divided by $g(x(t))$, we might have one of the (constant) solutions $x(t) = a_j$ where the a_j are the zeros of g (why?). But for one of these constants to be a solution, we must of course require that $a = a_j$ for some j. Alternatively, we might not fix the value of x at any particular point t_0. Then *all* these constant functions would be solutions.

Example 9.1.3 *Consider*

$$\begin{cases} \dot{x}(t) = e^t x(t) \\ x(1) = 1. \end{cases} \tag{9.15}$$

Using our method, we get

$$\int_1^{x(t)} \frac{dx}{x} = \int_1^t e^s ds, \tag{9.16}$$

which implies

$$\ln x(t) = e^t - e. \tag{9.17}$$

Hence the solution is

$$x(t) = \exp\left(e^t - e\right),$$

and it is not hard to confirm (do that!) that this really is a solution to (9.15).

9.1.2.2 Linear

Consider the scalar differential equation

$$\begin{cases} \dot{x}(t) = a(t)\, x(t) + b(t) \\ x(t_0) \quad \text{given} \end{cases} \tag{9.18}$$

To solve this, we use a rather ingenious trick. First reshuffle the equation so that we have

$$\dot{x}(t) - a(t)\, x(t) = b(t) \tag{9.19}$$

Then multiply each side by $\exp\left\{-\int_{t_0}^t a(s)\, ds\right\}$. This leads to an equivalent expression, since $\exp(x) \neq 0$ for any (real or complex) x. The result is

$$\begin{aligned} &\exp\left\{-\int_{t_0}^t a(s)\, ds\right\} \dot{x}(t) - \exp\left\{-\int_{t_0}^t a(s)\, ds\right\} a(t)\, x(t) = \\ &= \exp\left\{-\int_{t_0}^t a(s)\, ds\right\} b(t). \end{aligned} \tag{9.20}$$

Now we note that the left hand side is $\frac{d}{dt}\left[\exp\left\{-\int_{t_0}^t a(s)\, ds\right\} x(t)\right]$. Now we

want to integrate both sides of the equation from t_0 to t with respect to t. This doesn't quite make sense, but if you replace t by v and integrate from t_0 to t with respect to v, we get

$$\exp\left\{-\int_{t_0}^{t} a(s)\,ds\right\} x(t) - x(t_0) = \int_{t_0}^{t} \exp\left\{-\int_{t_0}^{s} a(u)\,du\right\} b(s)\,ds \tag{9.21}$$

Consequently

$$\begin{aligned} x(t) &= \exp\left\{\int_{t_0}^{t} a(s)\,ds\right\} x(t_0) + \\ &+ \exp\left\{\int_{t_0}^{t} a(s)\,ds\right\} \int_{t_0}^{t} \exp\left\{-\int_{t_0}^{s} a(u)\,du\right\} b(s)\,ds \end{aligned} \tag{9.22}$$

or, tidying up a bit,

$$x(t) = \exp\left\{\int_{t_0}^{t} a(s)\,ds\right\} x(t_0) + \int_{t_0}^{t} \exp\left\{\int_{s}^{t} a(u)\,du\right\} b(s)\,ds. \tag{9.23}$$

Given that we won't discuss the case where $x(t)$ is a vector and $a(t) = A(t)$ is a matrix-valued function of t (except when $A(t) \equiv A$; see below), it may be of some interest to know that this solution is valid in that case as well, provided we interpret exp as the matrix exponential function (see section 9.1.4.1).

Advice. If you find the solution (9.23) hard to remember, memorize the derivation instead!

Example 9.1.4 *Consider*

$$\begin{cases} \dot{x}(t) &= \dfrac{1}{t}x(t) + t; \quad t \neq 0 \\ x(1) &= 1. \end{cases} \tag{9.24}$$

The first step is to figure out $\int_1^t \frac{1}{s} ds = \ln t$. *The rest is easy. We get*

$$\begin{aligned} x(t) &= t \cdot 1 + \int_1^t \exp(\ln t - \ln s)\, s ds = \\ &= t + t \int_1^t \frac{s}{s} ds = t + t(t-1). \end{aligned} \tag{9.25}$$

Exercise 9.1.2 *Check that (9.23) really solves (9.18)! Hint: Either use Leibniz' rule or recall the penultimate version of the solution, i.e. note that* $\int_s^t a(s)\, ds = \int_{t_0}^t a(s)\, ds - \int_{t_0}^s a(s)\, ds$ *and factor out* $\exp\left\{\int_{t_0}^t a(s)\, ds\right\}$ *from the second integral. Then use the product rule and the fundamental theorem of calculus.*

Exercise 9.1.3 *Consider (9.19) and suppose we are interested in a solution defined on the closed interval* $[t_0, t_1]$. *Suppose the initial value* $x(t_0)$ *is unknown but that the endpoint value* $x(t_1)$ *is known. Derive the solution in terms of* $x(t_1)$.

9.1.2.3 Bernoulli

Consider

$$\dot{x}(t) = a(t)\, x(t) + b(t)\, x^{\alpha}(t) \tag{9.26}$$

where α is an arbitrary real number. If $\alpha = 1$, then we have a separable differential equation. So let's assume that $\alpha \neq 1$ (then we can divide by $(1-\alpha)$ later on). Also, let's consider only solution paths such that $x(t) > 0$ (otherwise we get into trouble with the $x^{\alpha}(t)$ term). Now divide by $x^{\alpha}(t)$ (which is non-zero by our assumption). We get

$$x^{-\alpha}(t)\, \dot{x}(t) = a(t)\, x^{1-\alpha}(t) + b(t) \tag{9.27}$$

and define

$$z(t) = x^{1-\alpha}(t). \tag{9.28}$$

It follows that

$$\dot{z}(t) = (1-\alpha)\, x^{-\alpha}(t)\, \dot{x}(t). \tag{9.29}$$

Hence

$$\dot{z}(t) = (1-\alpha)\, a(t)\, z(t) + (1-\alpha)\, b(t). \tag{9.30}$$

But this is a *linear* ODE and we know how to solve that!

Example 9.1.5 *(from Macro 1). Consider a continuous-time version of Solow's growth model.*

$$\begin{cases} \dot{k}(t) = sk^{\alpha}(t) - \delta k(t) \\ k(0) = k_0. \end{cases} \tag{9.31}$$

This is a Bernoulli differential equation. So define $z(t) = k^{1-\alpha}(t)$. *Then* $\mathbf{z}$ *solves*

$$\begin{cases} \dot{z}(t) + (1-\alpha)\,\delta z(t) = s(1-\alpha) \\ z(0) = k_0^{1-\alpha} \end{cases} \tag{9.32}$$

Now multiply by $e^{(1-\alpha)\delta t}$ *and we get*

$$\frac{d}{dt}\left[e^{(1-\alpha)\delta t} z(t)\right] = s(1-\alpha)\, e^{(1-\alpha)\delta t} \tag{9.33}$$

Integrating from 0 *to* t, *we get*

$$\begin{aligned} e^{(1-\alpha)\delta t} z(t) - k_0^{1-\alpha} &= s(1-\alpha) \int_0^t e^{(1-\alpha)\delta s} ds = \\ &= \frac{s}{\delta}\left[e^{(1-\alpha)\delta t} - 1\right]. \end{aligned} \tag{9.34}$$

Hence

$$z(t) = \frac{s}{\delta} + e^{-(1-\alpha)\delta t}\left(k_0^{1-\alpha} - \frac{s}{\delta}\right). \tag{9.35}$$

Consequently

$$k(t) = \left[\frac{s}{\delta} + e^{-(1-\alpha)\delta t}\left(k_0^{1-\alpha} - \frac{s}{\delta}\right)\right]^{\frac{1}{1-\alpha}}. \tag{9.36}$$

Notice that if we set the initial value to $k_0 = \left(\frac{s}{\delta}\right)^{\frac{1}{1-\alpha}}$, *then* $k(t)$ *remains at this level for all* $t \geq 0$. *Also, if* $0 < \alpha < 1$ *then* $k(t) \to \left(\frac{s}{\delta}\right)^{\frac{1}{1-\alpha}}$ *as* $t \to +\infty$ whatever *the initial value* k_0.

Qualitative properties of the solution (like the above convergence property) can sometimes be determined without solving the ODE explicitly. When we *can't* solve the ODE explicitly this qualitative analysis becomes important.

9.1.3 Introduction to qualitative analysis

In this section, we will always have in the background the following time-independent system of differential equations without initial or endpoint conditions.

$$\dot{x}(t) = f(x(t)). \tag{9.37}$$

We now define a *steady state* and also what we mean by *stability*.

Definition 9.1.3 *A* steady state *of (9.37) is a point* x^* *at which* $f(x^*) = 0$.

Proposition 9.1.3 *Let* x^* *be a steady state of (9.37) and let (9.37) have a unique solution. Then the unique solution to (9.37) with the condition* $x(t_0) = x^*$ *is the constant function* $x(t) = x^*$.

Proof. Obvious. ■

Definition 9.1.4 *A steady state* x^* *of (9.37) is said to be* stable *if for each* $\varepsilon > 0$ *there is a* δ_ε *such that every solution path* $x(t)$ *with* $\|x(0) - x^*\| < \delta_\varepsilon$ *satisfies* $\|x(t) - x^*\| < \varepsilon$ *for all* $t \geq 0$.

Definition 9.1.5 *A steady state x^* of (9.37) is said to be* asymptotically *stable if it is stable and there is a $\delta > 0$ such that every solution path $x(t)$ with $\|x(0) - x^*\| < \delta$ converges to x^* as $t \to +\infty$.*

In words, if you start at a steady state, you stay there forever. If the steady state is stable, then if you start close to it, you stay close to it. If it is asymptotically stable, then if you start close to it, you converge to it.

Notice that our definitions of stability are local in the sense that they only have something to say about the behavior of the system when you begin close to the steady state in question. In the linear (with constant coefficients) case, but not in general, if there is a unique steady state, then local asymptotic stability implies global asymptotic stability in the sense that we will converge to the unique steady state *wherever* we start out. (This claim will be justified in the next section.)

We now want to characterize a (locally) stable steady state, and the idea is to linearize the system around a steady state and hope that the local dynamics of the actual system are well approximated in some sense by the linearized system. (This hope turns out to be substantiated by Lyapunov's theorem.) We then analyze the stability of the linear system. So here it seems appropriate to pause for a rather long digression on linear systems, which are in any case very important in themselves.

9.1.4 First-order linear systems with constant coefficients

In this section we consider systems of the form

$$\begin{cases} \dot{x}(t) = Ax(t) + b \\ x(0) = x_0 \end{cases} \tag{9.38}$$

We will not treat linear systems with variable coefficients here. See [14] instead. We will also (initially at least) assume A to be invertible; this is equivalent to there being a unique steady state. Then we can consider the dynamics of the deviation from the steady state $x^* = -A^{-1}b$. Defining $y(t) = x(t) - x^*$, $y(t)$ solves

$$\begin{cases} \dot{y}(t) = Ay(t) \\ y(0) = y_0 \end{cases} \tag{9.39}$$

where, of course, $y_0 = x_0 - x^*$. Such a system is called *homogeneous*, and to the analysis of such systems we now turn. Inspired by the scalar case, it is tempting to say that the solution is

$$y(t) = \exp(At) \cdot y_0 \tag{9.40}$$

But what does it mean to take the exponential of a matrix? That question is answered in the following section.

N.B. In the following sections, we will be talking about homogeneous systems, and definitions and results will be stated that apply formally only to them. But of course we can easily extend (by looking at deviations from the steady state) all the results to non-homogeneous systems with a unique steady state. Keep that in mind as you read on.

9.1.4.1 The matrix exponential function

Again inspired by the scalar case, we offer the following definition.

Definition 9.1.6 *Let A be a square matrix. Then*

$$\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!} \tag{9.41}$$

where the convention is that $A^0 = I$ for every square matrix A.

Remark 9.1.6 *It may of course be confirmed that this series converges (uniformly) for every square matrix A.*

Warning. It may happen that

$$\exp(A)\exp(B) \neq \exp(A+B). \tag{9.42}$$

However, we do have the following proposition.

Proposition 9.1.4 *If $AB = BA$ then $\exp(A)\exp(B) = \exp(A+B)$.*

Proof. Exercise. ■

Corollary 9.1.3 *For all square matrices A and all scalars s, t, we have*

$$\exp(As)\exp(At) = \exp(A(s+t))$$

Corollary 9.1.4 *For any square matrix A, the matrix $\exp(A)$ is non-singular with inverse $\exp(-A)$.*

Proposition 9.1.5 *For any square matrix A,*

$$\frac{d}{dt}\exp(At) = A\exp(At). \tag{9.43}$$

Proof. Copy the proof of the scalar case. ■

Remark 9.1.7 *This shows that (9.40) really solves (9.39)!*

Remark 9.1.8 *Note that when we differentiate a matrix with respect to a* scalar, *we don't bother to vectorize it.*

Proposition 9.1.6 *If* $A = \operatorname{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ *is a diagonal matrix, then*

$$\exp(At) = \begin{bmatrix} e^{\lambda_1 t} & & & \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ & & & e^{\lambda_n t} \end{bmatrix} \tag{9.44}$$

where the blank spaces represent zeros.

Proposition 9.1.7 *If* $A = CBC^{-1}$ *then*

$$\exp(A) = C \exp(B) C^{-1} \tag{9.45}$$

Proof. Exercise. ■

Example 9.1.6 *Consider*

$$\begin{cases} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \\ \\ \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \end{cases} \tag{9.46}$$

Apparently A *is nilpotent, i.e.* $A^n = 0$ *for all* $n \geq 2$. *So we can calculate the matrix exponential explicitly! We get*

$$\exp(At) = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!} = I + At + 0 + ... = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \tag{9.47}$$

and hence the unique solution is

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_1 + a_2 t \\ a_2 \end{bmatrix} \tag{9.48}$$

9.1.4.2 Uncoupling by diagonalization

We have just solved our system (9.40). We found that

$$y(t) = \exp(At)\, y_0 \tag{9.49}$$

and consequently the solution to the corresponding non-homogeneous system is

$$x(t) = x^* + \exp(At)\,(x_0 - x^*)\,. \tag{9.50}$$

But when A isn't nilpotent, this usually isn't explicit enough. We know we could be more explicit if A were diagonal. But it is just as good if A is diagonalizable! So suppose A is diagonalizable, and let $A = \Omega\Lambda\Omega^{-1}$ be an eigenvalue/eigenvector decomposition. Then, of course

$$x(t) = x^* + \Omega \exp(\Lambda t)\,\Omega^{-1}\,(x_0 - x^*) \tag{9.51}$$

i.e.

$$x(t) = x^* + \Omega \begin{bmatrix} e^{\lambda_1 t} & & & \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ & & & e^{\lambda_n t} \end{bmatrix} \Omega^{-1}\,(x_0 - x^*)\,. \tag{9.52}$$

Calling $\Omega^{-1}(x_0 - x^*) = c$ and writing $\Omega = \begin{bmatrix} s_1 & s_2 & \cdots & s_n \end{bmatrix}$ where the s_i are the columns of Ω (and hence the eigenvectors of A!) we find that

$$x(t) = x^* + \sum_{k=1}^{n} c_k e^{\lambda_k t} s_k. \tag{9.53}$$

This equation illustrates an important fact about the solutions to ODEs: they are the steady state plus arbitrary linear combinations of basis solutions $e^{\lambda_k t} s_k$ to the corresponding homogeneous system, which are linearly independent and hence span the whole space of solutions. (Recall that the set of solutions to

the homogeneous system is a vector space!) Note that the columns of $\exp(At)$ always form a set of linearly independent solutions, since $\exp(At)$ is a non-singular matrix. To see that the basis solutions $e^{\lambda_k t} s_k$ are linearly independent, recall our assumption that $\Omega = \begin{bmatrix} s_1 & s_2 & \cdots & s_n \end{bmatrix}$ is non-singular.

9.1.4.3 Initial values and endpoints

As we have seen, each initial value $y_{k,0}$ imposes one linear restriction on the $n-$dimensional space of solutions. (Here $y_{k,t}$ is the value of the kth element of the vector y_t.) In the case of linear systems with constant coefficients, fixing the entire vector y_0 pins down the solution exactly. But, alternatively, we could fix endpoint values $y_{k,T}$ instead. Or we can combine initial and endpoint values. There is little to say theoretically about this except that it works as soon as you fix as many points as there are variables in the system.

Exercise 9.1.4 *Let*

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = c_1 e^t \begin{bmatrix} 0 \\ 1 \end{bmatrix} + c_2 e^{-t} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{9.54}$$

be a representation of the full set of solutions to a two-dimensional linear system of ODEs. Determine c_1 and c_2 to ensure that $x(0) = a$ and $y(1) = b$.

Exercise 9.1.5 *Consider the dynamic system*

$$\begin{bmatrix} \dot{x}(t) \\ \dot{\lambda}(t) \end{bmatrix} = A \begin{bmatrix} x(t) \\ \lambda(t) \end{bmatrix} \tag{9.55}$$

where $x(t)$ and $\lambda(t)$ are vectors of the same dimension. Let

$$\exp(At) = \begin{bmatrix} \mathcal{E}_{11}(t) & \mathcal{E}_{12}(t) \\ \mathcal{E}_{21}(t) & \mathcal{E}_{22}(t) \end{bmatrix} \tag{9.56}$$

be a partition into equal parts. Suppose $x(0) = x_0$ and $x(T) = x_T$. Find $\lambda(0)$. Do you need to make any invertibility assumptions?

9.1.4.4 Complex eigenvalues and real oscillatory solutions

Some or all of the eigenvalues may be complex, so that the representation (9.53) of the set of solutions may easily produce complex solutions even if you restrict the c vector to have real entries. Thus if you want to look only at the real solutions, the representation (9.53) is not convenient. What we would like is a representation so that any real-valued solution is a real-linear combination of a set of real-valued basis functions. (By a real-linear combination I mean a linear combination with real scalars.)

Before we go on to the solution of this problem, we note that there is actually a very easy way forward here: take a real initial value y_0; the solution then stays real. But such an approach misses the important conceptual point that complex eigenvalues, even if we focus only on the real solutions, give rise to oscillatory behavior. The essential reason for this is Euler's formula. Putting $\lambda_i = \alpha_i + i\beta_i$, Euler's formula says that $e^{\lambda_k t} = e^{\alpha_k t}(\cos \beta_k t + i \sin \beta_k t)$. So as soon as $\beta_k \neq 0$ we get oscillatory behavior, either in the form of damped ($\alpha_k < 0$) or explosive ($\alpha_k > 0$) cycles.

In any case, here comes the general solution and an explanation (but not quite a proof). Suppose A has real entries so that the complex eigenvalues and eigenvectors come in complex conjugate pairs. Let λ_k be an eigenvalue with a non-zero imaginary part and consider the pair of eigenvalues $\lambda_k, \overline{\lambda}_k$ and the associated eigenvectors s_k and $\overline{s}_k$. We now pause to state a useful lemma which says that if a complex-valued function solves a system of ODEs, then their real and imaginary parts do, too.

Theorem 9.1.2 *Let $y(t) = u(t) + iv(t)$ satisfy $\dot{y}(t) = Ay(t)$, where $u(t)$ and $v(t)$ are real-valued functions. Then so do $u(t)$ and $v(t)$.*

Proof. We are told that $y(t)$ satisfies $\dot{y}(t) = Ay(t)$. But then $\dot{u}(t) + i\dot{v}(t) = Au(t) + iAv(t)$. But this can only be true if $\dot{u}(t) = Au(t)$ and $\dot{v}(t) = Av(t)$. ■

Getting back on track, apparently two linearly independent solutions to our homogeneous system of ODEs are

$$y_k(t) = e^{\lambda_k t} s_k \tag{9.57}$$

and

$$y_{k+1}(t) = e^{\overline{\lambda}_k t} \overline{s}_k. \tag{9.58}$$

(Here the vector $y_k(t)$ is the value at t of the kth basis solution.) Now take the real and imaginary parts of, say, the first solution $y_k(t)$. Writing the eigenvalue λ_k and eigenvector s_k in Cartesian form with $\lambda_k = \alpha_k + i\beta_k$ and $s_k = a_k + ib_k$, and using Euler's formula, we get

$$\begin{aligned} y_k(t) = \; & e^{\alpha_k t}(a_k \cos(\beta_k t) - b_k \sin(\beta_k t)) + \\ & + ie^{\alpha_k t}(a_k \sin(\beta_k t) + b_k \cos(\beta_k t)) \end{aligned} \tag{9.59}$$

Writing $y_k(t) = u_k(t) + iv_k(t)$ and invoking Theorem 9.1.2, we have the following candidate basis solutions.

$$u_k(t) = e^{\alpha_k t}(a_k \cos(\beta_k t) - b_k \sin(\beta_k t)) \tag{9.60}$$

$$v_k(t) = e^{\alpha_k t}(a_k \sin(\beta_k t) + b_k \cos(\beta_k t)) \tag{9.61}$$

and these can be shown to be linearly independent.

This motivates the following general representation of the full set of real solutions to (9.40). Arrange the eigenvalues in an order so that complex conjugate pairs are juxtaposed. Go through the eigenvalues in order from $k = 1$ to n. For each real eigenvalue λ_k and eigenvector s_k, the corresponding basis solution is just

$$\varphi_k(t) = e^{\lambda_k t} s_k. \tag{9.62}$$

For each complex eigenvalue λ_k and eigenvector s_k, construct $u_k(t)$ and $v_k(t)$ as above and set

$$\begin{cases} \varphi_k(t) &= u_k(t) \\ \varphi_{k+1}(t) &= v_k(t). \end{cases} \tag{9.63}$$

For every complex eigenvalue λ_k you encounter, ignore the conjugate eigenvalue $\lambda_{k+1} = \overline{\lambda}_k$. Now the full set of solutions may be represented as

$$y(t) = \sum_{k=1}^{n} c_k \varphi_k(t) \tag{9.64}$$

Note that it is essential for this result that A have real entries and hence that the eigenvalues and eigenvectors come in reciprocal pairs. Otherwise we could not afford to ignore one out of every two complex eigenvalues. You may want to check that, in our case, looking at λ_k and $\overline{\lambda}_k$ yields exactly the same space of solutions.

Exercise 9.1.6 *Solve the following system of ODEs.*

$$\begin{cases} \dot{x}(t) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} x(t) \\ x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{cases} \tag{9.65}$$

Express you answer in terms of sines and cosines only (no complex exponentials). Illustrate the solution path in a picture. Hint: The solution $x(t)$ goes round and round the unit circle.

9.1.4.5 Stability

You may already have guessed that the stability of a system of ODEs is intimately connected with the eigenvalues of the coefficient matrix A. The notion of the norm

of a matrix turns out also to be important, and, what is more, the eigenvalues turn out to be closely related to the matrix norm. More precisely, we have the following definitions and propositions.

Definition 9.1.7 *Let B be a square matrix. Then its* spectrum $\lambda(B)$ *is defined via*

$$\lambda(B) = \{\lambda \in \mathbb{C} : \det(B - \lambda I) = 0\} \tag{9.66}$$

i.e. $\lambda(B)$ is the set of eigenvalues of B.

Proposition 9.1.8 *Let A be a square matrix. Then*

$$\|A\| = \max_{\lambda \in \lambda(A^T A)} \sqrt{|\lambda|}. \tag{9.67}$$

Corollary 9.1.5 *Let A be a square symmetric matrix. Then*

$$\|A\| = \max_{\lambda \in \lambda(A)} |\lambda|. \tag{9.68}$$

Proof. See [20]. ■

Proposition 9.1.9 *Let A and B be square matrices. Then*

$$\|AB\| \leq \|A\| \|B\| \tag{9.69}$$

Proof. See [17]. ■

Definition 9.1.8 *Let A be a square invertible matrix. Then its* condition number *is defined via*

$$c(A) = \|A\| \left\|A^{-1}\right\|.$$

Now let's take a look at the solution $y(t)$ of our homogeneous system and see under what circumstances it is stable and/or converges to zero. Recall that

$$y(t) = \Omega \exp(\Lambda t) \Omega^{-1} y_0. \tag{9.70}$$

By the definition of the matrix norm,

$$\|y(t)\| \leq \left\|\Omega \exp(\Lambda t)\,\Omega^{-1}\right\| \|y_0\| \tag{9.71}$$

By Propositions 9.1.9 and 9.1.8, and the fact that the exponential function is non-decreasing,

$$\begin{aligned} \|\Omega \exp(\Lambda t)\,\Omega^{-1}\| &\leq \|\Omega \exp(\Lambda t)\| \|\Omega^{-1}\| \leq \\ \leq \|\Omega\| \|\exp(\Lambda t)\| \|\Omega^{-1}\| &= c(\Omega) \|\exp(\Lambda t)\| = \\ = c(\Omega) \max_{\lambda \in \lambda(A)} &\left|e^{\lambda t}\right|. \end{aligned} \tag{9.72}$$

Hence

$$\|y(t)\| \leq c(\Omega) \max_{\lambda \in \lambda(A)} \left|e^{\lambda t}\right| \|y_0\|. \tag{9.73}$$

Now represent the maximizing eigenvalue $\lambda_{\max}$ in Cartesian form. Then $\lambda_{\max} = x + yi$ where $x = \mathrm{Re}(\lambda_{\max})$ and $y = \mathrm{Im}(\lambda_{\max})$ are real numbers and

$$\|y(t)\| \leq c(\Omega) \left|e^{xt}\right| \left|e^{yit}\right| \|y_0\| = \left|e^{xt}\right| c(\Omega) \|y_0\|.$$

All this suggests (and proves for diagonalizable matrices!) the following theorem.

Theorem 9.1.3 *Suppose the square matrix A is such that all the eigenvalues of A have strictly negative real parts. Then, and only then, the system (9.40) is globally asymptotically stable.*

Proof. If y is a real number, then $|e^{iyt}| = 1$ for all t, and if x is a negative real number, then $|e^{xt}| \to 0$ monotonically as $t \to \infty$. Note that the assumption of the theorem implies (why?) that A is invertible, and hence that there is a unique steady state. The only gap in the proof is that we assume A to be diagonalizable. The generalization to any square matrix is left to the reader. ■

Corollary 9.1.6 *For a linear system with constant coefficients, local asymptotic stability of the steady state is equivalent to global asymptotic stability.*

9.1.4.6 Saddle paths

Even if a homogeneous linear system is not globally stable, there may be initial values from which the solution does converge, or, to include a borderline case which we count as stable, at least remains bounded, i.e. there is an M such that $\|x(t)\| \leq M$ for all t. Consider the set of such initial values $S \subset \mathbb{R}^n$. We call S the *saddle path* for the system.

To characterize the set S we arrange Λ in a stable and an unstable part. Just rearrange the eigenvalues and eigenvectors so that the n_u unstable eigenvalues (the ones with positive real parts) come first, and the remaining n_s stable ones last. (We assume that there are no zero eigenvalues, i.e. that there is a unique steady state.) Partitioning, we have

$$\Lambda = \begin{bmatrix} \underset{n_u \times n_u}{\Lambda_1} & \\ & \underset{n_s \times n_s}{\Lambda_2} \end{bmatrix}. \tag{9.74}$$

We now want to characterize the set of initial vectors y_0 such that

$$\Omega \exp(\Lambda t) \Omega^{-1} y_0 \to 0. \tag{9.75}$$

This is of course the same set (why?) as the set of y_0 such that

$$\exp(\Lambda t) \Omega^{-1} y_0 \to 0. \tag{9.76}$$

Defining

$$\Omega^{-1} = \begin{bmatrix} \underset{n_u \times n}{\Omega^1} \\ \underset{n_s \times n}{\Omega^2} \end{bmatrix} \tag{9.77}$$

and writing the expression out in partitioned form, the requirement is

$$\begin{bmatrix} e^{\Lambda_1 t} & \\ & e^{\Lambda_2 t} \end{bmatrix} \begin{bmatrix} \Omega^1 \\ \Omega^2 \end{bmatrix} y_0 \to 0. \tag{9.78}$$

For this to be true, we need the coefficients multiplying the explosive factor $e^{\Lambda_1 t}$ to be zero. This means that

$$\Omega^1 y_0 = 0. \tag{9.79}$$

So the saddle path S is just the set of vectors $y_0 \in \mathbb{R}^n$ such that $\Omega^1 y_0 = 0$. Apparently this equation contains n_u (independent!) linear restrictions. The polar cases are the following. If there are $n_u = n$ unstable eigenvalues, then there are n independent linear restrictions; thus the only choice is $y_0 = 0$ and hence $S = \{0\}$. Conversely if $n_u = 0$, then $S = \mathbb{R}^n$.

Remark 9.1.9 *You may wonder whether saddle paths have anything to do with the sort of thing that John Wayne used to sit on. Well, they do. Consider a saddle, and consider dropping a ball on it and see what happens. Typically it will roll off, but there is a line that points in the same direction as where the horse is going such that the ball rolls down to roughly where Wayne's groin used to be and stays there. This line is the saddle path.*

More mathematically, consider the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ *and think of the set of points* $\mathcal{S} = \{(x, y, z) \in \mathbb{R}^3 : f(x, y) = z\}$ *as a surface in three-dimensional space. Now think of* $\mathbb{R}^3$ *as physical space and* $\mathcal{S}$ *as, say, a mountain range, covered with ice to remove the friction. Now make a snowball and let go of it at* $(x, y, f(x))$*. Then (why?) the ball will roll away in the direction* $-\nabla f(x, y)$ *(the minus sign says that it rolls down rather than climbs up). So, if we ignore the velocity and focus only on the direction, it is as if the ball's position* $(x(t), y(t))$

were governed by the dynamic system

$$\dot{x}(t) = -\nabla f(x, y). \tag{9.80}$$

Now let $f(x) = x^T(-A)x$ *and consequently* $\dot{x}(t) = -\nabla f(x(t), y(t)) = Ax(t)$. *We now have linear system of ODEs.*

From elementary calculus, we know that if A is a symmetric matrix with real entries, then if A is negative definite (all the eigenvalues are real and negative), the origin is a global minimum of f, and if A is positive definite (all the eigenvalues are real and positive), then the origin is a global maximum of f. Otherwise it is a saddle point, which in three-dimensional space is like a mountain pass. Extending this to non-symmetric matrices with real eigenvalues, we conclude that if all the of the eigenvalues of A are real and negative, f has a single global minimum at the origin, and our ball always rolls down into the valley and stays there. On the other hand, if the two eigenvalues are real and have opposite signs, the origin is a saddle point and the ball only settles down there if it starts from somewhere on a certain line (visualize this!).

In any case, S is always an n_s-dimensional subspace of $\mathbb{R}^n$. Actually, we can characterize this set in a very convenient way. It turns out that S is the subspace of $\mathbb{R}^n$ spanned by the stable eigenvectors (i.e. the eigenvectors associated with the stable eigenvalues). To see why, define

$$\Omega = \underset{n\times n_u}{\Omega_1} \quad \underset{n\times n_s}{\Omega_2} \tag{9.81}$$

so that Ω_2 is an $n \times n_s$ matrix of linearly independent eigenvectors associated with stable eigenvalues. Now, by the definition of an inverse,

$$\Omega^{-1}\Omega = I \tag{9.82}$$

and consequently

$$\begin{bmatrix} \Omega^1 \\ \Omega^2 \end{bmatrix} \begin{bmatrix} \Omega_1 & \Omega_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{9.83}$$

It follows that $\Omega^1\Omega_2 = \underset{n_s \times n_s}{0}$ and hence that any stable eigenvector is on the saddle path S. Indeed, let $y_0 = \Omega_2 h$ for any vector $h \in \mathbb{R}^{n_s}$. Then $\Omega^1 y_0 = 0$ and hence any linear combination of the column vectors of Ω_2 is on the saddle path. Moreover, since Ω^1 has full rank (it consists of linearly independent rows), if a vector is *not* a linear combination of the columns of Ω_2, then it is *not* on the saddle path. Consequently, S just *is* the column space of the matrix Ω_2. The result is the following.

Proposition 9.1.10 *Let S be the saddle path of (9.40). Let A be invertible and diagonalizable. Let Ω_2 be the matrix whose columns are linearly independent eigenvectors associated with the n_s stable eigenvalues of A. Then S is the column space of Ω_2, i.e.*

$$S = \{y_0 \in \mathbb{R}^n : y_0 = \Omega_2 h \ \text{ for some } h \in \mathbb{R}^{n_s}\}. \tag{9.84}$$

Proof. See above. ■

We now wrap up with a couple of definitions.

Definition 9.1.9 *Let S be the saddle path of (9.40). Let A be invertible so that the unique steady state is $y^* = 0$. If $S = \{0\}$ then (9.40) is unstable (not stable) and the steady state y^* is sometimes called an* unstable node *when the eigenvalues are real and an* unstable spiral *if at least one eigenvalue is complex. Conversely, if $S = \mathbb{R}^n$ then (9.40) is stable and the steady state y^* is sometimes called a* stable node *when the eigenvalues are real and a* stable spiral *if at least one eigenvalue is complex.*

In the intermediate case, the terminology is the following.

Definition 9.1.10 *Let S be the saddle path of (9.40). Let A be invertible so that the unique steady state is $y^* = 0$. Suppose $\{0\} \neq S \neq \mathbb{R}^n$. Then S will be called a* proper *saddle path of (9.40) and the steady state y^* will be called a* saddle point*. Sometimes we will be sloppy and refer to a proper saddle path as merely a saddle path.*

Remark 9.1.10 *According to our definition of stability, a saddle point is unstable. But, as we have seen, it is nevertheless possible to force the solution to be bounded or, if the real parts of the eigenvalues are all non-zero, convergent, by choosing the initial value in a suitable way.*

9.1.4.7 Two-dimensional case

The two-dimensional case is instructive since it is easy to draw, so we will spend a lot of time practicing on it.

In the two-dimensional case, it so happens that we can characterize the unstable case, the stable case and the (proper) saddle path case in terms of the trace and determinant of the matrix A. Let $\lambda(A) = \{\lambda_1, \lambda_2\}$. Recall that $\operatorname{tr}(A) = \lambda_1 + \lambda_2$ and $\det(A) = \lambda_1\lambda_2$. Note that if A has real entries, both the trace and the determinant are real numbers. We now have the following conditions for stability.

Proposition 9.1.11 *Let A be a 2×2 matrix with real entries. Suppose $\operatorname{tr}(A) < 0$ and $\det(A) > 0$. Then both eigenvalues of A have negative real parts, corresponding to a stable node.*

Proof. Let $\lambda(A) = \{\lambda_1, \lambda_2\}$. Since $\operatorname{tr}(A)$ is real, the imaginary parts of λ_1 and λ_2 sum to zero. So either they are both real or they are each others' complex

conjugates. Hence $\mathrm{tr}\,(A)$ is just the sum of the real parts of λ_1 and λ_2, and it is negative, so at least one of the eigenvalues has a negative real part. Now if both eigenvalues are real, then $\det(A)$ is just the product of the real parts, and since it is positive, the eigenvalues must have the same sign. Alternatively, if at least one of the eigenvalues is complex, then they are each others' complex conjugates. But then the real parts are the same, and hence, a fortiori, they have the same sign. So the real parts have the same sign, and at least one of them is negative. Hence they are both negative. ■

Proposition 9.1.12 *Let A be a 2×2 matrix with real entries. Suppose* $\det(A) < 0$. *Then the eigenvalues of A are real and have opposite signs, corresponding to a proper saddle path.*

Proof. Exercise. ■

Proposition 9.1.13 *Let A be a 2×2 matrix with real entries. Suppose* $\mathrm{tr}\,(A) > 0$ *and* $\det(A) > 0$. *Then both eigenvalues of A have positive real parts, corresponding to an unstable node.*

Proof. Exercise. ■

Applying our general result, we note that, in the 2×2 proper saddle path case, the saddle path is just a line through the origin in the direction of a stable eigenvector.

Exercise 9.1.7 *Show that the saddle path S of a homogeneous linear system of ODEs really is a subspace, i.e. that for any $x, y \in S$ and any scalar α, we have $\alpha(x+y) \in S$.*

9.1.4.8 When A is not diagonalizable

When the A matrix is not diagonalizable, we can still use the Schur form. The details of this approach can be found in the section on linear systems of difference equations.

9.1.4.9 When A is singular

When A is singular, it has a non-trivial nullspace and hence there are inifinitely many steady states. But we can still solve the system. The solution is

$$x(t) = \exp(At) x_0 + \int_0^t \exp(As) b ds. \tag{9.85}$$

9.1.5 Reducing a pth order system to a first-order one

Occasionally we want to solve homogeneous linear systems of higher than first order, i.e. systems of the form

$$x^{(p)}(t) = \sum_{k=0}^{p-1} A_k x^{(k)}(t). \tag{9.86}$$

where $x^{(k)}$ is the kth derivative of $x(t)$ with respect to t. This $n-$dimensional pth order system can be reduced to an $np-$dimensional first-order system in the following way. Define

$$\widetilde{x}(t) = \begin{bmatrix} x^{(p-1)}(t) \\ x^{(p-2)}(t) \\ \vdots \\ x(t) \end{bmatrix}. \tag{9.87}$$

We now write down an $np \times np$ first-order linear system for the $np-$dimensional vector function $\widetilde{x}(t)$ The first row will state our original system; the others will state the relationships between the elements in $\widetilde{x}(t)$ and $\dot{\widetilde{x}}(t)$ given our definition

of $\widetilde{x}(t)$. We have

$$\begin{bmatrix} x^{(p)}(t) \\ x^{(p-1)}(t) \\ \vdots \\ \dot{x}(t) \end{bmatrix} = \begin{bmatrix} A_{p-1} & A_{p-2} & \cdots & A_0 \\ I & & & \\ & \ddots & & \\ & & I & \end{bmatrix} \begin{bmatrix} x^{(p-1)}(t) \\ x^{(p-2)}(t) \\ \vdots \\ x(t) \end{bmatrix}. \tag{9.88}$$

Defining $\widetilde{A}$ in the obvious way, we have

$$\dot{\tilde{x}}(t) = \widetilde{A}\widetilde{x}(t). \tag{9.89}$$

Note that the space of solutions has dimension np.

9.1.5.1 A quick trick

In order to arrive faster at the solution, we may proceed as follows. Suppose $x(t)$ is scalar. It can then be shown that the solution has the form

$$x(t) = \sum_{k=1}^{n} c_k e^{\lambda_k t}. \tag{9.90}$$

Indeed, any function $x(t) = e^{\lambda t}$ is a solution provided λ is chosen among the relevant eigenvalues. In order to characterize these eigenvalues, we plug $x(t) = e^{\lambda t}$ this into (9.86) with $A_k = a_k$ and get

$$\lambda^p e^{\lambda t} = \sum_{k=0}^{p-1} a_k \lambda^k e^{\lambda t}. \tag{9.91}$$

Dividing by $e^{\lambda t}$, we find that λ must satisfy the polynomial equation

$$\lambda^p - \sum_{k=0}^{p-1} a_k \lambda^k = 0. \tag{9.92}$$

Example 9.1.7 *Consider*

$$\ddot{x}(t) - x(t) = 0. \tag{9.93}$$

Plugging in the candidate solution $x(t) = e^{\lambda t}$, *we get*

$$\lambda^2 - 1 = 0. \tag{9.94}$$

The two solutions are $\lambda_1 = 1$ *and* $\lambda_2 = -1$. *Hence the entire set of solutions can be written as*

$$x(t) = c_1 e^t + c_2 e^{-t}. \tag{9.95}$$

9.1.6 Lyapunov's theorem

Having dealt with linear systems at some length, we can now state and motivate a theorem that characterizes the local dynamics of the non-linear system (9.37).

Theorem 9.1.4 (Lyapunov) *Consider the dynamic system (9.37). Let* f *be differentiable at the steady state* x^* *and define*

$$A = f'(x^*). \tag{9.96}$$

Then x^* *is asymptotically stable if the eigenvalues of* A *are all stable (have negative real parts).*

Similarly, if $f'(x^*)$ has both stable and unstable eigenvalues, there is a sense in which the dynamics of the non-linear system is locally well approximated by a linear system with a proper saddle path. We won't be precise about this, but the idea is that there will be a neighborhood O of x^* in which the set of initial points $x_0 \in O$ such that the system converges to x^* is an n_s-dimensional surface whose tangent at x^* is the hyperplane spanned by the stable eigenvectors of $f'(x^*)$.

In the two-dimensional case, this n_s-dimensional surface is of course just a curve in $\mathbb{R}^2$.

9.1.7 Phase diagrams

The point of phase diagrams is to use pictures to enable us to visualize the dynamics of solution paths (called *trajectories*) to systems of ODEs. Note that there is one trajectory $x(t)$ for each initial value x_0.

9.1.7.1 One-dimensional case

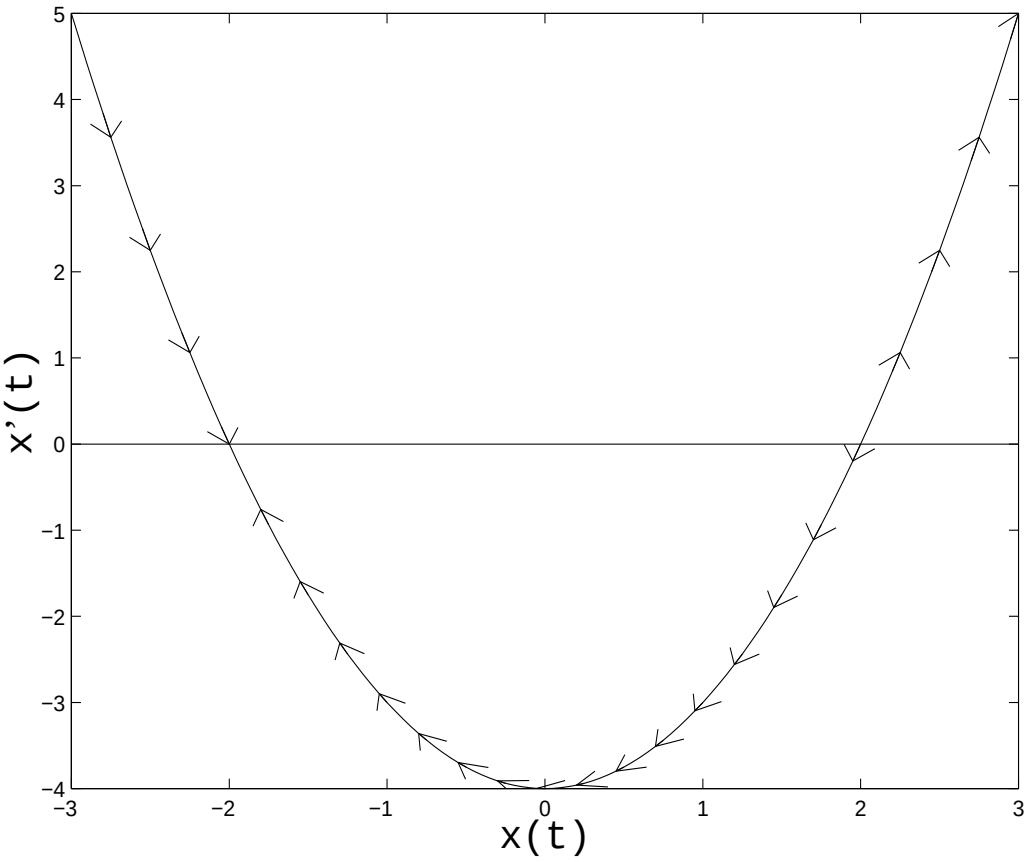
Example 9.1.8 *Consider the scalar ODE*

$$\dot{x}(t) = x^2(t) - 4. \tag{9.97}$$

Apparently, there are two steady states, $x_1^* = -2$ *and* $x_2^* = 2$. *Are any of them stable or asymptotically stable? This is most easily investigated not by using*

Lyapunov's theorem but by looking at a phase diagram. See Figure 9.1.7.1.

Figure 9.1.7.1



9.1.7.2 Two-dimensional case

Step 1. Characterize the steady state. Typically, there will be two equations which together determine the steady state, and the steady state will be the point of intersection between two curves. The first curve is the locus of points in the $x - y$ plane such that $\dot{x} = 0$ and the second is the locus of points in the $x - y$ plane such that $\dot{y} = 0$. Draw these loci.

Step 2. Check what happens to the rate of change of the two variables when you are not on either of these loci. Check the four different areas separated by the two loci. If the rates of change are determinate in sign, draw arrows indicating

the sign of the rate of change in each area (the options are northeast, northwest, southeast and southwest).

Step 3. If your system was not linear to begin with, linearize around the steady state. Use the trace and determinant conditions to check if you are dealing with a stable node, an unstable node, a proper saddle path, an unstable spiral or a stable spiral. See if this agrees with the arrows you've drawn.

Step 4. Draw asymptotes/saddle paths, i.e. the paths spanned by the eigenvectors. [Explain in more detail.]

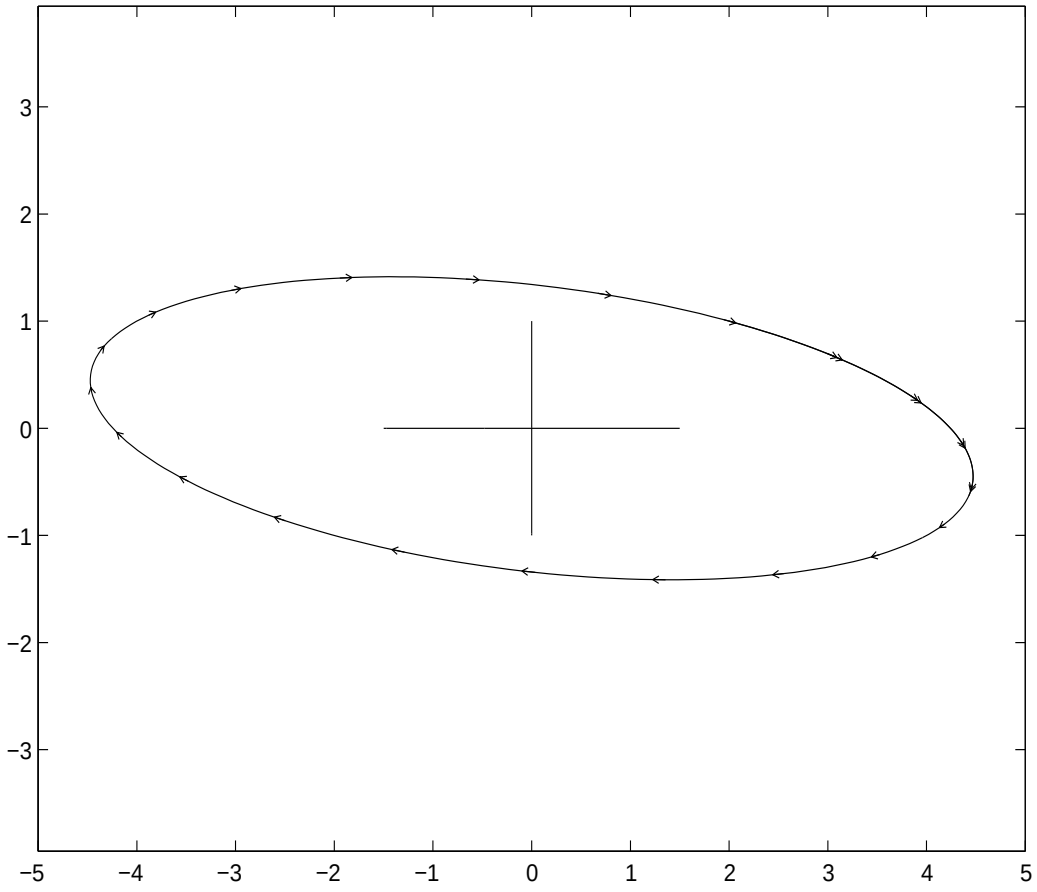
Step 4. Draw trajectories, i.e. solution paths for different initial values. Note that the trajectories cross the $\dot{x} = 0$ locus vertically and the $\dot{y} = 0$ locus horizontally. In the proper saddle path case, draw a saddle path. Draw trajectories on and off the saddle path with arrows to indicate in what direction the rate of change is as t increases. Note that (why?) if we begin on the saddle path or an asymptote, we never leave it.

Example 9.1.9 *(stable spiral). Consider the* $2-$*dimensional homogeneous linear system*

$$\begin{cases} \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} -0.1 & 1 \\ -1 & -0.1 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \end{cases} \tag{9.98}$$

Apparently the unique steady state is the origin. Looking at the eigenvalues, we see that they are stable and complex. So the solution trajectory is a convergent spiral. Figure 9.1.7.2.a depicts a single trajectory starting at $(1, 0)$*. In this case, it is as if we are going round an ever shrinking circle. For many other values of the coefficients in A, we go around a shrinking or growing* ellipse*.*

Figure 9.1.7.2.a.

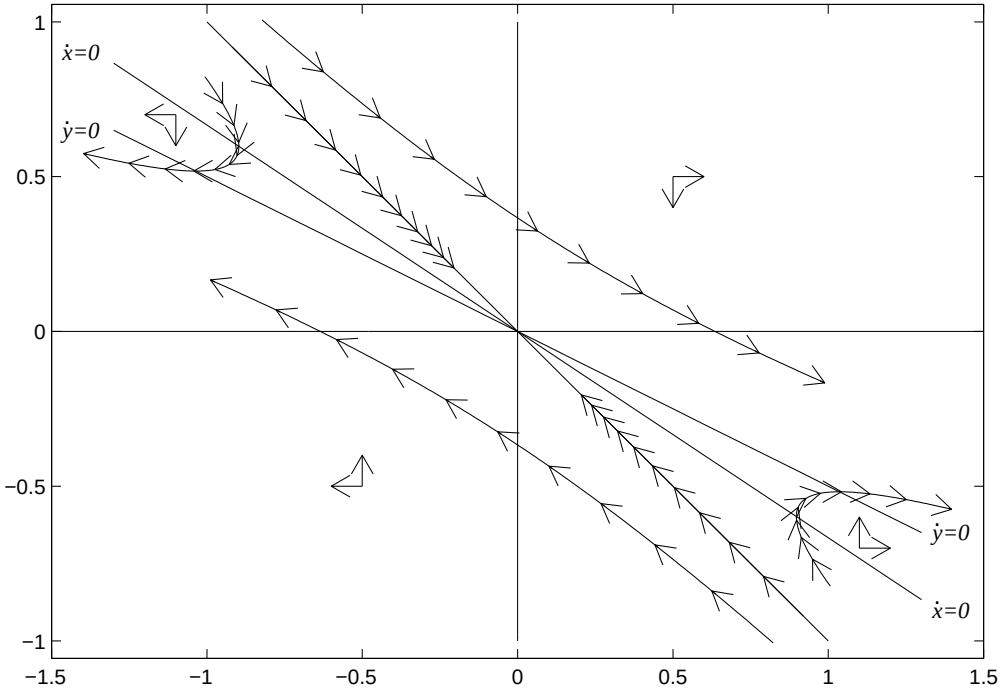


Example 9.1.10 *(proper saddle path). Consider the 2-dimensional homogeneous linear system*

$$\begin{cases} \begin{bmatrix} \dot{x}(t) \\ \dot{y}(t) \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \\ \\ \begin{bmatrix} x(0) \\ y(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \end{cases} \tag{9.99}$$

We find that the eigenvalues are real and of opposite signs, so we have a proper saddle path. Figure 9.1.7.2.b contains a phase diagram for the system. It contains everything that you are required to do when asked to draw a phase diagram.

Figure 9.1.7.2.b



9.2 Difference equations

9.2.1 Definition of problem and existence of solution

In this section time is discrete, i.e. $t = 0, 1, ...$ Generally speaking, we would like to be able to solve systems of the form $f(t, x_t, x_{t+1}) = 0$; $t = 0, 1, ...$ Ultimately it would be very nice indeed to be able to deal with such systems in general. But as an introduction, we will be considering time-invariant dynamic systems of the form

$$\begin{cases} x_{t+1} = f(x_t, z_t) \\ x_0 \ \text{ given.} \end{cases} \tag{9.100}$$

where $\langle z_t \rangle_{t=0}^{\infty}$ is an exogenously given sequence.. Existence and uniqueness of the solution is clearly no problem. Just fix x_0 and the given sequence $\langle z_t \rangle_{t=0}^{\infty}$ and keep on applying the function f. Indeed, this is a *constructive* proof so there seems little else to say about difference equations. Well, actually, there is more. Just as in the continuous time case, we can talk about and characterize stability. And in the linear case we will be able to solve for x_t in terms of t and the exogenous sequence $\langle z_t \rangle$. Nevertheless, for numerical purposes, it is often convenient to stick to the iterative technique of just applying f repeatedly so we will recommend it whenever appropriate.

We will not be exploring the non-linear case in any detail. So if you have a non-linear system and want to apply the methods of this section, linearize around a steady state and investigate the dynamics of the resulting linear system. Note that with an exogenous sequence $\langle z_t \rangle_{t=0}^{\infty}$, the notion of a steady state is typically not well-defined. However, if $z_t \equiv \overline{z}$, then the steady state x^* solves

$$x = f(x, \overline{z}). \tag{9.101}$$

9.2.2 Scalar linear difference equations with an exogenous driving sequence

Consider

$$x_{t+1} = ax_t + z_t; \quad t = 0, 1, 2, ... \tag{9.102}$$

where $\langle z_t \rangle_{t=0}^{\infty}$ is a given sequence.

However, let's suppose x_0 is not given so that we have a whole 1-dimensional space of solutions. Suppose also for the sake of argument that $a \neq 0$ (we need this assumption during the derivations but can drop it at the end). To find a representation of the entire class of solutions, we use the following ingenious trick. Introduce a new sequence $\langle w_t \rangle$ defined via

$$x_t = a^t w_t. \tag{9.103}$$

(This defines w_t uniquely since $a \neq 0$.) It follows that

$$a^{t+1} w_{t+1} = a^{t+1} w_t + z_t. \tag{9.104}$$

Hence

$$w_{t+1} - w_t = a^{-t-1} z_t. \tag{9.105}$$

Now sum both sides from 0 to $t-1$. We get

$$\sum_{k=0}^{t-1} (w_{k+1} - w_k) = \sum_{k=0}^{t-1} a^{-k-1} z_k \tag{9.106}$$

Now notice that

$$\sum_{k=0}^{t-1} (w_{k+1} - w_k) = w_t - w_{t-1} + w_{t-1} - w_{t-2} + \cdots - w_0 = w_t - w_0. \tag{9.107}$$

Hence

$$w_t = w_0 + \sum_{k=0}^{t-1} a^{-k-1} z_k \tag{9.108}$$

and it follows that (think about the last step!)

$$x_t = a^t x_0 + a^t \sum_{k=0}^{t-1} a^{-k-1} z_k = a^t x_0 + \sum_{k=0}^{t-1} a^k z_{t-k-1} \tag{9.109}$$

and note that this represents the whole class of solutions as x_0 ranges over all the real numbers. For this to make sense when $t = 0$, introduce the convention

$$\sum_{k=0}^{-1} h_k = \sum_{k=1}^{0} h_k = 0, \tag{9.110}$$

and for it to make sense (and give the right answer!) when $a = 0$, put $0^0 = 1$. We call (9.109) the finite sum representation of the class of all solutions.

We now investigate stability. First we define it.

Definition 9.2.1 *A sequence of complex numbers $\langle x_t \rangle_{t=0}^{\infty}$ is said to be* stable *if it is bounded, i.e. if there is an $M \geq 0$ such that $|x_t| \leq M$ for each $t = 0, 1, ...$*

When $|a| < 1$ the prospects for the existence of a stable solution seem particularly good. So suppose $|a| < 1$ and let $\langle z_t \rangle_{t=0}^{\infty}$ be bounded. Is this enough to guarantee that $\langle x_t \rangle_{t=0}^{\infty}$ is stable? Let's see if we prove that by constructing an upper bound for $|x_t|$. Let M be an upper bound for $|z_t|$. Then

$$\begin{aligned} |x_t| = \left| a^t x_0 + \sum_{k=0}^{t-1} a^k z_{t-k-1} \right| \leq |a|^t |x_0| + \sum_{k=0}^{t-1} |a|^k |z_{t-k-1}| \leq \\ \leq |x_0| + \sum_{k=0}^{t-1} |a|^k M \leq |x_0| + \sum_{k=0}^{\infty} |a|^k M = |x_0| + \frac{1}{1-|a|} M \end{aligned} \tag{9.111}$$

Thus if $|a| < 1$, *any* solution to (9.102) is bounded. So there is a one-dimensional set of stable solutions.

On the other hand, suppose $|a| > 1$ but maintain the assumption that $\langle z_t \rangle$ is bounded. Then $\langle x_t \rangle$ will fail to explode geometrically only for a very special value of x_0 (the saddle space is zero-dimensional and hence contains a single point

only). The easiest way to see what that value is is to represent the entire class of solutions in a new way - the forward representation. The idea is to go back to the stage when we noted that

$$w_t = w_0 + \sum_{k=0}^{t-1} a^{-k-1} z_k \tag{9.112}$$

where w_0 was an arbitrary number. Since it is arbitrary, we can add any constant to the representation and still represent the same class of solutions. So let's add $-\sum_{k=0}^{\infty} a^{-k-1} z_k$ which we have just assumed is a convergent sum and hence a real number. We get

$$w_t = c - \sum_{k=t}^{\infty} a^{-k-1} z_k \tag{9.113}$$

where c is a new arbitrary constant. We can now derive the so-called forward representation of the class of solutions as

$$x_t = a^t c - a^t \sum_{k=t}^{\infty} a^{-k-1} z_k = a^t c - \sum_{k=0}^{\infty} a^{-k-1} z_{t+k}. \tag{9.114}$$

Now if $|a| > 1$ it is easy to see that to avoid $\langle x_t \rangle$ being unstable, we must set $c = 0$ in which case

$$x_0 = -\sum_{k=0}^{\infty} a^{-k-1} z_k \tag{9.115}$$

and, more generally,

$$x_t = -\sum_{k=0}^{\infty} a^{-k-1} z_{t+k}. \tag{9.116}$$

We may confirm that this defines a stable solution if $\langle z_t \rangle$ is stable. The confirmation that (9.116) defines a solution is left to the reader. That it is stable follows from the following argument.

Let M be an upper bound for $\langle z_t \rangle_{t=0}^{\infty}$. Then

$$|x_t| = \left| \sum_{k=0}^{\infty} a^{-k-1} z_{t+k} \right| \leq \sum_{k=0}^{\infty} |a|^{-k-1} |z_{t+k}| \leq \sum_{k=0}^{\infty} |a|^{-k-1} M = \frac{1}{|a|-1} M.$$

Notice that there is a unique stable solution when $|a| > 1$; in this case the stability requirement replaces fixing the initial value in pinning down the solution. Notice also that in this case x_t depends only on the 'future' of $\langle z_t \rangle$. It is because economists (1) require non-explosiveness and (2) consider equations with $|a| > 1$ (we will see later why we so often do!) that we think that (expectations of) the future are important!

Exercise 9.2.1 *Consider*

$$x_{t+1} = 2x_t + t\rho^t \tag{9.117}$$

where $|\rho| < 1$. *Find the entire one-dimensional space of solutions. What is the unique stable solution? Hint: When* $|\rho| < 1$, $\sum_{t=0}^{\infty} t\rho^t = \frac{\rho}{(1-\rho)^2}$.

9.2.3 Sargent's metric space approach to scalar linear difference equations

[Omitted in this version.]

9.2.4 First-order linear systems with constant coefficients

Consider the first-order homogeneous linear system

$$\begin{cases} y_{t+1} = Ay_t \\ y_0 \text{ given.} \end{cases} \tag{9.118}$$

In this section, we study the solutions to homogeneous systems. To a very large extent, we will be able to copy the results form the corresponding section on homogeneous linear systems of ODEs, and, just as for ODEs, the extension to systems $x_{t+1} = Ax_t + b$ with a unique steady state is trivial. Note that in this case the steady state is $x^* = (I - A)^{-1} b$.

The main difference as compared with the continuous time case will be the definition of a stable eigenvalue. In discrete time, an eigenvalue λ is called stable if $|\lambda| < 1$, borderline stable if $|\lambda| \leq 1$ and unstable if it is not stable or borderline stable. We will see (if you have not already guessed) why this definition is appropriate.

Not surprisingly, the unique solution (9.118) is given by

$$y_t = A^t y_0 \tag{9.119}$$

and as before we uncouple the dynamics by factorizing A. So suppose A is diagonalizable, and let $A = \Omega\Lambda\Omega^{-1}$ be an eigenvalue/eigenvector decomposition of A. Note that $A^t = \Omega\Lambda^t\Omega^{-1}$ and that

$$\Lambda^t = \begin{bmatrix} \lambda_1^t & & & \\ & \lambda_2^t & & \\ & & \ddots & \\ & & & \lambda_n^t \end{bmatrix} \tag{9.120}$$

Hence our solution can be written as

$$y_t = \sum_{k=1}^{n} c_k \lambda_i^t s_k \tag{9.121}$$

where $c = \Omega^{-1} y_0 \in \mathbb{C}^n$ is arbitrary and $\Omega = \begin{bmatrix} s_1 & s_2 & \cdots & s_n \end{bmatrix}$.

To see why the stability of the solution is governed by whether the eigenvalues λ_k are inside or outside the unit circle (i.e. whether $|\lambda_k| \leq 1$ or not), we note that (why?) $\left|\lambda_k^t\right| = |\lambda_k|^t$. We can now conclude that the solution converges to the origin for any initial value y_0 if all the eigenvalues are strictly inside the unit circle in the complex plane. Even if not, there will be a set of initial values from which the solutions do converge, and again we will call this the saddle path. And again it turns out to be the case that the saddle path is spanned by the stable eigenvectors.

Sometimes we will want to use the requirement of (borderline) stability of the solution in conjunction with initial values to pin down a unique (borderline) stable solution. The idea is that stability and initial values together will impose n linear restrictions on the solution so that a single one is picked out. This idea is carried out in section 9.2.5.

9.2.4.1 Complex eigenvalues and real oscillatory solutions

For some purposes, we want to represent the whole class of solutions as an arbitrary real-linear combination of real-valued functions. Note that this is of little practical importance, since if the coefficient matrix A and the initial vector y_0 has real entries, then (why?) the solution remains real as time goes on. But the project still has relevance in that it enhances the conceptual understanding of the oscillatory (non-monotone) behavior introduced by negative and complex eigenvalues.

Again we note that, if A has real entries, the complex eigenvalues (those with a non-zero imaginary part) and eigenvectors appear in complex conjugate pairs. (For convenience, arrange the complex eigenvalues so that $\lambda_{k+1} = \overline{\lambda}_k$.) Actually, the real but negative eigenvalues also give rise to oscillatory behavior, so we will need a treatment of them as well. In what follows it will be important to be able to rewrite a complex number z from Cartesian form $x + iy$ to polar form $re^{i\theta}$. If you are unsure of how to do this, read section 3.2.5 carefully. Note that oscillatory behavior arises as soon as the argument θ is non-zero, which includes the case of z being real but negative.

Let's begin with the real negative eigenvalues. Let the eigenvalue λ_k be real and negative and let s_k be the corresponding eigenvector, which (why?) is also real since A is. Then (why?) $\lambda_k = r_k e^{i\pi} = r_k \cos \pi t$ and consequently $\lambda_k^t =$

$r_k^t \cos \pi t = r_k^t (-1)^t$. Note that it is crucial here that $t = 0, 1, 2, ...$ is an integer. So if λ_k is real and negative, one of the real-valued basis functions spanning the space of real-valued solutions will evidently be

$$\varphi_{k,t} = r_k^t \cos \pi t = r_k^t (-1)^t s_k.$$

Now consider the complex eigenvalues. Let $\lambda_k = r_k e^{i\theta_k}$ be complex. Then the corresponding complex-valued basis function is, as we know,

$$\psi_{k,t} = \lambda_k^t s_k. \tag{9.122}$$

Writing $s_k = u_k + iv_k$ and using Euler's formula, we find that the real and imaginary parts of this basis solution are

$$\varphi_{k,t} = r_k^t [\cos(\theta_k t) u_k - \sin(\theta_k t) v_k] \tag{9.123}$$

and

$$\varphi_{k+1,t} = r_k^t [\sin(\theta_k t) u_k + \cos(\theta_k t) v_k]. \tag{9.124}$$

and these functions are also (why?) solutions of our linear system, and in fact they are linearly independent so we can use them as basis functions for our set of real-valued solutions. We can ignore (why?) the basis solutions $\overline{\lambda}_k^t \overline{s}_k$.

Now construct n basis functions in this way. When λ_k is real and positive, just set $\varphi_{k,t} = \lambda_k^t s_k$. When it is negative, set $\varphi_{k,t} = r_k^t (-1)^t s_k$. Finally, when λ_k is complex, set $\varphi_{k,t} = r_k^t [\cos(\theta_k t) u_k - \sin(\theta_k t) v_k]$ and $\varphi_{k+1,t} [\sin(\theta_k t) u_k + \cos(\theta_k t) v_k]$. Ignore $\lambda_{k+1} = \overline{\lambda}_k$. We can now write any solution as

$$y_t = \sum_{k=1}^{n} c_k \varphi_{k,t} \tag{9.125}$$

where the real-valued functions φ_k span the set of real-valued solutions to $y_{t+1} = Ay_t$.

Exercise 9.2.2 *Consider the* 2×2 *system*

$$\begin{bmatrix} x_{1,t+1} \\ x_{2,t+1} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{4} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix}. \tag{9.126}$$

Find a basis $\{\varphi_{1,t}, \varphi_{2,t}\}$ *of real-valued functions for the real-valued solutions of this system.*

9.2.5 Linear systems with constant coefficients and an exogenous driving sequence

Consider the $n-$dimensional system

$$x_{t+1} = Ax_t + z_t. \tag{9.127}$$

It is not hard to believe that the class of all solutions can be represented as

$$x_t = A^t x_0 + \sum_{k=0}^{t-1} A^k z_{t-k-1}. \tag{9.128}$$

We now have several options if we want a unique solution. Either fix all elements of x_0 or fix some elements of x_0 and some elements of, say, x_T, making sure that you have a total of n fixed values. Another alternative, which it will take the remainder of this section to sort out in detail, is when we impose some initial values $x_{k,0}$ and the requirement that the solution *not explode* as $t \to \infty$ (we will define what this means exactly below). In auspicious circumstances, this restriction on the behavior of x at infinity can play the same role as the fixing of endpoint values $x_{k,T}$.

Just as before, we now factorize A to uncouple the dynamics and analyze stability. Suppose for simplicity that A is diagonalizable with eigenvalue/eigenvector factorization $A = \Omega\Lambda\Omega^{-1}$. Then $\Omega^{-1}A = \Lambda\Omega^{-1}$. To avoid cluttering the notation,

we now introduce the auxiliary sequence

$$y_t = \Omega^{-1} x_t \tag{9.129}$$

and notice that if we solve for y_t we have solved for x_t since Ω^{-1} is invertible. So as to facilitate the analysis of stability, arrange Λ and Ω so that the unstable eigenvalues and associated eigenvectors come first. Now premultiply our system by Ω^{-1} which creates an equivalent system since Ω is invertible. We get

$$\Omega^{-1} x_{t+1} = \Lambda \Omega^{-1} x_t + \Omega^{-1} z_t \tag{9.130}$$

and hence

$$y_{t+1} = \Lambda y_t + \Omega^{-1} z_t. \tag{9.131}$$

Notice that this is a diagonal system! In principle, we can solve this row by row as a scalar equation! But we won't do that here. Instead we will proceed block by block, trying to pick out the stable (bounded) solution(s) as we go along.

Assuming that $\langle z_t \rangle$ is stable, it is not hard to believe that the solution $x_t = \Omega y_t$ is stable for any x_0 provided all the eigenvalues of A are stable or borderline stable, i.e. $|\lambda_k| \leq 1$ for all $k = 1, 2, ..., n$. More generally, we have eigenvalues on both sides of the unit circle in the complex plane, i.e. we have a saddle path. To see what restrictions we need to impose to ensure a stable solution, we partition the system according to the absolute magnitude of the eigenvalues, putting the unstable (not borderline stable) ones first. (Notice that this requires that Ω to be arranged so that the unstable eigenvectors appear first also.)

$$\begin{bmatrix} y_{1,t+1} \\ y_{2,t+1} \end{bmatrix} = \begin{bmatrix} \underset{n_u \times n_u}{\Lambda_1} & \\ & \underset{n_s \times n_s}{\Lambda_2} \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} + \Omega^{-1} z_t \tag{9.132}$$

Consider the first row and solve forward for $y_{1,t}$. To ensure stability, we must set

$$y_{1,t} = -\sum_{k=0}^{\infty} \Lambda_1^{-k-1} \Omega^{-1} z_{t+k}. \tag{9.133}$$

As in the continuous time case, the stability condition imposes n_u linear restrictions on the solution, where n_u is the number of unstable eigenvalues. Let $n - n_u = n_s$ be the number of stable (or borderline stable) eigenvalues. We are now left with an n_s-dimensional space of stable solutions - the saddle space.

To reduce the solution to a singleton set, we need to impose another n_s independent restrictions. Ideally, we would like this to be done by fixing initial values of a certain predefined subset of the variables in x_t. So let's suppose that x_t is arranged in such a way that the initial values of the last n_k variables have given initial values $x_{2,0} \in \mathbb{R}^{n_k}$. A necessary condition for these initial values together with the stability condition to pin down exactly one solution is that $n_k = n_s$, i.e. that the number of variables with given initial values equal the number of stable eigenvalues. If $n_k > n_s$ there is no stable solution, and if $n_k < n_s$ there are infinitely many.

Now if we were assigning initial values to variables in y_t, or could assign initial values to any elements of x_t (not necessarily in the predefined subset), $n_k = n_s$ would be sufficient for there to be a unique stable solution. But it is more reasonable to suppose that what we have fixed is initial values for certain specific elements of x_t. For these initial values of x_t to translate into the required initial values of y_t we need another condition. To see what we need we will just go boldly ahead and try to solve for x_t. We will need the definitional equation for y_t which in partitioned form says

$$\begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} = \begin{bmatrix} \underset{n_u \times n_u}{\Omega_{11}} & \underset{n_u \times n_s}{\Omega_{12}} \\ \underset{n_s \times n_u}{\Omega_{21}} & \underset{n_s \times n_s}{\Omega_{22}} \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}. \tag{9.134}$$

The aim now will be to express $x_{1,t}$ in terms of $y_{1,t}$ (which we have just solved for) and $x_{2,t}$ (which we will be able to solve for recursively; see below). Assume that Ω_{22} is invertible (this assumption cannot be dropped, see exercise 9.2.3). Then

manipulation of the definition of x_t and y_t yields

$$x_{1,t} = \left[\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right] y_{1,t} + \Omega_{12}\Omega_{22}^{-1} x_{2t}. \tag{9.135}$$

Recall our original system and partition it. We have

$$\begin{bmatrix} x_{1,t+1} \\ x_{2,t+1} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} + \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix}. \tag{9.136}$$

We will now construct the solution recursively. We won't get a nice formula for the solution, but we will have a useful algorithm for simulating the solution on a computer.

Recall that we are given $x_{2,0}$ and $\langle z_t \rangle_{t=0}^{\infty}$ and have calculated $y_{1,t}$ in terms of $\langle z_t \rangle_{t=0}^{\infty}$. So take $x_{2,0}$ and use (9.135) to calculate $x_{1,0}$. Then use the second row of (9.136) to calculate $x_{2,1}$. Then use (9.135) again to calculate $x_{1,1}$. And so on.

Exercise 9.2.3 *Consider the two-dimensional system*

$$\begin{cases} \begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \\ \\ y_0 \text{ given.} \end{cases} \tag{9.137}$$

1. *There is no stable solution to this system. Why not?*

2. *Suppose instead that x_0 is given but that y_0 is free. Write down the unique stable solution.*

Exercise 9.2.4 (difficult) *Show that the solution presented in this section really is stable provided that $\langle z_t \rangle$ is.*

9.2.6 What to do when A is not diagonalizable

[Omitted in this version. The idea is to use the Schur form.]

9.2.7 Reducing a pth order system to a first order system

Occasionally we want to solve linear systems of higher than first order, i.e. systems of the form

$$x_{t+p} = \sum_{k=0}^{p-1} A_k x_{t+k} + z_t. \tag{9.138}$$

This $n-$dimensional pth order system can be reduced to an np-dimensional first-order system in the following way. Define

$$\widetilde{x}_t = \begin{bmatrix} x_{t+p-1} \\ x_{t+p-2} \\ \vdots \\ x_t \end{bmatrix}. \tag{9.139}$$

We now write down an $np-$dimensional first-order linear system for the $np-$dimensional sequence $\widetilde{x}_t$ The first row will state our original system; the others will state the relationships between the elements in $\widetilde{x}_t$ and $\widetilde{x}_{t+1}$ given our definition of $\widetilde{x}_t$. We have

$$\begin{bmatrix} x_{t+p} \\ x_{t+p-1} \\ \vdots \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} A_{p-1} & A_{p-2} & \cdots & A_0 \\ I & & & \\ & \ddots & & \\ & & I & \end{bmatrix} \begin{bmatrix} x_{t+p-1} \\ x_{t+p-2} \\ \vdots \\ x_t \end{bmatrix} + \begin{bmatrix} z_t \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{9.140}$$

Defining $\widetilde{A}$ and $\widetilde{z}_t$ in the obvious way, we have

$$\widetilde{x}_{t+1} = \widetilde{A}\widetilde{x}_t + \widetilde{z}_t. \tag{9.141}$$

Note that the space of solutions has dimension np.

9.2.7.1 A quick trick

In order to arrive faster at the solution, we may proceed as follows. Suppose x_t is scalar and that $z_t = 0$. (If z_t is a constant this does not involve much loss of

generality since one can usually analyze the deviations from the steady state.) It can then be shown that the solution has the form

$$x_t = \sum_{k=1}^{n} c_k \lambda_k^t. \tag{9.142}$$

Indeed, any function $x_t = \lambda^t$ is a solution provided λ is chosen among the relevant eigenvalues. In order to characterize these eigenvalues, we plug $x_t = \lambda^t$ this into (9.138) with $A_k = a_k$ and get

$$\lambda^{t+p} = \sum_{k=0}^{p-1} a_k \lambda^{t+k}. \tag{9.143}$$

Dividing by λ^t, we find that λ must satisfy the polynomial equation

$$\lambda^p - \sum_{k=0}^{p-1} a_k \lambda^k = 0. \tag{9.144}$$

Example 9.2.1 *Consider*

$$x_{t+2} - x_t = 0. \tag{9.145}$$

Plugging in the candidate solution $x_t = \lambda^t$, *we get*

$$\lambda^2 - 1 = 0. \tag{9.146}$$

The two solutions are $\lambda_1 = 1$ *and* $\lambda_2 = -1$. *Hence the entire set of solutions can be written as*

$$x(t) = c_1 + c_2 (-1)^t. \tag{9.147}$$

9.2.8 Expectational difference equations.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\boldsymbol{\xi}$ be a white noise process relative to its natural filtration $\mathcal{F}_t = \sigma\{\xi_s; s \leq t\}$. Now consider the expectational difference equation

$$E[x_{t+1} | \mathcal{F}_t] = Ax_t. \tag{9.148}$$

If x_0 were given exogenously and the expectations error $x_{t+1} - E[x_{t+1}|\mathcal{F}_t]$ were also given exogenously, then we would immediately have a recursive representation of the unique solution. This solution would be stable if A were a stable matrix.

More generally, partition x_t via

$$x_t = \begin{bmatrix} \lambda_t \\ k_t \\ {\scriptstyle n_k \times 1} \end{bmatrix} \tag{9.149}$$

where k_0 is given exogenously and the expectations error $k_{t+1} - E[k_{t+1}|\mathcal{F}_t] = \xi_{t+1}$ is also given exogenously. Corresponding to these assumptions, we need A to have at least n_k stable eigenvalues for there to be a stable solution. If A has precisely n_k stable eigenvalues, there is usually a unique stable solution (the exception is the pathological case in Exercise 9.2.3). So suppose that A has precisely n_k stable eigenvalues.

To find the unique solution, we use the Schur form of the matrix A. What we need is a unitary matrix Q and a lower triangular matrix T with diagonal entries of descending modulus such that $QA = TQ$. Having found these matrices, partition Q and T conformably with the partition of x_t. Do the same with Q^H, the Hermitian transpose of Q. Suppose the bottom right-hand block of Q^H is invertible. Call that bottom right-hand block Q_{22}^H. Define Q_{12}^H similarly. (Note the abuse of notation; Q_{22}^H is *not* the Hermitian transpose of the bottom right-hand block of Q). Then the unique stable solution is given recursively by

$$\lambda_t = Q_{12}^H \left(Q_{22}^H\right)^{-1} k_t \tag{9.150}$$

$$k_{t+1} = Q_{22}^H T_{22} \left(Q_{22}^H\right)^{-1} k_t + \xi_{t+1.} \tag{9.151}$$

Exercise 9.2.5 *Derive the result in this section.*

9.2.8.1 Singular difference equations

[Omitted in this version. See [29].]

Chapter 10

Dynamic optimization

10.1 Introductory remarks

The word 'dynamic' indicates that we are concerned with time, and the variable t below will indeed be thought of as time, reflecting the macro bias of the author. But the apparatus we will be looking at can be used more generally, e.g. in contract theory where we optimize over an outcome space rather than over time. For examples of this, see [43].

10.2 Continuous time

10.2.1 Introduction and statement of problem

You will recall from MatFU I that when we have a problem of the following type (where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^k$)

$$\begin{aligned} &\max_x f(x) \\ &\text{s.t. } g(x) = 0 \end{aligned} \tag{10.1}$$

then we have something like the following theorem.

Theorem 10.2.1 *Define*

$$L(x,\lambda) = f(x) + \lambda^T g(x) \tag{10.2}$$

and let x^* solve (10.1). (Also, assume some regularity conditions.) Then there exists a $\lambda^* \in \mathbb{R}^n$ such that

$$\frac{\partial L(x^*,\lambda^*)}{\partial x} = 0 \tag{10.3}$$

Remark 10.2.1 *If f and each g_i are concave and the λ_i^* are positive, then, x^* maximizes L. This will be interesting to us later.*

Now consider a more difficult problem.

$$\max_{\mathbf{u}} \int_0^T f(t, x(t), u(t))\, dt$$

$$\text{s.t.} \begin{cases} \dot{x}(t) = g(t, x(t), u(t)) \\ x(0) = a \\ x(T) = b \\ u(t) \in U \subset \mathbb{R}^k \text{ for all } t \in [0,T] \end{cases} \tag{10.4}$$

Remark 10.2.2 *The initial and final instants 0 and T are fixed. In engineering applications, these are sometimes free variables. E.g. take me to the moon using a minimal amount of fuel, arrival time up to you. Rare in economics.*

Again it is worth pondering a while over what we have just written down. We want to maximize a function with respect to a whole trajectory $\mathbf{u} : [0,T] \to \mathbb{R}$, i.e. with respect to another function. So a solution to the problem will be an entire function rather than just a number or even a vector. In an abstract sense, though, $\mathbf{u}$ may be thought of as a vector with components $u(t)$. But there are clearly infinitely many such components, and that is why what we are dealing with right

now is called *infinite-dimensional* optimization. (It is also called control theory; see below.)

Of course, infinite-dimensional optimization is conceptually a bit trickier than finite-dimensional optimization. But the good new is that our problem is significantly simplified by the assumption that the objective function is a (kind of) sum and that the time periods are tied together in a special way, viz. by a (system of) differential equation(s). One nice way to think about our problem is to say that we are dealing with a 'controlled' differential equation. The dynamics of the 'state' x can be affected by the 'control' u via the 'state equation' $\dot{x} = g(t, x, u)$. Notice that the function f and the state equation together express the trade-off that has to be resolved when changing $u(t)$: there is one immediate effect on the payoff via f and an effect on the future value of x via g.

Getting back to our problem, we now want to know whether there is an infinite-dimensional counterpart of $\frac{\partial L(x^*, \lambda^*)}{\partial x} = 0$. The answer: there sure is! Just as in finite-dimensional optimization, the two ideas at work are

1. If the control variable u is optimal, then the objective function $\int_0^T f dt$ should be stationary with respect to small changes in u.

2. Instead of choosing u and letting x be determined by the constraint $\dot{x} = g$, we choose both x *and* u treating $\dot{x} = g$ as a constraint associated with the shadow price λ. The hope is that this shadow price will take care of the 'planning ahead' (worrying about the future) aspect of our problem.

The next section uses these ideas in an exceedingly sloppy way (we will clean up our act later). Before we go on, though, we should know what to expect. The infinite-dimensional counterpart of the condition $\nabla_x L(x^*, \lambda^*) = 0$ will turn out to be a system of ODE:s, which we have just learned something about how to solve!

10.2.2 Pontryagin's maximum principle (PMP)

In this section, we will denote the Euclidean inner product in $\mathbb{R}^n$ by $x \cdot y$ rather than $x^T y$ since the letter T will have another meaning. This notation also has the advantage that when you read this section the first time, you can easily think of the scalar case and ignore the vector calculus so as to focus on the new ideas involved. In any case, the function f below is always real scalar-*valued* since it makes no sense to maximize a vector-valued function (or indeed a complex-valued function).

By analogy with the finite-dimensional case, we introduce a 'Lagrangian' as follows (assuming for the moment that $U = \mathbb{R}^k$ so that we can ignore the condition $u(t) \in U$).

$$L(\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = \int_0^T f(t, x(t), u(t))\, dt + \int_0^T \lambda(t) \cdot [g(t, x(t), u(t)) - \dot{x}(t)]\, dt \quad (10.5)$$

Now integrate by parts, and we get

$$L = \int_0^T \left[f + \lambda \cdot g + \dot{\lambda} \cdot x \right] dt + \lambda(T) \cdot \underset{\downarrow}{\overset{x(T)}{b}} - \lambda(T) \cdot \underset{\downarrow}{\overset{x(0)}{a}}. \quad (10.6)$$

Keeping in mind that our choice variables are $x(t)$ for each $t \in [0, T]$ and $u(t)$ for each $t \in [0, T]$, it does not seem unreasonable that a solution should satisfy

$$\text{`}\frac{\partial L}{\partial x(t)} = 0\text{'} \text{ for each } t \in [0, T] \quad (10.7)$$

and

$$\text{`}\frac{\partial L}{\partial u(t)} = 0\text{'} \text{ for each } t \in [0, T]. \quad (10.8)$$

Now introduce the following meaningless but intuitive rule of differentiating an integral with respect to a single value of the integrand.

$$\text{`}\frac{d}{df(s)} \int_0^T g(f(t))\, dt = \frac{dg(f(s))}{df(s)} ds\text{'} \quad (10.9)$$

If this seems puzzling, cf. what we would do with a discrete sum. Then

$$\frac{d}{dx_i}\sum_{k=1}^{n} g\left(x_k\right)\Delta x_k = \frac{dg\left(x_i\right)}{dx_i}\Delta x_i. \tag{10.10}$$

In any case, using this 'rule', and 'dividing' by ds, we get

$$\frac{\partial}{\partial x}\left[f + \lambda \cdot g\right] + \dot{\lambda} = 0 \tag{10.11}$$

and

$$\frac{\partial}{\partial u}\left[f + \lambda \cdot g\right] = 0 \tag{10.12}$$

Now suppose f is concave, the elements of λ are positive and the g_i are concave in u (or f is concave and the g_i are linear). Then our solution u^* will deliver the maximum of $[f + \lambda \cdot g]$. If the constraint set $U \neq \mathbb{R}^k$, then it is not hard to believe that our derivative condition (10.12) generalizes to the requirement that, for each t, $u\left(t\right)$ is that element of U which delivers the greatest value of $f + \lambda \cdot g$.

All this (useful) nonsense motivates the following theorem.

Theorem 10.2.2 (Pontryagin's maximum principle) *Define the* Hamiltonian function $H : \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^k\times\mathbb{R}^n \to \mathbb{R}$ *via*

$$H\left(t, x, u, \lambda\right) = f\left(t, x, u\right) + \lambda \cdot g\left(t, x, u\right). \tag{10.13}$$

Let x^*,u^* solve (10.4). Then, given some regularity conditions and a reasonable definition of what control functions u are admissible, there is a continuous and piecewise differentiable (i.e. differentiable except possibly at a finite number of points) function $\lambda : [0, T] \to \mathbb{R}$ such that

$$\frac{\partial H\left(t, x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right)}{\partial x} + \dot{\lambda}\left(t\right) = 0 \tag{10.14}$$

for all $t \in [0, T]$ at which u is continuous as a function of t and

$$H\left(t, x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right) = \sup_{u\in U} H\left(t, x^*\left(t\right), u, \lambda\left(t\right)\right) \tag{10.15}$$

for all $t \in [0, T]$.

Proof. See [14]. ■

Remark 10.2.3 *The point of PMP is that it gives us an ordinary finite-dimensional maximization problem to solve for each t, and we know from Mathematics 1 how to do that!*

Remark 10.2.4 *Amazingly, our rather wild derivations have led us to (a slightly vague version of) the right answer!*

Remark 10.2.5 *In this version of Pontryagin's maximum principle, we have assumed that the constraint set U (for the control u) is independent of the state, and that there are no restrictions on the state variables x. There are other versions of PMP that drop these assumptions. See section 10.2.4.2 and [10] for more details.*

Remark 10.2.6 *PMP was invented in the 1950s to deal with problems where the solution is such that u is sometimes on the boundary of U, a phenomenon that the 18th century calculus-of-variations approach could not deal with. This is particularly important in so-called bang-bang problems, where it is optimal to have u occasionally jump discontinuously from one boundary point of U to another (something that the calculus of variations approach explicitly forbids). Note that if u has a discontinuity at t then $x(t)$ fails to be differentiable at t. So the condition $\dot{x} = g$ is only required to hold 'almost everywhere'; for example, it suffices that it holds everywhere except at countably many points. More precisely, the constraint is, for $0 \leq t \leq T$,*

$$x(t) = x(0) + \int_0^t g(s, x(s), u(s))\, ds. \tag{10.16}$$

You will have noticed that our statement of PMP is not quite precise, and I refer those of you who yearn for precision either to [14] or to section 10.2.4, where

we will state and rigorously prove a theorem to the effect that a certain set of conditions are *sufficient* for a solution under suitable assumptions. (Given the existence problems; see below, sufficient conditions are in any case more useful.)

10.2.3 Some remarks about existence

Compact sets, closed balls in infinite-dimensional sets, Weierstrass not applicable. Example.

$$\begin{aligned} &\min_{u} \int_{0}^{1} x^{2}(t)\, dt \\ &\text{s.t} \begin{cases} \dot{x} = u \\ x(0) = x(1) = 1 \end{cases} \end{aligned} \tag{10.17}$$

Compactification by extending the notion of a function?

10.2.4 Mangasarian's sufficient conditions

Doing it the economic way; individual optimization and market equilibrium. Arrow-Debreu: trade in dated goods takes place in meta-time. Representative agent (everyone the same) and no distortions/externalities (the competitive equilibrium is then a Pareto optimal allocation) but these assumptions can be easily generalized. We will drop the assumption of Pareto optimality in section 10.3.2 (on dynamic optimization in discrete time). (Heterogeneous agents are beyond the scope of this course, but you mustn't think that this is an impossible topic, although it does face a few numerical problems. Just make sure every agent optimizes.) Constraint set possibly dependent on time (but not the state). Interpret the differential equation as an integral equation and allow finitely many points at which x is not differentiable.

To make the problem economic, write

$$\max_u \int_0^T f\left(t, x\left(t\right), u\left(t\right)\right) dt$$

$$\text{s.t.} \begin{cases} \dot{x}\left(t\right) = g\left(t, x\left(t\right), u\left(t\right)\right) \\ x\left(0\right) = x_0 \\ \text{Some suitable NPG condition} \\ u\left(t\right) \in U \subset \mathbb{R}^k \text{ for all } t \in \left[0, T\right]. \end{cases} \tag{10.18}$$

Note that the endpoint condition $x\left(T\right) = b$ has been replaced by the somewhat vague phrase 'some suitable NPG condition'. We will make this precise, but first let's motivate the precise definition. 'NPG condition' stands for 'no Ponzi-game condition'. A Ponzi game in our context means going into debt without ever paying back, or, in an infinite horizon setting, borrowing money to finance the interest payments on the previous loan and so on ad infinitum. So an NPG condition means that, at the endpoint T the agent should leave behind a portfolio of assets with a nonnegative value. Notice that for the value of a portfolio to be defined, we need asset prices. But these are determined in equilibrium, so the NPG condition cannot be stated when looking merely at one agent's problem and ignoring the market interaction. It can only be stated if we know the prices $\boldsymbol{\lambda}$. Given these prices, the NPG condition is

$$\lambda\left(T\right) \cdot x\left(T\right) \geq 0. \tag{10.19}$$

Note that this must hold not for any prices, only for the equilibrium prices $\boldsymbol{\lambda}$. Yet it must hold for any feasible ('admissible') $\mathbf{x}$. If $T = +\infty$ then the appropriate requirement is that $\lambda\left(t\right) \cdot x\left(t\right)$ eventually stop ever dipping below zero. The mathematical statement of this is

$$\liminf_{t \to \infty} \lambda\left(t\right) \cdot x\left(t\right) \geq 0. \tag{10.20}$$

Now let every agent in the economy solve (10.18), and, to motivate why a competitive equilibrium is the appropriate solution concept, suppose each agent is so small that her behavior cannot influence prices. We now consider sufficient conditions for the profile $(\mathbf{x}^*, \mathbf{u}^*)$ to be a competitive equilibrium allocation enforced by the equilibrium prices $\boldsymbol{\lambda}$. The sense in which $\boldsymbol{\lambda}$ is a set of prices is that $\lambda(t)$ will turn out to be the marginal value from the point of view of an individual of increasing $x(t)$. This idea will be stated more precisely in section 10.2.4.1. We are now ready to state (an economic version of) Mangasarian's theorem formally.

Nota bene As above, we will denote a function using bold type. Other symbols denote numbers or vectors.

Definition 10.2.1 *An* admissible allocation $(\mathbf{x}, \mathbf{u})$ *at market prices* $\boldsymbol{\lambda}$ *for the problem (10.18) is a pair of functions* $\mathbf{x} : [0,T] \to \mathbb{R}^n$ *and* $\mathbf{u} : [0,T] \to \mathbb{R}^m$ *such that*

1. $u(t) \in U$ *for each* $t \in [0,T]$,
2. $x(t) = x_0 + \int_0^t g(s, x(s), u(s))\, ds$ *for each* $t \in [0,T]$, *and*
3. $\lambda(T) \cdot x(T) \geq 0$ *if* $T < \infty$ *and* $\liminf_{t\to\infty} \lambda(t) \cdot x(t) \geq 0$ *if* $T = \infty$.

Definition 10.2.2 *The profile* $(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda})$ *is said to be a* competitive equilibrium *of the economy where all agents' preferences and constraints are given by (10.18) if* $\mathbf{u}^*$ *solves (10.18) and the prices* $\boldsymbol{\lambda}$ *clear the market for the 'assets'* $\mathbf{x}$ *clears in every period, i.e. if agents could trade assets freely at prices* $\boldsymbol{\lambda}$, *then each agent would demand the quantities* $\mathbf{x}^*$. *The profile* $(\mathbf{x}^*, \mathbf{u}^*)$ *is then called a competitive equilibrium* allocation *enforced by the equilibrium* prices $\boldsymbol{\lambda}$.

Remark 10.2.7 *This definition perhaps sounds a bit loose and non-mathematical. For example, we have not defined what it means to 'trade freely'. But we know that this can be done rigorously; see [50].*

Theorem 10.2.3 (Mangasarian) *Let* $(\mathbf{x}^*, \mathbf{u}^*)$ *be an admissible allocation of (10.18) at market prices* $\boldsymbol{\lambda}$*. Let* $\boldsymbol{\lambda} : [0,T] \to \mathbb{R}$ *be a continuous and (except possibly at countably many points) differentiable function. Moreover, suppose the set* $U \in \mathbb{R}^m$ *is convex. Now define the function* $H : [0,T] \times \mathbb{R}^n \times U \times \mathbb{R}^n \to \mathbb{R}$ *via*

$$H(t,x,u,\lambda) = f(t,x,u) + \lambda \cdot g(t,x,u). \tag{10.21}$$

Suppose now that H *is continuously differentiable with respect to* x *on its entire domain. Suppose also that* $H(t,x,u,\lambda(t))$ *is concave in* (x,u) *for each* $t \in [0,T]$*. Finally, suppose that*

1. $\dfrac{\partial H(t,x^*(t),u^*(t),\lambda(t))}{\partial x} + \dot{\lambda}(t) = 0$ *for all* $t \in [0,T]$ *except possibly at finitely many points,*

2. $u^*(t) \in \underset{u \in U}{\operatorname{argmax}} H(t,x^*(t),u,\lambda(t))$*, for all* $t \in [0,T]$ *and*

3. $\lambda(T) \cdot x^*(T) = 0$ *if* $T < \infty$ *and* $\lim_{t\to\infty} \lambda(t) \cdot x^*(t) = 0$ *if* $T = \infty$*.*

Then $(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda})$ *is a competitive equilibrium.*

Remark 10.2.8 *If all the prices* $\boldsymbol{\lambda}$ *are strictly positive,* $\lambda(T) \cdot x(T) = 0$ *implies* $x(T) = 0$*.*

Remark 10.2.9 *The NPG condition* $\lambda(T) \cdot x(T) \geq 0$ *or* $\liminf_{t\to\infty} \lambda(t) \cdot x(t) \geq 0$ *is a* constraint*. The condition* $\lambda(T) \cdot x^*(T) = 0$ *or* $\lim_{t\to\infty} \lambda(t) \cdot x^*(t) = 0$ *is called a* transversality *condition, and is not a constraint but an optimality condition. The intuitive meaning of the NPG condition is 'You mustn't leave any debt behind'*

and the intuitive meaning of the transversality condition is 'Since you can't leave any debt behind, and there is no point in leaving assets behind, set the net worth of your bequest to zero'.

Before the proof starts, we need some preliminaries. Eventually the strategy will be to show that $\mathbf{u}^*$ solves 10.18 by showing that it delivers a higher value of the objective than any admissible alternative. Then we will invoke an envelope theorem to show that $\boldsymbol{\lambda}$ clears the asset market. But for now, we introduce some definitions and results that we'll need in the proof.

Definition 10.2.3 *Let $A \subset \mathbb{R}^n$ be a convex set. A function $f : A \to \mathbb{R}$ is said to be* concave *if for all $x^0, x^1 \in A$ and all $\lambda \in [0,1]$ we have*

$$f\left(\lambda x^0 + (1-\lambda)\, x^1\right) \geq \lambda f\left(x_0\right) + (1-\lambda)\, f\left(x^1\right). \tag{10.22}$$

Example 10.2.1 *Let $x, x^1 \in A$. Define $\Delta x = x^1 - x$ and set $\lambda = 1 - \dfrac{1}{M}$ where $M > 0$. Then $\lambda x + (1-\lambda)\, x^1 = x + \dfrac{\Delta x}{M}$ and if f is concave, then*

$$f\left(x + \frac{\Delta x}{M}\right) \geq \frac{1}{M}\left[f\left(x + \Delta x\right) - f\left(x\right)\right]. \tag{10.23}$$

Lemma 10.2.1 (Mean value theorem) *Let $x, y \in \mathbb{R}^n$ and let*

$$A\left(x, y\right) \triangleq \left\{z \in \mathbb{R}^n : z = \lambda x + (1-\lambda)\, y \text{ for some } \lambda \in (0,1)\right\}. \tag{10.24}$$

Then $A\left(x, y\right)$ is the (open) line segment between x and y. Let $f : A\left(x, y\right) \to \mathbb{R}$ be continuously differentiable. Then there is a $w \in A\left(x, y\right)$ such that

$$f\left(x\right) - f\left(y\right) = \nabla f\left(w\right) \cdot \left(x - y\right). \tag{10.25}$$

Proof. See [7]. ■

Lemma 10.2.2 *Let H and U be as in Theorem (10.2.3). Let x^* and λ be arbitrary fixed vectors in $\mathbb{R}^n$ (and hence* not *functions), let $t \in [0,T]$ be fixed and let u^* be such that*

$$u^* \in \underset{u \in U}{\operatorname{argmax}} H(t, x^*, u, \lambda). \tag{10.26}$$

Then for any $u \in U$ and $x \in \mathbb{R}^n$ we have

$$H(t, x^*, u^*, \lambda) - H(t, x, u, \lambda) \geq \frac{\partial H(t, x^*, u^*, \lambda)}{\partial x} \cdot [x^* - x]. \tag{10.27}$$

Proof. Let $x \in \mathbb{R}^n$ and $u \in U$ be fixed vectors and define $\Delta x = x - x^*$, $\Delta u = u - u^*$. For each $M > 0$ we have

$$H\left(t, x^* + \frac{\Delta x}{M}, u^* + \frac{\Delta u}{M}, \lambda\right) \geq \{\text{concavity!}\} \geq$$

$$\geq H(t, x^*, u^*, \lambda^*) + \frac{1}{M}\left[H(t, x, u, \lambda) - H(t, x^*, u^*, \lambda)\right] \geq$$

$$\geq \left\{u^* \in \underset{u \in U}{\operatorname{argmax}} H(t, x^*, u, \lambda) \text{ and } \left(u^* + \frac{\Delta u}{M}\right) \in U \text{ since } U \text{ is convex}\right\} \geq$$

$$\geq H\left(t, x^*, u^* + \frac{\Delta u}{M}, \lambda\right) + \frac{1}{M}\left[H(t, x, u, \lambda) - H(t, x^*, u^*, \lambda)\right]. \tag{10.28}$$

Hence

$$H(t, x^*, u^*, \lambda) - H(t, x, u, \lambda) \geq$$

$$\geq M\left[H\left(t, x^*, u^* + \frac{\Delta u}{M}, \lambda\right) - H\left(t, x^* + \frac{\Delta x}{M}, u^* + \frac{\Delta u}{M}, \lambda\right)\right] = \tag{10.29}$$

$$= \{\text{Mean value theorem!}\}$$

$$= -\frac{\partial}{\partial x} H\left(t, \widehat{x}_M, u^* + \frac{\Delta u}{M}, \lambda\right) \cdot \Delta x$$

for some $\widehat{x}_M \in A\left(x^*, x^* + \frac{\Delta x}{M}\right)$ where $A\left(x^*, x^* + \frac{\Delta x}{M}\right)$ is the open line segment between x^* and $x^* + \frac{\Delta x}{M}$. Now let $M \to \infty$ and invoke the continuity of H_x. ■

The proof can now begin in earnest. Note first that

$$\int_0^T f\left(t, x\left(t\right), u\left(t\right)\right) dt = \int_0^T \left[H\left(t, x\left(t\right), u\left(t\right)\right) - \lambda\left(t\right) \dot{x}\left(t\right)\right] dt \tag{10.30}$$

Now consider the candidate optimal allocation $(\mathbf{x}^*, \mathbf{u}^*)$ which by assumption satisfies $u^*\left(t\right) \in \underset{u \in U}{\operatorname{argmax}} H\left(t, x^*\left(t\right), u, \lambda\left(t\right)\right)$ for each $t \in [0, T]$ and let (x, u) be an admissible allocation. We will now show that $(\mathbf{x}^*, \mathbf{u}^*)$ delivers a value of the objective function no smaller than does $(\mathbf{x}, \mathbf{u})$.

$$\int_0^T f\left(t, x^*\left(t\right), u^*\left(t\right)\right) dt - \int_0^T f\left(t, x\left(t\right), u\left(t\right)\right) dt =$$

$$= \int_0^T \left[H\left(t, x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right) - H\left(t, x\left(t\right), u\left(t\right), \lambda\left(t\right)\right)\right] dt - \int_0^T \lambda\left(t\right) \cdot \left[\dot{x}^*\left(t\right) - \dot{x}\left(t\right)\right] dt \geq$$

$$\geq \{\text{Lemma 10.2.2!}\} \geq$$

$$\geq \int_0^T \left\{H_x\left(t, x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right) \cdot \left[x^*\left(t\right) - x\left(t\right)\right]\right\} dt - \int_0^T \lambda^*\left(t\right) \cdot \left[\dot{x}^*\left(t\right) - \dot{x}\left(t\right)\right] dt =$$

$$= \{\text{Integration by parts!}\} =$$

$$= \int_0^T \left[H_x\left(t, x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right) + \dot{\lambda}\left(t\right)\right] \cdot \left[x^*\left(t\right) - x\left(t\right)\right] dt -$$

$$-\lambda\left(T\right)\left[x^*\left(T\right) - x\left(T\right)\right] + \lambda\left(0\right)\left[x^*\left(0\right) - x\left(0\right)\right]. \tag{10.31}$$

What we want to show, of course, is that the last expression is ≥ 0 under the assumption of our Theorem. The first term vanishes by the $H_x + \dot{\lambda} = 0$ condition. The final term vanishes since $x^*(0) - x(0) = x_0$ (recall that $\mathbf{x}$ and $\mathbf{x}^*$ are admissible!). This leaves us with the middle term

$$\lambda(T)\, x(T) - \lambda(T)\, x^*(T).$$

Consider the case $T < \infty$ first. Then the admissibility of x guarantees that $\lambda(T)\, x(T) \geq 0$. Meanwhile, the transversality condition says that $\lambda(T)\, x^*(T) = 0$. This finishes the proof that u^* is optimal for a finite T. On the other hand, if $T = \infty$, we consider the difference

$$\lambda(t)\, x(t) - \lambda(t)\, x^*(t) \tag{10.32}$$

and let $t \to \infty$. By the NPG condition, $\lambda(t)\, x(t)$ eventually stops ever dipping below zero as $t \to \infty$, and by the transversality condition, $\lambda(t)\, x^*(t) \to 0$ as $t \to \infty$. The final part of the proof shows that the prices $\boldsymbol{\lambda}$ really clear the market. This will be true if each agent's marginal values (in terms of her objective function) of the assets $\mathbf{x}$ are equal to the market prices $\boldsymbol{\lambda}$. This is the Envelope Theorem; see section 10.2.4.1. ■

10.2.4.1 The Envelope Theorem

Theorem 10.2.4 *Let $(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda})$ satisfy the sufficient conditions for being a competitive equilibrium of an economy described by (10.18). Suppose the sufficient conditions define u^* uniquely. [Add another technical condition here; $u^*(t)$ is locally bounded as a function of x.] Define the value function (indirect utility*

function) via

$$V(t,x) = \max_{\mathbf{u}} \int_t^T f(s, x(s), u(s))\, ds$$

$$s.t. \begin{cases} \dot{x}(s) = g(s, x(s), u(s)) \\ x(t) = x \\ NPG \text{ (relative to market clearing prices)} \\ u(s) \in U \subset \mathbb{R}^k \text{ for all } s \in [t,T]. \end{cases} \tag{10.33}$$

Then, for all $t = [0,T)$, *we have*

$$\frac{\partial V(t, x^*(t))}{\partial x} = \lambda(t). \tag{10.34}$$

and if we define $\dfrac{\partial V(T, x^*(T))}{\partial x} = \lim_{t \to T} \dfrac{\partial V(t, x^*(t))}{\partial x}$ *then the statement folds for* $t = T$ *as well.*

Proof. See [3]. ■

Remark 10.2.10 *The equation (10.34) is only valid aloong the optimal path* $\mathbf{x}^*$.

Example 10.2.2 *Consider the optimal consumption problem*

$$\max_c \int_0^1 \ln c(t)\, dt$$

$$s.t. \begin{cases} \dot{k}(t) = -c(t) \\ k(0) = 1 \\ NPG \end{cases} \tag{10.35}$$

The Hamiltonian is

$$H(t, x(t), \lambda(t)) = \ln c(t) - \lambda(t)\, c(t) \tag{10.36}$$

Our optimality conditions become

$$\dot{\lambda}(t) = 0 \tag{10.37}$$

and

$$\frac{1}{c(t)} = \lambda(t) \tag{10.38}$$

Hence $\lambda(t) \equiv \lambda$ *and* $c(t) \equiv c = \frac{1}{\lambda}$. *According to the law of motion for the capital stock,*

$$k(t) = 1 - \int_0^t c ds = 1 - tc. \tag{10.39}$$

The NPG condition now becomes

$$\frac{1-c}{c} = 0. \tag{10.40}$$

This implies that $c = 1$ *and we have solved the problem. Note that* $\lambda = 1$ *and* $k^*(t) = 1 - t$. *To find the value function, we now solve*

$$\max_c \int_t^1 \ln c(s)\, ds$$

$$s.t. \begin{cases} \dot{k}(s) = -c(s) \\ k(t) = k \\ NPG \end{cases} \tag{10.41}$$

In this case, the solution is $c(t) \equiv c = \frac{k}{1-t}$. *Hence the value function is*

$$V(t,k) = \int_t^1 \ln\left(\frac{k}{1-t}\right) ds = (1-t)\left[\ln k - \ln(1-t)\right] \tag{10.42}$$

and

$$\frac{\partial V(t,k)}{\partial k} = \frac{1-t}{k}. \tag{10.43}$$

Consequently

$$\frac{\partial V\left(t,k^{*}\left(t\right)\right)}{\partial k}=\frac{1-t}{1-t}=1. \tag{10.44}$$

and we have confirmed the envelope theorem in a special case.

10.2.4.2 Constraints involving the state variables

Sometimes we have constraints of the form $h\left(t,x\left(t\right),u\left(t\right)\right)\leq 0$.

Avoiding these constraints. Dealing with them if you can't.

10.2.4.3 Integral constraints

Sometimes we have constraints of the form

$$\int_{0}^{T}k\left(t,x\left(t\right),u\left(t\right)\right)dt=K \tag{10.45}$$

See [3].

10.2.4.4 Endpoint evaluation

Occasionally one comes across problems where the maximand has the form

$$\int_{0}^{T}f\left(t,x\left(t\right),u\left(t\right)\right)dt+h\left(x\left(T\right)\right). \tag{10.46}$$

where $x\left(T\right)$ is free. It is possible to transform this to a problem without endpoint evaluation, but it is often easier not to. The optimum conditions are then the same as before except that the appropriate transversality condition is

$$\frac{\partial h\left(x\left(T\right)\right)}{\partial x}-\lambda\left(T\right)=0. \tag{10.47}$$

10.2.5 Using the sufficient conditions to calculate the solution

Do our sufficient conditions 'usually' deliver a unique solution? [Subtlety: even if they do, there may be more than one solution to the optimization problem.] Well, suppose $u^*(t)$ is in the interior of U for each t. Then our sufficient conditions can be written (in shorthand notation) as

1. $H_x + \dot{\lambda} = 0$ (n differential equations)
2. $\dot{x} = g$ (n differential equations)
3. $H_u = 0$ (m algebraic equations for each t)
4. $x(0) = x_0$ (n initial conditions).

Now in well-behaved cases, we can use $H_u = 0$ to solve for the m elements of u in terms of λ and x. That leaves us with a $2n$-dimensional system of differential equations with n initial conditions. We need n more linear restrictions to pin down the solution. If $T < \infty$ and all the prices are strictly positive, the transversality condition implies $x(T) = 0$ which gives us another n linear restrictions, which is just what we need. If $T = \infty$, the condition $\lim_{t\to\infty} \lambda(t) x(t) = 0$ will typically imply that $x(t)$ is stable (or at least, which is what matters, be implied by x being stable). If there are precisely n unstable eigenvalues, this stability requirement imposes exactly n linear restrictions, and we have pinned down the solution.

It may seem unlikely that we should have exactly n unstable eigenvalues, but, in fact, it is a mathematical theorem that linear-quadratic control problems generically have this property! See section 10.3.8.1 for a discussion of the discrete time case. Economics of explosive roots: bootstrap property of asset pricing; the

higher the price of an asset tomorrow, the more valuable it is today. This positive feedback leads to an explosive eigenvalue for each of the n 'assets' in x.

So the problem of solving a dynamic optimization problem boils down to solving a system of ODEs with boundary values. Sometimes we'll be able to solve this system exactly, but often we won't. Then there are many numerical methods of finding an approximate solution. One particularly simple one is to linearize the system around its steady state, and then solve the resulting linear system. The solution will then be accurate close to the steady state.

If our only objective is to find a qualitative characterization of the solution (steady state, stability) it also makes sense to linearize and invoke Lyapunov's theorem. Having derived the approximate dynamics around a steady state, we can draw a rough phase diagram indicating the approximate behavior of the trajectories.

In many cases, however, it won't be obvious how to deal with the sufficient conditions. Ingenuity is often required, and ingenuity can only be acquired through practice.

Example 10.2.3 *Suppose we have a beehive which is Darwinianly selected to maximize the number of queens and drones at the end of the season. Let $x(t)$ be the number of sterile workers and $y(t)$ the number of queens and drones at time t. Thus the maximand is $y(T)$. [We can rewrite this problem without endpoint evaluation but it's easier not to.] Meanwhile, suppose the beehive faces the*

following constraints

$$\begin{cases} \dot{x}(t) = bu(t)\,x(t) - \gamma x(t) \\ \dot{y}(t) = c(1-u(t))\,x(t) \\ x(0) = 1 \\ y(0) = 0, \\ u(t) \in [0,1] \text{ for all } t \in [0,T] \end{cases} \tag{10.48}$$

where $b > \gamma$. We now proceed to solve this problem. We will take the objective to We will discover that the solution has a bang-bang nature. The Hamiltonian is

$$H = 1 + \lambda(t)\left[bu(t)\,x(t) - \gamma x(t)\right] + \mu(t)\left[c(1-u(t))\,x(t)\right] \tag{10.49}$$

where $\lambda(t)$ and $\mu(t)$ are the shadow prices of the constraints. Notice that H is linear in $u(t)$ with slope coefficient $\lambda(t)\,bx(t) - c\mu(t)\,x(t)$. Since $x(t)$ is always greater than zero (this follows from the constraints), the sign of this slope is the same as that of the "switch function"

$$\sigma(t) = b\lambda(t) - c\mu(t). \tag{10.50}$$

When $\sigma(t) > 0$, the Hamiltonian is maximized when $u(t) = 1$, and when $\sigma(t) < 0$ it is maximized when $u(t) = 0$. When $\sigma(t) = 0$, it is optimal to switch from 0 to 1 or vice versa. Note that PMP guarantees the continuity of $\lambda(t)$ and $\mu(t)$, so that the only switch times are those such that $\sigma(t) = 0$.

Now we look at the optimum conditions. Recall that we have endpoint evaluation with

$$h(x(T), y(T)) = y(T) \tag{10.51}$$

so that

$$1 - \mu(T) = 0$$

which implies that

$$\mu(T) = 1.$$

Summarizing, the sufficient conditions for an optimum are as follows.

$$\begin{cases} H_x = \lambda(t)[bu(t) - \gamma] + \mu(t)\, c(1 - u(t)) = -\dot{\lambda} \\ H_y = 0 = -\dot{\mu} \\ \lambda(T) = 0 \\ \mu(T) = 1. \end{cases} \tag{10.52}$$

We may immediately conclude that $\mu(t) \equiv 1$ *and hence that the switch function can be written as*

$$\sigma(t) = b\lambda(t) - c. \tag{10.53}$$

Also, we can now rewrite the differential equation for λ *as*

$$\begin{cases} \dot{\lambda}(t) = [\gamma - bu(t)]\lambda(t) - c(1 - u(t)) \\ \lambda(T) = 0. \end{cases} \tag{10.54}$$

We now ask how many switches there will be between 0 *and* T*. To answer that, we check the sign of* $\dot{\sigma}$ *at* $\sigma = 0$*. We get*

$$\dot{\sigma} = b\gamma\lambda - b^2 u\lambda - bc(1 - u) = \gamma c - bcu - bc + bcu = c(\gamma - b) < 0 \tag{10.55}$$

so that there is at most one switch. If σ *ever reaches zero, it forever stays below zero after that. The structure of the solution, then is*

$$u(t) = \begin{cases} 1 & \text{when } t \in [0, s) \\ 0 & \text{when } t \in (s, T] \end{cases} \tag{10.56}$$

(the value of u *at* $t = s$ *doesn't matter). Notice that we have not excluded the possibility* $s = 0$ *(the possibility* $s > T$ *is an awkward on which we'll ignore for now). But what is the value of* s*? To find out, we consider the differential equation*

for λ from s to T, where we know that $u = 0$. We get

$$\begin{cases} \dot{\lambda}(t) = \gamma\lambda(t) - c \\ \lambda(T) = 0. \end{cases} \tag{10.57}$$

Using the formula for solving a linear ODE (keeping in mind that switching around the limits of an integral switches the sign), we get

$$\lambda(t) = \int_t^T \exp\left\{\int_u^t \gamma dv\right\} c du = \frac{c}{\gamma}\left[1 - e^{\gamma(t-T)}\right]. \tag{10.58}$$

Now recall that the time s is defined via $\sigma(s) = 0$ and hence $\lambda(s)\, b = c$. This means that

$$\frac{c}{\gamma}\left[1 - e^{\gamma(s-T)}\right] b = c \tag{10.59}$$

and consequently

$$s = T + \frac{1}{\gamma}\ln\left(1 - \frac{\gamma}{b}\right). \tag{10.60}$$

Since $b > \gamma$, we have $s \leq T$. Thus it is never optimal to spend the whole season just rearing sterile workers. However, there is no guarantee that $s \geq 0$. If $s < 0$ according to our formula, then of course it is optimal never to switch and set $u(t) \equiv 0$ for all $t \in [0, T]$ and hence devote the entire season to rearing queens and drones.

Apparently, field biologists have been able to confirm that this is indeed how many beehives behave. Indeed, there has even been some quantitative predictive success with calibrated models of the kind presented here.

10.2.6 Current value costate

Sometimes the time-dependence of a dynamic optimization problem takes the form of geometric discounting only, i.e. we have a problem of the form

$$\max_{u} \int_{0}^{T} e^{-\rho t} f\left(x\left(t\right), u\left(t\right)\right) dt$$

$$\text{s.t.} \begin{cases} \dot{x}\left(t\right) = g\left(x\left(t\right), u\left(t\right)\right) \\ x\left(0\right) = x_0 \\ \text{NPG} \\ u\left(t\right) \in U \subset \mathbb{R}^k \text{ for all } t \in \left[0, T\right]. \end{cases} \tag{10.61}$$

In this case it is often useful to redefine the costate $\lambda\left(t\right)$ as the current rather than the present value of $x\left(t\right)$ (we'll see in a moment precisely what this means). The definition is

$$\lambda^c\left(t\right) = e^{\rho t}\lambda^p\left(t\right) \tag{10.62}$$

where the 'present value' costate $\lambda^p\left(t\right)$ is just the costate that we have been working with so far. The sense in which this is the current value is that it is the derivative of the following 'current' value function

$$V\left(t, x\right) = \max_{\mathbf{u}} \int_{t}^{T} e^{\rho\left(s-t\right)} f\left(x\left(s\right), u\left(s\right)\right) ds$$

$$\text{s.t.} \begin{cases} \dot{x}\left(s\right) = g\left(x\left(s\right), u\left(s\right)\right) \\ x\left(t\right) = x \\ \text{NPG} \\ u\left(s\right) \in U \subset \mathbb{R}^k \text{ for all } s \in \left[t, T\right]. \end{cases} \tag{10.63}$$

This means that $\lambda^c\left(t\right)$ can be interpreted as the prices at which the assets x are traded at the instant t in actual time rather than the prices of $x\left(t\right)$ in the Arrow-Debreu market in meta-time (or, if you like, at the instant $t = 0$).

Apart from the nice interpretation, the current value approach has benefits in terms of calculation. One thing is that it makes the first order conditions time-independent, which is crucial if we are looking for a time-independent feedback rule (see section 10.3.4). Another is that is simplifies the calculations regardless of how we want to represent the solution.

Given our new definition of the costate, we can define new optimality conditions in terms of a current value Hamiltonian in the following way. The old-fashioned (present-value) Hamiltonian as applied to our discounted case is of course

$$H^p\left(t, x\left(t\right), u\left(t\right), \lambda\left(t\right)\right) = e^{-\rho t} f\left(x\left(t\right), u\left(t\right)\right) + \lambda^p\left(t\right) \cdot g\left(x\left(t\right), u\left(t\right)\right). \quad (10.64)$$

Rewriting in terms of the current value, we get

$$H^p\left(t, x\left(t\right), u\left(t\right), \lambda\left(t\right)\right) = e^{-\rho t} f\left(x\left(t\right), u\left(t\right)\right) + e^{-\rho t}\lambda^c\left(t\right) \cdot g\left(x\left(t\right), u\left(t\right)\right) \quad (10.65)$$

and maximizing this with respect to $u\left(t\right)$ is of course the same as maximizing the current value Hamiltonian

$$H^c\left(x\left(t\right), u\left(t\right), \lambda\left(t\right)\right) = f\left(x\left(t\right), u\left(t\right)\right) + \lambda^c\left(t\right) \cdot g\left(x\left(t\right), u\left(t\right)\right). \quad (10.66)$$

The condition $H_x + \dot{\lambda} = 0$ is a little bit trickier. The correct condition is of course $H_x^p + \dot{\lambda}^p = 0$. Since $H_x^p = e^{-\rho t} H_x^c$ and $\dot{\lambda}^p = -\rho e^{-\rho t}\lambda^c + e^{-\rho t}\dot{\lambda}^c$, this condition becomes $H_x^c + \dot{\lambda}^c - \rho\lambda^c$ in terms of the current value Hamiltonian and costate (which is independent of time!). In summary, our optimality conditions become

$$\begin{cases} \dfrac{\partial H^c\left(x^*\left(t\right), u^*\left(t\right), \lambda\left(t\right)\right)}{\partial x} + \dot{\lambda}^c\left(t\right) - \rho\lambda^c\left(t\right) = 0 \\ u^*\left(t\right) \in \underset{u \in U}{\operatorname{argmax}} H^c\left(x^*\left(t\right), u, \lambda\left(t\right)\right) \\ \lambda^c\left(T\right) \cdot x\left(T\right) = 0 \text{ if } T < \infty \text{ and } \lim\limits_{t \to \infty} e^{-\rho t}\lambda^c\left(t\right) \cdot x\left(t\right) = 0 \text{ if } T = \infty \end{cases} \quad (10.67)$$

where the current value Hamiltonian is defined via

$$H^c\left(x\left(t\right),u\left(t\right),\lambda\left(t\right)\right)=f\left(x\left(t\right),u\left(t\right)\right)+\lambda^c\left(t\right)\cdot g\left(x\left(t\right),u\left(t\right)\right). \tag{10.68}$$

Provided one knows what one is doing, the notation $\lambda^c\left(t\right)=\lambda\left(t\right)$ is perfectly permissible. In any case, the usefulness of the current value approach can best be demonstrated in examples, and to such an example we now turn. [To be written.]

Example 10.2.4 *Installation costs. From last year's problem sets.*

Example 10.2.5 *Qualitative reasoning; example from B&F, pp. 122-123.*

10.2.7 Linear-quadratic control problems

The sufficient conditions are linear. This case will be discussed in detail in the discrete time case. Here we will just give an example.

Example 10.2.6 *Consider a firm which maximizes the discounted value of profits. It maximizes*

$$\int_0^\infty e^{-rt}\left[k\left(t\right)-\frac{1}{2}k^2\left(t\right)-i\left(t\right)-\frac{a}{2}\left(i\left(t\right)-\delta k\left(t\right)\right)^2\right]dt \tag{10.69}$$

subject to

$$\dot{k}\left(t\right)=-\delta k\left(t\right)+i\left(t\right) \tag{10.70}$$

where we think of $k\left(t\right)-\frac{1}{2}k^2\left(t\right)$ *as output and* $i\left(t\right)+\frac{a}{2}\left(i\left(t\right)-\delta k\left(t\right)\right)^2$ *as the costs of investment. Notice that there is a convex cost of changing the capital stock whose significance increases in* a*. We assume that* $a>0$ *and* $r>0$*.*

The current value Hamiltonian is

$$H=k\left(t\right)-\frac{1}{2}k^2\left(t\right)-i\left(t\right)-\frac{a}{2}\left(i\left(t\right)-\delta k\left(t\right)\right)^2+\mu\left(t\right)\left[-\delta k\left(t\right)+i\left(t\right)\right] \tag{10.71}$$

and the optimum conditions are

$$\begin{cases} H_k + \dot{\mu} - r\mu = 1 - k(t) + a\delta(i(t) - \delta k(t)) - \delta\mu(t) + \dot{\mu}(t) - r\mu(t) = 0 \\ H_i = -1 - a(i(t) - \delta k(t)) + \mu(t) = 0 \end{cases} \tag{10.72}$$

Using the $H_i = 0$ *condition to solve for the control* $i(t)$*, we get*

$$i(t) = \delta k(t) + \frac{1}{a}(\mu(t) - 1) \tag{10.73}$$

Substituting into the $H_k + \dot{\mu} - r\mu = 0$ *condition and the law of motion* $\dot{k}(t) = -\delta k(t) + i(t)$*, we get*

$$\begin{cases} \dot{\mu}(t) = k(t) + r\mu(t) - (1 - \delta) \\ \dot{k}(t) = \frac{1}{a}(\mu(t) - 1) \end{cases} \tag{10.74}$$

or, in matrix format,

$$\begin{bmatrix} \dot{\mu}(t) \\ \dot{k}(t) \end{bmatrix} = \begin{bmatrix} r & 1 \\ \frac{1}{a} & 0 \end{bmatrix} \begin{bmatrix} \mu(t) \\ k(t) \end{bmatrix} + \begin{bmatrix} -(1-\delta) \\ -\frac{1}{a} \end{bmatrix}. \tag{10.75}$$

First we find the steady state, which is

$$\begin{bmatrix} \mu^* \\ k^* \end{bmatrix} = \begin{bmatrix} r & 1 \\ \frac{1}{a} & 0 \end{bmatrix}^{-1} \begin{bmatrix} (1-\delta) \\ \frac{1}{a} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 - \delta - r \end{bmatrix} \tag{10.76}$$

Notice that the marginal product of capital in the steady state is $r + \delta$*. To analyze the dynamics around the steady state, we investigate the eigenvalues and eigenvectors of the coefficient matrix. We find that*

$$\begin{vmatrix} r & 1 \\ \frac{1}{a} & 0 \end{vmatrix} = -\frac{1}{a} < 0 \tag{10.77}$$

so that the solution is a saddle path. Denote the stable (negative) eigenvalue by λ_1*. The corresponding eigenvector* x *satisfies* $\frac{1}{a}x_1 = \lambda_1 x_2$ *so that, along the saddle path, we have*

$$\mu(t) - \mu^* = a\lambda_1(k(t) - k^*). \tag{10.78}$$

It follows that

$$i(t) = \delta k(t) + \lambda_1 (k(t) - k^*) \tag{10.79}$$

and consequently

$$\dot{k}(t) = \lambda_1 (k(t) - k^*) \tag{10.80}$$

so that the rate of convergence towards the steady state is given by the stable eigenvalue.

On intuitive grounds, it seems reasonable to suppose that as $a \to 0$, *we have immediate convergence so that* $\lambda_1 \to -\infty$. *Conversely, when* $a \to \infty$, *capital should be constant since it is infinitely costly to change it and hence* $\lambda_1 \to 0$. *Using the formula for the roots of a quadratic equation, it is not hard to confirm these conjectures. We have*

$$\lambda_1 = \frac{r}{2} - \sqrt{\left(\frac{r}{2}\right)^2 + \frac{1}{a}} \tag{10.81}$$

and the conjectures follow almost immediately. In between these extremes, it would be nice to verify that

$$\frac{\partial \lambda_1}{\partial a} > 0. \tag{10.82}$$

To show this, note that

$$\begin{cases} \lambda_1 + \lambda_2 = r \\ \lambda_1 \lambda_2 = -\dfrac{1}{a} \end{cases} \tag{10.83}$$

from which it follows that

$$\frac{\partial \lambda_1}{\partial a} = \frac{1}{a(\lambda_2 - \lambda_1)} > 0 \tag{10.84}$$

since, by definition, $\lambda_1 < 0 < \lambda_2$.

10.3 Discrete time

10.3.1 Definition of problem

Straight to distorted market equilibrium in a representative agent economy.

Finite horizon: we did this in MatFU 1, implicitly! But special form, let's exploit that.

Infinite horizon: we need to learn more. Rather than going through a Pontryagin-style argument, we go straight to the sufficient conditions. What to expect: a system of difference equations.

10.3.2 Mangasarian-style sufficient conditions

Here we will consider a competitive economy with identical agents but (possibly) with externalities. Formally, we let the economy consist of uncountably many small agents, each of measure zero but of total measure one. Let the set of agents be a measure space $(I, \mathcal{F}, \mu)$ where $\mu(i) = 0$ for each $i \in I$ but $\mu(i) = 1$. Now let, say, the asset holdings of an agent i at time t be $x^i(t)$. We then define the aggregate asset holdings by

$$x(t) = \int_I x^i(t)\, d\mu(i). \tag{10.85}$$

and similarly for other variables.

Since all agents are small and alike in all respectss (a philosopher would say 'qualitatively identical but not numerically identical' when we say 'alike in all respects'), prices depend only on aggregate variables. We model the externality by saying that prices are determined by a market clearing condition involving only aggregate variables as follows.

$$m(x_t, u_t, p_t) = 0 \tag{10.86}$$

where $p_t \in \mathbb{R}^k$ is a vector of prices (other than the prices of the assets x), z_t is an exogenous sequence and $m : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^k$ is a vector of market clearing conditions and aggregate resource constraints. An example of a market clearing condition is that the wage is equal to the marginal product of labor. An example of an aggregate resource constraint is the GDP is equal to the sum of consumption, investment and net exports. Now suppose each agent i solves

$$\max_{\mathbf{u}^i} \sum_{t=0}^{T} f\left(t, x_t^i, u_t^i, p_t^i\right)$$

$$\text{s.t.} \begin{cases} x_{t+1}^i = g\left(t, x_t^i, u_t^i, p_t\right) & (10.87) \\ u_t^i \in U \subset \mathbb{R}^m \\ x_0^i = a \\ \text{NPG} & (10.88) \end{cases}$$

where possibly $T = \infty$. Look carefully at what we have allowed to be individual-specific and what not! In addition to the constraints, the agent is also aware of the aggregate constraints $m\left(x_t, u_t, p_t\right) = 0$ and $x_{t+1} = g\left(t, x_t, u_t, p_t\right)$ and has rational expectations about the present and future behavior of all the other agents in the economy (which in this case means perfect foresight).

Hint. When writing down market clearing conditions, you always get one more than you need. By Walras' law, you can drop one of them without losing any information.

We now state sufficient conditions for $(\mathbf{x}^*, \mathbf{u}^*)$ to be a competitive equilibrium allocation enforced by the prices $(\boldsymbol{\lambda}, \mathbf{p})$. Given the close analogy with the continuous case and the finite-dimensional case, they are not hard to believe.

To understand the optimum conditions intuitively, it is useful to think of there being a single Lagrangian of the form

$$\mathcal{L} = \sum_{t=0}^{T} \left[f_t + \lambda_t \cdot (g_t - x_{t+1}) \right] \tag{10.89}$$

Note that, in period t, we choose u_t and x_{t+1}, x_t being given by history. Now fix a particular t and consider the terms involving the choice variables u_t and x_{t+1}. We have

$$f_t + \lambda_t \cdot [g_t - x_{t+1}] + f_{t+1} + \lambda_{t+1} \cdot [g_{t+1} - x_{t+2}] \tag{10.90}$$

where the term g_{t+1} involves x_{t+1}. Now maximize this for each t! This yields precisely the optimum conditions below.

Nota bene. As above, we will denote a sequence using bold type. Other symbols denote numbers or vectors.

Definition 10.3.1 *An* admissible allocation $(\mathbf{x}, \mathbf{u})$ *at market prices* $\boldsymbol{\lambda}$ *and* $\mathbf{p}$ *for the economy described by (10.87) is a pair of sequences* $\langle x_t \rangle_{t=0}^{T+1}$ *and* $\langle u_t \rangle_{t=0}^{T}$ *such that*

1. $u_t \in U$ *for each* $t = 0, 1, ..., T$,
2. $x_0 = a$
3. $x_t = g(t, x_t, u_t, p_t)$ *for each* $t = 0, 1, ..., T$,
4. $m(x_t, u_t, p_t) = 0$ *for each* $t = 0, 1, ..., T$, *and*
5. $\lambda_T x_{T+1} = 0$ *if* $T < \infty$ *and* $\lim_{t \to \infty} \lambda_t x_{t+1} = 0$ *if* $T = \infty$

Theorem 10.3.1 *Let* $(\mathbf{x}^*, \mathbf{u}^*)$ *be an admissible allocation of the economy described by (10.87). Let* $\boldsymbol{\lambda} = \langle \lambda_t \rangle_{t=0}^{T-1}$ *be an* $n-$*dimensional sequence. Moreover, suppose the set* $U \in \mathbb{R}^m$ *is convex. Now define the function* H *via*

$$H(t, x, u, p, \lambda) = f(t, x, u, p) + \lambda \cdot g(t, x, u, p) \tag{10.91}$$

for $t = 0, 1, ..., T-1$ *and, if* $T < \infty$,

$$H(T, x, u, p, \lambda) = f(T, x, u, p). \tag{10.92}$$

Suppose now that H *is continuously differentiable with respect to* x *on its entire domain. Suppose also that* $H(t, x, u, p_t, \lambda_t)$ *is concave in* (x, u) *for each* $t = 0, 1, ..., T$. *(Note that the concavity property only has to hold for our particular choice of* $\mathbf{p}$, *and* $\boldsymbol{\lambda}$, *not generally.) Finally, suppose that*

1. $\dfrac{\partial H\left(t+1, x^*_{t+1}, u^*_{t+1}, p_{t+1}, \lambda_{t+1}\right)}{\partial x} - \lambda_t = 0$ *for all* $t = 0, 1, ..., T-1$,
2. $u^*_t \in \underset{u \in U}{\operatorname{argmax}} H(t, x^*_t, u, p_t, \lambda_t)$, *for all* $t = 0, 1, ..., T$ *and*[1]
3. $\lambda_T \cdot x_{T+1} = 0$ *if* $T < \infty$ *and* $\lim_{t \to \infty} \lambda_t \cdot x_{t+1} = 0$ *if* $T = \infty$.

Then $(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}, \mathbf{p})$ *is a competitive equilibrium.*

Remark 10.3.1 *If all the prices* λ_T *are strictly positive,* $\lambda_T \cdot x_{T+1} = 0$ *implies* $x_{T+1} = 0$.

Proof. See [3]. ■

Remark 10.3.2 *Note that Harald and Chow (and nearly everybody else) have a different timing convention for* λ_t *than I do. They associate* λ_{t+1} *with the constraint* $x_{t+1} = g_t$, *whereas I associate* λ_t *with that constraint. Both conventions have their pros and cons but of course yield the same solution for* $\mathbf{x}$ *and* $\mathbf{u}$. *The advantage of my convention is that* u_t *becomes a function of* λ_t *rather than* λ_{t+1} *or, in the stochastic case,* $E[\lambda_{t+1}|\mathcal{F}_t]$. *You will see advantages of Harald's and Chow's convention below.*

[1] In contrast to the continuous time case, this condition is not necessary. If H is convex in (x, u), there may well be a solution that does not maximize the Hamiltonian. See Sydsæter et al: "Further mathematics for economic analysis."

Hint. It is much easier to remember maximizing

$$f_t + \lambda_t \cdot [g_t - x_{t+1}] + f_{t+1} + \lambda_{t+1} \cdot [g_{t+1} - x_{t+2}] \tag{10.93}$$

than to memorize the above Theorem. The initial and boundary values (transversality conditions) are most easily remembered as x_0 given and $x_{T+1} = 0$ if $T < \infty$ and $(\mathbf{x}, \boldsymbol{\lambda})$ stable if $T = \infty$. Notice also that the sign convention (writing $f_t + \lambda_t \cdot [g_t - x_{t+1}]$ or $f_t + \lambda_t \cdot [x_{t+1} - g_t]$) is only important if you care about the sign of λ_t; usually you don't.

Example 10.3.1 *Equilibrium prices.* $w_t = f_{L,t}$; $r_t = f_{K,t}$

10.3.3 The envelope theorem

Theorem 10.3.2 *Let* $(\mathbf{x}^*, \mathbf{u}^*, \boldsymbol{\lambda}, \mathbf{p})$*be a competitive equilibrium of an economy described by (10.87). [Add technical conditions here as in continuous case.] Define the value function (indirect utility function) via*

$$V(t, x) = \max_{\mathbf{u}} \sum_{s=t}^{T} f(s, x_s^i, u_s^i, p_s)$$

$$s.t. \begin{cases} x_{s+1}^i = g(s, x_s^i, u_s^i, p_s) \\ u_s^i \in U \subset \mathbb{R}^m \\ x_t^i = x \\ \textit{NPG (relative to } \boldsymbol{\lambda}\textit{)} \end{cases} \tag{10.94}$$

Then, for all $t = 0, 1, ..., T$ *we have*

$$\frac{\partial V(t, x_t^*)}{\partial x} = \lambda_{t-1}. \tag{10.95}$$

Proof. See [3]. ■

Remark 10.3.3 *Note that with Harald's or Chow's timing convention, the Envelope Theorem says*

$$\frac{\partial V\left(t, x_t^*\right)}{\partial x} = \lambda_t. \tag{10.96}$$

10.3.4 Feedback representation of the solution

Often it is practical to represent the solution as $u_t = d\left(t, x_t\right)$. This is called a *feedback, recursive* or *Markov* representation (or *decision rule*) and has great practical value, especially in stochastic problems. In stochastic problems, the solution is a whole stochastic process, and a recursive representation of the solution is the only manageable alternative. In particularly auspicious circumstances, the dependence on t vanishes, and we can write $u_t = d\left(x_t\right)$.

Exercise 10.3.1 *Let $\alpha, \beta \in (0,1)$ and consider the consumption/saving problem*

$$\max_{\mathbf{c}} \sum_{t=0}^{T} \beta^t \ln c_t$$

$$s.t. \begin{cases} k_{t+1} = k_t^\alpha - c_t \\ k_0 \text{ given} \\ NPG. \end{cases} \tag{10.97}$$

Verify that the feedback rule

$$c_t = \frac{1-\alpha\beta}{1-(\alpha\beta)^{T-t+1}} k_t^\alpha \tag{10.98}$$

solves our problem when $T < \infty$. Verify that the (time-independent!) feedback rule

$$c_t = \left(1-\alpha\beta\right) k_t^\alpha \tag{10.99}$$

solves the problem when $T = \infty$. What is the intuitive reason for the time independence of the solution when $T = \infty$?

10.3.5 Constraints of the form $h(t, x_t, u_t) \leq 0$

Avoiding these constraints. Dealing with them if you can't. Not too hard: associate the constraint with a Lagrange multiplier and use Kuhn-Tucker to maximize the Hamiltonian.

10.3.6 Current value costate

Consider the discounted control problem.

$$\begin{aligned} &\max_{\mathbf{u}} \sum_{t=0}^{\infty} \beta^t f(x_t, u_t) \\ &\text{s.t.} \begin{cases} x_{t+1} = g(x_t, u_t) \\ x_0 \text{ given} \\ \text{NPG} \end{cases} \end{aligned} \tag{10.100}$$

In problems like these, it makes sense to redefine the costate λ in an analogous way to what we did in section 10.2.6. The reasons for this are the same as we discussed there. Specifically, the redefinition is

$$\lambda_t^c \triangleq \beta^{-t} \lambda_t^p. \tag{10.101}$$

In practice, this means that, for each t, we maximize

$$f_t + \lambda_t \cdot [g_t - x_{t+1}] + \beta f_{t+1} + \beta \lambda_{t+1} \cdot [g_{t+1} - x_{t+2}] \tag{10.102}$$

with respect to u_t and x_t, keeping in mind that the transversality condition becomes $\lim_{t \to \infty} \beta^t \lambda_t x_{t+1} = 0$ when $T = \infty$. Notice that the first-order conditions are time-independent, so that, if linearized, it becomes a linear system of equations with constant coefficients.

10.3.7 Using the sufficient conditions to find the solution

In principle, we proceed as in section 10.2.5, and the essential task is of course to solve a system of difference equations. Usually, of course, this is difficult; more specifically, the difficulty lies in picking out the *stable* solution(s). In the linear case this problem is not too hard, and sometimes not much precision is lost by linearizing he first order conditions. We can then proceed as in chapter 9. In an interesting class of cases the sufficient conditions actually *are* linear, and to that class of cases we now turn.

10.3.8 The deterministic LQ control problem

In this section, we will denote the transpose of A by A'.

Consider a Pareto optimal economy where the representative agent solves

$$\begin{aligned} &\max_{\mathbf{u}} \left\{ \frac{1}{2} \sum_{t=0}^{T} x_t' Q x_t + u_t' R u_t \right\} \\ &s.t. \begin{cases} x_{t+1} = Ax_t + Bu_t \\ x_0 \text{ given.} \end{cases} \end{aligned} \tag{10.103}$$

This is a linear-quadratic problem in the sense that the objective function is a quadratic form and the constraint is a system of linear equations. Without loss of generality Q and R are symmetric.[2] Notice also that there is no discounting and no cross-products between states and controls. This in fact involves no loss of generality; we can transform a problem with discounting and cross-products to one without any of these; see [5]. Nevertheless, this transformation is not practical in concrete cases; here we exploit it only to make a theoretical point.

[2] If, say, Q should not be symmetric, just replace Q by $\frac{1}{2}(Q + Q')$ and we would have the same quadratic form with the symmetric matrix $\frac{1}{2}(Q + Q')$.

If you want to solve a particular problem, use the methods described in section 10.3.9.1 instead.

Checking concavity of Hamiltonian. Q, R negative semidefinite.

Now to find the optimality conditions, we maximize, for each $t = 0, 1, ..., T-1$,

$$\begin{aligned} &\frac{1}{2}x_t' Q x_t + \frac{1}{2}u_t' R u_t + \lambda_t' \left[x_{t+1} - Ax_t - Bu_t\right] + \\ &+\frac{1}{2}x_{t+1}' Q x_{t+1} + \frac{1}{2}u_{t+1}' R u_{t+1} + \lambda_{t+1}' \left[x_{t+2} - Ax_{t+1} - Bu_{t+1}\right] \end{aligned} \tag{10.104}$$

with respect to u_t and x_{t+1}. Differentiating with respect to u_t (rather than u_t') and x_{t+1} (rather than x_{t+1}'), we find that the first order conditions are

$$\begin{cases} Ru_t - B'\lambda_t &= 0 \\ \lambda_t + Qx_{t+1} - A'\lambda_{t+1} &= 0 \end{cases} \tag{10.105}$$

Assuming that R is invertible, the first block of equations says that

$$u_t = R^{-1}B'\lambda_t. \tag{10.106}$$

Now substitute this into the second block and into the constraint. We get

$$\begin{cases} -A'\lambda_{t+1} + Qx_{t+1} &= -\lambda_t \\ x_{t+1} &= BR^{-1}B'\lambda_t + Ax_t. \end{cases} \tag{10.107}$$

Combining these conditions, we have

$$\begin{bmatrix} -A' & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} \lambda_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} -I & 0 \\ BR^{-1}B' & A \end{bmatrix} \begin{bmatrix} \lambda_t \\ x_t. \end{bmatrix} \tag{10.108}$$

or

$$M \begin{bmatrix} \lambda_{t+1} \\ x_{t+1} \end{bmatrix} = N \begin{bmatrix} \lambda_t \\ x_t \end{bmatrix} \tag{10.109}$$

Now if M is invertible ($\Longleftrightarrow$ A is invertible), this can be solved as described in section 9.2.4. Just use the given value of x_0, and, if $T < \infty$, the transversality condition $x_{T+1} = 0$. If A is not invertible, use the methods described in section 8.5.

10.3.8.1 Uniqueness of the solution when $T = \infty$

Proposition 10.3.1 *Let A be invertible, let B be arbitrary, let Q, R be symmetric. Then the matrix $M^{-1}N$ defined via*

$$M^{-1}N = \begin{bmatrix} -A' & Q \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} -I & 0 \\ BR^{-1}B' & A \end{bmatrix} \tag{10.110}$$

has the same eigenvalues as a symplectic matrix. Note that with Harald's or Chow's timing convention, $M^{-1}N$ would have been *symplectic.*

Proof. Use the formula for the inverse of a partitioned matrix in [25] and then invoke the relevant result in section 8.4.

Corollary 10.3.1 *The eigenvalues of $M^{-1}N$ appear in reciprocal pairs, i.e. if λ_k is an eigenvalue, so is $\frac{1}{\lambda_k}$. It follows that there are just as many eigenvalues inside the unit circle as there are outside. Generically, half are strictly inside and half are strictly outside the unit circle (the only exception is if there are unit eigenvalues).*

So, generically, the system of difference equations that comes out of an LQ control problem has a unique stable solution! Indeed, we can solve for the feedback solution in a remarkably simple way. The idea will be to use the fact that we are always on the saddle path to determine a relationship between x_t and λ_t. That will then define the feedback rule!

Let $M^{-1}N$ be diagonalizable, and let Ω be a matrix of linearly independent eigenvectors of Ω, ordered so that the eigenvectors associated with unstable eigenvalues comes first. Now partition the $2n \times 2n$ matrix Ω^{-1} into $n \times n$ blocks as follows.

$$\Omega^{-1} = \begin{bmatrix} \Omega^{11} & \Omega^{12} \\ \Omega^{21} & \Omega^{22} \end{bmatrix}. \tag{10.111}$$

We know (see section 9.1.4.6) that any point $\begin{bmatrix} \lambda_t \\ x_t \end{bmatrix}$ on the saddle path satisfies

$$\Omega^{11}\lambda_t + \Omega^{12}x_t = 0. \tag{10.112}$$

Hence if Ω^{11} is invertible, we have

$$\lambda_t = -\left(\Omega^{11}\right)^{-1}\Omega^{12}x_t \tag{10.113}$$

and hence the (time independent!) feedback representation of the solution to our dynamic optimization problem is

$$u_t = -R^{-1}B'\left(\Omega^{11}\right)^{-1}\Omega^{12}x_t. \tag{10.114}$$

Note that this solution can be extended to the stochastic case as well. In that case, we will use slightly different methods (see section 10.3.9.1), but the intuition remains the same: the feedback solution comes from the requirement of always being on the saddle path, even when that path jumps around as it does in the stochastic case.

Example 10.3.2 *This example shows how far it is possible to get without solving a problem explicitly. Consider a firm that faces a given fixed interest rate r and maximizes the discounted sum of present and future profits. Suppose initial capital k_0 is given and that the problem of the firm is to maximize*

$$\sum_{t=0}^{\infty}(1+r)^{-t}\pi_t \tag{10.115}$$

where

$$\pi_t = f(k_t) - g(k_t, i_t) \tag{10.116}$$

and

$$k_{t+1} = (1-\delta)k_t + i_t. \tag{10.117}$$

The production function f is defined via

$$f(k) = k - \frac{1}{2}k^2. \tag{10.118}$$

This means that production peaks at $k = 1$, so it is pointless to accumulate any more capital than that. Meanwhile, the investment cost function g is given by

$$g(k, i) = i + \frac{a}{2}(i - \delta k)^2. \tag{10.119}$$

The second term of the right hand side expresses the assumption that it is costly to change the capital stock, and that this cost is convex in the size of the change. Note that the extent to which it is costly to change the capital stock is determined by the parameter a. Obviously, we assume that $a > 0$ and $r > 0$.

We will now investigate the properties of the solution to this problem. We begin by maximizing

$$\mathcal{L}_t = \pi_t + \mu_t\left[(1-\delta)k_t + i_t - k_{t+1}\right] + (1+r)^{-1}\pi_{t+1} + (1+r)^{-1}\mu_{t+1}\left[(1-\delta)k_{t+1} + i_{t+1} - k_{t+2}\right]. \tag{10.120}$$

The first order conditions are

$$\begin{cases} -\mu_t + (1+r)^{-1}\left[1 - k_{t+1} + a\delta(i_{t+1} - \delta k_{t+1}) + (1-\delta)\mu_{t+1}\right] = 0 \\ -1 - a(i_t - \delta k_t) + \mu_t = 0. \end{cases} \tag{10.121}$$

The second row allows us to solve for the control i_t. We get

$$i_t = \delta k_t + \frac{1}{a}(\mu_t - 1) \tag{10.122}$$

which means that investment is equal to replacement plus $(1/a)$ times the deviation of the shadow price of capital from 1. So the magnitude of the change in the capital stock depends negatively on a, which expresses the cost of changing the capital stock. Moreover, investment is increasing in the shadow price of capital.

Substituting the expression for i_t into the first row and into the law of motion $k_{t+1} = (1-\delta)\, k_t + i_t$, we get, in matrix format,

$$\begin{bmatrix} (1+r)^{-1} & -(1+r)^{-1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{t+1} \\ k_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{a} & 1 \end{bmatrix} \begin{bmatrix} \mu_t \\ k_t \end{bmatrix} + \begin{bmatrix} -(1+r)^{-1}(1-\delta) \\ -\frac{1}{a} \end{bmatrix}. \tag{10.123}$$

Our first goal is to find the steady state. The logic is to drop the time subscripts, notice that our equation then has the form $Ax = Bx + b$ and calculate $x^* = (A-B)^{-1} b$. We get

$$\begin{bmatrix} \mu^* \\ k^* \end{bmatrix} = \begin{bmatrix} (1+r)^{-1} - 1 & -(1+r)^{-1} \\ -\frac{1}{a} & 0 \end{bmatrix}^{-1} \begin{bmatrix} -(1+r)^{-1}(1-\delta) \\ -\frac{1}{a} \end{bmatrix} = \begin{bmatrix} 1 \\ 1-r-\delta \end{bmatrix}. \tag{10.124}$$

It follows that

$$i^* = \delta k^* \tag{10.125}$$

so that investment in the steady state is just enough to maintain the capital stock constant. Also, the steady state capital stock k^* is such that $f'(k^*) = r + \delta$, which is a natural result. The rewards of saving are then just equal to the costs, r being the opportunity cost of omitting to lend abroad, and δ being the depreciation rate.

The next step is to analyze the dynamics around the steady state, and those dynamics are determined by the matrix

$$M = \begin{bmatrix} (1+r)^{-1} & -(1+r)^{-1} \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ \frac{1}{a} & 1 \end{bmatrix} = \begin{bmatrix} 1+r+\frac{1}{a} & 1 \\ \frac{1}{a} & 1 \end{bmatrix}. \tag{10.126}$$

We now want to establish two facts: (1) the solution is a saddle path, and (2), the speed of convergence is decreasing in a (or, equivalently, the stable eigenvalue is increasing in a).

The first step is to show that both eigenvalues are real and positive. To see

this, note that they solve the characteristic equation

$$\lambda^2 - \mathrm{tr}\,(M) \cdot \lambda + \det(M) = 0. \tag{10.127}$$

Hence the solutions are both real if

$$\mathrm{tr}^{\,2}(M) - 4\det(M) > 0. \tag{10.128}$$

In our case,

$$\mathrm{tr}\,(M) = 2 + r + \frac{1}{a} \tag{10.129}$$

and

$$\det(M) = 1 + r. \tag{10.130}$$

Consequently

$$\mathrm{tr}^{\,2}(M) - 4\det(M) = r^2 + \frac{1}{a^2} + \frac{4}{a} + \frac{2r}{a} > 0. \tag{10.131}$$

Hence our eigenvalues are both real. To see that they are positive, note that

$$\lambda_1 + \lambda_2 = \mathrm{tr}\,(M) = 2 + r + \frac{1}{a} > 0 \tag{10.132}$$

and

$$\lambda_1 \lambda_2 = 1 + r > 0 \tag{10.133}$$

so both their sum and their product are positive. Hence they are both positive. Now to see that one is stable and one unstable, note that if they are both positive, the stability of an eigenvalue (whether $|\lambda| < 1$) is determined by whether it is greater or less than $+1$. To show that the eigenvalues are on opposite sides of $+1$, note that

$$(1 - \lambda_1)(1 - \lambda_2) = 1 - \mathrm{tr}\,(M) + \det(M) = -\frac{1}{a} < 0. \tag{10.134}$$

So far, then, we have shown that the solution is a saddle path. The final step is to investigate the speed of convergence, and this is of course governed by the stable

eigenvalue, which we know to be between 0 *and* +1*. Denote the stable eigenvalue by* λ_1 *and the unstable eigenvalue by* λ_2*. We know that*

$$\begin{cases} \lambda_1 + \lambda_2 = 2 + r + \dfrac{1}{a} \\ \lambda_1 \lambda_2 = 1 + r. \end{cases} \tag{10.135}$$

It follows that

$$\frac{\partial \lambda_1}{\partial a} = \frac{\lambda_1}{\lambda_2 - \lambda_1} \cdot \frac{1}{a^2} > 0 \tag{10.136}$$

so that convergence is slower the greater the cost of changing the capital stock is.

Also, it is tempting to conjecture that as $a \to 0$*, convergence is immediate, so that* $k_{t+1} = 0 \cdot k_t + k^*$*. Similarly, as* $a \to \infty$*, we should have* $k_{t+1} = k_t + 0 \cdot k^*$ *so that the capital stock is constant when it is infinitely costly to change it. It is actually not hard to confirm these conjectures. To do it, we note that*

$$\lambda_1 = \frac{2 + r + \dfrac{1}{a}}{2} - \sqrt{\left(\frac{2 + r + \dfrac{1}{a}}{2} \right)^2 - (1 + r)} \tag{10.137}$$

Letting $a \to \infty$*, the* $\dfrac{1}{a}$ *terms disappear, and we get*

$$\lim_{a \to \infty} \lambda_1 = \frac{2 + r}{2} - \sqrt{\left(\frac{2 + r}{2} \right)^2 - (1 + r)} = 1. \tag{10.138}$$

Conversely, when $a \to 0$*, the* $\dfrac{1}{a}$ *terms become dominant (and under the root sign, the* $\dfrac{1}{a^2}$ *terms become dominant). Hence*

$$\lim_{a \to 0} \lambda_1 = \frac{1/a}{2} - \sqrt{\left(\frac{1/a}{2} \right)^2} = 0. \tag{10.139}$$

Finally, let's write down the solution in terms of the stable eigenvalue. A stable eigenvector x *satisfies* $\dfrac{1}{a} x_1 + x_2 = \lambda_1 x_2$*. Thus, along a stable eigenvector* x*, we have* $x_1 = -a(1 - \lambda_1) x_2$*. Hence* $\mu_t - \mu^* = -a(1 - \lambda_1)(k_t - k^*)$ *for all* $t = 0, 1, ...$ *Recalling that* $\mu^* = 1$*, it follows that*

$$\mu_t = 1 - a(1 - \lambda_1)(k_t - k^*) \tag{10.140}$$

and consequently

$$i_t = \delta k_t - (1 - \lambda_1)(k_t - k^*) \tag{10.141}$$

which expresses the fact that investment is equal to depreciation plus a change in the capital stock in the direction of the steady state and at the rate determined by the stable eigenvalue. The equilibrium dynamics of the capital stock now becomes

$$k_{t+1} = (1 - \delta) k_t + \delta k_t - (1 - \lambda_1)(k_t - k^*) = \lambda_1 k_t + (1 - \lambda_1) k^* \tag{10.142}$$

so that the rate of convergence is given by the stable eigenvalue.

Exercise 10.3.2 *Confirm that if Q and R are negative semidefinite, then the Hamiltonian associated with (10.103) is concave.*

Exercise 10.3.3 *Verify that, with Harald's or Chow's timing convention for the costate, the first-order conditions of a discrete-time linear-quadratic dynamic optimization problem without discounting and without cross-products between states and controls yield a dynamic system with a symplectic coefficient matrix.*

Exercise 10.3.4 *Consider*

$$\max_{\mathbf{u}} \sum_{t=0}^{\infty} \left[-\frac{1}{4} x_t^2 - \frac{1}{2} u_t^2 \right] \tag{10.143}$$

$$\tag{10.144}$$

$$s.t. \begin{cases} x_{t+1} = x_t + u_t \\ x_0 \text{ given} \\ NPG. \end{cases} \tag{10.145}$$

Write down the first order conditions as a dynamic system either in λ_t and x_t or in u_t and x_t. Find the saddle path of this system (spanned by the stable eigenvector!). Hence verify that the unique stable solution to the first-order conditions is characterized by

$$u_t = -\frac{1}{2} x_t. \tag{10.146}$$

Verify that this feedback rule really solves the problem by checking that it satisfies the transversality condition.

10.3.9 Stochastic case

The remarkable thing about the stochastic case is how simple it is: all you do is stick in a conditional expectation; otherwise the optimality conditions are the same.

Let $\left(\Omega, \mathcal{F}, P, \langle \mathcal{F}_t \rangle_{t=0}^{T}\right)$ be a filtered probability space, and let $\langle z_t \rangle_{t=0}^{\infty}$ be an adapted stochastic process on this space. Now consider

$$\max_{\mathbf{u}} E\left[\sum_{t=0}^{T} f\left(t, x_t, u_t, z_t\right)\right]$$

$$\text{s.t.} \begin{cases} x_{t+1} = g\left(t, x_t, u_t, z_t\right) \\ u_t \in \mathcal{F}_t \\ x_0 \text{ given} \\ \text{NPG} \end{cases} \tag{10.147}$$

There are a couple of things that are new here.

In the first place, the maximand is an unconditionally expected value. Note that taking the expectation means that we are maximizing a real-valued function. Without the expectations operator, the maximand would take values in the set of stochastic variables, and the optimization problem would make no sense. Note that it is conventional to maximize not the unconditionally expected value but the conditionally expected value $E_0\left[\sum_{t=0}^{T} f\left(t, x_t, u_t, z_t\right)\right] = E\left[\sum_{t=0}^{T} f\left(t, x_t, u_t, z_t\right) | \mathcal{F}_0\right]$. This convention is slightly strange, since unless $\mathcal{F}_0 = \{\emptyset, \Omega\}$, we would then be trying to maximize a function that does not take values in $\mathbb{R}$.

There is a very important sense in which it does make sense to maximize a

conditionally expected value, however! Suppose for example that our information set $\mathcal{G}$ is generated by a finite partition $\mathbb{P} = \{P_1, P_2, ..., P_n\}$. Let Z be a random variable. Then solving the maximization problem

$$\max_X E\left[f\left(X, Z\right)|\mathcal{G}\right] \tag{10.148}$$

should be interpreted as solving, for each P_k, the problem

$$\max_{x_k} f\left(x_k, z_k\right) \tag{10.149}$$

where z_k is chosen such that $Z\left(\omega\right) = z_k$ for each $\omega \in P_k$. Note that the solution is a random variable X which is measurable with respect to $\mathcal{G}$. In fact, this random variable solves the problem

$$\max_{X \in \mathcal{G}} E\left[f\left(X, Z\right)\right]. \tag{10.150}$$

We are now in a position to *define* what it means to maximize a conditionally expected value even when our information set is not generated by a partition.

Definition 10.3.2 *Let $\mathcal{G} \subset \mathcal{F}$ be a $\sigma-$algebra and consider the maximization problem*

$$\max E\left[f\left(X, Z\right)|\mathcal{G}\right]. \tag{10.151}$$

The solution to this problem is defined as the random variable X which solves

$$\max_{X \in \mathcal{G}} E\left[f\left(X, Z\right)\right]. \tag{10.152}$$

To illustrate the usefulness of the language of maximizing a conditionally expected value, we note that this is a good description of what we do in practice when we solve problems like (10.152) and $\mathcal{G}$ is generated not by a partition but a random variable Y.

Proposition 10.3.2 *Let Y be a random variable and let $\mathcal{G} = \sigma(Y)$. Now consider the maximization problem*

$$\max_{X \in \mathcal{G}} E[f(X, Z)] \tag{10.153}$$

Now consider the random variable

$$E[f(X, Z) | \mathcal{G}] \tag{10.154}$$

For each fixed random variable X, this random variable is a function of Y alone. Hence it is a function of X and Y alone, and we may write $E[f(X, Z) | \mathcal{G}] = g(X, Y)$. Similarly X is a function of Y alone since it is $\mathcal{G}-$measurable, and we write $X = d(Y)$. Now let the function d be such that, for each real number y, the real number $x = d(y)$ maximizes the real-valued function $g(x, y)$. Then $X = d(Y)$ solves (10.153).

Proof. That X is $\mathcal{G}$-measurable is guaranteed by definition. The remainder of the proof is easy if Y is simple (and hence $\mathcal{G}$ is generated by a partition) and is left as an exercise. The case when Y is not simple is not so easy and is omitted here. ■

Another piece of news is the informational restriction $u_t \in \mathcal{F}_t$. The notation here is of course abusive; formally, we mean that, for each t, u_t is measurable with respect to $\mathcal{F}_t$, i.e. that the stochastic process $\langle u_t \rangle_{t=0}^T$ is adapted to the filtration $\langle \mathcal{F}_t \rangle_{t=0}^T$. Substantively, this is a restriction that prevents the agent from basing her decisions on things she doesn't yet know, e.g. the obviously (?) infeasible stock trading strategy 'buy low, sell high'. Note that this restriction is usually omitted, but that such an omission is a serious mistake that does not become less serious by being common practice among economists.

Also, we need to state what we mean by the NPG condition. Clearly it is not enough that it should hold in unconditional expectation; there shouldn't

be states of the world where we leave behind debt. Instead we require that it hold either almost surely or in $\mathcal{L}^2$. Note that Harald instead requires it to hold in conditional expectation with respect to $\mathcal{F}_\tau$ for each τ. Requiring it to hold almost surely implies this.

Finally, note that the problem is set up in such a way that x_t is predetermined (why?), i.e. x_0 is given and its one-period-ahead prediction error $x_{t+1} - E[x_{t+1}|\mathcal{F}_t]$ is zero.

It is now time to write down the sufficient conditions for an equilibrium.

Definition 10.3.3 *Let* $\left(\Omega, \mathcal{F}, P, \langle \mathcal{F}_t \rangle_{t=0}^T\right)$ *be the filtered probability space associated with (10.87). An* admissible allocation $(\mathbf{x}, \mathbf{u})$ *at market prices* $\boldsymbol{\lambda}$ *for the economy described by (10.87) is a pair of vector-valued stochastic processes sequences* $\langle x_t \rangle_{t=0}^{T+1}$ *and* $\langle u_t \rangle_{t=0}^T$ *such that*

1. $u_t \in U$ *for each* $t = 0, 1, ..., T$ *and all* $\omega \in \Omega$.
2. $x_{t+1} = g(t, x_t, u_t, z_t)$ *for each* $t = 0, 1, ..., T$ *and all* $\omega \in \Omega$.
3. *Almost surely (P) we have* $\lambda_T x_{T+1} = 0$ *if* $T < \infty$ *and* $\lim_{t \to \infty} \lambda_t x_{t+1} = 0$ *if* $T = \infty$.

Theorem 10.3.3 *Let* $\left(\Omega, \mathcal{F}, P, \langle \mathcal{F}_t \rangle_{t=0}^T\right)$ *be the filtered probability space associated with (10.87). Let* $(\mathbf{x}^*, \mathbf{u}^*)$ *be an admissible allocation of the economy described by (10.87). Let* $\boldsymbol{\lambda} = \langle \lambda_t \rangle_{t=0}^{T-1}$ *be an* $n-$*dimensional adapted stochastic process. Moreover, suppose the set* $U \in \mathbb{R}^m$ *is convex. Now define the function* H *via*

$$H(t, x, u, z, \lambda) = f(t, x, u, z) + \lambda \cdot g(t, x, u, z) \tag{10.155}$$

for $t = 0, 1, ..., T-1$ *and, if* $T < \infty$,

$$H(T, x, u, z, \lambda) = f(T, x, u, z). \tag{10.156}$$

Suppose now that H is continuously differentiable with respect to x on its entire domain. Suppose also that $H(t, x, u, z_t, \lambda_t)$ is concave in (x, u) for each $t = 0, 1, ..., T$. (Note that the concavity property only has to hold for our particular choice of $\mathbf{z}$, and $\boldsymbol{\lambda}$, not generally.) Finally, suppose that

1. $E\left[\left.\frac{\partial H\left(t+1, x_{t+1}^*, u_{t+1}^*, p_{t+1}, \lambda_{t+1}\right)}{\partial x}\right| \mathcal{F}_t\right] + \lambda_t = 0$ *for all* $t = 0, 1, ..., T-1$,
2. $u_t^* \in \underset{u \in U}{\operatorname{argmax}} H(t, x_t^*, u, z_t, \lambda_t)$, *for all* $t = 0, 1, ..., T$ *and all* $\omega \in \Omega$ *and*
3. *Almost surely (P),* $\lambda_T \cdot x_{T+1} = 0$ *if* $T < \infty$ *and* $\lim_{t \to \infty} \lambda_t \cdot x_{t+1} = 0$ *if* $T = \infty$.

Then $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{z}, \boldsymbol{\lambda})$ is a competitive equilibrium.

Remark 10.3.4 *Note that our notion of competitive equilibrium is that of Arrow and Debreu: trade in goods distinguished by physical properties, date and contingency takes place in meta-time.*

Hint. It is much easier to remember maximizing

$$E\left[f_t + \lambda_t \cdot [g_t - x_{t+1}] + f_{t+1} + \lambda_{t+1} \cdot [g_{t+1} - x_{t+2}] \middle| \mathcal{F}_t\right] \tag{10.157}$$

than to remember the above theorem.

Remark 10.3.5 *In concrete cases, we always let the filtration be generated by the driving process $\langle z_t \rangle_{t=0}^T$. This means that we can calculate the conditional expectation with respect to $\mathcal{F}_t$ by just treating all variables with a subscript t or smaller (and x_{t+1}!) as deterministic.*

When calculating the solution to a stochastic dynamic optimization problem, the only sensible way of representing the solution is as a feedback rule. Otherwise we would have to write down not just a function of t but of ω as well. However,

such a recursive representation does not always exist. Why might such a representation fail to exist? Well, let's take a silly example. Let $\Omega = \{1, 2\}$, let $\mathcal{F} = \mathcal{F}_0 = \mathcal{F}_1 = 2^\Omega$, let $P(\{1\}) = \frac{1}{2}$, let $z_0 \equiv 0$ and let $z_1(1) = 1$ $z_1(2) = 0$. Now suppose we want to solve

$$\begin{aligned} &\max E\left[-(x_1 - z_1)^2\right] \\ &\text{s.t.} \begin{cases} x_{t+1} = u_t \\ x_0 \text{ given} \\ x_2 = 0 \end{cases} \end{aligned} \tag{10.158}$$

Clearly the solution is $u_0(1) = 1$, $u_0(2) = 0$ and $u_1 \equiv 0$. This yields a maximand of zero. But u_t is not a function of (t, x_t, z_t). In particular, u_0 uses the information in $\mathcal{F}_0$ which is not contained in $\sigma(x_0, z_0)$. Clearly this kind of problem would disappear if we set $\langle \mathcal{F}_t \rangle_{t=0}^T$ to be the filtration generated by the (possibly random) x_0 and $\langle z_t \rangle$ so that $\mathcal{F}_t = \sigma(\{x_0, z_0, z_1, ..., z_t\})$. But it may still be the case that there is no feedback representation, since we may need information about the whole history of z_t to calculate the conditional expectation of future values of z_t.

A feedback representation of the solution does, however, exist when $\langle \mathcal{F}_t \rangle$ is generated by $\langle z_t \rangle$ and $\langle z_t \rangle$ is a *Markov process* with respect to this natural filtration. Then the solution has a representation $u_t = d(t, x_t, z_t)$. In particularly auspicious cases, we can drop the dependence on t, and if we amalgamate x_t and z_t into a single state vector s_t, the solution can be written as $u_t = d(s_t)$. This will happen in infinite-horizon problems driven by time-homogeneous Markov processes where the time dependence in f (if any) takes the form of geometric discounting and g is time-independent. A famous example is given by the infinite-horizon stochastic LQ control problem.

10.3.9.1 The stochastic LQ control problem

Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space. Let $\mathbf{z}$ be a stochastic process and let $\underline{\mathcal{F}}$ be the filtration generated by $\mathbf{z}$. For simplicity, let z_0 be deterministic so that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. We write $\mathbf{u} \in \underline{\mathcal{F}}$ to mean that $\mathbf{u}$ is adapted to $\underline{\mathcal{F}}$. Now consider

$$\begin{aligned} &\max_{\mathbf{u}\in\underline{\mathcal{F}}} E\left[\frac{1}{2}\sum_{t=0}^{\infty}\beta^t \begin{bmatrix} x_t' & u_t' & z_t' \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}' & Q_{22} & Q_{23} \\ Q_{13}' & Q_{23}' & Q_{33} \end{bmatrix} \begin{bmatrix} x_t \\ u_t \\ z_t \end{bmatrix}\right] \\ &\text{s.t.} \begin{cases} x_{t+1} = Ax_t + Bu_t + Cz_t \\ z_{t+1} = \Phi z_t + \varepsilon_{t+1} \\ x_0, z_0 \text{ given} \\ \text{NPG} \end{cases} \end{aligned} \tag{10.159}$$

where Φ is a stable matrix, $\boldsymbol{\varepsilon}$ is a vector-valued martingale difference (white noise) process with respect to $(P, \underline{\mathcal{F}})$ and x_0 is an exogenous deterministic vector. Note that the unconditional expected value of z_t is 0; this is without loss of generality since we can always consider its deviation from the mean. Note that $E[z_{t+1}|\mathcal{F}_t] = \Phi z_t$.

Defining λ_t as the current value, the optimality conditions are

$$\begin{cases} Q_{22}u_t + Q_{12}'x_t + Q_{23}z_t + B'\lambda_t = 0 \\ E\left[-\lambda_t + \beta Q_{11}x_{t+1} + \beta Q_{12}u_{t+1} + \beta Q_{13}z_{t+1} + \beta A'\lambda_{t+1}|\mathcal{F}_t\right] = 0 \end{cases} \tag{10.160}$$

Now suppose Q_{22} is invertible. Then the first row says

$$u_t = -Q_{22}^{-1}\left[Q_{12}'x_t + Q_{23}z_t + B'\lambda_t\right] \tag{10.161}$$

Substituting this into the second row and the constraint, we get the system of

linear expectational difference equations

$$\begin{bmatrix} \beta A' - \beta Q_{12}Q_{22}^{-1}B' & \beta Q_{11} - \beta Q_{12}Q_{22}^{-1}Q_{12}' \\ 0 & I \end{bmatrix} E\left[\left[\begin{array}{c} \lambda_{t+1} \\ x_{t+1} \end{array}\right]\middle|\mathcal{F}_t\right] =$$

$$= \begin{bmatrix} -I & 0 \\ BQ_{22}^{-1}B' & BQ_{22}^{-1}Q_{12}' - A \end{bmatrix} \begin{bmatrix} \lambda_t \\ x_t \end{bmatrix} + \begin{bmatrix} -\beta Q_{12}^{-1}Q_{23}\Phi + \beta Q_{13}\Phi \\ BQ_{22}^{-1}Q_{23} - C \end{bmatrix} z_t. \tag{10.162}$$

or

$$ME\left[\left[\begin{array}{c} \lambda_{t+1} \\ x_{t+1} \end{array}\right]\middle|\mathcal{F}_t\right] = N\begin{bmatrix} \lambda_t \\ x_t \end{bmatrix} + Lz_t \tag{10.163}$$

and provided M is invertible, this can be solved for the unique stable solution using the (recursive!) methods in section 9.2.8. When checking the transversality condition, note that λ_t is the current value so that the transversality condition becomes $\lim_{t\to\infty} \beta^t \lambda_t' x_{t+1} = 0$ a.s. (P). Now if the vector process $\begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{x} \end{bmatrix}$ is stable (i.e. is bounded in $\mathcal{L}^2$), then it is not hard to believe that this holds. The proof looks at $E\left[\left|\beta^t \lambda_t' x_{t+1}\right|\right] = \beta^t E\left[|\lambda_t' x_{t+1}|\right]$, notes that this tends to zero since $E\left[|\lambda_t' x_{t+1}|\right]$ is bounded.

Note the certainty equivalence result: the decision rule is the same regardless of the variance of z_t. This means that LQ approximation/linearization of optimality conditions is unsuitable for the analysis of problems where the point is to analyze how risk aversion affects decision-making and asset-pricing.

Alternative to the approach here: Drop z_t and let $x_t = Ax_t + Bu_t + C\varepsilon_{t+1}$ where $\langle \varepsilon_t \rangle$ is exogenous $(P, \underline{\mathcal{F}})-$white noise. Then the initial value x_0 is given and the prediction error of x_t is not zero but exogenous, which is what matters for there to be a unique solution (see [29]).

If you should run into invertibility or diagonalizability problems, see [29].

Indeed, the methods described there always work, so you can actually ignore section 9.2.8 if you want.

10.3.10 Bellman's approach

10.3.10.1 Introduction and motivation

The only relevance of Bellman's approach to dynamic optimization is to understand what other people are doing; I don't recommend ever using it in practice.[3] Possibly (but even this is dubious), it might also enhance conceptual understanding.

It is widely believed that there is a conceptual link between a Markov (feedback, recursive) representation of the solution to a dynamic optimization problem and Bellman's equation (see below). However, as we have seen, it is possible to discuss and prove the existence of optimal Markov decision rules with no reference to Bellman.

Apparently the Lagrange approach can be used even in a class of cases that the Bellman approach seems to be tailor-made for: calculating the subgame-perfect equilibrium of a dynamic game. Nevertheless, it should be conceded that in this case the Bellman equation *is* important conceptually in that it is the natural language to use when *defining* the notion of a subgame perfect equilibrium. Dynamic games are discussed in [12], [4] and [11] but won't be discussed further here. We will confine ourselves to competitive economies, i.e. economies where each optimizing agent is small.

[3] Bellmans metod påminner en hel del om Bellmans upptåg i Bellman-historier: den är originell och infallsrik, men i praktiken mindre bra. Min bestämda rekommendation är alltså att hålla sig till franskens (Lagranges) metod.

10.3.10.2 The principle of optimality

10.3.10.2.1 General case Let $(\Omega, \mathcal{F}, P, \underline{\mathcal{F}})$ be a filtered probability space. Consider first the stochastic dynamic optimization problem

$$\begin{aligned} &\max E\left[\sum_{t=0}^{T} f\left(t, x_t, u_t, z_t\right)\right] \\ &s.t. \begin{cases} x_{t+1} = g\left(t, x_t, u_t, z_t\right) \\ x_0 \text{ given} \\ \text{NPG} \end{cases} \end{aligned} \tag{10.164}$$

As in the Lagrange multiplier approach, the idea behind Bellman's approach is to solve a separate low-dimensional problem for each t rather than a single high-dimensional (or, if $T = \infty$, infinite-dimensional) problem. The final step, as before, is to tie these problems together by a kind of difference equation.

Definition 10.3.4 *Define a sequence of value functions $\langle V_t \rangle$ via*

$$\begin{aligned} V_t(x) &= \max_{\mathbf{u}} E\left[\sum_{s=t}^{T} f\left(s, x_s, u_s, z_s\right) \middle| \mathcal{F}_t\right] \\ s.t &\begin{cases} x_{s+1} = g\left(s, x_s, u_s\right) \\ x_t = x \end{cases} \end{aligned} \tag{10.165}$$

Remark 1 *With this definition of the value function, we have that, for each fixed x, $\langle V_t(x) \rangle$ is an $\underline{\mathcal{F}}$-adapted stochastic process. If $\langle x_t \rangle$ is an $\underline{\mathcal{F}}$-adapted stochastic process, then $\langle V_t(x_t) \rangle$ is another $\underline{\mathcal{F}}$-adapted stochastic process, and $\left\langle \frac{\partial}{\partial x} V_t(x_t) \right\rangle$ is yet another $\underline{\mathcal{F}}$-adapted stochastic process.*

Theorem 10.3.4 (Bellman's principle of optimality) *Note that, in this theorem, x' is not the transpose of x, but just another variable. Consider the problem 10.164 and its associated value functions $\langle V_t \rangle$. We have, for each $t = 0, 1, ..., T$,*

$$\begin{aligned} V_t(x) \equiv \sup_u &\left\{ f\left(t, x, u, z_t\right) + E\left[V_{t+1}\left(x'\right) | \mathcal{F}_t\right] \right\} \\ &s.t.\ x' = g\left(t, x, u, z_t\right). \end{aligned} \tag{10.166}$$

Remark 10.3.6 *Bellman's equation is a functional equation. The equality holds for each x.*

Remark 10.3.7 *Think about what Bellman's equation means intuitively until you believe it.*

We now notice that Bellman's equation says no more than what we already know. To see this, consider the right-hand-side maximization problem. The FOC is

$$f_u(t, x_t, u_t, z_t) + E\left[\frac{\partial}{\partial x'} V_{t+1}(x_{t+1}) \cdot g_x(t, x_t, u_t, z_t) \,|\mathcal{F}_t\right] = 0. \tag{10.167}$$

Inspired by a stochastic version of the envelope theorem, we write

$$E\left[\frac{\partial}{\partial x'} V_{t+1}(x_{t+1}) \,|\mathcal{F}_t\right] = \lambda_t \tag{10.168}$$

and this FOC becomes

$$f_u(t, x_t, u_t, z_t) + \lambda_t \cdot g_x(t, x_t, u_t, z_t) \tag{10.169}$$

which is just what we established before (for an interior solution). Moreover, since Bellman's equation is an identity, we can differentiate it with respect to x, and it then remains true. Evaluating at the optimum, we drop the sup operator. We get, using our λ notation and stepping forward one step,

$$\lambda_t = E\left[f_x(t+1, x_{t+1}, u_{t+1}, z_{t+1}) + \lambda_{t+1} \cdot g_x((t+1, x_{t+1}, u_{t+1}, z_{t+1})) \,|\mathcal{F}_t\right] \tag{10.170}$$

which again is a result we have seen before. Lagrange takes the essential element of Bellman! All we need is the derivative of the value function, not the value function itself!

10.3.10.2.2 Markov processes Let $\mathbf{z}$ be a Markov process and let $\underline{\mathcal{F}}$ be the filtration generated by $\mathbf{z}$. We can now make the value-function deterministic by baking all the dependence on chance into dependence on $z_t = z$.

$$\begin{aligned} V_t(x,z) = \max_{\mathbf{u}} E\left[\sum_{s=t}^{T} f(s,x_s,u_s,z_s)\middle| z_t = z\right] \\ s.t \begin{cases} x_{s+1} = g(s,x_s,u_s,z_s) \\ x_t = x \end{cases} \end{aligned} \tag{10.171}$$

In this case, Bellman's principle of optimality becomes

$$\begin{aligned} V_t(x,z) = \sup_u \{f(t,x,u,z) + E[V_{t+1}(x',z_{t+1})|z_t = z]\} \\ \text{s.t. } x' = g(t,x,u,z) \end{aligned} \tag{10.172}$$

The solution will now be representable by a Markov decision rule. The equilibrium allocation (u_t^*, x_t^*, z_t) becomes a Markov process as well.

10.3.10.2.3 Time homogeneous Markov processes Infinite horizon; time-independence except for geometric discounting. Stationarizing value function by defining it as the current value.

$$\begin{aligned} V(x,z) = \sup_u \{f(x,u,z) + \beta E[V(x',z')|z]\} \\ \text{s.t. } x' = g(x,u,z) \end{aligned} \tag{10.173}$$

where the slightly cryptic notation $E[V(x',z')|z]$ is the conditional expectation of $V(x_{t+1}, z_{t+1})$ given $z_t = z$; the reason for the z and z' notation is to stress the fact that this is the same function of z for each t.

Time independent Markov decision rules. $u_t^* = d(x_t, z_t)$. The equilibrium allocation (u_t^*, x_t^*, z_t) becomes a time homogeneous Markov process as well.

Existence of unique value function. Banach's theorem at work.

Chapter 11

Some numerical methods

11.1 Solving linear systems

The purpose of this section is to say something about how to solve a system of equations of the form

$$Ax = b. \tag{11.1}$$

In primary school, we learned how to solve systems of linear equations by using Gaussian elimination. Later in life, we learned how to do it by inverting a matrix. Numerically, it is worth going back to primary school because Gaussian elimination is the quickest and most precise method. The command in Matlab is $x = A\backslash b$, and in Gauss it is $x = b/A$. Avoid $x = \text{inv}\,(A) * b$. Even when you really do want to calculate the inverse of a matrix, Gaussian elimination is preferable. Write $A\backslash eye(n)$ in Matlab and similarly in Gauss.

If A is close to being singular, trying to solve (11.1) is not a good idea. Nor is it a good idea to check whether A is close to being singular by checking if its determinant is close to zero. For example, the matrix $A = 0.1I_{80}$ has the ridiculously small determinant 10^{-80}. Yet it can (obviously) be reliably inverted,

since $A^{-1} = 10I_{80}$. Instead, you might check the rank of the matrix by using the function rank in Gauss or Matlab. This function calculates the numerical rank of a matrix by calculating the number of so-called singular values of A whose modulus is greater than the computer's precision (roughly 10^{-13}), and the numerical rank of a matrix in this sense really does tell you whether solving (11.1) is something that can be reliably done. Getting even more directly to the point, you might check the condition number of the matrix as defined in section 9.1.4.5. It can be calculated by using the function cond in Matlab and if it is large (say greater than 10^{13}), then the matrix is close to being singular (it is then said to be *ill-conditioned*).

11.1.1 Solving sparse linear systems

Often we want to solve a high-dimensional linear system where the A matrix has lots of zeros in it. We then say that A is *sparse,* and it is important to exploit this sparseness, or you'll waste a lot of time and memory capacity. (The theory of this is discussed in [21].) The way forward in Matlab is to use the command sparse. Suppose your matrix A has n non-zero components at positions $\langle (i_1, j_1), (i_2, j_2), \cdots, (i_n, j_n) \rangle$ and that the values of those components are $\langle c_1, c_2, \cdots, c_n \rangle$. Then create A by writing

$$A = \text{sparse}\,(i, j, c) \tag{11.2}$$

and solve for x by writing $x = A \backslash b$.

As far as I know, Gauss has no *general* way to deal with sparse matrices, but you will find a method that works in a special case in the exercises.

Exercise 11.1.1 (The Hodrick-Prescott (HP) filter) *Suppose you have the empirical time series* $\langle x_t \rangle_{t=1}^{T}$ *and let's say you want to decompose this series into*

a trend and cyclical component in the manner of Hodrick and Prescott. Call the HP trend $\langle \tilde{x}_t \rangle_{t=1}^T$. *By definition, the HP trend solves*

$$\min_{\langle \tilde{x}_t \rangle_{t=1}^T} \left\{ \sum_{t=1}^{T} (x_t - \tilde{x}_t)^2 + \lambda \sum_{t=2}^{T-1} ((\tilde{x}_{t+1} - \tilde{x}_t) - (\tilde{x}_t - \tilde{x}_{t-1}))^2 \right\}. \tag{11.3}$$

The HP filter formalizes the trade-off between (1) the trend $\tilde{x}_t$ *tracking* x_t *closely (it should be the trend of* $\langle x_t \rangle_{t=1}^T$, *not of some other series) and (2)* $\tilde{x}_t$ *being smooth in the sense of having a near-constant rate of change (it should be a* trend, *after all). The parameter* λ *measures the relative weight attached to smoothness as against close tracking.*

1. *Show, by writing down the first order conditions for solving the above minimization problem, that these conditions constitute a pentadiagonal linear system. (A pentadiagonal system has non-zero coefficients in five adjacent diagonal bands, the middle one being the main diagonal.)*

2. *(Gauss or Matlab exercise) Take an arbitrary time series* $\langle x_t \rangle_{t=1}^T$ *(you can draw random numbers if you want) with* $T = 500$. *Set* $\lambda = 1600$. *Compare the amount of time it takes to solve for the HP trend when*

   (a) *you exploit the sparseness of the matrix and*

   (b) *you don't.*

   Report your results and your code.

   Hint for Matlab users: You can download sparseness-exploiting code from Ellen McGrattan's ftp site ([47]). Or just download the file hptrend2.m from my homepage.

   Hint for Gauss users: You can download sparseness-exploiting code (written by Simon van Norden) from an ftp site at the Université du Québec à Montréal. A link can be found on [45]. Or just download the file hpfilter.prg from my homepage.

Hint for everyone: One way of getting code that does not exploit sparseness is to modify the code that does.

11.2 Solving non-linear systems and optimizing

Being able to solve $f(x) = 0$ where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a non-linear function is often useful in itself. For example, the steady state conditions of a typical dynamic optimization problem turn out to be a finite-dimensional non-linear system of equations.It also has instrumental value when we want to solve an optimization problem, since the first-order conditions of a non-quadratic finite-dimensional optimization problem becomes a finite-dimensional non-linear system of the form $f(x) = 0$.

In any case, the most popular approach to this type of problem (but far from the only one) is the Newton-Rhapson method. It is used by Matlab's fsolve and Gauss's nlsys. The idea is the following. Call the solution x^*. Use Taylor's formula and write

$$f(x^*) \approx f(x_0) + f'(x_0) \cdot (x^* - x_0) \tag{11.4}$$

where, by definition, $f'(x_0)$ is an $n \times n$ matrix and $\cdot$ is the Euclidean inner product in $\mathbb{R}^n$. Note that this formula holds exactly when $x_0 = x^*$. By definition, $f(x^*) = 0$ so we may write

$$0 \approx f(x_0) + f'(x_0) \cdot (x^* - x_0). \tag{11.5}$$

Fiddling a bit with the formula, we find that

$$x^* - x_0 \approx -[f'(x_0)]^{-1} f(x_0) \tag{11.6}$$

or

$$x^* \approx x_0 - [f'(x_0)]^{-1} f(x_0). \tag{11.7}$$

Inspired by this formula, define the mapping $H : \mathbb{R}^n \to \mathbb{R}^n$ via

$$H(x) = x - [f'(x)]^{-1} f(x) \tag{11.8}$$

and note that our solution x^* is a fixed point of H. If we are in luck, H is a contraction, and we can use the constructive proof of Banach's fixed point theorem to suggest a good algorithm for approximating x^*. Simply take an arbitrary x_0 and keep on applying the function H.

Warning. Things can easily go wrong. For example, consider the scalar case and suppose there is a point x between x_0 and x^* at which $f'(x) = 0$. Then the Newton-Rhapson method carries you off in the wrong direction. (Illustrate geometrically!) A related problem is that we may hit a point x such that $f'(x)$ is singular along the way, so that the calculation of $[f'(x)]^{-1}$ crashes.

However, when you *are* close to a solution, Newton-Rhapson delivers quadratic convergence, i.e. there is a $c \geq 0$ such that

$$\|x_{k+1} - x^*\| \leq c \|x_k - x^*\|^2 . \tag{11.9}$$

Moral. 1. Choose an initial guess as close as possible to the solution!

2. If one x_0 leads to disaster, try another one!

You will have noticed that Newton-Rhapson requires you to calculate the derivatives in the matrix $f'(x)$. Indeed, suppose you are optimizing, say minimizing the scalar valued function $g(x)$ with respect to x. Then the gradient $f'(x)$ is the *Hessian* $g''(x)$.

Maximum speed and precision is achieved if you calculate the derivatives in $f'(x)$ analytically. If that is hard, however, you may want to approximate $f'(x)$ by finite difference quotients. That is the topic of the next section. But before

going on, it is worth issuing another warning. A natural way to proceed is to iterate on Newton-Rhapson until the difference

$$|x_{t+1} - x_t| \tag{11.10}$$

falls below some specified tolerance level and conclude that when it has, we are close in some sense to the solution. However, consider the counterexample

$$\sqrt[3]{x} \exp\left(-x^2\right) = 0. \tag{11.11}$$

The unique solution is $x^* = 0$, and to find it, Newton-Rhapson suggests iterating on

$$x_{t+1} = x_t + \frac{x_t}{2x_t^2 - \frac{1}{3}}. \tag{11.12}$$

However, this generates a divergent sequence for any $x_0 \neq 0$. Yet the increment $x_{t+1} - x_t$ converges to zero.

Exercise 11.2.1 (Solving for a steady state) *Consider the dynamic optimization problem*

$$\max_{\langle c_t, h_t \rangle_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \left[\alpha \ln c_t + (1-\alpha) \ln (1 - h_t)\right] \tag{11.13}$$

subject to

$$c_t + k_{t+1} = (1-\delta) k_t + k_t^{\theta} h_t^{1-\theta} \tag{11.14}$$

and a suitable no-Ponzi-scheme condition. Let k_0 be given.

1. *Derive the necessary/sufficient conditions for an optimum. Drop the t subscripts to find a system of equations characterizing a steady state.*

2. *(Gauss or Matlab exercise) Set $\alpha = 0.3$, $\beta = 0.99$, $\delta = 0.025$ and $\theta = 0.36$. Use Gauss or Matlab to find the steady state values of c_t, k_t and h_t.*

Report your results and your code.

11.3 Numerical derivatives

Often we want to differentiate functions which are so messy that pencil-and-paper differentiation of them is a nightmare. So instead we do it numerically.

The idea is to take a small $h > 0$ and calculate ratios of the form

$$\frac{\partial f_i(x)}{\partial x_j} \approx \frac{f_i(x + he_j) - f_i(x)}{h} \tag{11.15}$$

where e_j is the jth unit vector, i.e. a vector with zeros everywhere except at the jth position, where there is a one. These are forward differences. If you are really scrupulous, you might want to consider trying $h < 0$ (backward differences) or the central difference

$$\frac{\partial f_i(x)}{\partial x_j} \approx \frac{f_i(x + he_j) - f_i(x - he_j)}{h} \tag{11.16}$$

and see if it makes any difference (no pun intended).

In Gauss, use the function gradp . In Matlab, download the file grad.m from my homepage at [46]. You may also want to consider the gradient functions on Ellen McGrattan's ftp site at [47].

As for numerical Hessians, you are strongly recommended to avoid them if you can. Their precision is not always reliable. However, if you can't avoid them, use hessp in Gauss. Or if you are in the business of optimizing, go straight for the Gauss function optmum which minimizes a function by calculating numerical Hessians. In Matlab, check out Ellen McGrattan's ftp site and look for the file numder.m which calculates both numerical gradients and numerical Hessians. Or use Ellen's minimization function uncmin.m or something like it in the Optimization Toolbox.

Actually, since optimization functions are often more robust to silly initial guesses and ill-behaved functions than just doing Newton-Rhapson on the first-order conditions, you might want to use a minimization routine for solving systems

of non-linear equations, even if they are not the first-order conditions of any optimization problem. Or at least you might if speed is not your number one priority but getting a reasonable solution at all is. Just minimize the scalar-valued function $\|f(x)\|$.

Exercise 11.3.1 *(Gauss or Matlab exercise) Consider the dynamic optimization problem*

$$\max_{\langle c_t \rangle_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \ln c_t \tag{11.17}$$

subject to

$$c_t + k_{t+1} = (1-\delta) k_t + k_t^{\theta} \tag{11.18}$$

and a suitable no-Ponzi-scheme condition. Let k_0 be given. Set $\beta = 0.99$, $\delta = 0.025$ and $\theta = 0.36$.

(a) Use numerical derivatives to linearize the first order conditions around the steady state. To improve precision, linearize around the natural logarithms of the variables, i.e. let the linearized system be linear in the deviations of the natural logs from their steady states.

(b) Find the eigenvalues and eigenvectors of the relevant matrix numerically. Solve for an approximate decision rule by looking at the saddle path of the linearized dynamic system.

(c) When $\delta = 1$ there is an exact solution given by $\ln c_t - \ln c^* = \theta (\ln k_t - \ln k^*)$. Verify that your algorithm comes close to this solution when you set $\delta = 1$.

Bibliography

[1] Harald Lang: Comments on difference equations, http://www.math.kth.se/~lang/

[2] Harald Lang: Systems of differential equations, http://www.math.kth.se/~lang/

[3] Harald Lang: Dynamic Optimization, http://www.math.kth.se/~lang/

[4] Thomas J. Sargent: Dynamic macroeconomic theory

[5] Lars Peter Hansen and Thomas J. Sargent: Recursive Models of Dynamic Linear Economies, http://riffle.stanford.edu/hansen.html.

[6] Knut Sydsæter: Matematisk Analyse, bind I

[7] Knut Sydsæter: Matematisk Analyse, bind II

[8] Knut Sydsæter (and Peter Hammond): Mathematics for economic analysis

[9] Robert E. Lucas and Nancy Stokey: Recursive Methods in Economic Dynamics

[10] Atle Seierstad and Knut Sydsæter: Optimal Control Theory with Economic Applications

[11] Chow: Dynamic Economics

[12] Basar & Olsder: Dynamic non-cooperative game theory

[13] Alpha Chiang: Elements of Dynamic Optimization

[14] Leitmann: The Calculus of Variations and Optimal Control

[15] Kamien and Schwarz: Dynamic Optimization

[16] Luenberger: Optimization by vector space methods

[17] Walter Rudin: Principles of Mathematical Analysis

[18] Walter Rudin: Real and Complex Analysis

[19] Walter Rudin: Functional Analysis

[20] Gilbert Strang: Linear Algebra and its Applications

[21] Datta: Numerical Linear Algebra and Applications

[22] Nobel and Daniel: Applied Linear Algebra

[23] Golub & van Loan: Matrix Computations

[24] Tomas Björk: Stokastisk kalkyl och kapitalmarknadsteori

[25] Sune Karlsson: Matrisalgebra

[26] Jan R. Magnus and Heinz Neudecker: Matrix Differential Calculus with Applications in Statistics and Econometrics

[27] Andersson et al: Ordinära differentialekvationer

[28] Boyce and diPrima: Elementary Differential Equations and Boundary Value Problems

[29] Klein: Using the generalized Schur form to solve a system of linear expectational difference equations, http://www.iies.su.se/data/home/kleinp/homepage.htm.

[30] Kai Lai Chung: A course in probability theory

[31] P. Billingsley: Probability and Measure

[32] Brockwell and Davis: Time Series, Theory and Methods

[33] G. R. Grimmett and D. R. Stirzaker: Probability and Random Processes

[34] David Williams: Probability with Martingales

[35] Cinlar: Introduction to stochastic processes.

[36] Numerical recipes online, http://cfatab.harvard.edu/nr/nronline.html

[37] Burden & Faires: Numerical Analysis

[38] Frennemo et al: Elementär analys i en dimension.

[39] Wunsch: Complex Analysis with Applications

[40] Royden: Real Analysis

[41] Soo Bong Chae: Lebesgue Integration

[42] Jörgen Weibull: Lecture notes on mathematics for economics

[43] Laffont & Tirole: A Theory of Incentives in Procurement and Regulation

[44] Duffie: Dynamic Asset Pricing Theory

[45] Christian Zimmerman's code page, http://www.er.uqam.ca/nobel/r14160/rbc/codes.html

[46] Paul Klein's homepage, http://www.iies.su.se/data/home/kleinp/homepage.htm

[47] Ellen McGrattan's Matlab codes, ftp://res.mpls.frb.fed.us/pub/research/mcgrattan/mfiles

[48] Ellen McGrattan's lecture notes, various postscript files under ftp://res.mpls.frb.fed.us/pub/research/mcgrattan

[49] Priestley: The Spectral Analysis of Time Series

[50] Gérard Debreu: Theory of Value

Index