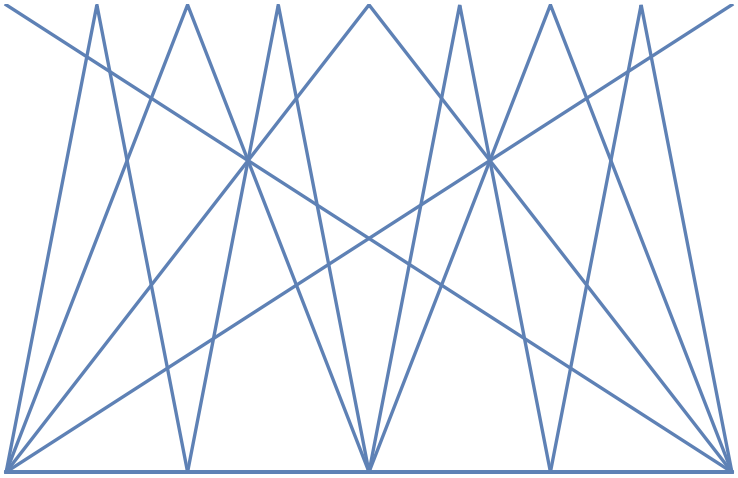
Topics in Real and Functional Analysis

Gerald Teschl



Graduate Studies
in Mathematics
Volume (to appear)

American Mathematical Society
Providence, Rhode Island

Gerald Teschl
Fakultät für Mathematik
Oskar-Mogenstern-Platz 1
Universität Wien
1090 Wien, Austria

E-mail: Gerald.Teschl@univie.ac.at
URL: http://www.mat.univie.ac.at/~gerald/

2010 Mathematics subject classification. 46-01, 28-01, 46E30, 47H10, 47H11, 58Fxx, 76D05

Abstract. This manuscript provides a brief introduction to Real and (linear and nonlinear) Functional Analysis. It covers basic Hilbert and Banach space theory as well as basic measure theory including Lebesgue spaces and the Fourier transform.

Keywords and phrases. Functional Analysis, Banach space, Hilbert space, Measure theory, Lebesgue spaces, Fourier transform, Mapping degree, fixed point theorems, differential equations, Navier–Stokes equation.

Typeset by $\mathcal{AMS}$-LaTeX and Makeindex.
Version: March 1, 2019

Contents

Preface

The present manuscript was written for my course *Functional Analysis* given at the University of Vienna in winter 2004 and 2009. It was adapted and extended for a course *Real Analysis* given in summer 2011, 2013, and 2018. The last part are the notes for my course *Nonlinear Functional Analysis* held at the University of Vienna in Summer 1998, 2001, and 2018. The three parts are essentially independent. In particular, the first part does not assume any knowledge from measure theory (at the expense of hardly mentioning L^p spaces).

It is updated whenever I find some errors and extended from time to time. Hence you might want to make sure that you have the most recent version, which is available from

`http://www.mat.univie.ac.at/~gerald/ftp/book-fa/`

Please do not redistribute this file or put a copy on your personal webpage but link to the page above.

Goals

The main goal of the present book is to give students a concise introduction which gets to some interesting results without much ado while using a sufficiently general approach suitable for further studies. Still I have tried to always start with some interesting special cases and then work my way up to the general theory. While this unavoidably leads to some duplications, it usually provides much better motivation and implies that the core material always comes first (while the more general results are then optional). Moreover, this book is *not* written under the assumption that it will be

read linearly starting with the first chapter and ending with the last. Consequently, I have tried to separate core and optional materials as much as possible while keeping the optional parts as independent as possible.

Furthermore, my aim is not to present an encyclopedic treatment but to provide a student with a versatile toolbox for further study. Moreover, in contradistinction to many other books, I do not have a particular direction in mind and hence I am trying to give a broad introduction which should prepare you for diverse fields such as spectral theory, partial differential equations, or probability theory. This is related to the fact that I am working in mathematical physics, an area where you never know what mathematical theory you will need next.

I have tried to keep a balance between verbosity and clarity in the sense that I have tried to provide sufficient detail for being able to follow the arguments but without drowning the key ideas in boring details. In particular, you will find a *show this* from time to time encouraging the reader to check the claims made (these tasks typically involve only simple routine calculations). Moreover, to make the presentation student friendly, I have tried to include many worked out examples within the main text. Some of them are standard counterexamples pointing out the limitations of theorems (and explaining why the assumptions are important). Others show how to use the theory in the investigation of practical examples.

Preliminaries

The present manuscript is intended to be gentle when it comes to required background. Of course I assume basic familiarity with analysis (real and complex numbers, limits, differentiation, basic (Riemann) integration, open sets) and linear algebra (finite dimensional vector spaces, matrices).

Apart from this natural assumptions I also expect some familiarity with metric spaces and point set topology. However, only a few basic things are required to begin with. This and much more is collected in the Appendix and I will refer you there from time to time such that you can refresh your memory should need arise. Moreover, you can always go there if you are unsure about a certain term (using the extensive index) or if there should be a need to clarify notation or conventions. I prefer this over referring you to several other books which might not alway be readily available. For convenience the Appendix contains full proofs in case one needs to fill some gaps. As some things are only outlined (or outsourced to exercises) it will require extra effort in case you see all this for the first time.

On the other hand I do not assume familiarity with Lebesgue integration and consequently L^p spaces will only be briefly mentioned as the completion

of continuous functions with respect to the corresponding integral norms in the first part. At a few places I also assume some basic results from complex analysis but it will be sufficient to just believe them.

Similarly, the second part on real analysis only requires a similar background and is essentially independent on the first part. Of course here you should already know what a Banach/Hilbert space is, but Chapter 1 will be sufficient to get you started.

Finally, the last part of course requires basic familiarity with functional analysis and measure theory (Lebesgue and Sobolev spaces). But apart from this it is again independent form the first two parts.

Content

Below follows a short description of each chapter together with some hints which parts can be skipped.

Chapter 1. The first part starts with Fourier's treatment of the heat equation which lead to the theory of Fourier analysis as well as the development of spectral theory which drove much of the development of functional analysis around the turn of the last century. In particular, the first chapter tries to introduce and motivate some of the key concepts and should be covered in detail except for Section 1.8 which introduces some interesting examples for later use.

Chapter 2 discusses basic Hilbert space theory and should be considered core material except for the last section discussing applications to Fourier series. They will only be used in some examples and could be skipped in case they are covered in a different course.

Chapter 3 develops basic spectral theory for compact self-adjoint operators. The first core result is the spectral theorem for compact symmetric (self-adjoint) operators which is then applied to Sturm–Liouville problems. Of course this application could be skipped, but this would reduce the didactical concept to absurdity. Nevertheless it is clearly possible to shorten the material as non of it (including the follow-up section which touches upon some more tools from spectral theory) will be required in later chapters. The last two sections on singular value decompositions as well as Hilbert–Schmidt and trace class operators cover important topics for applications, but will again not be required later on (except for Section 10.5, where applications to integral operators are discussed).

Chapter 4 discusses what is typically considered as the core results from Banach space theory. In order to keep the topological requirements to a minimum some advanced topics are shifted to the following chapters.

Except for possibly the last section, which discusses some application to minimizing nonlinear functionals, nothing should be skipped here.

Chapter 5 contains selected more advanced stuff. Except for the geometric Hahn–Banach theorem, which is a prerequisite for the other sections, the remaining sections are independent of each other.

Chapter 6 develops spectral theory for bounded self-adjoint operators via the framework of C^* algebras. Again a bottom-up approach is used such that the core results are in the first two sections and the rest is optional. Again the remaining four sections are independent of each other except for the fact that Section 6.3, which contains the spectral theorem for compact operators, and hence the Fredholm alternative for compact perturbations of the identity, is of course used to identify compact perturbations of the identity as premier examples of Fredholm operators.

Chapter 7 finally gives a brief introduction to operator semigroups, which can be considered optional.

In a similar vein, the second part tries to give a succinct introduction to measure theory.

Chapter 8 introduces the concept of a measure and constructs Borel measures on $\mathbb{R}^n$ via distribution functions (the case of $n = 1$ is done first) which should meet the needs of partial differential equations, spectral theory, and probability theory. I have chosen the Carathéodory approach because I feel that it is the most versatile one.

Chapter 9 discusses the core results of integration theory including the change of variables formula, surface measure and the Gauss–Green theorem. The latter two are only done in a smooth (C^1) setting, but these topics are needed in the chapter on Sobolev spaces and I wanted to be self-contained here. Two optional appendices discuss transforming one-dimensional measures (which should be useful in both spectral theory and probability theory) and the connection with the Riemann integral.

Chapter 10 contains the core material on L^p spaces including basic inequalities and ends with an optional section on integral operators.

Chapter 11 collects some further results. Except for the first section, the results are mostly optional and independent of each other. Lebesgue points discussed in the second section are also used at some places later on. There is also a final section on functions of bounded variation and absolutely continuous functions.

Chapter 12 discusses the dual space of L^p for $1 \leq p < \infty$ and $p = \infty$ as well as some variants of the Riesz–Markov representation theorem.

Chapter 13 contains some core material on Sobolev spaces. I feel that it should be sufficient as a background for applications to partial differential equations (PDE).

Chapter 14 covers the Fourier transform on $\mathbb{R}^n$. It also gives an independent approach to L^2 based Sobolev spaces on $\mathbb{R}^n$. Of course the Fourier transform is vital in the treatment of constant coefficient PDE. However, in many introductory courses some of the technical details are swept under the carpet. I try to discuss these examples with full rigor.

Chapter 15 finally discusses two basic interpolation techniques.

Finally, there is a part on nonlinear functional analysis.

Chapter 16 discusses analysis in Banach spaces (with a view towards applications in the calculus of variations and infinite dimensional dynamical systems).

Chapter 17 and 18 cover degree theory and fixed point theorems in finite and infinite dimensional spaces. Several applications to integral equations, ordinary differential equations and to the stationary Navier–Stokes equation are given.

Chapter 19 provides some basics about monotone maps.

Sometimes also the historic development of the subject is of interest. This is however not covered in the present book and we reefer to [**26, 39, 40**] as good starting points.

To the teacher

There are a couple of courses to be taught from this book. First of all there is of course a basic functional analysis course: Chapters 1 to 4 (skipping some optional material as discussed above) and perhaps adding some material from Chapter 5 or 6. If one wants to cover Lebesgue spaces, this can be easily done by including Chapters 8, 9, and 10 from the second part. In this case one could cover Section 8.2 (Section 8.1 contains just motivation), give an outline of Section 8.3 (by covering Dynkin's π-λ theorem, the uniqueness theorem for measures ,and then quoting the existence theorem for Lebesgue measure), cover Section 8.5. The core material from Chapter 9 are the first two sections and from Chapter 10 the first three sections. I think that this gives a well-balanced introduction to functional analysis which contains several optional topics to choose from depending on personal preferences and time constraints.

The remaining material from the first part could then be used for a course on advanced functional analysis. Typically one could also add some further topics from the second part or some material from unbounded operators in

Hilbert spaces following [**43**] (where one can start with Chapter 2) or from unbounded operators in Banach spaces following the book by Kato [**23**] (e.g. Sections 3.4, 3.5 and 4.1).

The second part can either serve as a succinct introduction to the Lebesgue integral (as explained before) or as a full course focusing either on measure theory (covering Chapters 8 to 12 and possibly also Chapter 13) or on real analysis (skipping Chapter 11 except Section 11.1, 11.8 and including Chapters 14 and 15).

The third part gives a short basis for a course on nonlinear functional analysis.

Problems relevant for the main text are marked with a "*". A Solutions Manual will be available electronically for instructors only.

Acknowledgments

I wish to thank my readers, Kerstin Ammann, Phillip Bachler, Batuhan Bayır, Alexander Beigl, Ho Boon Suan, Peng Du, Christian Ekstrand, Damir Ferizović, Michael Fischer, Raffaello Giulietti, Melanie Graf, Josef Greilhuber, Matthias Hammerl, Jona Marie Hassenbach, Nobuya Kakehashi, Jerzy Knopik, Nikolas Knotz, Florian Kogelbauer, Helge Krüger, Reinhold Küstner, Oliver Leingang, Juho Leppäkangas, Joris Mestdagh, Alice Mikikits-Leitner, Claudiu Mîndrilă, Jakob Möller, Caroline Moosmüller, Matthias Ostermann, Piotr Owczarek, Martina Pflegpeter, Mateusz Piorkowski, Tobias Preinerstorfer, Maximilian H. Ruep, Tidhar Sariel, Christian Schmid, Laura Shou, Vincent Valmorin, David Wallauch, Richard Welke, David Wimmesberger, Song Xiaojun, Markus Youssef, Rudolf Zeidler, and colleagues Nils C. Framstad, Fritz Gesztesy, Heinz Hanßmann, Günther Hörmann, Aleksey Kostenko, Wallace Lam, Daniel Lenz, Johanna Michor, Viktor Qvarfordt, Alex Strohmaier, David C. Ullrich, Hendrik Vogt, Marko Stautz, Maxim Zinchenko who have pointed out several typos and made useful suggestions for improvements. I am also grateful to Volker Enß for making his lecture notes on nonlinear Functional Analysis available to me.

Finally, no book is free of errors. So if you find one, or if you have comments or suggestions (no matter how small), please let me know.

Gerald Teschl

Vienna, Austria
January, 2019

Part 1

Functional Analysis

Chapter 1

A first look at Banach and Hilbert spaces

Functional analysis is an important tool in the investigation of all kind of problems in pure mathematics, physics, biology, economics, etc.. In fact, it is hard to find a branch in science where functional analysis is not used.

The main objects are (infinite dimensional) vector spaces with different concepts of convergence. The classical theory focuses on linear operators (i.e., functions) between these spaces but nonlinear operators are of course equally important. However, since one of the most important tools in investigating nonlinear mappings is linearization (differentiation), linear functional analysis will be our first topic in any case.

1.1. Introduction: Linear partial differential equations

Rather than listing an overwhelming number of classical examples I want to focus on one: linear partial differential equations. We will use this example as a guide throughout our first three chapters and will develop all necessary tools for a successful treatment of our particular problem.

In his investigation of heat conduction Fourier was lead to the (one dimensional) **heat** or diffusion equation

$$\frac{\partial}{\partial t}u(t,x) = \frac{\partial^2}{\partial x^2}u(t,x). \tag{1.1}$$

Here $u : \mathbb{R} \times [0,1] \to \mathbb{R}$ is the temperature distribution in a thin rod at time $t \in \mathbb{R}$ at the point $x \in [0,1]$. It is usually assumed, that the temperature at $x = 0$ and $x = 1$ is fixed, say $u(t,0) = a$ and $u(t,1) = b$. By considering $u(t,x) \to u(t,x) - a - (b-a)x$ it is clearly no restriction to assume $a = b = 0$.

Moreover, the initial temperature distribution $u(0,x) = u_0(x)$ is assumed to be known as well.

Since finding the solution seems at first sight unfeasable, we could try to find at least some solutions of (1.1). For example, we could make an ansatz for $u(t,x)$ as a product of two functions, each of which depends on only one variable, that is,

$$u(t,x) := w(t)y(x). \tag{1.2}$$

Plugging this ansatz into the heat equation we arrive at

$$\dot{w}(t)y(x) = y''(x)w(t), \tag{1.3}$$

where the dot refers to differentiation with respect to t and the prime to differentiation with respect to x. Bringing all t, x dependent terms to the left, right side, respectively, we obtain

$$\frac{\dot{w}(t)}{w(t)} = \frac{y''(x)}{y(x)}. \tag{1.4}$$

Accordingly, this ansatz is called **separation of variables**.

Now if this equation should hold for all t and x, the quotients must be equal to a constant $-\lambda$ (we choose $-\lambda$ instead of λ for convenience later on). That is, we are lead to the equations

$$-\dot{w}(t) = \lambda w(t) \tag{1.5}$$

and

$$-y''(x) = \lambda y(x), \qquad y(0) = y(1) = 0, \tag{1.6}$$

which can easily be solved. The first one gives

$$w(t) = c_1 \mathrm{e}^{-\lambda t} \tag{1.7}$$

and the second one

$$y(x) = c_2 \cos(\sqrt{\lambda}x) + c_3 \sin(\sqrt{\lambda}x). \tag{1.8}$$

However, $y(x)$ must also satisfy the boundary conditions $y(0) = y(1) = 0$. The first one $y(0) = 0$ is satisfied if $c_2 = 0$ and the second one yields (c_3 can be absorbed by $w(t)$)

$$\sin(\sqrt{\lambda}) = 0, \tag{1.9}$$

which holds if $\lambda = (\pi n)^2$, $n \in \mathbb{N}$ (in the case $\lambda < 0$ we get $\sinh(\sqrt{-\lambda}) = 0$, which cannot be satisfied and explains our choice of sign above). In summary, we obtain the solutions

$$u_n(t,x) := c_n \mathrm{e}^{-(\pi n)^2 t} \sin(n\pi x), \qquad n \in \mathbb{N}. \tag{1.10}$$

So we have found a large number of solutions, but we still have not dealt with our initial condition $u(0,x) = u_0(x)$. This can be done using the superposition principle which holds since our equation is linear: Any finite linear combination of the above solutions will again be a solution. Moreover,

under suitable conditions on the coefficients we can even consider infinite linear combinations. In fact, choosing

$$u(t,x) := \sum_{n=1}^{\infty} c_n \mathrm{e}^{-(\pi n)^2 t} \sin(n\pi x), \tag{1.11}$$

where the coefficients c_n decay sufficiently fast (e.g. absolutely summable), we obtain further solutions of our equation. Moreover, these solutions satisfy

$$u(0,x) = \sum_{n=1}^{\infty} c_n \sin(n\pi x) \tag{1.12}$$

and expanding the initial conditions into a Fourier sine series

$$u_0(x) = \sum_{n=1}^{\infty} \hat{u}_{0,n} \sin(n\pi x), \tag{1.13}$$

we see that the solution of our original problem is given by (1.11) if we choose $c_n = \hat{u}_{0,n}$.

Of course for this last statement to hold we need to ensure that the series in (1.11) converges and that we can interchange summation and differentiation. You are asked to do so in Problem 1.1.

In fact, many equations in physics can be solved in a similar way:

• **Reaction-Diffusion equation**:

$$\begin{aligned} &\frac{\partial}{\partial t} u(t,x) - \frac{\partial^2}{\partial x^2} u(t,x) + q(x) u(t,x) = 0, \\ &u(0,x) = u_0(x), \\ &u(t,0) = u(t,1) = 0. \end{aligned} \tag{1.14}$$

Here $u(t,x)$ could be the density of some gas in a pipe and $q(x) > 0$ describes that a certain amount per time is removed (e.g., by a chemical reaction).

• **Wave equation**:

$$\begin{aligned} &\frac{\partial^2}{\partial t^2} u(t,x) - \frac{\partial^2}{\partial x^2} u(t,x) = 0, \\ &u(0,x) = u_0(x), \quad \frac{\partial u}{\partial t}(0,x) = v_0(x) \\ &u(t,0) = u(t,1) = 0. \end{aligned} \tag{1.15}$$

Here $u(t,x)$ is the displacement of a vibrating string which is fixed at $x = 0$ and $x = 1$. Since the equation is of second order in time, both the initial displacement $u_0(x)$ and the initial velocity $v_0(x)$ of the string need to be known.

• **Schrödinger equation**:

$$\begin{aligned} &\mathrm{i}\frac{\partial}{\partial t}u(t,x) = -\frac{\partial^2}{\partial x^2}u(t,x) + q(x)u(t,x),\\ &u(0,x) = u_0(x),\\ &u(t,0) = u(t,1) = 0. \end{aligned} \tag{1.16}$$

Here $|u(t,x)|^2$ is the probability distribution of a particle trapped in a box $x \in [0,1]$ and $q(x)$ is a given external potential which describes the forces acting on the particle.

All these problems (and many others) lead to the investigation of the following problem

$$Ly(x) = \lambda y(x), \qquad L := -\frac{d^2}{dx^2} + q(x), \tag{1.17}$$

subject to the **boundary conditions**

$$y(a) = y(b) = 0. \tag{1.18}$$

Such a problem is called a **Sturm–Liouville boundary value problem**. Our example shows that we should prove the following facts about Sturm–Liouville problems:

(i) The Sturm–Liouville problem has a countable number of eigenvalues E_n with corresponding eigenfunctions u_n, that is, u_n satisfies the boundary conditions and $Lu_n = E_n u_n$.

(ii) The eigenfunctions u_n are complete, that is, any *nice* function u can be expanded into a generalized Fourier series

$$u(x) = \sum_{n=1}^{\infty} c_n u_n(x).$$

This problem is very similar to the eigenvalue problem of a matrix and we are looking for a generalization of the well-known fact that every symmetric matrix has an orthonormal basis of eigenvectors. However, our linear operator L is now acting on some space of functions which is not finite dimensional and it is not at all clear what (e.g.) orthogonal should mean in this context. Moreover, since we need to handle infinite series, we need convergence and hence we need to define the distance of two functions as well.

Hence our program looks as follows:

- What is the distance of two functions? This automatically leads us to the problem of convergence and completeness.

- If we additionally require the concept of orthogonality, we are lead to Hilbert spaces which are the proper setting for our eigenvalue problem.
- Finally, the spectral theorem for compact symmetric operators will be the solution of our above problem.

Problem 1.1. *Suppose* $\sum_{n=1}^{\infty} |c_n| < \infty$. *Show that* (1.11) *is continuous for* $(t, x) \in [0, \infty) \times [0, 1]$ *and solves the heat equation for* $(t, x) \in (0, \infty) \times [0, 1]$. *(Hint: Weierstraß M-test. When can you interchange the order of summation and differentiation?)*

1.2. The Banach space of continuous functions

Our point of departure will be the set of continuous functions $C(I)$ on a compact interval $I := [a, b] \subset \mathbb{R}$. Since we want to handle both real and complex models, we will formulate most results for the more general complex case only. In fact, most of the time there will be no difference but we will add a remark in the rare case where the real and complex case do indeed differ.

One way of declaring a distance, well-known from calculus, is the **maximum norm** of a function $f \in C(I)$:

$$\|f\|_\infty := \max_{x \in I} |f(x)|. \tag{1.19}$$

It is not hard to see that with this definition $C(I)$ becomes a normed vector space:

A **normed vector space** X is a vector space X over $\mathbb{C}$ (or $\mathbb{R}$) with a nonnegative function (the **norm**) $\|.\| : X \to [0, \infty)$ such that

- $\|f\| > 0$ for $f \neq 0$ (**positive definiteness**),
- $\|\alpha f\| = |\alpha| \, \|f\|$ for all $\alpha \in \mathbb{C}$, $f \in X$ (**positive homogeneity**), and
- $\|f + g\| \leq \|f\| + \|g\|$ for all $f, g \in X$ (**triangle inequality**).

If positive definiteness is dropped from the requirements, one calls $\|.\|$ a **seminorm**.

From the triangle inequality we also get the **inverse triangle inequality** (Problem 1.2)

$$|\|f\| - \|g\|| \leq \|f - g\|, \tag{1.20}$$

which shows that the norm is continuous.

Let me also briefly mention that norms are closely related to convexity. To this end recall that a subset $C \subseteq X$ is called **convex** if for every $x, y \in C$ we also have $\lambda x + (1 - \lambda) y \in C$ whenever $\lambda \in (0, 1)$. Moreover, a mapping

$f : C \to \mathbb{R}$ is called **convex** if $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$ whenever $\lambda \in (0,1)$ and in our case the triangle inequality plus homogeneity imply that every norm is convex:

$$\|\lambda x + (1-\lambda)y\| \le \lambda\|x\| + (1-\lambda)\|y\|, \qquad \lambda \in [0,1]. \tag{1.21}$$

Moreover, choosing $\lambda = \frac{1}{2}$ we get back the triangle inequality upon using homogeneity. In particular, the triangle inequality could be replaced by convexity in the definition.

Once we have a norm, we have a **distance** $d(f,g) := \|f-g\|$ and hence we know when a sequence of vectors f_n **converges** to a vector f (namely if $d(f_n, f) \to 0$). We will write $f_n \to f$ or $\lim_{n\to\infty} f_n = f$, as usual, in this case. Moreover, a mapping $F : X \to Y$ between two normed spaces is called **continuous** if for every convergent sequence $f_n \to f$ from X we have $F(f_n) \to F(f)$ (with respect to the norm of X and Y, respectively). In fact, the norm, vector addition, and multiplication by scalars are continuous (Problem 1.3).

In addition to the concept of convergence, we also have the concept of a **Cauchy sequence** and hence the concept of completeness: A normed space is called **complete** if every Cauchy sequence has a limit. A complete normed space is called a **Banach space**.

Example 1.1. By completeness of the real numbers, $\mathbb{R}$ as well as $\mathbb{C}$ with the absolute value as norm are Banach spaces. ⋄

Example 1.2. The space $\ell^1(\mathbb{N})$ of all complex-valued sequences $a = (a_j)_{j=1}^{\infty}$ for which the norm

$$\|a\|_1 := \sum_{j=1}^{\infty} |a_j| \tag{1.22}$$

is finite is a Banach space.

To show this, we need to verify three things: (i) $\ell^1(\mathbb{N})$ is a vector space, that is, closed under addition and scalar multiplication, (ii) $\|.\|_1$ satisfies the three requirements for a norm, and (iii) $\ell^1(\mathbb{N})$ is complete.

First of all, observe

$$\sum_{j=1}^{k} |a_j + b_j| \le \sum_{j=1}^{k} |a_j| + \sum_{j=1}^{k} |b_j| \le \|a\|_1 + \|b\|_1$$

for every finite k. Letting $k \to \infty$, we conclude that $\ell^1(\mathbb{N})$ is closed under addition and that the triangle inequality holds. That $\ell^1(\mathbb{N})$ is closed under scalar multiplication together with homogeneity as well as definiteness are straightforward. It remains to show that $\ell^1(\mathbb{N})$ is complete. Let $a^n = (a_j^n)_{j=1}^{\infty}$ be a Cauchy sequence; that is, for given $\varepsilon > 0$ we can find an N_ε such that $\|a^m - a^n\|_1 \le \varepsilon$ for $m, n \ge N_\varepsilon$. This implies, in particular, $|a_j^m - a_j^n| \le \varepsilon$ for

every fixed j. Thus a_j^n is a Cauchy sequence for fixed j and, by completeness of $\mathbb{C}$, it has a limit: $a_j := \lim_{n\to\infty} a_j^n$. Now consider $\sum_{j=1}^k |a_j^m - a_j^n| \le \varepsilon$ and take $m \to \infty$:

$$\sum_{j=1}^{k} |a_j - a_j^n| \le \varepsilon.$$

Since this holds for all finite k, we even have $\|a-a^n\|_1 \le \varepsilon$. Hence $(a-a^n) \in \ell^1(\mathbb{N})$ and since $a^n \in \ell^1(\mathbb{N})$, we finally conclude $a = a^n + (a-a^n) \in \ell^1(\mathbb{N})$. By our estimate $\|a-a^n\|_1 \le \varepsilon$, our candidate a is indeed the limit of a^n. $\diamond$

Example 1.3. The previous example can be generalized by considering the space $\ell^p(\mathbb{N})$ of all complex-valued sequences $a = (a_j)_{j=1}^\infty$ for which the norm

$$\|a\|_p := \left(\sum_{j=1}^{\infty} |a_j|^p\right)^{1/p}, \qquad p \in [1,\infty), \tag{1.23}$$

is finite. By $|a_j + b_j|^p \le 2^p \max(|a_j|,|b_j|)^p = 2^p \max(|a_j|^p, |b_j|^p) \le 2^p(|a_j|^p + |b_j|^p)$ it is a vector space, but the triangle inequality is only easy to see in the case $p = 1$. (It is also not hard to see that it fails for $p < 1$, which explains our requirement $p \ge 1$. See also Problem 1.14.)

To prove the triangle inequality we need Young's inequality (Problem 1.7)

$$\alpha^{1/p}\beta^{1/q} \le \frac{1}{p}\alpha + \frac{1}{q}\beta, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad \alpha,\beta \ge 0, \tag{1.24}$$

which implies **Hölder's inequality**

$$\|ab\|_1 \le \|a\|_p \|b\|_q \tag{1.25}$$

for $a \in \ell^p(\mathbb{N})$, $b \in \ell^q(\mathbb{N})$. In fact, by homogeneity of the norm it suffices to prove the case $\|a\|_p = \|b\|_q = 1$. But this case follows by choosing $\alpha = |a_j|^p$ and $\beta = |b_j|^q$ in (1.24) and summing over all j. (A different proof based on convexity will be given in Section 10.2.)

Now using $|a_j + b_j|^p \le |a_j|\,|a_j + b_j|^{p-1} + |b_j|\,|a_j + b_j|^{p-1}$, we obtain from Hölder's inequality (note $(p-1)q = p$)

$$\begin{aligned}\|a+b\|_p^p &\le \|a\|_p \|(a+b)^{p-1}\|_q + \|b\|_p \|(a+b)^{p-1}\|_q \\ &= (\|a\|_p + \|b\|_p)\|a+b\|_p^{p-1}.\end{aligned}$$

Hence $\ell^p(\mathbb{N})$ is a normed space. That it is complete can be shown as in the case $p = 1$ (Problem 1.8).

The unit ball with respect to these norms in $\mathbb{R}^2$ is depicted in Figure 1.1. One sees that for $p < 1$ the unit ball is not convex (explaining once more our restriction $p \ge 1$). Moreover, for $1 < p < \infty$ it is even strictly convex (that is, the line segment joining two distinct points is always in the interior).

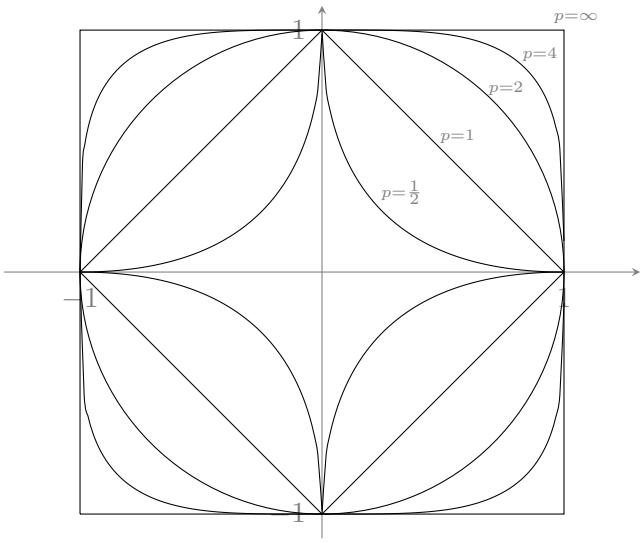


Figure 1.1. Unit balls for $\|.\|_p$ in $\mathbb{R}^2$

This is related to the question of equality in the triangle inequality and will be discussed in Problems 1.11 and 1.12. ⋄

Example 1.4. The space $\ell^\infty(\mathbb{N})$ of all complex-valued bounded sequences $a = (a_j)_{j=1}^\infty$ together with the norm

$$\|a\|_\infty := \sup_{j\in\mathbb{N}} |a_j| \tag{1.26}$$

is a Banach space (Problem 1.9). Note that with this definition, Hölder's inequality (1.25) remains true for the cases $p = 1$, $q = \infty$ and $p = \infty$, $q = 1$. The reason for the notation is explained in Problem 1.13. ⋄

Example 1.5. Every closed subspace of a Banach space is again a Banach space. For example, the space $c_0(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$ of all sequences converging to zero is a closed subspace. In fact, if $a \in \ell^\infty(\mathbb{N})\backslash c_0(\mathbb{N})$, then $\limsup_{j\to\infty} |a_j| = \varepsilon > 0$ and thus $\|a - b\|_\infty \geq \varepsilon$ for every $b \in c_0(\mathbb{N})$. ⋄

Now what about completeness of $C(I)$? A sequence of functions f_n converges to f if and only if

$$\lim_{n\to\infty} \|f - f_n\|_\infty = \lim_{n\to\infty} \max_{x\in I} |f(x) - f_n(x)| = 0. \tag{1.27}$$

That is, in the language of real analysis, f_n converges uniformly to f. Now let us look at the case where f_n is only a Cauchy sequence. Then $f_n(x)$ is clearly a Cauchy sequence of complex numbers for every fixed $x \in I$. In particular, by completeness of $\mathbb{C}$, there is a limit $f(x)$ for each x. Thus we

get a limiting function $f(x)$. Moreover, letting $m \to \infty$ in

$$|f_m(x) - f_n(x)| \le \varepsilon \qquad \forall m, n > N_\varepsilon,\ x \in I, \tag{1.28}$$

we see

$$|f(x) - f_n(x)| \le \varepsilon \qquad \forall n > N_\varepsilon,\ x \in I; \tag{1.29}$$

that is, $f_n(x)$ converges uniformly to $f(x)$. However, up to this point we do not know whether it is in our vector space $C(I)$, that is, whether it is continuous. Fortunately, there is a well-known result from real analysis which tells us that the uniform limit of continuous functions is again continuous: Fix $x \in I$ and $\varepsilon > 0$. To show that f is continuous we need to find a δ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$. Pick n so that $\|f_n - f\|_\infty < \varepsilon/3$ and δ so that $|x - y| < \delta$ implies $|f_n(x) - f_n(y)| < \varepsilon/3$. Then $|x - y| < \delta$ implies

$$|f(x)-f(y)| \le |f(x)-f_n(x)|+|f_n(x)-f_n(y)|+|f_n(y)-f(y)| < \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3} = \varepsilon$$

as required. Hence $f \in C(I)$ and thus every Cauchy sequence in $C(I)$ converges. Or, in other words,

Theorem 1.1. *Let $I \subset \mathbb{R}$ be a compact interval, then the continuous functions $C(I)$ with the maximum norm is a Banach space.*

For finite dimensional vector spaces the concept of a basis plays a crucial role. In the case of infinite dimensional vector spaces one could define a basis as a maximal set of linearly independent vectors (known as a **Hamel basis**; Problem 1.6). Such a basis has the advantage that it only requires finite linear combinations. However, the price one has to pay is that such a basis will be way too large (typically uncountable, cf. Problems 1.5 and 4.1). Since we have the notion of convergence, we can handle countable linear combinations and try to look for *countable bases*. We start with a few definitions.

The set of all finite linear combinations of a set of vectors $\{u_n\}_{n\in\mathcal{N}} \subset X$ is called the **span** of $\{u_n\}_{n\in\mathcal{N}}$ and denoted by

$$\operatorname{span}\{u_n\}_{n\in\mathcal{N}} := \{\sum_{j=1}^{m} \alpha_j u_{n_j} | n_j \in \mathcal{N}, \alpha_j \in \mathbb{C}, m \in \mathbb{N}\}. \tag{1.30}$$

A set of vectors $\{u_n\}_{n\in\mathcal{N}} \subset X$ is called **linearly independent** if every finite subset is. If $\{u_n\}_{n=1}^N \subset X$, $N \in \mathbb{N} \cup \{\infty\}$, is countable, we can throw away all elements which can be expressed as linear combinations of the previous ones to obtain a subset of linearly independent vectors which have the same span.

We will call a countable sequence of vectors $(u_n)_{n=1}^N \subset X$ a **Schauder basis** if every element $f \in X$ can be uniquely written as a countable linear

combination of the basis elements:

$$f = \sum_{n=1}^{N} \alpha_n u_n, \qquad \alpha_n = \alpha_n(f) \in \mathbb{C}, \tag{1.31}$$

where the sum has to be understood as a limit if $N = \infty$ (the sum is not required to converge unconditionally and hence the order of the basis elements is important). Since we have assumed the coefficients $\alpha_n(f)$ to be uniquely determined, the vectors are necessarily linearly independent. Moreover, one can show that the coordinate functionals $f \mapsto \alpha_n(f)$ are continuous (cf. Problem 4.6). A Schauder basis and its corresponding coordinate functionals $u_n^* : X \to \mathbb{C}$, $f \mapsto \alpha_n(f)$ form a so-called **biorthogonal system**: $u_m^*(u_n) = \delta_{m,n}$, where

$$\delta_{n,m} := \begin{cases} 1, & n = m, \\ 0, & n \neq m, \end{cases} \tag{1.32}$$

is the **Kronecker delta**.

Example 1.6. The sequence of vectors $\delta^n = (\delta_m^n = \delta_{n,m})_{m\in\mathbb{N}}$ is a Schauder basis for the Banach space $\ell^p(\mathbb{N})$, $1 \leq p < \infty$.

Let $a = (a_j)_{j=1}^{\infty} \in \ell^p(\mathbb{N})$ be given and set $a^m := \sum_{n=1}^{m} a_n \delta^n$. Then

$$\|a - a^m\|_p = \left(\sum_{j=m+1}^{\infty} |a_j|^p \right)^{1/p} \to 0$$

since $a_j^m = a_j$ for $1 \leq j \leq m$ and $a_j^m = 0$ for $j > m$. Hence

$$a = \sum_{n=1}^{\infty} a_n \delta^n$$

and $(\delta^n)_{n=1}^{\infty}$ is a Schauder basis (uniqueness of the coefficients is left as an exercise).

Note that $(\delta^n)_{n=1}^{\infty}$ is also Schauder basis for $c_0(\mathbb{N})$ but not for $\ell^\infty(\mathbb{N})$ (try to approximate a constant sequence). ⋄

A set whose span is dense is called **total**, and if we have a countable total set, we also have a countable dense set (consider only linear combinations with rational coefficients — show this). A normed vector space containing a countable dense set is called **separable**.

Warning: Some authors use the term total in a slightly different way — see the warning on page 117.

Example 1.7. Every Schauder basis is total and thus every Banach space with a Schauder basis is separable (the converse puzzled mathematicians for quite some time and was eventually shown to be false by Per Enflo). In particular, the Banach space $\ell^p(\mathbb{N})$ is separable for $1 \leq p < \infty$.

However, $\ell^\infty(\mathbb{N})$ is not separable (Problem 1.10)! $\diamond$

While we will not give a Schauder basis for $C(I)$ (Problem 1.15), we will at least show that $C(I)$ is separable. We will do this by showing that every continuous function can be approximated by polynomials, a result which is of independent interest. But first we need a lemma.

Lemma 1.2 (Smoothing)**.** *Let u_n be a sequence of nonnegative continuous functions on $[-1,1]$ such that*

$$\int_{|x|\le 1} u_n(x)dx = 1 \quad \text{and} \quad \int_{\delta\le|x|\le 1} u_n(x)dx \to 0, \quad \delta > 0. \tag{1.33}$$

(In other words, u_n has mass one and concentrates near $x = 0$ as $n \to \infty$.)

Then for every $f \in C[-\frac{1}{2}, \frac{1}{2}]$ which vanishes at the endpoints, $f(-\frac{1}{2}) = f(\frac{1}{2}) = 0$, we have that

$$f_n(x) := \int_{-1/2}^{1/2} u_n(x-y)f(y)dy \tag{1.34}$$

converges uniformly to $f(x)$.

Proof. Since f is uniformly continuous, for given ε we can find a $\delta < 1/2$ (independent of x) such that $|f(x) - f(y)| \le \varepsilon$ whenever $|x - y| \le \delta$. Moreover, we can choose n such that $\int_{\delta\le|y|\le 1} u_n(y)dy \le \varepsilon$. Now abbreviate $M := \max_{x\in[-1/2,1/2]}\{1, |f(x)|\}$ and note

$$|f(x) - \int_{-1/2}^{1/2} u_n(x-y)f(x)dy| = |f(x)|\,|1 - \int_{-1/2}^{1/2} u_n(x-y)dy| \le M\varepsilon.$$

In fact, either the distance of x to one of the boundary points $\pm\frac{1}{2}$ is smaller than δ and hence $|f(x)| \le \varepsilon$ or otherwise $[-\delta, \delta] \subset [x - 1/2, x + 1/2]$ and the difference between one and the integral is smaller than ε.

Using this, we have

$$\begin{aligned}
|f_n(x) - f(x)| \le& \int_{-1/2}^{1/2} u_n(x-y)|f(y) - f(x)|dy + M\varepsilon \\
=& \int_{|y|\le 1/2, |x-y|\le\delta} u_n(x-y)|f(y) - f(x)|dy \\
&+ \int_{|y|\le 1/2, |x-y|\ge\delta} u_n(x-y)|f(y) - f(x)|dy + M\varepsilon \\
\le& \varepsilon + 2M\varepsilon + M\varepsilon = (1 + 3M)\varepsilon,
\end{aligned}$$

which proves the claim. $\square$

Note that f_n will be as smooth as u_n, hence the title smoothing lemma. Moreover, f_n will be a polynomial if u_n is. The same idea is used to approximate noncontinuous functions by smooth ones (of course the convergence will no longer be uniform in this case).

Now we are ready to show:

Theorem 1.3 (Weierstraß)**.** *Let $I \subset \mathbb{R}$ be a compact interval. Then the set of polynomials is dense in $C(I)$.*

Proof. Let $f \in C(I)$ be given. By considering $f(x) - f(a) - \frac{f(b)-f(a)}{b-a}(x-a)$ it is no loss to assume that f vanishes at the boundary points. Moreover, without restriction, we only consider $I = [-\frac{1}{2}, \frac{1}{2}]$ (why?).

Now the claim follows from Lemma 1.2 using the **Landau kernel**

$$u_n(x) := \frac{1}{I_n}(1-x^2)^n,$$

where (using integration by parts)

$$\begin{aligned} I_n &:= \int_{-1}^{1} (1-x^2)^n dx = \frac{n}{n+1}\int_{-1}^{1} (1-x)^{n-1}(1+x)^{n+1} dx \\ &= \cdots = \frac{n!}{(n+1)\cdots(2n+1)} 2^{2n+1} = \frac{(n!)^2 2^{2n+1}}{(2n+1)!} = \frac{n!}{\frac{1}{2}(\frac{1}{2}+1)\cdots(\frac{1}{2}+n)}. \end{aligned}$$

Indeed, the first part of (1.33) holds by construction, and the second part follows from the elementary estimate

$$\frac{1}{\frac{1}{2}+n} < I_n < 2,$$

which shows $\int_{\delta \le |x| \le 1} u_n(x) dx \le 2u_n(\delta) < (2n+1)(1-\delta^2)^n \to 0$. □

Corollary 1.4. *The monomials are total and hence C(I) is separable.*

Note that while the proof of Theorem 1.3 provides an explicit way of constructing a sequence of polynomials $f_n(x)$ which will converge uniformly to $f(x)$, this method still has a few drawbacks from a practical point of view: Suppose we have approximated f by a polynomial of degree n but our approximation turns out to be insufficient for the intended purpose. First of all, since our polynomial will not be optimal in general, we could try to find another polynomial of the same degree giving a better approximation. However, as this is by no means straightforward, it seems more feasible to simply increase the degree. However, if we do this, all coefficients will change and we need to start from scratch. This is in contradistinction to a Schauder basis where we could just add one new element from the basis (and where it suffices to compute one new coefficient).

In particular, note that this shows that the monomials are no Schauder basis for $C(I)$ since adding monomials incrementally to the expansion gives a uniformly convergent power series whose limit must be analytic. This observation emphasizes that a Schauder basis is more than a set of linearly independent vectors whose span is dense.

We will see in the next section that the concept of orthogonality resolves these problems.

Problem* 1.2. *Show that* $|\|f\| - \|g\|| \le \|f - g\|$.

Problem* 1.3. *Let X be a Banach space. Show that the norm, vector addition, and multiplication by scalars are continuous. That is, if $f_n \to f$, $g_n \to g$, and $\alpha_n \to \alpha$, then $\|f_n\| \to \|f\|$, $f_n + g_n \to f + g$, and $\alpha_n g_n \to \alpha g$.*

Problem 1.4. *Let X be a Banach space. Show that $\sum_{j=1}^{\infty} \|f_j\| < \infty$ implies that*

$$\sum_{j=1}^{\infty} f_j = \lim_{n\to\infty} \sum_{j=1}^{n} f_j$$

*exists. The series is called **absolutely convergent** in this case. Conversely, show that a normed space is complete if every absolutely convergent series converges.*

Problem 1.5. *While $\ell^1(\mathbb{N})$ is separable, it still has room for an uncountable set of linearly independent vectors. Show this by considering vectors of the form*

$$a^\alpha = (1, \alpha, \alpha^2, \dots), \qquad \alpha \in (0, 1).$$

(Hint: Recall the Vandermonde determinant. See Problem 4.1 for a generalization.)

Problem 1.6. *A **Hamel basis** is a maximal set of linearly independent vectors. Show that every vector space X has a Hamel basis $\{u_\alpha\}_{\alpha\in A}$. Show that given a Hamel basis, every $x \in X$ can be written as a finite linear combination $x = \sum_{j=1}^{n} c_j u_{\alpha_j}$, where the vectors u_{α_j} and the constants c_j are uniquely determined. (Hint: Use Zorn's lemma, see Appendix A, to show existence.)*

Problem* 1.7. *Prove Young's inequality* (1.24). *Show that equality occurs precisely if $\alpha = \beta$. (Hint: Take logarithms on both sides.)*

Problem* 1.8. *Show that $\ell^p(\mathbb{N})$, $1 \le p < \infty$, is complete.*

Problem* 1.9. *Show that $\ell^\infty(\mathbb{N})$ is a Banach space.*

Problem* 1.10. *Show that $\ell^\infty(\mathbb{N})$ is not separable. (Hint: Consider sequences which take only the value one and zero. How many are there? What is the distance between two such sequences?)*

Problem* 1.11. *Show that there is equality in the Hölder inequality* (1.25) *for $1 < p < \infty$ if and only if either $a = 0$ or $|b_j|^p = \alpha |a_j|^q$ for all $j \in \mathbb{N}$. Show that we have equality in the triangle inequality for $\ell^1(\mathbb{N})$ if and only if $a_j b_j^* \geq 0$ for all $j \in \mathbb{N}$ (here the '$*$' denotes complex conjugation). Show that we have equality in the triangle inequality for $\ell^p(\mathbb{N})$ for $1 < p < \infty$ if and only if $a = 0$ or $b = \alpha a$ with $\alpha \geq 0$.*

Problem* 1.12. *Let X be a normed space. Show that the following conditions are equivalent.*

(i) *If $\|x + y\| = \|x\| + \|y\|$ then $y = \alpha x$ for some $\alpha \geq 0$ or $x = 0$.*

(ii) *If $\|x\| = \|y\| = 1$ and $x \neq y$ then $\|\lambda x + (1 - \lambda)y\| < 1$ for all $0 < \lambda < 1$.*

(iii) *If $\|x\| = \|y\| = 1$ and $x \neq y$ then $\frac{1}{2}\|x + y\| < 1$.*

(iv) *The function $x \mapsto \|x\|^2$ is strictly convex.*

*A norm satisfying one of them is called **strictly convex**.*

Show that $\ell^p(\mathbb{N})$ is strictly convex for $1 < p < \infty$ but not for $p = 1, \infty$.

Problem 1.13. *Show that $p_0 \leq p$ implies $\ell^{p_0}(\mathbb{N}) \subset \ell^p(\mathbb{N})$ and $\|a\|_p \leq \|a\|_{p_0}$. Moreover, show*

$$\lim_{p \to \infty} \|a\|_p = \|a\|_\infty.$$

Problem 1.14. *Formally extend the definition of $\ell^p(\mathbb{N})$ to $p \in (0, 1)$. Show that $\|.\|_p$ does not satisfy the triangle inequality. However, show that it is a **quasinormed space**, that is, it satisfies all requirements for a normed space except for the triangle inequality which is replaced by*

$$\|a + b\| \leq K(\|a\| + \|b\|)$$

with some constant $K \geq 1$. Show, in fact,

$$\|a + b\|_p \leq 2^{1/p-1}(\|a\|_p + \|b\|_p), \qquad p \in (0, 1).$$

Moreover, show that $\|.\|_p^p$ satisfies the triangle inequality in this case, but of course it is no longer homogeneous (but at least you can get an honest metric $d(a, b) = \|a - b\|_p^p$ which gives rise to the same topology). (Hint: Show $\alpha + \beta \leq (\alpha^p + \beta^p)^{1/p} \leq 2^{1/p-1}(\alpha + \beta)$ for $0 < p < 1$ and $\alpha, \beta \geq 0$.)

Problem 1.15. *Show that the following set of functions is a Schauder basis for $C[0, 1]$: We start with $u_1(t) = t$, $u_2(t) = 1 - t$ and then split $[0, 1]$ into 2^n intervals of equal length and let $u_{2^n+k+1}(t)$, for $1 \leq k \leq 2^n$, be a piecewise linear peak of height 1 supported in the k'th subinterval: $u_{2^n+k+1}(t) = \max(0, 1 - |2^{n+1}t - 2k + 1|)$ for $n \in \mathbb{N}_0$ and $1 \leq k \leq 2^n$.*

1.3. The geometry of Hilbert spaces

So it looks like $C(I)$ has all the properties we want. However, there is still one thing missing: How should we define orthogonality in $C(I)$? In Euclidean space, two vectors are called **orthogonal** if their scalar product vanishes, so we would need a scalar product:

Suppose $\mathfrak{H}$ is a vector space. A map $\langle .,..\rangle : \mathfrak{H} \times \mathfrak{H} \to \mathbb{C}$ is called a **sesquilinear form** if it is conjugate linear in the first argument and linear in the second; that is,

$$\begin{aligned} \langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle &= \alpha_1^* \langle f_1, g\rangle + \alpha_2^* \langle f_2, g\rangle, \\ \langle f, \alpha_1 g_1 + \alpha_2 g_2\rangle &= \alpha_1 \langle f, g_1\rangle + \alpha_2 \langle f, g_2\rangle, \end{aligned} \qquad \alpha_1, \alpha_2 \in \mathbb{C}, \tag{1.35}$$

where '$*$' denotes complex conjugation. A symmetric

$$\langle f, g\rangle = \langle g, f\rangle^* \qquad \text{(symmetry)}$$

sesquilinear form is also called a **Hermitian form** and a positive definite

$$\langle f, f\rangle > 0 \text{ for } f \neq 0 \qquad \text{(positive definite)},$$

Hermitian form is called an **inner product** or **scalar product**. Note that positivity already implies symmetry in the complex case (Problem 1.19). Associated with every scalar product is a norm

$$\|f\| := \sqrt{\langle f, f\rangle}. \tag{1.36}$$

Only the triangle inequality is nontrivial. It will follow from the Cauchy–Schwarz inequality below. Until then, just regard (1.36) as a convenient short hand notation.

Warning: There is no common agreement whether a sesquilinear form (scalar product) should be linear in the first or in the second argument and different authors use different conventions.

The pair $(\mathfrak{H}, \langle .,..\rangle)$ is called an **inner product space**. If $\mathfrak{H}$ is complete (with respect to the norm (1.36)), it is called a **Hilbert space**.

Example 1.8. Clearly, $\mathbb{C}^n$ with the usual scalar product

$$\langle a, b\rangle := \sum_{j=1}^{n} a_j^* b_j \tag{1.37}$$

is a (finite dimensional) Hilbert space. ⋄

Example 1.9. A somewhat more interesting example is the Hilbert space $\ell^2(\mathbb{N})$, that is, the set of all complex-valued sequences

$$\Big\{(a_j)_{j=1}^{\infty} \Big| \sum_{j=1}^{\infty} |a_j|^2 < \infty\Big\} \tag{1.38}$$

with scalar product

$$\langle a, b\rangle := \sum_{j=1}^{\infty} a_j^* b_j. \tag{1.39}$$

To see that this sum is (absolutely) convergent (and thus well-defined) for $a, b \in \ell^2(\mathbb{N})$ follows from Hölder's inequality (1.25) in the case $p = q = 2$.

Observe that the norm $\|a\| = \sqrt{\langle a, a\rangle}$ is identical to the norm $\|a\|_2$ defined in the previous section. In particular, $\ell^2(\mathbb{N})$ is complete and thus indeed a Hilbert space. ⋄

A vector $f \in \mathfrak{H}$ is called **normalized** or a **unit vector** if $\|f\| = 1$. Two vectors $f, g \in \mathfrak{H}$ are called **orthogonal** or **perpendicular** ($f \perp g$) if $\langle f, g\rangle = 0$ and **parallel** if one is a multiple of the other.

If f and g are orthogonal, we have the **Pythagorean theorem**:

$$\|f + g\|^2 = \|f\|^2 + \|g\|^2, \qquad f \perp g, \tag{1.40}$$

which is one line of computation (do it!).

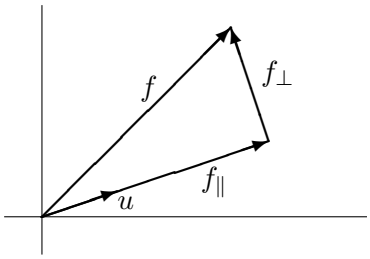
Suppose u is a unit vector. Then the projection of f in the direction of u is given by

$$f_\parallel := \langle u, f\rangle u, \tag{1.41}$$

and $f_\perp$, defined via

$$f_\perp := f - \langle u, f\rangle u, \tag{1.42}$$

is perpendicular to u since $\langle u, f_\perp\rangle = \langle u, f - \langle u, f\rangle u\rangle = \langle u, f\rangle - \langle u, f\rangle\langle u, u\rangle = 0$.



Taking any other vector parallel to u, we obtain from (1.40)

$$\|f - \alpha u\|^2 = \|f_\perp + (f_\parallel - \alpha u)\|^2 = \|f_\perp\|^2 + |\langle u, f\rangle - \alpha|^2 \tag{1.43}$$

and hence $f_\parallel$ is the unique vector parallel to u which is closest to f.

As a first consequence we obtain the **Cauchy–Bunyakovsky–Schwarz** inequality:

Theorem 1.5 (Cauchy–Bunyakovsky–Schwarz)**.** *Let $\mathfrak{H}_0$ be an inner product space. Then for every $f, g \in \mathfrak{H}_0$ we have*

$$|\langle f, g\rangle| \le \|f\|\,\|g\| \tag{1.44}$$

with equality if and only if f and g are parallel.

Proof. It suffices to prove the case $\|g\| = 1$. But then the claim follows from $\|f\|^2 = |\langle g, f\rangle|^2 + \|f_\perp\|^2$. □

We will follow common practice and refer to (1.44) simply as Cauchy–Schwarz inequality. Note that the Cauchy–Schwarz inequality implies that the scalar product is continuous in both variables; that is, if $f_n \to f$ and $g_n \to g$, we have $\langle f_n, g_n\rangle \to \langle f, g\rangle$.

As another consequence we infer that the map $\|.\|$ is indeed a norm. In fact,

$$\|f + g\|^2 = \|f\|^2 + \langle f, g\rangle + \langle g, f\rangle + \|g\|^2 \le (\|f\| + \|g\|)^2. \tag{1.45}$$

But let us return to $C(I)$. Can we find a scalar product which has the maximum norm as associated norm? Unfortunately the answer is no! The reason is that the maximum norm does not satisfy the parallelogram law (Problem 1.18).

Theorem 1.6 (Jordan–von Neumann)**.** *A norm is associated with a scalar product if and only if the* ***parallelogram law***

$$\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2 \tag{1.46}$$

holds.

In this case the scalar product can be recovered from its norm by virtue of the ***polarization identity***

$$\langle f, g\rangle = \frac{1}{4}\left(\|f + g\|^2 - \|f - g\|^2 + \mathrm{i}\|f - \mathrm{i}g\|^2 - \mathrm{i}\|f + \mathrm{i}g\|^2\right). \tag{1.47}$$

Proof. If an inner product space is given, verification of the parallelogram law and the polarization identity is straightforward (Problem 1.19).

To show the converse, we define

$$s(f, g) := \frac{1}{4}\left(\|f + g\|^2 - \|f - g\|^2 + \mathrm{i}\|f - \mathrm{i}g\|^2 - \mathrm{i}\|f + \mathrm{i}g\|^2\right).$$

Then $s(f, f) = \|f\|^2$ and $s(f, g) = s(g, f)^*$ are straightforward to check. Moreover, another straightforward computation using the parallelogram law shows

$$s(f, g) + s(f, h) = 2s(f, \frac{g + h}{2}).$$

Now choosing $h = 0$ (and using $s(f, 0) = 0$) shows $s(f, g) = 2s(f, \frac{g}{2})$ and thus $s(f, g)+s(f, h) = s(f, g+h)$. Furthermore, by induction we infer $\frac{m}{2^n}s(f, g) = s(f, \frac{m}{2^n}g)$; that is, $\alpha\, s(f, g) = s(f, \alpha g)$ for a dense set of positive rational numbers α. By continuity (which follows from continuity of the norm) this holds for all $\alpha \ge 0$ and $s(f, -g) = -s(f, g)$, respectively, $s(f, \mathrm{i}g) = \mathrm{i}\, s(f, g)$, finishes the proof. □

In the case of a real Hilbert space, the polarization identity of course simplifies to $\langle f, g\rangle = \frac{1}{4}(\|f+g\|^2 - \|f-g\|^2)$.

Note that the parallelogram law and the polarization identity even hold for sesquilinear forms (Problem 1.19).

But how do we define a scalar product on $C(I)$? One possibility is

$$\langle f, g\rangle := \int_a^b f^*(x)g(x)dx. \tag{1.48}$$

The corresponding inner product space is denoted by $\mathcal{L}^2_{cont}(I)$. Note that we have

$$\|f\| \le \sqrt{|b-a|}\|f\|_\infty \tag{1.49}$$

and hence the maximum norm is stronger than the $\mathcal{L}^2_{cont}$ norm.

Suppose we have two norms $\|.\|_1$ and $\|.\|_2$ on a vector space X. Then $\|.\|_2$ is said to be **stronger** than $\|.\|_1$ if there is a constant $m > 0$ such that

$$\|f\|_1 \le m\|f\|_2. \tag{1.50}$$

It is straightforward to check the following.

Lemma 1.7. *If $\|.\|_2$ is stronger than $\|.\|_1$, then every $\|.\|_2$ Cauchy sequence is also a $\|.\|_1$ Cauchy sequence.*

Hence if a function $F : X \to Y$ is continuous in $(X, \|.\|_1)$, it is also continuous in $(X, \|.\|_2)$, and if a set is dense in $(X, \|.\|_2)$, it is also dense in $(X, \|.\|_1)$.

In particular, $\mathcal{L}^2_{cont}$ is separable since the polynomials are dense. But is it also complete? Unfortunately the answer is no:

Example 1.10. Take $I = [0, 2]$ and define

$$f_n(x) := \begin{cases} 0, & 0 \le x \le 1 - \frac{1}{n}, \\ 1 + n(x-1), & 1 - \frac{1}{n} \le x \le 1, \\ 1, & 1 \le x \le 2. \end{cases}$$

Then $f_n(x)$ is a Cauchy sequence in $\mathcal{L}^2_{cont}$, but there is no limit in $\mathcal{L}^2_{cont}$! Clearly, the limit should be the step function which is 0 for $0 \le x < 1$ and 1 for $1 \le x \le 2$, but this step function is discontinuous (Problem 1.22)! ⋄

Example 1.11. The previous example indicates that we should consider (1.48) on a larger class of functions, for example on the class of Riemann integrable functions

$$\mathcal{R}(I) := \{f : I \to \mathbb{C} | f \text{ is Riemann integrable}\}$$

such that the integral makes sense. While this seems natural it implies another problem: Any function which vanishes outside a set which is negligible for the integral (e.g. finitely many points) has norm zero! That is,

$\|f\|_2 := (\int_I |f(x)|^2 dx)^{1/2}$ is only a seminorm on $\mathcal{R}(I)$ (Problem 1.21). To get a norm we consider $\mathcal{N}(I) := \{f \in \mathcal{R}(I) | \, \|f\|_2 = 0\}$. By homogeneity and the triangle inequality $\mathcal{N}(I)$ is a subspace and we can consider equivalence classes of functions which differ by a negligible function from $\mathcal{N}(I)$:

$$\mathcal{L}^2_{Ri} := \mathcal{R}(I)/\mathcal{N}(I).$$

Since $\|f\|_2 = \|g\|_2$ for $f - g \in \mathcal{N}(I)$ we have a norm on $\mathcal{L}^2_{Ri}$. Moreover, since this norm inherits the parallelogram law we even have an inner product space. However, this space will not be complete unless we replace the Riemann by the Lebesgue integral. Hence we will not pursue this further until we have the Lebesgue integral at our disposal. ⋄

This shows that in infinite dimensional vector spaces, different norms will give rise to different convergent sequences. In fact, the key to solving problems in infinite dimensional spaces is often finding the right norm! This is something which cannot happen in the finite dimensional case.

Theorem 1.8. *If X is a finite dimensional vector space, then all norms are equivalent. That is, for any two given norms $\|.\|_1$ and $\|.\|_2$, there are positive constants m_1 and m_2 such that*

$$\frac{1}{m_2}\|f\|_1 \le \|f\|_2 \le m_1\|f\|_1. \tag{1.51}$$

Proof. Choose a basis $\{u_j\}_{1\le j\le n}$ such that every $f \in X$ can be written as $f = \sum_j \alpha_j u_j$. Since equivalence of norms is an equivalence relation (check this!), we can assume that $\|.\|_2$ is the usual Euclidean norm: $\|f\|_2 := \|\sum_j \alpha_j u_j\|_2 = (\sum_j |\alpha_j|^2)^{1/2}$. Then by the triangle and Cauchy–Schwarz inequalities,

$$\|f\|_1 \le \sum_j |\alpha_j| \|u_j\|_1 \le \sqrt{\sum_j \|u_j\|_1^2}\, \|f\|_2$$

and we can choose $m_2 = \sqrt{\sum_j \|u_j\|_1^2}$.

In particular, if f_n is convergent with respect to $\|.\|_2$, it is also convergent with respect to $\|.\|_1$. Thus $\|.\|_1$ is continuous with respect to $\|.\|_2$ and attains its minimum $m > 0$ on the unit sphere $S := \{u | \|u\|_2 = 1\}$ (which is compact by the Heine–Borel theorem, Theorem B.22). Now choose $m_1 = 1/m$. □

Finally, I remark that a real Hilbert space can always be embedded into a complex Hilbert space. In fact, if $\mathfrak{H}$ is a real Hilbert space, then $\mathfrak{H} \times \mathfrak{H}$ is a complex Hilbert space if we define

$$(f_1, f_2)+(g_1, g_2) = (f_1+g_1, f_2+g_2), \quad (\alpha+\mathrm{i}\beta)(f_1, f_2) = (\alpha f_1-\beta f_2, \alpha f_2+\beta f_1) \tag{1.52}$$

and

$$\langle (f_1, f_2), (g_1, g_2) \rangle = \langle f_1, g_1 \rangle + \langle f_2, g_2 \rangle + \mathrm{i}(\langle f_1, g_2 \rangle - \langle f_2, g_1 \rangle). \tag{1.53}$$

Here you should think of (f_1, f_2) as $f_1 + \mathrm{i} f_2$. Note that we have a conjugate linear map $C : \mathfrak{H} \times \mathfrak{H} \to \mathfrak{H} \times \mathfrak{H}$, $(f_1, f_2) \mapsto (f_1, -f_2)$ which satisfies $C^2 = \mathbb{I}$ and $\langle Cf, Cg \rangle = \langle g, f \rangle$. In particular, we can get our original Hilbert space back if we consider $\mathrm{Re}(f) = \frac{1}{2}(f + Cf) = (f_1, 0)$.

Problem 1.16. *Show that the norm in a Hilbert space satisfies $\|f + g\| = \|f\| + \|g\|$ if and only if $f = \alpha g$, $\alpha \geq 0$, or $g = 0$. Hence Hilbert spaces are strictly convex (cf. Problem 1.12).*

Problem 1.17 (Generalized parallelogram law)**.** *Show that, in a Hilbert space,*

$$\sum_{1 \leq j < k \leq n} \|x_j - x_k\|^2 + \| \sum_{1 \leq j \leq n} x_j \|^2 = n \sum_{1 \leq j \leq n} \|x_j\|^2$$

for every $n \in \mathbb{N}$. The case $n = 2$ is (1.46)*.*

Problem 1.18. *Show that the maximum norm on $C[0, 1]$ does not satisfy the parallelogram law.*

Problem* 1.19. *Suppose $\mathfrak{Q}$ is a complex vector space. Let $s(f, g)$ be a sesquilinear form on $\mathfrak{Q}$ and $q(f) := s(f, f)$ the associated quadratic form. Prove the* ***parallelogram law***

$$q(f + g) + q(f - g) = 2q(f) + 2q(g) \tag{1.54}$$

and the ***polarization identity***

$$s(f, g) = \frac{1}{4} \left(q(f + g) - q(f - g) + \mathrm{i}\, q(f - \mathrm{i} g) - \mathrm{i}\, q(f + \mathrm{i} g) \right). \tag{1.55}$$

Show that $s(f, g)$ is symmetric if and only if $q(f)$ is real-valued.

Note, that if $\mathfrak{Q}$ is a real vector space, then the parallelogram law is unchanged but the polarization identity in the form $s(f, g) = \frac{1}{4}(q(f+g) - q(f-g))$ will only hold if $s(f, g)$ is symmetric.

Problem 1.20. *A sesquilinear form on a complex inner product space is called* ***bounded*** *if*

$$\|s\| := \sup_{\|f\| = \|g\| = 1} |s(f, g)|$$

is finite. Similarly, the associated quadratic form q is ***bounded*** *if*

$$\|q\| := \sup_{\|f\| = 1} |q(f)|$$

is finite. Show

$$\|q\| \leq \|s\| \leq 2\|q\|.$$

(Hint: Use the polarization identity from the previous problem.)

Problem* 1.21. *Suppose $\mathfrak{Q}$ is a vector space. Let $s(f,g)$ be a sesquilinear form on $\mathfrak{Q}$ and $q(f) := s(f,f)$ the associated quadratic form. Show that the Cauchy–Schwarz inequality*

$$|s(f,g)| \le q(f)^{1/2} q(g)^{1/2}$$

holds if $q(f) \ge 0$. In this case $q(.)^{1/2}$ satisfies the triangle inequality and hence is a seminorm.

(Hint: Consider $0 \le q(f+\alpha g) = q(f) + 2\mathrm{Re}(\alpha\, s(f,g)) + |\alpha|^2 q(g)$ and choose $\alpha = t\, s(f,g)^*/|s(f,g)|$ with $t \in \mathbb{R}$.)

Problem* 1.22. *Prove the claims made about f_n in Example 1.10.*

1.4. Completeness

Since $\mathcal{L}^2_{cont}$ is not complete, how can we obtain a Hilbert space from it? Well, the answer is simple: take the **completion**.

If X is an (incomplete) normed space, consider the set of all Cauchy sequences $\mathcal{X}$. Call two Cauchy sequences equivalent if their difference converges to zero and denote by $\bar{X}$ the set of all equivalence classes. It is easy to see that $\bar{X}$ (and $\mathcal{X}$) inherit the vector space structure from X. Moreover,

Lemma 1.9. *If x_n is a Cauchy sequence in X, then $\|x_n\|$ is also a Cauchy sequence and thus converges.*

Consequently, the norm of an equivalence class $[(x_n)_{n=1}^\infty]$ can be defined by $\|[(x_n)_{n=1}^\infty]\| := \lim_{n\to\infty} \|x_n\|$ and is independent of the representative (show this!). Thus $\bar{X}$ is a normed space.

Theorem 1.10. *$\bar{X}$ is a Banach space containing X as a dense subspace if we identify $x \in X$ with the equivalence class of all sequences converging to x.*

Proof. (Outline) It remains to show that $\bar{X}$ is complete. Let $\xi_n = [(x_{n,j})_{j=1}^\infty]$ be a Cauchy sequence in $\bar{X}$. Without loss of generality (by dropping terms) we can choose the representatives $x_{n,j}$ such that $|x_{n,j} - x_{n,k}| \le \frac{1}{n}$ for $j,k \ge n$. Then it is not hard to see that $\xi = [(x_{j,j})_{j=1}^\infty]$ is its limit. □

Let me remark that the completion $\bar{X}$ is unique. More precisely, every other complete space which contains X as a dense subset is isomorphic to $\bar{X}$. This can for example be seen by showing that the identity map on X has a unique extension to $\bar{X}$ (compare Theorem 1.17 below).

In particular, it is no restriction to assume that a normed vector space or an inner product space is complete (note that by continuity of the norm the parallelogram law holds for $\bar{X}$ if it holds for X).

Example 1.12. The completion of the space $\mathcal{L}^2_{cont}(I)$ is denoted by $L^2(I)$. While this defines $L^2(I)$ uniquely (up to isomorphisms) it is often inconvenient to work with equivalence classes of Cauchy sequences. A much more convenient characterization can be given with the help of the Lebesgue integral (see Chapter 10 if you are familiar with basic Lebesgue integration).

Similarly, we define $L^p(I)$, $1 \le p < \infty$, as the completion of $C(I)$ with respect to the norm

$$\|f\|_p := \left(\int_a^b |f(x)|^p dx\right)^{1/p}.$$

The only requirement for a norm which is not immediate is the triangle inequality (except for $p = 1, 2$) but this can be shown as for ℓ^p (cf. Problem 1.25). ⋄

Problem 1.23. *Provide a detailed proof of Theorem 1.10.*

Problem 1.24. *For every $f \in L^1(I)$ we can define its integral*

$$\int_c^d f(x)dx$$

as the (unique) extension of the corresponding linear functional from $C(I)$ to $L^1(I)$ (by Theorem 1.17 below). Show that this integral is linear and satisfies

$$\int_c^e f(x)dx = \int_c^d f(x)dx + \int_d^e f(x)dx, \qquad \left|\int_c^d f(x)dx\right| \le \int_c^d |f(x)|dx.$$

Problem* 1.25. *Show the* ***Hölder inequality***

$$\|fg\|_1 \le \|f\|_p \|g\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad 1 \le p, q \le \infty,$$

for $f \in L^p(I)$, $g \in L^q(I)$ and conclude that $\|.\|_p$ is a norm on $C(I)$. Also conclude that $L^p(I) \subseteq L^1(I)$.

1.5. Compactness

In finite dimensions relatively compact sets are easily identified as they are precisely the bounded sets by the Heine–Borel theorem (Theorem B.22). In the infinite dimensional case the situation is more complicated. Before we look into this, please recall that for a subset U of a Banach space the following are equivalent (see Corollary B.20 and Lemma B.26):

- U is relatively compact (i.e. its closure is compact)
- every sequence from U has a convergent subsequence
- U is totally bounded (i.e. it has a finite ε-cover for every $\varepsilon > 0$)

Example 1.13. Consider the bounded sequence $(\delta^n)_{n=1}^\infty$ in $\ell^p(\mathbb{N})$. Since $\|\delta^n - \delta^m\|_p = 1$ for $n \neq m$, there is no way to extract a convergent subsequence. ⋄

In particular, the Heine–Borel theorem fails for $\ell^p(\mathbb{N})$. In fact, it fails in any infinite dimensional space.

Theorem 1.11. *The closed unit ball in a Banach space X is compact if and only if X is finite dimensional.*

Proof. If X is finite dimensional, then by Theorem 1.8 we can assume $X = \mathbb{C}^n$ and the closed unit ball is compact by the Heine–Borel theorem.

Conversely, suppose $S = \{x \in X | \, \|x\| = 1\}$ is compact. Then $\{X \setminus \operatorname{Ker}(\ell)\}_{\ell \in X^*}$ is an open cover since for every $x \in S$ there is some $\ell \in X^*$ with $\ell(x) \neq 0$. This is easy to see for the Banach spaces we have dealt with so far and it will follow in general by the Hahn–Banach theorem (see Corollary 4.15). This cover has a finite subcover, $S \subset \bigcup_{j=1}^n X \setminus \operatorname{Ker}(\ell_j) = X \setminus \bigcap_{j=1}^n \operatorname{Ker}(\ell_j)$. Hence $\bigcap_{j=1}^n \operatorname{Ker}(\ell_j) = \{0\}$ and the map $X \to \mathbb{C}^n$, $x \mapsto (\ell_1(x), \cdots, \ell_n(x))$ is injective, that is, $\dim(X) \leq n$. □

Hence one needs criteria when a given subset is relatively compact. Our strategy will be based on total boundedness and can be outlined as follows: Project the original set to some finite dimensional space such that the information loss can be made arbitrarily small (by increasing the dimension of the finite dimensional space) and apply Heine–Borel to the finite dimensional space. This idea is formalized in the following lemma.

Lemma 1.12. *Let X be a metric space and K some subset. Assume that for every $\varepsilon > 0$ there is a metric space Y_n, a surjective map $P_n : X \to Y_n$, and some $\delta > 0$ such that $P_n(K)$ is totally bounded and $d(x, y) < \varepsilon$ whenever $x, y \in K$ with $d(P_n(x), P_n(y)) < \delta$. Then K is totally bounded.*

In particular, if X is a Banach space the claim holds if P_n can be chosen a linear map onto a finite dimensional subspace Y_n such that $\|P_n\| \leq C$, $P_n K$ is bounded, and $\|(1 - P_n)x\| \leq \varepsilon$ for $x \in K$.

Proof. Fix $\varepsilon > 0$. Then by total boundedness of $P_n(K)$ we can find a δ-cover $\{B_\delta(y_j)\}_{j=1}^m$ for $P_n(K)$. Now if we choose $x_j \in P_n^{-1}(\{y_j\}) \cap K$, then $\{B_\varepsilon(x_j)\}_{j=1}^n$ is an ε-cover for K since $P_n^{-1}(B_\delta(y_j)) \cap K \subseteq B_\varepsilon(x_j)$.

For the last claim take P_n corresponding to $\varepsilon/3$ and note that $\|x - y\| \leq \|(1 - P_n)x\| + \|P_n(x - y)\| + \|(1 - P_n)y\| < \varepsilon$ for $\delta := \varepsilon/3$. □

The first application will be to $\ell^p(\mathbb{N})$.

Theorem 1.13 (Fréchet)**.** *Consider $\ell^p(\mathbb{N})$, $1 \le p < \infty$, and let $P_n a = (a_1, \dots, a_n, 0, \dots)$ be the projection onto the first n components. A subset $\mathcal{K} \subseteq \ell^p(\mathbb{N})$ is relatively compact if and only if*

(i) *it is pointwise bounded, $\sup_{a\in\mathcal{K}} |a_j| \le M_j$ for all $j \in \mathbb{N}$, and*

(ii) *for every $\varepsilon > 0$ there is some n such that $\|(1 - P_n)a\|_p \le \varepsilon$ for all $a \in \mathcal{K}$.*

In the case $p = \infty$ conditions (i) and (ii) still imply that $\mathcal{K}$ is relatively compact, but the converse only holds for $\mathcal{K} \subseteq c_0(\mathbb{N})$.

Proof. Clearly (i) and (ii) is what is needed for Lemma 1.12.

Conversely, if $\mathcal{K}$ is relatively compact it is bounded. Moreover, given δ we can choose a finite δ-cover $\{B_\delta(a^j)\}_{j=1}^m$ for $\mathcal{K}$ and some n such that $\|(1 - P_n)a^j\|_p \le \delta$ for all $1 \le j \le m$. Now given $a \in \mathcal{K}$ we have $a \in B_\delta(a^j)$ for some j and hence $\|(1-P_n)a\|_p \le \|(1-P_n)(a-a^j)\|_p + \|(1-P_n)a^j\|_p \le 2\delta$ as required. □

Example 1.14. Fix $a \in \ell^p(\mathbb{N})$ if $1 \le p < \infty$ or $a \in c_0(\mathbb{N})$ else. Then $\mathcal{K} := \{b\,|\,|b_j| \le |a_j|\} \subset \ell^p(\mathbb{N})$ is compact. ⋄

The second application will be to $C(I)$. A family of functions $F \subset C(I)$ is called (pointwise) **equicontinuous** if for every $\varepsilon > 0$ and every $x \in I$ there is a $\delta > 0$ such that

$$|f(y) - f(x)| \le \varepsilon \quad \text{whenever} \quad |y - x| < \delta, \quad \forall f \in F. \tag{1.56}$$

That is, in this case δ is required to be independent of the function $f \in F$.

Theorem 1.14 (Arzelà–Ascoli)**.** *Let $F \subset C(I)$ be a family of continuous functions. Then every sequence from F has a uniformly convergent subsequence if and only if F is equicontinuous and the set $\{f(x_0)|f \in F\}$ is bounded for one $x_0 \in I$. In this case F is even bounded.*

Proof. Suppose F is equicontinuous and pointwise bounded. Fix $\varepsilon > 0$. By compactness of I there are finitely many points $x_1, \dots, x_n \in I$ such that the balls $B_{\delta_j}(x_j)$ cover I, where δ_j is the δ corresponding to x_j as in the definition of equicontinuity. Now first of all note that, since I is connected and since $x_0 \in B_{\delta_j}(x_j)$ for some j, we see that F is bounded: $|f(x)| \le \sup_{f\in F} |f(x_0)| + n\varepsilon$.

Next consider $P : C[0, 1] \to \mathbb{C}^n$, $P(f) = (f(x_1), \dots, f(x_n))$. Then $P(F)$ is bounded and $\|f - g\|_\infty \le 3\varepsilon$ whenever $\|P(f) - P(g)\|_\infty < \varepsilon$. Indeed, just note that for every x there is some j such that $x \in B_{\delta_j}(x_j)$ and thus $|f(x) - g(x)| \le |f(x) - f(x_j)| + |f(x_j) - g(x_j)| + |g(x_j) - g(x)| \le 3\varepsilon$. Hence F is relatively compact by Lemma 1.12.

Conversely, suppose F is relatively compact. Then F is totally bounded and hence bounded. To see equicontinuity fix $x \in I$, $\varepsilon > 0$ and choose a corresponding ε-cover $\{B_\varepsilon(f_j)\}_{j=1}^n$ for F. Pick $\delta > 0$ such that $y \in B_\delta(x)$ implies $|f_j(y) - f_j(x)| < \varepsilon$ for all $1 \le j \le n$. Then $f \in B_\varepsilon(f_j)$ for some j and hence $|f(y) - f(x)| \le |f(y) - f_j(y)| + |f_j(y) - f_j(x)| + |f_j(x) - f(x)| \le 3\varepsilon$, proving equicontinuity. □

Example 1.15. Consider the solution $y_n(x)$ of the initial value problem

$$y' = \sin(ny), \qquad y(0) = 1.$$

(Assuming this solution exists — it can in principle be found using separation of variables.) Then $|y_n'(x)| \le 1$ and hence the mean value theorem shows that the family $\{y_n\} \subseteq C([0,1])$ is equicontinuous. Hence there is a uniformly convergent subsequence. ⋄

Problem 1.26. *Show that a subset $F \subset c_0(\mathbb{N})$ is relatively compact if and only if there is a nonnegative sequence $a \in c_0(\mathbb{N})$ such that $|b_n| \le a_n$ for all $n \in \mathbb{N}$ and all $b \in F$.*

Problem 1.27. *Find a family in $C[0,1]$ that is equicontinuous but not bounded.*

Problem 1.28. *Which of the following families are relatively compact in $C[0,1]$?*

(i) $F = \{f \in C^1[0,1] | \, \|f\|_\infty \le 1\}$
(ii) $F = \{f \in C^1[0,1] | \, \|f'\|_\infty \le 1\}$
(iii) $F = \{f \in C^1[0,1] | \, \|f\|_\infty \le 1, \|f'\|_2 \le 1\}$

1.6. Bounded operators

Given two normed spaces X and Y A linear map

$$A : \mathfrak{D}(A) \subseteq X \to Y \tag{1.57}$$

will be called a **(linear) operator**. The linear subspace $\mathfrak{D}(A)$ on which A is defined is called the **domain** of A and is frequently required to be dense. The **kernel** (also **null space**)

$$\mathrm{Ker}(A) := \{f \in \mathfrak{D}(A) | Af = 0\} \subseteq X \tag{1.58}$$

and **range**

$$\mathrm{Ran}(A) := \{Af | f \in \mathfrak{D}(A)\} = A\mathfrak{D}(A) \subseteq Y \tag{1.59}$$

are again linear subspaces. Note that a linear map A will be continuous if and only it is continuous at 0, that is, $x_n \in \mathfrak{D}(A) \to 0$ implies $Ax_n \to 0$.

The operator A is called **bounded** if the operator norm

$$\|A\| := \sup_{f\in\mathfrak{D}(A),\|f\|_X=1} \|Af\|_Y \tag{1.60}$$

is finite. This says that A is bounded if the image of the closed unit ball $\bar{B}_1(0) \subset X$ is contained in some closed ball $\bar{B}_r(0) \subset Y$ of finite radius r (with the smallest radius being the operator norm). Hence A is bounded if and only if it maps bounded sets to bounded sets.

Note that if you replace the norm on X or Y then the operator norm will of course also change in general. However, if the norms are equivalent so will be the operator norms.

By construction, a bounded operator satisfies

$$\|Af\|_Y \le \|A\|\|f\|_X, \qquad f \in \mathfrak{D}(A), \tag{1.61}$$

and hence is is Lipschitz continuous, that is, $\|Af - Ag\|_Y \le \|A\|\|f - g\|_X$ for $f, g \in \mathfrak{D}(A)$. In particular, it is continuous. The converse is also true:

Theorem 1.15. *A linear operator A is bounded if and only if it is continuous.*

Proof. Suppose A is continuous but not bounded. Then there is a sequence of unit vectors $u_n \in \mathfrak{D}(A)$ such that $\|Au_n\_Y| \ge n$. Then $f_n := \frac{1}{n}u_n$ converges to 0 but $\|Af_n\|_Y \ge 1$ does not converge to 0. □

If X is finite dimensional, then every operator is bounded.

Lemma 1.16. *Let X, Y be normed spaces with X finite dimensional. Then every linear operator $A : X \to Y$ is bounded.*

Proof. Choose a basis $\{x_j\}_{j=1}^n$ for X such that every $x \in X$ can be written as $x = \sum_{j=1}^n \alpha_j x_j$. By Theorem 1.8 we can assume $\|x\|_X = (\sum_{j=1}^n |\alpha_j|^2)^{1/2}$ without loss of generality. Then

$$\|Ax\|_Y \le \sum_{j=1}^n |\alpha_j|\|Ax_j\|_Y \le \sqrt{\sum_{j=1}^n \|Ax_j\|_Y^2}\,\|x\|$$

and thus $\|A\| \le (\sum_{j=1}^n \|Ax_j\|_Y^2)^{1/2}$. □

In the infinite dimensional case an operator can be unbounded. Moreover, one and the same operation might be bounded (i.e. continuous) or unbounded, depending on the norm chosen.

Example 1.16. Let $X := \ell^p(\mathbb{N})$ and $a \in \ell^\infty(\mathbb{N})$. Consider the multiplication operator $A : X \to X$ defined by

$$(Ab)_j := a_j b_j.$$

Then $|(Ab)_j| \le \|a\|_\infty |b_j|$ shows $\|A\| \le \|a\|_\infty$. In fact, we even have $\|A\| = \|a\|_\infty$ (show this). If a is unbounded we need a domain $\mathfrak{D}(A) := \{b \in \ell^p(\mathbb{N}) | (a_j b_j)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})\}$ and A will be unbounded (show this). ⋄

Example 1.17. Consider the vector space of differentiable functions $X := C^1[0,1]$ and equip it with the norm (cf. Problem 1.31)

$$\|f\|_{\infty,1} := \max_{x\in[0,1]} |f(x)| + \max_{x\in[0,1]} |f'(x)|.$$

Let $Y := C[0,1]$ and observe that the differential operator $A = \frac{d}{dx} : X \to Y$ is bounded since

$$\|Af\|_\infty = \max_{x\in[0,1]} |f'(x)| \le \max_{x\in[0,1]} |f(x)| + \max_{x\in[0,1]} |f'(x)| = \|f\|_{\infty,1}.$$

However, if we consider $A = \frac{d}{dx} : \mathfrak{D}(A) \subseteq Y \to Y$ defined on $\mathfrak{D}(A) = C^1[0,1]$, then we have an unbounded operator. Indeed, choose $u_n(x) := \sin(n\pi x)$ which is normalized, $\|u_n\|_\infty = 1$, and observe that

$$Au_n(x) = u_n'(x) = n\pi \cos(n\pi x)$$

is unbounded, $\|Au_n\|_\infty = n\pi$. Note that $\mathfrak{D}(A)$ contains the set of polynomials and thus is dense by the Weierstraß approximation theorem (Theorem 1.3). ⋄

If A is bounded and densely defined, it is no restriction to assume that it is defined on all of X.

Theorem 1.17. *Let $A : \mathfrak{D}(A) \subseteq X \to Y$ be a bounded linear operator between a normed space X and a Banach space Y. If $\mathfrak{D}(A)$ is dense, there is a unique (continuous) extension of A to X which has the same operator norm.*

Proof. Since a bounded operator maps Cauchy sequences to Cauchy sequences, this extension can only be given by

$$\overline{A}f := \lim_{n\to\infty} Af_n, \qquad f_n \in \mathfrak{D}(A), \quad f \in X.$$

To show that this definition is independent of the sequence $f_n \to f$, let $g_n \to f$ be a second sequence and observe

$$\|Af_n - Ag_n\| = \|A(f_n - g_n)\| \le \|A\|\|f_n - g_n\| \to 0.$$

Since for $f \in \mathfrak{D}(A)$ we can choose $f_n = f$, we see that $\overline{A}f = Af$ in this case, that is, $\overline{A}$ is indeed an extension. From continuity of vector addition and scalar multiplication it follows that $\overline{A}$ is linear. Finally, from continuity of the norm we conclude that the operator norm does not increase. □

The set of all bounded linear operators from X to Y is denoted by $\mathscr{L}(X,Y)$. If $X = Y$, we write $\mathscr{L}(X) := \mathscr{L}(X,X)$. An operator in $\mathscr{L}(X,\mathbb{C})$ is called a **bounded linear functional**, and the space $X^* := \mathscr{L}(X,\mathbb{C})$ is

called the **dual space** of X. The dual space takes the role of coordinate functions in a Banach space.

Example 1.18. Let X be a finite dimensional space and $\{x_j\}_{j=1}^n$ a basis. Then every $x \in X$ can be uniquely written as $x = \sum_{j=1}^n \alpha_j x_j$ and we can define linear functionals via $\ell_j(x) := \alpha_j$ for $1 \le j \le n$. The functionals $\{\ell_j\}_{j=1}^n$ are called a **dual basis** since $\ell_k(x_j) = \delta_{j,k}$ and since any other linear functional $\ell \in X^*$ can be written as $\ell = \sum_{j=1}^n \ell(x_j)\ell_j$. In particular, X and X^* have the same dimension. ⋄

Example 1.19. Let $X := \ell^p(\mathbb{N})$. Then the coordinate functions

$$\ell_j(a) := a_j$$

are bounded linear functionals: $|\ell_j(a)| = |a_j| \le \|a\|_p$ and hence $\|\ell_j\| = 1$. More general, let $b \in \ell^q(\mathbb{N})$ where $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\ell_b(a) := \sum_{j=1}^n b_j a_j$$

is a bounded linear functional satisfying $\|\ell_b\| \le \|b\|_q$ by Hölder's inequality. In fact, we even have $\|\ell_b\| = \|b\|_q$ (Problem 4.19). Note that the first example is a special case of the second one upon choosing $b = \delta^j$. ⋄

Example 1.20. Consider $X := C(I)$. Then for every $x_0 \in I$ the point evaluation $\ell_{x_0}(f) := f(x_0)$ is a bounded linear functional. In fact, $\|\ell_{x_0}\| = 1$ (show this).

However, note that ℓ_{x_0} is unbounded on $\mathcal{L}^2_{cont}(I)$! To see this take $f_n(x) := \sqrt{\frac{3n}{2}} \max(0, 1 - n|x - x_0|)$ which is a triangle shaped peak supported on $[x_0 - n^{-1}, x_0 + n^{-1}]$ and normalized according to $\|f_n\|_2 = 1$ for n sufficiently large such that the support is contained in I. Then $\ell_{x_0}(f) = f_n(x_0) = \sqrt{\frac{3n}{2}} \to \infty$. This implies that ℓ_{x_0} cannot be extended to the completion of $\mathcal{L}^2_{cont}(I)$ in a natural way and reflects the fact that the integral cannot see individual points (changing the value of a function at one point does not change its integral). ⋄

Example 1.21. Consider $X := C(I)$ and let g be some continuous function. Then

$$\ell_g(f) := \int_a^b g(x) f(x) dx$$

is a linear functional with norm $\|\ell_g\| = \|g\|_1$. Indeed, first of all note that

$$|\ell_g(f)| \le \int_a^b |g(x) f(x)| dx \le \|f\|_\infty \int_a^b |g(x)| dx$$

shows $\|\ell_g\| \le \|g\|_1$. To see that we have equality consider $f_\varepsilon = g^*/(|g| + \varepsilon)$ and note

$$|\ell_g(f_\varepsilon)| = \int_a^b \frac{|g(x)|^2}{|g(x)| + \varepsilon} dx \ge \int_a^b \frac{|g(x)|^2 - \varepsilon^2}{|g(x)| + \varepsilon} dx = \|g\|_1 - (b-a)\varepsilon.$$

Since $\|f_\varepsilon\| \le 1$ and $\varepsilon > 0$ is arbitrary this establishes the claim. ⋄

Theorem 1.18. *The space $\mathscr{L}(X, Y)$ together with the operator norm* (1.60) *is a normed space. It is a Banach space if Y is.*

Proof. That (1.60) is indeed a norm is straightforward. If Y is complete and A_n is a Cauchy sequence of operators, then $A_n f$ converges to an element g for every f. Define a new operator A via $Af = g$. By continuity of the vector operations, A is linear and by continuity of the norm $\|Af\| = \lim_{n\to\infty} \|A_n f\| \le (\lim_{n\to\infty} \|A_n\|)\|f\|$, it is bounded. Furthermore, given $\varepsilon > 0$, there is some N such that $\|A_n - A_m\| \le \varepsilon$ for $n, m \ge N$ and thus $\|A_n f - A_m f\| \le \varepsilon \|f\|$. Taking the limit $m \to \infty$, we see $\|A_n f - Af\| \le \varepsilon \|f\|$; that is, $A_n \to A$. □

The Banach space of bounded linear operators $\mathscr{L}(X)$ even has a multiplication given by composition. Clearly, this multiplication is distributive

$$(A+B)C = AC + BC, \quad A(B+C) = AB + BC, \quad A, B, C \in \mathscr{L}(X) \tag{1.62}$$

and associative

$$(AB)C = A(BC), \qquad \alpha\,(AB) = (\alpha A)B = A\,(\alpha B), \quad \alpha \in \mathbb{C}. \tag{1.63}$$

Moreover, it is easy to see that we have

$$\|AB\| \le \|A\|\|B\|. \tag{1.64}$$

In other words, $\mathscr{L}(X)$ is a so-called **Banach algebra**. However, note that our multiplication is not commutative (unless X is one-dimensional). We even have an **identity**, the identity operator $\mathbb{I}$, satisfying $\|\mathbb{I}\| = 1$.

Problem 1.29. *Consider $X = \mathbb{C}^n$ and let $A \in \mathscr{L}(X)$ be a matrix. Equip X with the norm (show that this is a norm)*

$$\|x\|_\infty := \max_{1\le j\le n} |x_j|$$

and compute the operator norm $\|A\|$ with respect to this norm in terms of the matrix entries. Do the same with respect to the norm

$$\|x\|_1 := \sum_{1\le j\le n} |x_j|.$$

Problem 1.30. *Show that the integral operator*

$$(Kf)(x) := \int_0^1 K(x, y) f(y) dy,$$

where $K(x,y) \in C([0,1]\times[0,1])$, defined on $\mathfrak{D}(K) := C[0,1]$, is a bounded operator both in $X := C[0,1]$ (max norm) and $X := \mathcal{L}^2_{cont}(0,1)$. Show that the norm in the $X = C[0,1]$ case is given by

$$\|K\| = \max_{x\in[0,1]} \int_0^1 |K(x,y)|dy.$$

Problem* 1.31. *Let I be a compact interval. Show that the set of differentiable functions $C^1(I)$ becomes a Banach space if we set $\|f\|_{\infty,1} := \max_{x\in I}|f(x)| + \max_{x\in I}|f'(x)|$.*

Problem* 1.32. *Show that $\|AB\| \le \|A\|\|B\|$ for every $A, B \in \mathscr{L}(X)$. Conclude that the multiplication is continuous: $A_n \to A$ and $B_n \to B$ imply $A_nB_n \to AB$.*

Problem 1.33. *Let $A \in \mathscr{L}(X)$ be a bijection. Show*

$$\|A^{-1}\|^{-1} = \inf_{f\in X, \|f\|=1} \|Af\|.$$

Problem* 1.34. *Suppose $B \in \mathscr{L}(X)$ with $\|B\| < 1$. Then $\mathbb{I}+B$ is invertible with*

$$(\mathbb{I}+B)^{-1} = \sum_{n=0}^{\infty} (-1)^n B^n.$$

Consequently for $A, B \in \mathscr{L}(X,Y)$, $A+B$ is invertible if A is invertible and $\|B\| < \|A^{-1}\|^{-1}$.

Problem* 1.35. *Let*

$$f(z) := \sum_{j=0}^{\infty} f_j z^j, \qquad |z| < R,$$

be a convergent power series with radius of convergence $R > 0$. Suppose X is a Banach space and $A \in \mathscr{L}(X)$ is a bounded operator with $\limsup_n \|A^n\|^{1/n} < R$ (note that by $\|A^n\| \le \|A\|^n$ the limsup is finite). Show that

$$f(A) := \sum_{j=0}^{\infty} f_j A^j$$

exists and defines a bounded linear operator. Moreover, if f and g are two such functions and $\alpha \in \mathbb{C}$, then

$$(f+g)(A) = f(A) + g(A), \quad (\alpha f)(A) = \alpha f(a), \quad (f\,g)(A) = f(A)g(A).$$

(Hint: Problem 1.4.)

Problem* 1.36. *Show that a linear map $\ell : X \to \mathbb{C}$ is continuous if and only if its kernel is closed. (Hint: If ℓ is not continuous, we can find a sequence of normalized vectors x_n with $|\ell(x_n)| \to \infty$ and a vector y with $\ell(y) = 1$.)*

1.7. Sums and quotients of Banach spaces

Given two Banach spaces X_1 and X_2 we can define their **(direct) sum** $X := X_1 \oplus X_2$ as the Cartesian product $X_1 \times X_2$ together with the norm $\|(x_1, x_2)\| := \|x_1\| + \|x_2\|$. Clearly X is again a Banach space and a sequence in X converges if and only if the components converge in X_1 and X_2, respectively. In fact, since all norms on $\mathbb{R}^2$ are equivalent (Theorem 1.8), we could as well take $\|(x_1, x_2)\| := (\|x_1\|^p + \|x_2\|^p)^{1/p}$ or $\|(x_1, x_2)\| := \max(\|x_1\|, \|x_2\|)$. We will write $X_1 \oplus_p X_2$ if we want to emphasize the norm used. In particular, in the case of Hilbert spaces the choice $p = 2$ will ensure that X is again a Hilbert space.

Note that X_1 and X_2 can be regarded as closed subspaces of $X_1 \times X_2$ by virtue of the obvious embeddings $x_1 \hookrightarrow (x_1, 0)$ and $x_2 \hookrightarrow (0, x_2)$. It is straightforward to generalize this concept to finitely many spaces (Problem 1.37).

If $A_j : \mathfrak{D}(A_j) \subseteq X_j \to Y_j$, $j = 1, 2$, are linear operators, then $A_1 \oplus A_2 : \mathfrak{D}(A_1) \times \mathfrak{D}(A_2) \subseteq X_1 \oplus X_2 \to Y_1 \oplus Y_2$ is defined as $A_1 \oplus A_2(x_1, x_2) = (A_1 x_1, A_2 x_2)$. Clearly $A_1 \oplus A_2$ will be bounded if and only if both A_1 and A_2 are bounded and $\|A_1 \oplus A_2\| = \max(\|A_1\|, \|A_2\|)$.

Note that if $A_j : X_j \to Y$, $j = 1, 2$, there is another natural way of defining an associated operator $X_1 \oplus X_2 \to Y$ given by $A_1 \hat{\oplus} A_2(x_1, x_2) := A_1 x_1 + A_2 x_2$. In particular, in the case $Y = \mathbb{C}$ we get that $(X_1 \oplus_p X_2)^* \cong X_1^* \oplus_q X_2^*$ for $\frac{1}{p} + \frac{1}{q} = 1$ via the identification $(\ell_1, \ell_2) \in X_1^* \oplus_q X_2^* \mapsto \ell_1 \hat{\oplus} \ell_2 \in (X_1 \oplus_p X_2)^*$. Clearly this identification is bijective and preserves the norm (to see this relate it to Hölder's inequality in $\mathbb{C}^2$ and note that equality is attained).

Given two subspaces $M, N \subseteq X$ of a vector space, we can define their sum as usual: $M + N := \{x + y | x \in M,\ y \in N\}$. In particular, the decomposition $x + y$ with $x \in M$, $y \in N$ is unique iff $M \cap N = \{0\}$ and we will write $M \dot{+} N$ in this case. It is important to observe, that $M \dot{+} N$ is in general different from $M \oplus N$ since both have different norms. In fact, $M \dot{+} N$ might not even be closed (no problems occur if one of the spaces is finite dimensional — see Corollary 1.20 below).

Example 1.22. Consider $X := \ell^p(\mathbb{N})$. Let $M = \{a \in X | a_{2n} = 0\}$ and $N = \{a \in X | a_{2n+1} = n^3 a_{2n}\}$. Then both subspaces are closed and $M \cap N = \{0\}$. Moreover, $M \dot{+} N$ is dense since it contains all sequences with finite support. However, it is not all of X since $a_n = \frac{1}{n^2} \notin M \dot{+} N$. Indeed, if we could write $a = b + c \in M \dot{+} N$, then $c_{2n} = \frac{1}{4n^2}$ and hence $c_{2n+1} = \frac{n}{4}$ contradicting $c \in N \subseteq X$. ⋄

A closed subspace M is called **complemented** if we can find another closed subspace N with $M \cap N = \{0\}$ and $M \dot{+} N = X$. In this case every

$x \in X$ can be uniquely written as $x = x_1 + x_2$ with $x_1 \in M$, $x_2 \in N$ and we can define a projection $P : X \to M$, $x \mapsto x_1$. By definition $P^2 = P$ and we have a complementary projection $Q := \mathbb{I} - P$ with $Q : X \to N$, $x \mapsto x_2$. Moreover, it is straightforward to check $M = \operatorname{Ker}(Q) = \operatorname{Ran}(P)$ and $N = \operatorname{Ker}(P) = \operatorname{Ran}(Q)$. Of course one would like P (and hence also Q) to be continuous. If we consider the linear operator $\phi : M \oplus N \to X$, $(x_1, x_2) \to x_1 + x_2$ then this is equivalent to the question if ϕ^{-1} is continuous. By the triangle inequality ϕ is continuous with $\|\phi\| \leq 1$ and the inverse mapping theorem (Theorem 4.6) will answer this question affirmative.

It is important to emphasize, that it is precisely the requirement that N is closed which makes P continuous (conversely observe that $N = \operatorname{Ker}(P)$ is closed if P is continuous). Without this requirement we can always find N by a simple application of Zorn's lemma (order the subspaces which have trivial intersection with M by inclusion and note that a maximal element has the required properties). Moreover, the question which closed subspaces can be complemented is a highly nontrivial one. If M is finite (co)dimensional, then it can be complemented (see Problems 1.43 and 4.25).

Given a subspace M of a linear space X we can define the **quotient space** X/M as the set of all equivalence classes $[x] = x + M$ with respect to the equivalence relation $x \equiv y$ if $x - y \in M$. It is straightforward to see that X/M is a vector space when defining $[x] + [y] = [x + y]$ and $\alpha[x] = [\alpha x]$ (show that these definitions are independent of the representative of the equivalence class). In particular, for a linear operator $A : X \to Y$ the linear space $\operatorname{Coker}(A) := Y/\operatorname{Ran}(A)$ is know as the **cokernel** of A. The dimension of X/M is known as the **codimension** of M.

Lemma 1.19. *Let M be a closed subspace of a Banach space X. Then X/M together with the norm*

$$\|[x]\| := \operatorname{dist}(x, M) = \inf_{y \in M} \|x + y\| \tag{1.65}$$

is a Banach space.

Proof. First of all we need to show that (1.65) is indeed a norm. If $\|[x]\| = 0$ we must have a sequence $y_j \in M$ with $y_j \to -x$ and since M is closed we conclude $x \in M$, that is $[x] = [0]$ as required. To see $\|\alpha[x]\| = |\alpha|\|[x]\|$ we use again the definition

$$\begin{aligned} \|\alpha[x]\| = \|[\alpha x]\| &= \inf_{y \in M} \|\alpha x + y\| = \inf_{y \in M} \|\alpha x + \alpha y\| \\ &= |\alpha| \inf_{y \in M} \|x + y\| = |\alpha|\|[x]\|. \end{aligned}$$

The triangle inequality follows with a similar argument and is left as an exercise.

Thus (1.65) is a norm and it remains to show that X/M is complete. To this end let $[x_n]$ be a Cauchy sequence. Since it suffices to show that some subsequence has a limit, we can assume $\|[x_{n+1}]-[x_n]\| < 2^{-n}$ without loss of generality. Moreover, by definition of (1.65) we can chose the representatives x_n such that $\|x_{n+1}-x_n\| < 2^{-n}$ (start with x_1 and then chose the remaining ones inductively). By construction x_n is a Cauchy sequence which has a limit $x \in X$ since X is complete. Moreover, by $\|[x_n]-[x]\| = \|[x_n-x]\| \leq \|x_n-x\|$ we see that $[x]$ is the limit of $[x_n]$. □

Observe that $\|[x]\| = \operatorname{dist}(x, M) = 0$ whenever $x \in \overline{M}$ and hence we only get a semi-norm if M is not closed.

Example 1.23. If $X := C[0,1]$ and $M := \{f \in X | f(0) = 0\}$ then $X/M \cong \mathbb{C}$. ⋄

Example 1.24. If $X := c(\mathbb{N})$, the convergent sequences and $M := c_0(\mathbb{N})$ the sequences converging to 0, then $X/M \cong \mathbb{C}$. In fact, note that every sequence $x \in c(\mathbb{N})$ can be written as $x = y + \alpha e$ with $y \in c_0(\mathbb{N})$, $e = (1, 1, 1, \dots)$, and $\alpha \in \mathbb{C}$ its limit. ⋄

Note that by $\|[x]\| \leq \|x\|$ the quotient map $\pi : X \to X/M$, $x \mapsto [x]$ is bounded with norm at most one. As a small application we note:

Corollary 1.20. *Let X be a Banach space and let $M, N \subseteq X$ be two closed subspaces with one of them, say N, finite dimensional. Then $M + N$ is also closed.*

Proof. If $\pi : X \to X/M$ denotes the quotient map, then $M + N = \pi^{-1}(\pi(N))$. Moreover, since $\pi(N)$ is finite dimensional it is closed and hence $\pi^{-1}(\pi(N))$ is closed by continuity. □

Problem* 1.37. *Let X_j, $j = 1, \dots, n$, be Banach spaces. Let $X := \bigoplus_{p,j=1}^n X_j$ be the Cartesian product $X_1 \times \cdots \times X_n$ together with the norm*

$$\|(x_1, \dots, x_n)\|_p := \begin{cases} \left(\sum_{j=1}^n \|x_j\|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{j=1,\dots,n} \|x_j\|, & p = \infty. \end{cases}$$

Show that X is a Banach space. Show that all norms are equivalent and that this product is associative $(X_1 \oplus_p X_2) \oplus_p X_3 = X_1 \oplus_p (X_2 \oplus_p X_3)$.

Problem 1.38. *Let X_j, $j \in \mathbb{N}$, be Banach spaces. Let $X := \bigoplus_{p,j\in\mathbb{N}} X_j$ be the set of all elements $x = (x_j)_{j\in\mathbb{N}}$ of the Cartesian product for which the norm*

$$\|x\|_p := \begin{cases} \left(\sum_{j\in\mathbb{N}} \|x_j\|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{j\in\mathbb{N}} \|x_j\|, & p = \infty, \end{cases}$$

is finite. Show that X is a Banach space. Show that for $1 \le p < \infty$ the elements with finitely many nonzero terms are dense and conclude that X is separable if all X_j are.

Problem 1.39. *Let ℓ be a nontrivial linear functional. Then its kernel has codimension one.*

Problem 1.40. *Compute $\|[e]\|$ in $\ell^\infty(\mathbb{N})/c_0(\mathbb{N})$, where $e = (1, 1, 1, \dots)$.*

Problem 1.41 (Complexification)**.** *Given a real normed space X its **complexification** is given by $X_\mathbb{C} := X \times X$ together with the (complex) scalar multiplication $\alpha(x, y) = (\mathrm{Re}(\alpha)x - \mathrm{Im}(\alpha)y, \mathrm{Re}(\alpha)y + \mathrm{Im}(\alpha)x)$. By virtue of the embedding $x \hookrightarrow (x, 0)$ you should of course think of (x, y) as $x + \mathrm{i}y$.*

Show that

$$\|x + \mathrm{i}y\|_\mathbb{C} := \max_{0 \le t \le \pi} \|\cos(t)x + \sin(t)y\|,$$

defines a norm on $X_\mathbb{C}$ which satisfies $\|x\|_\mathbb{C} = \|x\|$ and

$$\max(\|x\|, \|y\|) \le \|x + \mathrm{i}y\|_\mathbb{C} \le (\|x\|^2 + \|y^2\|)^{1/2}$$

In particular, this norm is equivalent to the product norm on $X \oplus X$.

If X is a Hilbert space, then the above norm will in general not give rise to a scalar product. However, any bilinear form $s : X \times X \to \mathbb{R}$ gives rise to a sesquilinear form $s_\mathbb{C}(x_1 + \mathrm{i}y_1, x_2 + \mathrm{i}y_2) := s(x_1, x_2) + s(y_1, y_2) + \mathrm{i}\big(s(x_1, y_2) - s(y_1, x_2)\big)$. If s is symmetric or positive definite, so will be $s_\mathbb{C}$. The corresponding norm satisfies $\langle x + \mathrm{i}y, x + \mathrm{i}y\rangle_\mathbb{C} = \|x\|^2 + \|y\|^2$ and is equivalent to the above one since $\frac{1}{2}(\|x\|^2 + \|y\|^2) \le \|x + \mathrm{i}y\|_\mathbb{C}^2 \le \|x\|^2 + \|y\|^2$.

Given two real normed spaces X, Y, every linear operator $A : X \to Y$ gives rise to a linear operator $A_\mathbb{C} : X_\mathbb{C} \to Y_\mathbb{C}$ via $A_\mathbb{C}(x + \mathrm{i}y) = Ax + \mathrm{i}Ay$. Show $\|A_\mathbb{C}\| = \|A\|$.

Problem* 1.42. *Suppose $A \in \mathscr{L}(X, Y)$. Show that $\mathrm{Ker}(A)$ is closed. Suppose $M \subseteq \mathrm{Ker}(A)$ is a closed subspace. Show that the induced map $\tilde{A} : X/M \to Y$, $[x] \mapsto Ax$ is a well-defined operator satisfying $\|\tilde{A}\| = \|A\|$ and $\mathrm{Ker}(\tilde{A}) = \mathrm{Ker}(A)/M$. In particular, $\tilde{A}$ is injective for $M = \mathrm{Ker}(A)$.*

Problem* 1.43. *Show that if a closed subspace M of a Banach space X has finite codimension, then it can be complemented. (Hint: Start with a basis $\{[x_j]\}$ for X/M and choose a corresponding dual basis $\{\ell_k\}$ with $\ell_k([x_j]) = \delta_{j,k}$.)*

1.8. Spaces of continuous and differentiable functions

In this section we introduce a few further sets of continuous and differentiable functions which are of interest in applications.

First, for any set $U \subseteq \mathbb{R}^m$ the set of all bounded continuous functions $C_b(U)$ together with the sup norm

$$\|f\|_\infty := \sup_{x\in U} |f(x)| \tag{1.66}$$

is a Banach space as can be shown as in Section 1.2 (or use Corollary B.35). The space of continuous functions with compact support $C_c(U) \subseteq C_b(U)$ is in general not dense and its closure will be denoted by $C_0(U)$. If U is open it can be interpreted as the functions in $C_b(U)$ which vanish at the boundary

$$C_0(U) := \{f \in C(U) | \forall \varepsilon > 0, \exists K \subseteq U \text{ compact} : |f(x)| < \varepsilon,\ x \in U \setminus K\}. \tag{1.67}$$

Of course $\mathbb{R}^m$ could be replaced by any topological space up to this point.

Moreover, the above norm can be augmented to handle differentiable functions by considering the space $C_b^1(U)$ of all continuously differentiable functions for which the following norm

$$\|f\|_{\infty,1} := \|f\|_\infty + \sum_{j=1}^m \|\partial_j f\|_\infty \tag{1.68}$$

is finite, where $\partial_j = \frac{\partial}{\partial x_j}$. Note that $\|\partial_j f\|$ for one j (or all j) is not sufficient as it is only a seminorm (it vanishes for every constant function). However, since the sum of seminorms is again a seminorm (Problem 1.45) the above expression defines indeed a norm. It is also not hard to see that $C_b^1(U)$ is complete. In fact, let f^k be a Cauchy sequence, then $f^k(x)$ converges uniformly to some continuous function $f(x)$ and the same is true for the partial derivatives $\partial_j f^k(x) \to g_j(x)$. Moreover, since $f^k(x) = f^k(c, x_2, \dots, x_m) + \int_c^{x_1} \partial_j f^k(t, x_2, \dots, x_m) dt \to f(x) = f(c, x_2, \dots, x_m) + \int_c^{x_1} g_j(t, x_2, \dots, x_m) dt$ we obtain $\partial_j f(x) = g_j(x)$. The remaining derivatives follow analogously and thus $f^k \to f$ in $C_b^1(U)$.

To extend this approach to higher derivatives let $C^k(U)$ be the set of all complex-valued functions which have partial derivatives of order up to k. For $f \in C^k(U)$ and $\alpha \in \mathbb{N}_0^n$ we set

$$\partial_\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \qquad |\alpha| = \alpha_1 + \cdots + \alpha_n. \tag{1.69}$$

An element $\alpha \in \mathbb{N}_0^n$ is called a **multi-index** and $|\alpha|$ is called its **order**. With this notation the above considerations can be easily generalized to higher order derivatives:

Theorem 1.21. *Let $U \subseteq \mathbb{R}^m$ be open. The space $C_b^k(U)$ of all functions whose partial derivatives up to order k are bounded and continuous form a*

Banach space with norm

$$\|f\|_{\infty,k} := \sum_{|\alpha|\le k} \sup_{x\in U} |\partial_\alpha f(x)|. \tag{1.70}$$

An important subspace is $C_0^k(\mathbb{R}^m)$, the set of all functions in $C_b^k(\mathbb{R}^m)$ for which $\lim_{|x|\to\infty} |\partial_\alpha f(x)| = 0$ for all $|\alpha| \le k$. For any function f not in $C_0^k(\mathbb{R}^m)$ there must be a sequence $|x_j| \to \infty$ and some α such that $|\partial_\alpha f(x_j)| \ge \varepsilon > 0$. But then $\|f - g\|_{\infty,k} \ge \varepsilon$ for every g in $C_0^k(\mathbb{R}^m)$ and thus $C_0^k(\mathbb{R}^m)$ is a closed subspace. In particular, it is a Banach space of its own.

Note that the space $C_b^k(U)$ could be further refined by requiring the highest derivatives to be Hölder continuous. Recall that a function $f : U \to \mathbb{C}$ is called uniformly **Hölder continuous** with exponent $\gamma \in (0, 1]$ if

$$[f]_\gamma := \sup_{x\ne y\in U} \frac{|f(x) - f(y)|}{|x - y|^\gamma} \tag{1.71}$$

is finite. Clearly, any Hölder continuous function is uniformly continuous and, in the special case $\gamma = 1$, we obtain the **Lipschitz continuous** functions. Note that for $\gamma = 0$ the Hölder condition boils down to boundedness and also the case $\gamma > 1$ is not very interesting (Problem 1.44).

Example 1.25. By the mean value theorem every function $f \in C_b^1(U)$ is Lipschitz continuous with $[f]_\gamma \le \|\partial f\|_\infty$, where $\partial f = (\partial_1 f, \dots, \partial_m f)$ denotes the gradient. ⋄

Example 1.26. The prototypical example of a Hölder continuous function is of course $f(x) = x^\gamma$ on $[0, \infty)$ with $\gamma \in (0, 1]$. In fact, without loss of generality we can assume $0 \le x < y$ and set $t = \frac{x}{y} \in [0, 1)$. Then we have

$$\frac{y^\gamma - x^\gamma}{(y - x)^\gamma} \le \frac{1 - t^\gamma}{(1 - t)^\gamma} \le \frac{1 - t}{1 - t} = 1.$$

From this one easily gets further examples since the composition of two Hölder continuous functions is again Hölder continuous (the exponent being the product). ⋄

It is easy to verify that this is a seminorm and that the corresponding space is complete.

Theorem 1.22. *Let $U \subseteq \mathbb{R}^m$ be open. The space $C_b^{k,\gamma}(U)$ of all functions whose partial derivatives up to order k are bounded and Hölder continuous with exponent $\gamma \in (0, 1]$ form a Banach space with norm*

$$\|f\|_{\infty,k,\gamma} := \|f\|_{\infty,k} + \sum_{|\alpha|=k} [\partial_\alpha f]_\gamma. \tag{1.72}$$

As already noted before, in the case $\gamma = 0$ we get a norm which is equivalent to $\|f\|_{\infty,k}$ and we will set $C_b^{k,0}(U) := C_b^k(U)$ for notational convenience later on.

Note that by the mean value theorem all derivatives up to order lower than k are automatically Lipschitz continuous. Moreover, every Hölder continuous function is uniformly continuous and hence has a unique extension to the closure $\overline{U}$ (cf. Theorem B.38). In this sense, the spaces $C_b^{0,\gamma}(U)$ and $C_b^{0,\gamma}(\overline{U})$ are naturally isomorphic. Consequently, we can also understand $C_b^{k,\gamma}(\overline{U})$ in this fashion since for a function from $C_b^{k,\gamma}(U)$ all derivatives have a continuous extension to $\overline{U}$. For a function in $C_b^k(U)$ this only works for the derivatives of order up to $k-1$ and hence we define $C_b^k(\overline{U})$ as the functions from $C_b^k(U)$ for which all derivatives have a continuous extensions to $\overline{U}$. Note that with this definition $C_b^k(\overline{U})$ is still a Banach space (since $C_b(\overline{U})$ is a closed subspace of $C_b(U)$).

Theorem 1.23. *Suppose $U \subset \mathbb{R}^m$ is bounded. Then $C_b^{0,\gamma_2}(\overline{U}) \subseteq C_b^{0,\gamma_1}(\overline{U}) \subseteq C(\overline{U})$ for $0 < \gamma_1 < \gamma_2 \leq 1$ with the embeddings being compact.*

Proof. That we have continuous embeddings follows since $|x-y|^{-\gamma_1} = |x-y|^{-\gamma_2+(\gamma_2-\gamma_1)} \leq (2r)^{\gamma_2-\gamma_1}|x-y|^{-\gamma_2}$ if $U \subseteq B_r(0)$. Moreover, that the embedding $C_b^{0,\gamma_1}(U) \subseteq C_b(U)$ is compact follows from the Arzelà–Ascoli theorem (Theorem B.39). To see the remaining claim let f_m be a bounded sequence in $C_b^{0,\gamma_1}(\overline{U})$, explicitly $\|f_m\|_\infty \leq C$ and $[f]_{\gamma_1} \leq C$. Hence by the Arzelà–Ascoli theorem we can assume that f_m converges uniformly to some $f \in C(\overline{U})$. Moreover, taking the limit in $|f_m(x) - f_m(y)| \leq C|x-y|^{\gamma_1}$ we see that we even have $f \in C^{0,\gamma_1}(\overline{U})$. To see that f is the limit of f_m in $C^{0,\gamma_2(\overline{U})}$ we need to show $[g_m]_{\gamma_2} \to 0$, where $g_m := f_m - f$. Now observe that

$$\begin{aligned}[g_m]_{\gamma_2} &= \sup_{x\neq y\in U:|x-y|\geq\varepsilon} \frac{|g_m(x)-g_m(y)|}{|x-y|^{\gamma_2}} + \sup_{x\neq y\in U:|x-y|<\varepsilon} \frac{|g_m(x)-g_m(y)|}{|x-y|^{\gamma_2}} \\ &\leq 2\|g_m\|_\infty \varepsilon^{-\gamma_2} + [g_m]_{\gamma_1}\varepsilon^{\gamma_1-\gamma_2} \leq 2\|g_m\|_\infty\varepsilon^{-\gamma_2} + 2C\varepsilon^{\gamma_1-\gamma_2},\end{aligned}$$

implying $\limsup_{m\to\infty}[g_m]_{\gamma_2} \leq 2C\varepsilon^{\gamma_1-\gamma_2}$ and since $\varepsilon > 0$ is arbitrary this establishes the claim. □

As pointed out in the example before, the embedding $C_b^1(U) \subseteq C_b^{0,1}(U)$ is continuous and combining this with the previous result immediately gives

Corollary 1.24. *Suppose $U \subset \mathbb{R}^m$ is bounded, $k_1, k_2 \in \mathbb{N}_0$, and $0 \leq \gamma_1, \gamma_2 \leq 1$. Then $C_b^{k_2,\gamma_2}(\overline{U}) \subseteq C_b^{k_1,\gamma_1}(\overline{U})$ for $k_1 + \gamma_1 \leq k_2 + \gamma_2$ with the embeddings being compact if the inequality is strict.*

While the above spaces are able to cover a wide variety of situations, there are still cases where the above definitions are not suitable. In fact, for

some of these cases one cannot define a suitable norm and we will postpone this to Section 5.4.

Note that in all the above spaces we could replace complex-valued by $\mathbb{C}^n$-valued functions.

For now continuous functions on subsets of $\mathbb{R}^m$ will be sufficient for our purpose. However, once we delve deeper into the subject we will also need continuous functions on topological spaces X. Luckily most of the results extend to this case in a more or less straightforward way. If you are not familiar with these extensions you can find them in Section B.8.

Problem 1.44. *Let $U \subseteq \mathbb{R}^m$ be open. Suppose $f : U \to \mathbb{C}$ is Hölder continuous with exponent $\gamma > 1$. Show that f is constant on every connected component of U.*

Problem* 1.45. *Suppose X is a vector space and $\|.\|_j$, $1 \le j \le n$, is a finite family of seminorms. Show that $\|x\| := \sum_{j=1}^n \|x\|_j$ is a seminorm. It is a norm if and only if $\|x\|_j = 0$ for all j implies $x = 0$.*

Problem* 1.46. *Let $U \subseteq \mathbb{R}^m$. Show that $C_b(U)$ is a Banach space when equipped with the sup norm. Show that $\overline{C_c(U)} = C_0(U)$. (Hint: The function $m_\varepsilon(z) = \operatorname{sign}(z)\max(0, |z| - \varepsilon) \in C(\mathbb{C})$ might be useful.)*

Problem 1.47. *Let $U \subseteq \mathbb{R}^m$. Show that the product of two bounded Hölder continuous functions is again Hölder continuous with*

$$[fg]_\gamma \le \|f\|_\infty [g]_\gamma + [f]_\gamma \|g\|_\infty.$$

Chapter 2

Hilbert spaces

The additional geometric structure of Hilbert spaces allows for an intuitive geometric solution of many problems. In fact, in many situations, e.g. in Quantum Mechanics, Hilbert spaces occur naturally. This makes them the weapon of choice whenever possible. Throughout this chapter $\mathfrak{H}$ will be a (complex) Hilbert space.

2.1. Orthonormal bases

In this section we will investigate orthonormal series and you will notice hardly any difference between the finite and infinite dimensional cases. As our first task, let us generalize the projection into the direction of one vector.

A set of vectors $\{u_j\}$ is called an **orthonormal set** if $\langle u_j, u_k\rangle = 0$ for $j \neq k$ and $\langle u_j, u_j\rangle = 1$. Note that every orthonormal set is linearly independent (show this).

Lemma 2.1. *Suppose $\{u_j\}_{j=1}^n$ is a finite orthonormal set in a Hilbert space $\mathfrak{H}$. Then every $f \in \mathfrak{H}$ can be written as*

$$f = f_\parallel + f_\perp, \qquad f_\parallel := \sum_{j=1}^{n} \langle u_j, f\rangle u_j, \tag{2.1}$$

where $f_\parallel$ and $f_\perp$ are orthogonal. Moreover, $\langle u_j, f_\perp\rangle = 0$ for all $1 \leq j \leq n$. In particular,

$$\|f\|^2 = \sum_{j=1}^{n} |\langle u_j, f\rangle|^2 + \|f_\perp\|^2. \tag{2.2}$$

Moreover, every $\hat{f}$ in the span of $\{u_j\}_{j=1}^n$ satisfies

$$\|f - \hat{f}\| \geq \|f_\perp\| \tag{2.3}$$

with equality holding if and only if $\hat{f} = f_\parallel$. In other words, $f_\parallel$ is uniquely characterized as the vector in the span of $\{u_j\}_{j=1}^n$ closest to f.

Proof. A straightforward calculation shows $\langle u_j, f - f_\parallel\rangle = 0$ and hence $f_\parallel$ and $f_\perp := f - f_\parallel$ are orthogonal. The formula for the norm follows by applying (1.40) iteratively.

Now, fix a vector $\hat{f} := \sum_{j=1}^n \alpha_j u_j$ in the span of $\{u_j\}_{j=1}^n$. Then one computes

$$\begin{aligned}\|f - \hat{f}\|^2 &= \|f_\parallel + f_\perp - \hat{f}\|^2 = \|f_\perp\|^2 + \|f_\parallel - \hat{f}\|^2 \\ &= \|f_\perp\|^2 + \sum_{j=1}^n |\alpha_j - \langle u_j, f\rangle|^2\end{aligned}$$

from which the last claim follows. □

From (2.2) we obtain **Bessel's inequality**

$$\sum_{j=1}^n |\langle u_j, f\rangle|^2 \le \|f\|^2 \tag{2.4}$$

with equality holding if and only if f lies in the span of $\{u_j\}_{j=1}^n$.

Of course, since we cannot assume $\mathfrak{H}$ to be a finite dimensional vector space, we need to generalize Lemma 2.1 to arbitrary orthonormal sets $\{u_j\}_{j\in J}$. We start by assuming that J is countable. Then Bessel's inequality (2.4) shows that

$$\sum_{j\in J} |\langle u_j, f\rangle|^2 \tag{2.5}$$

converges absolutely. Moreover, for any finite subset $K \subset J$ we have

$$\|\sum_{j\in K} \langle u_j, f\rangle u_j\|^2 = \sum_{j\in K} |\langle u_j, f\rangle|^2 \tag{2.6}$$

by the Pythagorean theorem and thus $\sum_{j\in J}\langle u_j, f\rangle u_j$ is a Cauchy sequence if and only if $\sum_{j\in J} |\langle u_j, f\rangle|^2$ is. Now let J be arbitrary. Again, Bessel's inequality shows that for any given $\varepsilon > 0$ there are at most finitely many j for which $|\langle u_j, f\rangle| \ge \varepsilon$ (namely at most $\|f\|/\varepsilon$). Hence there are at most countably many j for which $|\langle u_j, f\rangle| > 0$. Thus it follows that

$$\sum_{j\in J} |\langle u_j, f\rangle|^2 \tag{2.7}$$

is well defined (as a countable sum over the nonzero terms) and (by completeness) so is

$$\sum_{j\in J} \langle u_j, f\rangle u_j. \tag{2.8}$$

Furthermore, it is also independent of the order of summation.

In particular, by continuity of the scalar product we see that Lemma 2.1 can be generalized to arbitrary orthonormal sets.

Theorem 2.2. *Suppose $\{u_j\}_{j\in J}$ is an orthonormal set in a Hilbert space $\mathfrak{H}$. Then every $f \in \mathfrak{H}$ can be written as*

$$f = f_\parallel + f_\perp, \qquad f_\parallel := \sum_{j\in J} \langle u_j, f\rangle u_j, \tag{2.9}$$

where $f_\parallel$ and $f_\perp$ are orthogonal. Moreover, $\langle u_j, f_\perp\rangle = 0$ for all $j \in J$. In particular,

$$\|f\|^2 = \sum_{j\in J} |\langle u_j, f\rangle|^2 + \|f_\perp\|^2. \tag{2.10}$$

Furthermore, every $\hat{f} \in \overline{\operatorname{span}\{u_j\}_{j\in J}}$ satisfies

$$\|f - \hat{f}\| \geq \|f_\perp\| \tag{2.11}$$

with equality holding if and only if $\hat{f} = f_\parallel$. In other words, $f_\parallel$ is uniquely characterized as the vector in $\overline{\operatorname{span}\{u_j\}_{j\in J}}$ closest to f.

Proof. The first part follows as in Lemma 2.1 using continuity of the scalar product. The same is true for the last part except for the fact that every $f \in \overline{\operatorname{span}\{u_j\}_{j\in J}}$ can be written as $f = \sum_{j\in J} \alpha_j u_j$ (i.e., $f = f_\parallel$). To see this, let $f_n \in \operatorname{span}\{u_j\}_{j\in J}$ converge to f. Then $\|f-f_n\|^2 = \|f_\parallel - f_n\|^2 + \|f_\perp\|^2 \to 0$ implies $f_n \to f_\parallel$ and $f_\perp = 0$. □

Note that from Bessel's inequality (which of course still holds), it follows that the map $f \to f_\parallel$ is continuous.

Of course we are particularly interested in the case where every $f \in \mathfrak{H}$ can be written as $\sum_{j\in J}\langle u_j, f\rangle u_j$. In this case we will call the orthonormal set $\{u_j\}_{j\in J}$ an **orthonormal basis** (ONB).

If $\mathfrak{H}$ is separable it is easy to construct an orthonormal basis. In fact, if $\mathfrak{H}$ is separable, then there exists a countable total set $\{f_j\}_{j=1}^N$. Here $N \in \mathbb{N}$ if $\mathfrak{H}$ is finite dimensional and $N = \infty$ otherwise. After throwing away some vectors, we can assume that f_{n+1} cannot be expressed as a linear combination of the vectors $f_1, \dots, f_n$. Now we can construct an orthonormal set as follows: We begin by normalizing f_1:

$$u_1 := \frac{f_1}{\|f_1\|}. \tag{2.12}$$

Next we take f_2 and remove the component parallel to u_1 and normalize again:

$$u_2 := \frac{f_2 - \langle u_1, f_2\rangle u_1}{\|f_2 - \langle u_1, f_2\rangle u_1\|}. \tag{2.13}$$

Proceeding like this, we define recursively

$$u_n := \frac{f_n - \sum_{j=1}^{n-1} \langle u_j, f_n \rangle u_j}{\| f_n - \sum_{j=1}^{n-1} \langle u_j, f_n \rangle u_j \|}. \tag{2.14}$$

This procedure is known as **Gram–Schmidt orthogonalization**. Hence we obtain an orthonormal set $\{u_j\}_{j=1}^N$ such that $\operatorname{span}\{u_j\}_{j=1}^n = \operatorname{span}\{f_j\}_{j=1}^n$ for any finite n and thus also for $n = N$ (if $N = \infty$). Since $\{f_j\}_{j=1}^N$ is total, so is $\{u_j\}_{j=1}^N$. Now suppose there is some $f = f_\parallel + f_\perp \in \mathfrak{H}$ for which $f_\perp \neq 0$. Since $\{u_j\}_{j=1}^N$ is total, we can find a $\hat{f}$ in its span such that $\|f - \hat{f}\| < \|f_\perp\|$, contradicting (2.11). Hence we infer that $\{u_j\}_{j=1}^N$ is an orthonormal basis.

Theorem 2.3. *Every separable Hilbert space has a countable orthonormal basis.*

Example 2.1. The vectors $\{\delta^n\}_{n\in\mathbb{N}}$ form an orthonormal basis for $\ell^2(\mathbb{N})$. ⋄

Example 2.2. In $\mathcal{L}^2_{cont}(-1,1)$, we can orthogonalize the monomials $f_n(x) = x^n$ (which are total by the Weierstraß approximation theorem — Theorem 1.3). The resulting polynomials are up to a normalization known as **Legendre polynomials**

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{3\,x^2 - 1}{2}, \quad \dots \tag{2.15}$$

(which are normalized such that $P_n(1) = 1$). ⋄

Example 2.3. The set of functions

$$u_n(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}nx}, \qquad n \in \mathbb{Z}, \tag{2.16}$$

forms an orthonormal basis for $\mathfrak{H} = \mathcal{L}^2_{cont}(0, 2\pi)$. The corresponding orthogonal expansion is just the ordinary Fourier series. We will discuss this example in detail in Section 2.5. ⋄

The following equivalent properties also characterize a basis.

Theorem 2.4. *For an orthonormal set $\{u_j\}_{j\in J}$ in a Hilbert space $\mathfrak{H}$, the following conditions are equivalent:*

(i) *$\{u_j\}_{j\in J}$ is a maximal orthogonal set.*

(ii) *For every vector $f \in \mathfrak{H}$ we have*

$$f = \sum_{j\in J} \langle u_j, f \rangle u_j. \tag{2.17}$$

(iii) *For every vector $f \in \mathfrak{H}$ we have* ***Parseval's relation***

$$\|f\|^2 = \sum_{j\in J} |\langle u_j, f \rangle|^2. \tag{2.18}$$

(iv) $\langle u_j, f\rangle = 0$ *for all* $j \in J$ *implies* $f = 0$.

Proof. We will use the notation from Theorem 2.2.
(i) $\Rightarrow$ (ii): If $f_\perp \neq 0$, then we can normalize $f_\perp$ to obtain a unit vector $\tilde{f}_\perp$ which is orthogonal to all vectors u_j. But then $\{u_j\}_{j\in J} \cup \{\tilde{f}_\perp\}$ would be a larger orthonormal set, contradicting the maximality of $\{u_j\}_{j\in J}$.
(ii) $\Rightarrow$ (iii): This follows since (ii) implies $f_\perp = 0$.
(iii) $\Rightarrow$ (iv): If $\langle f, u_j\rangle = 0$ for all $j \in J$, we conclude $\|f\|^2 = 0$ and hence $f = 0$.
(iv) $\Rightarrow$ (i): If $\{u_j\}_{j\in J}$ were not maximal, there would be a unit vector g such that $\{u_j\}_{j\in J} \cup \{g\}$ is a larger orthonormal set. But $\langle u_j, g\rangle = 0$ for all $j \in J$ implies $g = 0$ by (iv), a contradiction. □

By continuity of the norm it suffices to check (iii), and hence also (ii), for f in a dense set. In fact, by the inverse triangle inequality for $\ell^2(\mathbb{N})$ and the Bessel inequality we have

$$\left|\sum_{j\in J} |\langle u_j, f\rangle|^2 - \sum_{j\in J} |\langle u_j, g\rangle|^2\right| \leq \sqrt{\sum_{j\in J} |\langle u_j, f-g\rangle|^2}\sqrt{\sum_{j\in J} |\langle u_j, f+g\rangle|^2}$$
$$\leq \|f-g\|\|f+g\| \tag{2.19}$$

implying $\sum_{j\in J} |\langle u_j, f_n\rangle|^2 \to \sum_{j\in J} |\langle u_j, f\rangle|^2$ if $f_n \to f$.

It is not surprising that if there is one countable basis, then it follows that every other basis is countable as well.

Theorem 2.5. *In a Hilbert space* $\mathfrak{H}$ *every orthonormal basis has the same cardinality.*

Proof. Let $\{u_j\}_{j\in J}$ and $\{v_k\}_{k\in K}$ be two orthonormal bases. We first look at the case where one of them, say the first, is finite dimensional: $J = \{1, \dots, n\}$. Suppose the other basis has at least n elements $\{1, \dots, n\} \subseteq K$. Then $v_k = \sum_{j=1}^n U_{k,j} u_j$, where $U_{k,j} = \langle u_j, v_k\rangle$. By $\delta_{j,k} = \langle v_j, v_k\rangle = \sum_{l=1}^n U_{j,l}^* U_{k,l}$ we see $u_j = \sum_{k=1}^n U_{k,j}^* v_k$ showing that K cannot have more than n elements.

Now let us turn to the case where both J and K are infinite. Set $K_j = \{k \in K | \langle v_k, u_j\rangle \neq 0\}$. Since these are the expansion coefficients of u_j with respect to $\{v_k\}_{k\in K}$, this set is countable (and nonempty). Hence the set $\tilde{K} = \bigcup_{j\in J} K_j$ satisfies $|\tilde{K}| \leq |J \times \mathbb{N}| = |J|$ (Theorem A.9) But $k \in K \setminus \tilde{K}$ implies $v_k = 0$ and hence $\tilde{K} = K$. So $|K| \leq |J|$ and reversing the roles of J and K shows $|K| = |J|$. □

The cardinality of an orthonormal basis is also called the Hilbert space **dimension** of $\mathfrak{H}$.

It even turns out that, up to unitary equivalence, there is only one separable infinite dimensional Hilbert space:

A bijective linear operator $U \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is called **unitary** if U preserves scalar products:

$$\langle Ug, Uf\rangle_2 = \langle g, f\rangle_1, \qquad g, f \in \mathfrak{H}_1. \tag{2.20}$$

By the polarization identity, (1.47) this is the case if and only if U preserves norms: $\|Uf\|_2 = \|f\|_1$ for all $f \in \mathfrak{H}_1$ (note a norm preserving linear operator is automatically injective). The two Hilbert spaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are called **unitarily equivalent** in this case.

Let $\mathfrak{H}$ be a separable infinite dimensional Hilbert space and let $\{u_j\}_{j\in\mathbb{N}}$ be any orthogonal basis. Then the map $U : \mathfrak{H} \to \ell^2(\mathbb{N})$, $f \mapsto (\langle u_j, f\rangle)_{j\in\mathbb{N}}$ is unitary. Indeed by Theorem 2.4 (iii) it is norm preserving and hence injective. To see that it is onto, let $a \in \ell^2(\mathbb{N})$ and observe that by $\|\sum_{j=m}^n a_j u_j\|^2 = \sum_{j=m}^n |a_j|^2$ the vector $f := \sum_{j\in\mathbb{N}} a_j u_j$ is well defined and satisfies $a_j = \langle u_j, f\rangle$. In particular,

Theorem 2.6. *Any separable infinite dimensional Hilbert space is unitarily equivalent to $\ell^2(\mathbb{N})$.*

Of course the same argument shows that every finite dimensional Hilbert space of dimension n is unitarily equivalent to $\mathbb{C}^n$ with the usual scalar product.

Finally we briefly turn to the case where $\mathfrak{H}$ is not separable.

Theorem 2.7. *Every Hilbert space has an orthonormal basis.*

Proof. To prove this we need to resort to Zorn's lemma (see Appendix A): The collection of all orthonormal sets in $\mathfrak{H}$ can be partially ordered by inclusion. Moreover, every linearly ordered chain has an upper bound (the union of all sets in the chain). Hence Zorn's lemma implies the existence of a maximal element, that is, an orthonormal set which is not a proper subset of every other orthonormal set. □

Hence, if $\{u_j\}_{j\in J}$ is an orthogonal basis, we can show that $\mathfrak{H}$ is unitarily equivalent to $\ell^2(J)$ and, by prescribing J, we can find a Hilbert space of any given dimension. Here $\ell^2(J)$ is the set of all complex-valued functions $(a_j)_{j\in J}$ where at most countably many values are nonzero and $\sum_{j\in J} |a_j|^2 < \infty$.

Example 2.4. Define the set of **almost periodic functions** $AP(\mathbb{R})$ as the closure of the set of trigonometric polynomials

$$f(t) = \sum_{k=1}^n \alpha_k \mathrm{e}^{\mathrm{i}\theta_k t}, \qquad \alpha_k \in \mathbb{C},\ \theta_k \in \mathbb{R},$$

with respect to the sup norm. In particular $AP(\mathbb{R}) \subset C_b(\mathbb{R})$ is a Banach space when equipped with the sup norm. Since the trigonometric polynomials form an algebra, it is even a Banach algebra. Using the Stone–Weierstraß theorem one can verify that every periodic function is almost periodic (make the approximation on one period and note that you get the rest of $\mathbb{R}$ for free from periodicity) but the converse is not true (e.g. $e^{it}+e^{i\sqrt{2}t}$ is not periodic).

It is not difficult to show that every almost periodic function has a mean value

$$M(f) := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)dt$$

and one can show that

$$\langle f, g\rangle := M(f^* g)$$

defines a scalar product on $AP(\mathbb{R})$ (only positivity is nontrivial and it will not be shown here). Note that $\|f\| \le \|f\|_\infty$. Abbreviating $e_\theta(t) = e^{i\theta t}$ one computes $M(e_\theta) = 0$ if $\theta \neq 0$ and $M(e_0) = 1$. In particular, $\{e_\theta\}_{\theta\in\mathbb{R}}$ is an uncountable orthonormal set and

$$f(t) \mapsto \hat{f}(\theta) := \langle e_\theta, f\rangle = M(e_{-\theta} f)$$

maps $AP(\mathbb{R})$ isometrically (with respect to $\|.\|$) into $\ell^2(\mathbb{R})$. This map is however not surjective (take e.g. a Fourier series which converges in mean square but not uniformly — see later). ◇

Problem* 2.1. *Given some vectors $f_1, \dots, f_n$ we define their* ***Gram determinant*** *as*

$$\Gamma(f_1, \dots, f_n) := \det\left(\langle f_j, f_k\rangle\right)_{1\le j,k\le n}.$$

Show that the Gram determinant is nonzero if and only if the vectors are linearly independent. Moreover, show that in this case

$$\operatorname{dist}(g, \operatorname{span}\{f_1, \dots, f_n\})^2 = \frac{\Gamma(f_1, \dots, f_n, g)}{\Gamma(f_1, \dots, f_n)}$$

and

$$\Gamma(f_1, \dots, f_n) \le \prod_{j=1}^{n} \|f_j\|^2.$$

with equality if the vectors are orthogonal. (Hint: How does Γ change when you apply the Gram–Schmidt procedure?)

Problem 2.2. *Let $\{u_j\}$ be some orthonormal basis. Show that a bounded linear operator A is uniquely determined by its matrix elements $A_{jk} := \langle u_j, Au_k\rangle$ with respect to this basis.*

Problem 2.3. *Give an example of a nonempty closed bounded subset of a Hilbert space which does not contain an element with minimal norm. Can this happen in finite dimensions? (Hint: Look for a discrete set.)*

2.2. The projection theorem and the Riesz lemma

Let $M \subseteq \mathfrak{H}$ be a subset. Then $M^\perp := \{f | \langle g, f\rangle = 0, \forall g \in M\}$ is called the **orthogonal complement** of M. By continuity of the scalar product it follows that $M^\perp$ is a closed linear subspace and by linearity that $(\overline{\operatorname{span}(M)})^\perp = M^\perp$. For example, we have $\mathfrak{H}^\perp = \{0\}$ since any vector in $\mathfrak{H}^\perp$ must be in particular orthogonal to all vectors in some orthonormal basis.

Theorem 2.8 (Projection theorem)**.** *Let M be a closed linear subspace of a Hilbert space $\mathfrak{H}$. Then every $f \in \mathfrak{H}$ can be uniquely written as $f = f_\parallel + f_\perp$ with $f_\parallel \in M$ and $f_\perp \in M^\perp$, where $f_\parallel$ is uniquely characterized as the vector in M closest to f. One writes*

$$M \oplus M^\perp = \mathfrak{H} \tag{2.21}$$

in this situation.

Proof. Since M is closed, it is a Hilbert space and has an orthonormal basis $\{u_j\}_{j\in J}$. Hence the existence part follows from Theorem 2.2. To see uniqueness, suppose there is another decomposition $f = \tilde f_\parallel + \tilde f_\perp$. Then $f_\parallel - \tilde f_\parallel = \tilde f_\perp - f_\perp \in M \cap M^\perp = \{0\}$ (since $g \in M \cap M^\perp$ implies $\|g\|^2 = \langle g, g\rangle = 0$). □

Corollary 2.9. *Every orthogonal set $\{u_j\}_{j\in J}$ can be extended to an orthogonal basis.*

Proof. Just add an orthogonal basis for $(\{u_j\}_{j\in J})^\perp$. □

The operator $P_M f := f_\parallel$ is called the **orthogonal projection** corresponding to M. Note that we have

$$P_M^2 = P_M \qquad \text{and} \qquad \langle P_M g, f\rangle = \langle g, P_M f\rangle \tag{2.22}$$

since $\langle P_M g, f\rangle = \langle g_\parallel, f_\parallel\rangle = \langle g, P_M f\rangle$. Clearly we have $P_{M^\perp} f = f - P_M f = f_\perp$. Furthermore, (2.22) uniquely characterizes orthogonal projections (Problem 2.6).

Moreover, if M is a closed subspace, we have $P_{M^{\perp\perp}} = \mathbb{I} - P_{M^\perp} = \mathbb{I} - (\mathbb{I} - P_M) = P_M$; that is, $M^{\perp\perp} = M$. If M is an arbitrary subset, we have at least

$$M^{\perp\perp} = \overline{\operatorname{span}(M)}. \tag{2.23}$$

Note that by $\mathfrak{H}^\perp = \{0\}$ we see that $M^\perp = \{0\}$ if and only if M is total.

Finally we turn to **linear functionals**, that is, to operators $\ell : \mathfrak{H} \to \mathbb{C}$. By the Cauchy–Schwarz inequality we know that $\ell_g : f \mapsto \langle g, f\rangle$ is a bounded linear functional (with norm $\|g\|$). In turns out that, in a Hilbert space, every bounded linear functional can be written in this way.

Theorem 2.10 (Riesz lemma)**.** *Suppose ℓ is a bounded linear functional on a Hilbert space $\mathfrak{H}$. Then there is a unique vector $g \in \mathfrak{H}$ such that $\ell(f) = \langle g, f\rangle$ for all $f \in \mathfrak{H}$.*

In other words, a Hilbert space is equivalent to its own dual space $\mathfrak{H}^* \cong \mathfrak{H}$ via the map $f \mapsto \langle f, .\rangle$ which is a conjugate linear isometric bijection between $\mathfrak{H}$ and $\mathfrak{H}^*$.

Proof. If $\ell \equiv 0$, we can choose $g = 0$. Otherwise $\mathrm{Ker}(\ell) = \{f|\ell(f) = 0\}$ is a proper subspace and we can find a unit vector $\tilde{g} \in \mathrm{Ker}(\ell)^\perp$. For every $f \in \mathfrak{H}$ we have $\ell(f)\tilde{g} - \ell(\tilde{g})f \in \mathrm{Ker}(\ell)$ and hence

$$0 = \langle \tilde{g}, \ell(f)\tilde{g} - \ell(\tilde{g})f\rangle = \ell(f) - \ell(\tilde{g})\langle \tilde{g}, f\rangle.$$

In other words, we can choose $g = \ell(\tilde{g})^*\tilde{g}$. To see uniqueness, let g_1, g_2 be two such vectors. Then $\langle g_1 - g_2, f\rangle = \langle g_1, f\rangle - \langle g_2, f\rangle = \ell(f) - \ell(f) = 0$ for every $f \in \mathfrak{H}$, which shows $g_1 - g_2 \in \mathfrak{H}^\perp = \{0\}$. □

In particular, this shows that $\mathfrak{H}^*$ is again a Hilbert space whose scalar product (in terms of the above identification) is given by $\langle\langle f, .\rangle, \langle g, .\rangle\rangle_{\mathfrak{H}^*} = \langle f, g\rangle^*$.

We can even get a unitary map between $\mathfrak{H}$ and $\mathfrak{H}^*$ but such a map is not unique. To this end note that every Hilbert space has a conjugation C which generalizes taking the complex conjugate of every coordinate. In fact, choosing an orthonormal basis (and different choices will produce different maps in general) we can set

$$Cf := \sum_{j\in J} \langle u_j, f\rangle^* u_j = \sum_{j\in J} \langle f, u_j\rangle u_j.$$

Then C is conjugate linear, isometric $\|Cf\| = \|f\|$, and idempotent $C^2 = \mathbb{I}$. Note also $\langle Cf, Cg\rangle = \langle f, g\rangle^*$. As promised, the map $f \to \langle Cf, .\rangle$ is a unitary map from $\mathfrak{H}$ to $\mathfrak{H}^*$.

Finally, we remark that projections can not only be defined for subspaces but also for closed convex sets (of course they will no longer be linear in this case).

Theorem 2.11 (Hilbert projection theorem)**.** *Let $\mathfrak{H}$ be a Hilbert space and K a nonempty closed convex subset. Then for every $f \in \mathfrak{H} \setminus K$ there is a unique $P_K(f)$ such that $\|P_K(f) - f\| = \inf_{g\in K} \|f - g\|$. If we extend $P_K : \mathfrak{H} \to K$ by setting $P(g) = g$ for $g \in K$ then P_K will be Lipschitz continuous: $\|P_K(f) - P_K(g)\| \le \|f - g\|$, $f, g \in \mathfrak{H}$.*

Proof. Fix $f \in \mathfrak{H} \setminus K$ and choose a sequence $f_n \in K$ with $\|f_n - f\| \to d := \inf_{g\in K} \|f - g\|$. Then applying the parallelogram law to the vectors $f_n - f$

and $f_m - f$ we obtain

$$\begin{aligned}\|f_n - f_m\|^2 &= 2(\|f - f_n\|^2 + \|f - f_m\|^2) - 4\|f - \tfrac{1}{2}(f_n + f_m)\|^2 \\ &\leq 2(\|f - f_n\|^2 + \|f - f_m\|^2) - 4d^2,\end{aligned}$$

which shows that f_n is Cauchy and hence converges to some point which we call $P(f)$. By construction $\|P(f) - f\| = d$. If there would be another point $\tilde{P}(f)$ with the same property, we could apply the parallelogram law to $P(f) - f$ and $\tilde{P}(f) - f$ giving $\|P(f) - \tilde{P}(f)\|^2 \leq 0$ and hence $P(f)$ is uniquely defined.

Next, let $f \in \mathfrak{H}$, $g \in K$ and consider $\tilde{g} = (1-t)P(f) + t\, g \in K$, $t \in [0,1]$. Then

$$0 \geq \|f - P(f)\|^2 - \|f - \tilde{g}\|^2 = 2t\mathrm{Re}(\langle f - P(f), g - P(f)\rangle) - t^2\|g - P(f)\|^2$$

for arbitrary $t \in [0,1]$ shows $\mathrm{Re}(\langle f - P(f), P(f) - g\rangle) \geq 0$. Consequently we have $\mathrm{Re}(\langle f - P(f), P(f) - P(g)\rangle) \geq 0$ for all $f, g \in \mathfrak{H}$. Now reverse to roles of f, g and add the two inequalities to obtain $\|P(f) - P(g)\|^2 \leq \mathrm{Re}\langle f - g, P(f) - P(g)\rangle \leq \|f - g\|\|P(f) - P(g)\|$. Hence Lipschitz continuity follows. □

If K is a closed subspace then this projection will of course coincide with the orthogonal projection defined before.

Problem 2.4. *Suppose $U : \mathfrak{H} \to \mathfrak{H}$ is unitary and $M \subseteq \mathfrak{H}$. Show that $UM^\perp = (UM)^\perp$.*

Problem 2.5. *Show that an orthogonal projection $P_M \neq 0$ has norm one.*

Problem* 2.6. *Suppose $P \in \mathscr{L}(\mathfrak{H})$ satisfies*

$$P^2 = P \qquad \text{and} \qquad \langle Pf, g\rangle = \langle f, Pg\rangle$$

and set $M = \mathrm{Ran}(P)$. Show

- *$Pf = f$ for $f \in M$ and M is closed,*
- *$g \in M^\perp$ implies $Pg \in M^\perp$ and thus $Pg = 0$,*

and conclude $P = P_M$. In particular

$$\mathfrak{H} = \mathrm{Ker}(P) \oplus \mathrm{Ran}(P), \qquad \mathrm{Ker}(P) = (\mathbb{I} - P)\mathfrak{H},\ \mathrm{Ran}(P) = P\mathfrak{H}.$$

2.3. Operators defined via forms

One of the key results about linear maps is that they are uniquely determined once we know the images of some basis vectors. In fact, the matrix elements with respect to some basis uniquely determine a linear map. Clearly this raises the question how this results extends to the infinite dimensional setting. As a first result we show that the Riesz lemma, Theorem 2.10,

implies that a bounded operator A is uniquely determined by its associated sesquilinear form $\langle g, Af\rangle$. In fact, there is a one-to-one correspondence between bounded operators and bounded sesquilinear forms:

Lemma 2.12. *Suppose $s : \mathfrak{H}_2 \times \mathfrak{H}_1 \to \mathbb{C}$ is a bounded sesquilinear form; that is,*

$$|s(g, f)| \leq C\|g\|_{\mathfrak{H}_2} \|f\|_{\mathfrak{H}_1}. \tag{2.24}$$

Then there is a unique bounded operator $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ such that

$$s(g, f) = \langle g, Af\rangle_{\mathfrak{H}_2}. \tag{2.25}$$

Moreover, the norm of A is given by

$$\|A\| = \sup_{\|g\|_{\mathfrak{H}_2}=\|f\|_{\mathfrak{H}_1}=1} |\langle g, Af\rangle_{\mathfrak{H}_2}| \leq C. \tag{2.26}$$

Proof. For every $f \in \mathfrak{H}_1$ we have an associated bounded linear functional $\ell_f(g) := s(g, f)^*$ on $\mathfrak{H}_2$. By Theorem 2.10 there is a corresponding $h \in \mathfrak{H}_2$ (depending on f) such that $\ell_f(g) = \langle h, g\rangle_{\mathfrak{H}_2}$, that is $s(g, f) = \langle g, h\rangle_{\mathfrak{H}_2}$ and we can define A via $Af := h$. It is not hard to check that A is linear and from

$$\|Af\|^2_{\mathfrak{H}_2} = \langle Af, Af\rangle_{\mathfrak{H}_2} = s(Af, f) \leq C\|Af\|_{\mathfrak{H}_2}\|f\|_{\mathfrak{H}_1}$$

we infer $\|Af\|_{\mathfrak{H}_2} \leq C\|f\|_{\mathfrak{H}_1}$, which shows that A is bounded with $\|A\| \leq C$. Equation (2.26) is left as an exercise (Problem 2.9). □

Note that if $\{u_k\}_{k\in K} \subseteq \mathfrak{H}_1$ and $\{v_j\}_{j\in J} \subseteq \mathfrak{H}_2$ are some orthogonal bases, then the matrix elements $A_{j,k} := \langle v_j, Au_k\rangle_{\mathfrak{H}_2}$ for all $(j, k) \in J \times K$ uniquely determine $\langle g, Af\rangle_{\mathfrak{H}_2}$ for arbitrary $f \in \mathfrak{H}_1$, $g \in \mathfrak{H}_2$ (just expand f, g with respect to these bases) and thus A by our theorem.

Example 2.5. Consider $\ell^2(\mathbb{N})$ and let $A \in \mathscr{L}(\ell^2(\mathbb{N}))$ be some bounded operator. Let $A_{jk} = \langle \delta^j, A\delta^k\rangle$ be its matrix elements such that

$$(Aa)_j = \sum_{k=1}^{\infty} A_{jk}a_k.$$

Here the sum converges in $\ell^2(\mathbb{N})$ and hence, in particular, for every fixed j. Moreover, choosing $a^n_k = \alpha_n A_{jk}$ for $k \leq n$ and $a^n_k = 0$ for $k > n$ with $\alpha_n = (\sum_{j=1}^n |A_{jk}|^2)^{1/2}$ we see $\alpha_n = |(Aa^n)_j| \leq \|A\|\|a^n\| = \|A\|$. Thus $\sum_{j=1}^{\infty} |A_{jk}|^2 \leq \|A\|^2$ and the sum is even absolutely convergent. ⋄

Moreover, for $A \in \mathscr{L}(\mathfrak{H})$ the polarization identity (Problem 1.19) implies that A is already uniquely determined by its quadratic form $q_A(f) := \langle f, Af\rangle$.

As a first application we introduce the **adjoint operator** via Lemma 2.12 as the operator associated with the sesquilinear form $s(f, g) := \langle Af, g\rangle_{\mathfrak{H}_2}$.

Theorem 2.13. *For every bounded operator $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ there is a unique bounded operator $A^* \in \mathscr{L}(\mathfrak{H}_2, \mathfrak{H}_1)$ defined via*

$$\langle f, A^* g\rangle_{\mathfrak{H}_1} = \langle Af, g\rangle_{\mathfrak{H}_2}. \tag{2.27}$$

A bounded operator $A \in \mathscr{L}(\mathfrak{H})$ satisfying $A^* = A$ is called **self-adjoint**. Note that $q_{A^*}(f) = \langle Af, f\rangle = q_A(f)^*$ and hence a bounded operator is self-adjoint if and only if its quadratic form is real-valued.

Example 2.6. If $\mathfrak{H} := \mathbb{C}^n$ and $A := (a_{jk})_{1\le j,k\le n}$, then $A^* = (a^*_{kj})_{1\le j,k\le n}$. ⋄

Example 2.7. If $\mathbb{I} \in \mathscr{L}(\mathfrak{H})$ is the identity, then $\mathbb{I}^* = \mathbb{I}$. ⋄

Example 2.8. Consider the linear functional $\ell : \mathfrak{H} \to \mathbb{C}$, $f \mapsto \langle g, f\rangle$. Then by the definition $\langle f, \ell^*\alpha\rangle = \ell(f)^*\alpha = \langle f, \alpha g\rangle$ we obtain $\ell^* : \mathbb{C} \to \mathfrak{H}$, $\alpha \mapsto \alpha g$. ⋄

Example 2.9. Let $\mathfrak{H} := \ell^2(\mathbb{N})$, $a \in \ell^\infty(\mathbb{N})$ and consider the multiplication operator

$$(Ab)_j := a_j b_j.$$

Then

$$\langle Ab, c\rangle = \sum_{j=1}^{\infty} (a_j b_j)^* c_j = \sum_{j=1}^{\infty} b_j^* (a_j^* c_j) = \langle b, A^* c\rangle$$

with $(A^*c)_j = a_j^* c_j$, that is, A^* is the multiplication operator with a^*. ⋄

Example 2.10. Let $\mathfrak{H} := \ell^2(\mathbb{N})$ and consider the shift operators defined via

$$(S^{\pm} a)_j := a_{j\pm 1}$$

with the convention that $a_0 = 0$. That is, S^- shifts a sequence to the right and fills up the left most place by zero and S^+ shifts a sequence to the left dropping the left most place:

$$S^-(a_1, a_2, a_3, \cdots) = (0, a_1, a_2, \cdots), \qquad S^+(a_1, a_2, a_3, \cdots) = (a_2, a_3, a_4, \cdots).$$

Then

$$\langle S^- a, b\rangle = \sum_{j=2}^{\infty} a_{j-1}^* b_j = \sum_{j=1}^{\infty} a_j^* b_{j+1} = \langle a, S^+ b\rangle,$$

which shows that $(S^-)^* = S^+$. Using symmetry of the scalar product we also get $\langle b, S^- a\rangle = \langle S^+ b, a\rangle$, that is, $(S^+)^* = S^-$.

Note that S^+ is a left inverse of S^-, $S^+S^- = \mathbb{I}$, but not a right inverse as $S^-S^+ \neq \mathbb{I}$. This is different from the finite dimensional case, where a left inverse is also a right inverse and vice versa. ⋄

Example 2.11. Suppose $U \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is unitary. Then $U^* = U^{-1}$. This follows from Lemma 2.12 since $\langle f, g\rangle_{\mathfrak{H}_1} = \langle Uf, Ug\rangle_{\mathfrak{H}_2} = \langle f, U^*Ug\rangle_{\mathfrak{H}_1}$ implies $U^*U = \mathbb{I}_{\mathfrak{H}_1}$. Since U is bijective we can multiply this last equation from the right with U^{-1} to obtain the claim. Of course this calculation shows that

the converse is also true, that is $U \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is unitary if and only if $U^* = U^{-1}$. ⋄

A few simple properties of taking adjoints are listed below.

Lemma 2.14. *Let $A, B \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$, $C \in \mathscr{L}(\mathfrak{H}_2, \mathfrak{H}_3)$, and $\alpha \in \mathbb{C}$. Then*

(i) $(A+B)^* = A^* + B^*, \qquad (\alpha A)^* = \alpha^* A^*,$

(ii) $A^{**} = A,$

(iii) $(CA)^* = A^* C^*,$

(iv) $\|A^*\| = \|A\|$ *and* $\|A\|^2 = \|A^* A\| = \|A A^*\|.$

Proof. (i) is obvious. (ii) follows from $\langle g, A^{**} f\rangle_{\mathfrak{H}_2} = \langle A^* g, f\rangle_{\mathfrak{H}_1} = \langle g, Af\rangle_{\mathfrak{H}_2}$. (iii) follows from $\langle g, (CA) f\rangle_{\mathfrak{H}_3} = \langle C^* g, Af\rangle_{\mathfrak{H}_2} = \langle A^* C^* g, f\rangle_{\mathfrak{H}_1}$. (iv) follows using (2.26) from

$$\begin{aligned} \|A^*\| &= \sup_{\|f\|_{\mathfrak{H}_1}=\|g\|_{\mathfrak{H}_2}=1} |\langle f, A^* g\rangle_{\mathfrak{H}_1}| = \sup_{\|f\|_{\mathfrak{H}_1}=\|g\|_{\mathfrak{H}_2}=1} |\langle Af, g\rangle_{\mathfrak{H}_2}| \\ &= \sup_{\|f\|_{\mathfrak{H}_1}=\|g\|_{\mathfrak{H}_2}=1} |\langle g, Af\rangle_{\mathfrak{H}_2}| = \|A\| \end{aligned}$$

and

$$\begin{aligned} \|A^* A\| &= \sup_{\|f\|_{\mathfrak{H}_1}=\|g\|_{\mathfrak{H}_2}=1} |\langle f, A^* Ag\rangle_{\mathfrak{H}_1}| = \sup_{\|f\|_{\mathfrak{H}_1}=\|g\|_{\mathfrak{H}_2}=1} |\langle Af, Ag\rangle_{\mathfrak{H}_2}| \\ &= \sup_{\|f\|_{\mathfrak{H}_1}=1} \|Af\|^2 = \|A\|^2, \end{aligned}$$

where we have used that $|\langle Af, Ag\rangle_{\mathfrak{H}_2}|$ attains its maximum when Af and Ag are parallel (compare Theorem 1.5). □

Note that $\|A\| = \|A^*\|$ implies that taking adjoints is a continuous operation. For later use also note that (Problem 2.11)

$$\operatorname{Ker}(A^*) = \operatorname{Ran}(A)^{\perp}. \tag{2.28}$$

For the remainder of this section we restrict to the case of one Hilbert space. A sesquilinear form $s : \mathfrak{H} \times \mathfrak{H} \to \mathbb{C}$ is called nonnegative if $s(f, f) \geq 0$ and we will call $A \in \mathscr{L}(\mathfrak{H})$ **nonnegative**, $A \geq 0$, if its associated sesquilinear form is. We will write $A \geq B$ if $A - B \geq 0$. Observe that nonnegative operators are self-adjoint (as their quadratic forms are real-valued — here it is important that the underlying space is complex).

Example 2.12. For any operator A the operators $A^* A$ and $A A^*$ are both nonnegative. In fact $\langle f, A^* Af\rangle = \langle Af, Af\rangle = \|Af\|^2 \geq 0$ and similarly $\langle f, AA^* f\rangle = \|A^* f\|^2 \geq 0$. ⋄

Lemma 2.15. *Suppose $A \in \mathscr{L}(\mathfrak{H})$ satisfies $|\langle f, Af\rangle| \geq \varepsilon \|f\|^2$ for some $\varepsilon > 0$. Then A is a bijection with bounded inverse, $\|A^{-1}\| \leq \frac{1}{\varepsilon}$.*

Proof. By assumption $\varepsilon\|f\|^2 \le |\langle f, Af\rangle| \le \|f\|\|Af\|$ and thus $\varepsilon\|f\| \le \|Af\|$. In particular, $Af = 0$ implies $f = 0$ and thus for every $g \in \operatorname{Ran}(A)$ there is a unique $f = A^{-1}g$. Moreover, by $\|A^{-1}g\| = \|f\| \le \varepsilon^{-1}\|Af\| = \varepsilon^{-1}\|g\|$ the operator A^{-1} is bounded. So if $g_n \in \operatorname{Ran}(A)$ converges to some $g \in \mathfrak{H}$, then $f_n = A^{-1}g_n$ converges to some f. Taking limits in $g_n = Af_n$ shows that $g = Af$ is in the range of A, that is, the range of A is closed. To show that $\operatorname{Ran}(A) = \mathfrak{H}$ we pick $h \in \operatorname{Ran}(A)^\perp$. Then $0 = \langle h, Ah\rangle \ge \varepsilon\|h\|^2$ shows $h = 0$ and thus $\operatorname{Ran}(A)^\perp = \{0\}$. □

Combining the last two results we obtain the famous Lax–Milgram theorem which plays an important role in theory of elliptic partial differential equations.

Theorem 2.16 (Lax–Milgram)**.** *Let $s : \mathfrak{H} \times \mathfrak{H} \to \mathbb{C}$ be a sesquilinear form which is*

- *bounded, $|s(f, g)| \le C\|f\|\,\|g\|$, and*
- *coercive, $|s(f, f)| \ge \varepsilon\|f\|^2$ for some $\varepsilon > 0$.*

Then for every $g \in \mathfrak{H}$ there is a unique $f \in \mathfrak{H}$ such that

$$s(h, f) = \langle h, g\rangle, \qquad \forall h \in \mathfrak{H}. \tag{2.29}$$

Proof. Let A be the operator associated with s. Then A is a bijection and $f = A^{-1}g$. □

Example 2.13. Consider $\mathfrak{H} = \ell^2(\mathbb{N})$ and introduce the operator

$$(Aa)_j := -a_{j+1} + 2a_j - a_{j-1}$$

which is a discrete version of a second derivative (discrete one-dimensional Laplace operator). Here we use the convention $a_0 = 0$, that is, $(Aa)_1 = -a_2 + 2a_1$. In terms of the shift operators $S^\pm$ we can write

$$A = -S^+ + 2 - S^- = (S^+ - 1)(S^- - 1)$$

and using $(S^\pm)^* = S^\mp$ we obtain

$$s_A(a, b) = \langle (S^- - 1)a, (S^- - 1)b\rangle = \sum_{j=1}^{\infty} (a_{j-1} - a_j)^*(b_{j-1} - b_j).$$

In particular, this shows $A \ge 0$. Moreover, we have $|s_A(a, b)| \le 4\|a\|_2\|b\|_2$ or equivalently $\|A\| \le 4$.

Next, let

$$(Qa)_j = q_j a_j$$

for some sequence $q \in \ell^\infty(\mathbb{N})$. Then

$$s_Q(a, b) = \sum_{j=1}^{\infty} q_j a_j^* b_j$$

and $|s_Q(a,b)| \le \|q\|_\infty \|a\|_2 \|b\|_2$ or equivalently $\|Q\| \le \|q\|_\infty$. If in addition $q_j \ge \varepsilon > 0$, then $s_{A+Q}(a,b) = s_A(a,b) + s_Q(a,b)$ satisfies the assumptions of the Lax–Milgram theorem and

$$(A+Q)a = b$$

has a unique solution $a = (A+Q)^{-1}b$ for every given $b \in \ell^2(\mathbb{Z})$. Moreover, since $(A+Q)^{-1}$ is bounded, this solution depends continuously on b. ⋄

Problem* 2.7. *Let $\mathfrak{H}_1$, $\mathfrak{H}_2$ be Hilbert spaces and let $u \in \mathfrak{H}_1$, $v \in \mathfrak{H}_2$. Show that the operator*

$$Af := \langle u, f\rangle v$$

is bounded and compute its norm. Compute the adjoint of A.

Problem 2.8. *Show that under the assumptions of Problem 1.35 one has $f(A)^* = f^\#(A^*)$ where $f^\#(z) = f(z^*)^*$.*

Problem* 2.9. *Prove* (2.26). *(Hint: Use $\|f\| = \sup_{\|g\|=1} |\langle g, f\rangle|$ — compare Theorem 1.5.)*

Problem 2.10. *Suppose $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ has a bounded inverse $A^{-1} \in \mathscr{L}(\mathfrak{H}_2, \mathfrak{H}_1)$. Show $(A^{-1})^* = (A^*)^{-1}$.*

Problem* 2.11. *Show* (2.28).

Problem* 2.12. *Show that every operator $A \in \mathscr{L}(\mathfrak{H})$ can be written as the linear combination of two self-adjoint operators $\mathrm{Re}(A) := \frac{1}{2}(A + A^*)$ and $\mathrm{Im}(A) := \frac{1}{2\mathrm{i}}(A - A^*)$. Moreover, every self-adjoint operator can be written as a linear combination of two unitary operators. (Hint: For the last part consider $f_\pm(z) = z \pm \mathrm{i}\sqrt{1-z^2}$ and Problems 1.35, 2.8.)*

Problem 2.13 (Abstract Dirichlet problem)**.** *Show that the solution of* (2.29) *is also the unique minimizer of*

$$h \mapsto \mathrm{Re}\Big(\frac{1}{2}s(h,h) - \langle h, g\rangle\Big).$$

if s satisfies $s(w,w) \ge \varepsilon\|w\|^2$ for all $w \in \mathfrak{H}$.

2.4. Orthogonal sums and tensor products

Given two Hilbert spaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$, we define their **orthogonal sum** $\mathfrak{H}_1 \oplus \mathfrak{H}_2$ to be the set of all pairs $(f_1, f_2) \in \mathfrak{H}_1 \times \mathfrak{H}_2$ together with the scalar product

$$\langle (g_1, g_2), (f_1, f_2)\rangle := \langle g_1, f_1\rangle_{\mathfrak{H}_1} + \langle g_2, f_2\rangle_{\mathfrak{H}_2}. \tag{2.30}$$

It is left as an exercise to verify that $\mathfrak{H}_1 \oplus \mathfrak{H}_2$ is again a Hilbert space. Moreover, $\mathfrak{H}_1$ can be identified with $\{(f_1, 0) | f_1 \in \mathfrak{H}_1\}$, and we can regard $\mathfrak{H}_1$ as a subspace of $\mathfrak{H}_1 \oplus \mathfrak{H}_2$, and similarly for $\mathfrak{H}_2$. With this convention we have $\mathfrak{H}_1^\perp = \mathfrak{H}_2$. It is also customary to write $f_1 \oplus f_2$ instead of (f_1, f_2).

In the same way we can define the orthogonal sum $\bigoplus_{j=1}^n \mathfrak{H}_j$ of any finite number of Hilbert spaces.

Example 2.14. For example we have $\bigoplus_{j=1}^n \mathbb{C} = \mathbb{C}^n$ and hence we will write $\bigoplus_{j=1}^n \mathfrak{H} =: \mathfrak{H}^n$. ◇

More generally, let $\mathfrak{H}_j$, $j \in \mathbb{N}$, be a countable collection of Hilbert spaces and define

$$\bigoplus_{j=1}^\infty \mathfrak{H}_j := \{\bigoplus_{j=1}^\infty f_j | \, f_j \in \mathfrak{H}_j, \, \sum_{j=1}^\infty \|f_j\|_{\mathfrak{H}_j}^2 < \infty\}, \tag{2.31}$$

which becomes a Hilbert space with the scalar product

$$\langle \bigoplus_{j=1}^\infty g_j, \bigoplus_{j=1}^\infty f_j \rangle := \sum_{j=1}^\infty \langle g_j, f_j \rangle_{\mathfrak{H}_j}. \tag{2.32}$$

Example 2.15. $\bigoplus_{j=1}^\infty \mathbb{C} = \ell^2(\mathbb{N})$. ◇

Similarly, if $\mathfrak{H}$ and $\tilde{\mathfrak{H}}$ are two Hilbert spaces, we define their tensor product as follows: The elements should be products $f \otimes \tilde{f}$ of elements $f \in \mathfrak{H}$ and $\tilde{f} \in \tilde{\mathfrak{H}}$. Hence we start with the set of all finite linear combinations of elements of $\mathfrak{H} \times \tilde{\mathfrak{H}}$

$$\mathcal{F}(\mathfrak{H}, \tilde{\mathfrak{H}}) := \{\sum_{j=1}^n \alpha_j (f_j, \tilde{f}_j) | (f_j, \tilde{f}_j) \in \mathfrak{H} \times \tilde{\mathfrak{H}}, \, \alpha_j \in \mathbb{C}\}. \tag{2.33}$$

Since we want $(f_1 + f_2) \otimes \tilde{f} = f_1 \otimes \tilde{f} + f_2 \otimes \tilde{f}$, $f \otimes (\tilde{f}_1 + \tilde{f}_2) = f \otimes \tilde{f}_1 + f \otimes \tilde{f}_2$, and $(\alpha f) \otimes \tilde{f} = f \otimes (\alpha \tilde{f}) = \alpha (f \otimes \tilde{f})$ we consider $\mathcal{F}(\mathfrak{H}, \tilde{\mathfrak{H}}) / \mathcal{N}(\mathfrak{H}, \tilde{\mathfrak{H}})$, where

$$\mathcal{N}(\mathfrak{H}, \tilde{\mathfrak{H}}) := \operatorname{span}\{\sum_{j,k=1}^n \alpha_j \beta_k (f_j, \tilde{f}_k) - (\sum_{j=1}^n \alpha_j f_j, \sum_{k=1}^n \beta_k \tilde{f}_k)\} \tag{2.34}$$

and write $f \otimes \tilde{f}$ for the equivalence class of $(f, \tilde{f})$. By construction, every element in this quotient space is a linear combination of elements of the type $f \otimes \tilde{f}$.

Next, we want to define a scalar product such that

$$\langle f \otimes \tilde{f}, g \otimes \tilde{g} \rangle = \langle f, g \rangle_{\mathfrak{H}} \langle \tilde{f}, \tilde{g} \rangle_{\tilde{\mathfrak{H}}} \tag{2.35}$$

holds. To this end we set

$$s(\sum_{j=1}^n \alpha_j (f_j, \tilde{f}_j), \sum_{k=1}^n \beta_k (g_k, \tilde{g}_k)) = \sum_{j,k=1}^n \alpha_j^* \beta_k \langle f_j, g_k \rangle_{\mathfrak{H}} \langle \tilde{f}_j, \tilde{g}_k \rangle_{\tilde{\mathfrak{H}}}, \tag{2.36}$$

which is a symmetric sesquilinear form on $\mathcal{F}(\mathfrak{H},\tilde{\mathfrak{H}})$. Moreover, one verifies that $s(f,g)=0$ for arbitrary $f\in\mathcal{F}(\mathfrak{H},\tilde{\mathfrak{H}})$ and $g\in\mathcal{N}(\mathfrak{H},\tilde{\mathfrak{H}})$ and thus

$$\langle\sum_{j=1}^{n}\alpha_j\, f_j\otimes\tilde{f}_j,\sum_{k=1}^{n}\beta_k\, g_k\otimes\tilde{g}_k\rangle=\sum_{j,k=1}^{n}\alpha_j^*\beta_k\langle f_j,g_k\rangle_{\mathfrak{H}}\langle\tilde{f}_j,\tilde{g}_k\rangle_{\tilde{\mathfrak{H}}} \tag{2.37}$$

is a symmetric sesquilinear form on $\mathcal{F}(\mathfrak{H},\tilde{\mathfrak{H}})/\mathcal{N}(\mathfrak{H},\tilde{\mathfrak{H}})$. To show that this is in fact a scalar product, we need to ensure positivity. Let $f=\sum_i\alpha_i f_i\otimes\tilde{f}_i\neq 0$ and pick orthonormal bases u_j, $\tilde{u}_k$ for $\operatorname{span}\{f_i\}$, $\operatorname{span}\{\tilde{f}_i\}$, respectively. Then

$$f=\sum_{j,k}\alpha_{jk}u_j\otimes\tilde{u}_k,\quad \alpha_{jk}=\sum_i\alpha_i\langle u_j,f_i\rangle_{\mathfrak{H}}\langle\tilde{u}_k,\tilde{f}_i\rangle_{\tilde{\mathfrak{H}}} \tag{2.38}$$

and we compute

$$\langle f,f\rangle=\sum_{j,k}|\alpha_{jk}|^2>0. \tag{2.39}$$

The completion of $\mathcal{F}(\mathfrak{H},\tilde{\mathfrak{H}})/\mathcal{N}(\mathfrak{H},\tilde{\mathfrak{H}})$ with respect to the induced norm is called the **tensor product** $\mathfrak{H}\otimes\tilde{\mathfrak{H}}$ of $\mathfrak{H}$ and $\tilde{\mathfrak{H}}$.

Lemma 2.17. *If u_j, $\tilde{u}_k$ are orthonormal bases for $\mathfrak{H}$, $\tilde{\mathfrak{H}}$, respectively, then $u_j\otimes\tilde{u}_k$ is an orthonormal basis for $\mathfrak{H}\otimes\tilde{\mathfrak{H}}$.*

Proof. That $u_j\otimes\tilde{u}_k$ is an orthonormal set is immediate from (2.35). Moreover, since $\operatorname{span}\{u_j\}$, $\operatorname{span}\{\tilde{u}_k\}$ are dense in $\mathfrak{H}$, $\tilde{\mathfrak{H}}$, respectively, it is easy to see that $u_j\otimes\tilde{u}_k$ is dense in $\mathcal{F}(\mathfrak{H},\tilde{\mathfrak{H}})/\mathcal{N}(\mathfrak{H},\tilde{\mathfrak{H}})$. But the latter is dense in $\mathfrak{H}\otimes\tilde{\mathfrak{H}}$. □

Note that this in particular implies $\dim(\mathfrak{H}\otimes\tilde{\mathfrak{H}})=\dim(\mathfrak{H})\dim(\tilde{\mathfrak{H}})$.

Example 2.16. We have $\mathfrak{H}\otimes\mathbb{C}^n=\mathfrak{H}^n$. ⋄

Example 2.17. We have $\ell^2(\mathbb{N})\otimes\ell^2(\mathbb{N})=\ell^2(\mathbb{N}\times\mathbb{N})$ by virtue of the identification $(a_{jk})\mapsto\sum_{jk}a_{jk}\delta^j\otimes\delta^k$ where δ^j is the standard basis for $\ell^2(\mathbb{N})$. In fact, this follows from the previous lemma as in the proof of Theorem 2.6. ⋄

It is straightforward to extend the tensor product to any finite number of Hilbert spaces. We even note

$$(\bigoplus_{j=1}^{\infty}\mathfrak{H}_j)\otimes\mathfrak{H}=\bigoplus_{j=1}^{\infty}(\mathfrak{H}_j\otimes\mathfrak{H}), \tag{2.40}$$

where equality has to be understood in the sense that both spaces are unitarily equivalent by virtue of the identification

$$(\sum_{j=1}^{\infty}f_j)\otimes f=\sum_{j=1}^{\infty}f_j\otimes f. \tag{2.41}$$

Problem 2.14. *Show that $f\otimes\tilde{f}=0$ if and only if $f=0$ or $\tilde{f}=0$.*

Problem 2.15. *We have* $f \otimes \tilde{f} = g \otimes \tilde{g} \neq 0$ *if and only if there is some* $\alpha \in \mathbb{C} \setminus \{0\}$ *such that* $f = \alpha g$ *and* $\tilde{f} = \alpha^{-1}\tilde{g}$.

Problem* 2.16. *Show* (2.40).

2.5. Applications to Fourier series

We have already encountered the Fourier sine series during our treatment of the heat equation in Section 1.1. Given an integrable function f we can define its **Fourier series**

$$S(f)(x) := \frac{a_0}{2} + \sum_{k\in\mathbb{N}} \big(a_k \cos(kx) + b_k \sin(kx)\big), \tag{2.42}$$

where the corresponding Fourier coefficients are given by

$$a_k := \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(kx) f(x) dx, \qquad b_k := \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kx) f(x) dx. \tag{2.43}$$

At this point (2.42) is just a formal expression and it was (and to some extend still is) a fundamental question in mathematics to understand in what sense the above series converges. For example, does it converge at a given point (e.g. at every point of continuity of f) or when does it converge uniformly? We will give some first answers in the present section and then come back later to this when we have further tools at our disposal.

For our purpose the complex form

$$S(f)(x) = \sum_{k\in\mathbb{Z}} \hat{f}_k \mathrm{e}^{\mathrm{i}kx}, \qquad \hat{f}_k := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}ky} f(y) dy \tag{2.44}$$

will be more convenient. The connection is given via $\hat{f}_{\pm k} = \frac{a_k \mp b_k}{2}$, $k \in \mathbb{N}_0$ (with the convention $b_0 = 0$). In this case the n'th partial sum can be written as

$$S_n(f)(x) := \sum_{k=-n}^{n} \hat{f}_k \mathrm{e}^{\mathrm{i}kx} = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(x-y) f(y) dy, \tag{2.45}$$

where

$$D_n(x) = \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}kx} = \frac{\sin((n+1/2)x)}{\sin(x/2)} \tag{2.46}$$

is known as the **Dirichlet kernel** (to obtain the second form observe that the left-hand side is a geometric series). Note that $D_n(-x) = -D_n(x)$ and that $|D_n(x)|$ has a global maximum $D_n(0) = 2n+1$ at $x = 0$. Moreover, by $S_n(1) = 1$ we see that $\int_{-\pi}^{\pi} D_n(x) dx = 1$.

Since

$$\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}kx} \mathrm{e}^{\mathrm{i}lx} dx = 2\pi\delta_{k,l} \tag{2.47}$$

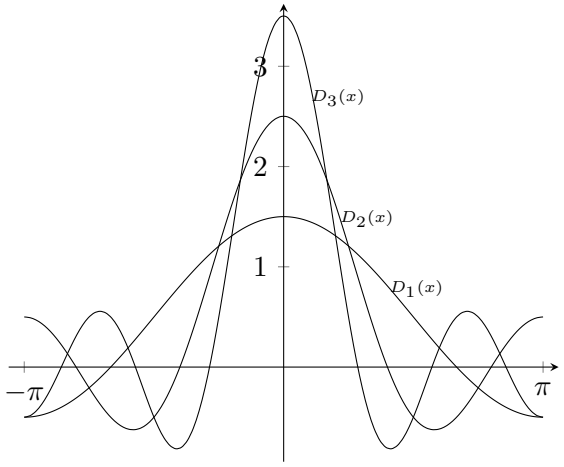


Figure 2.1. The Dirichlet kernels D_1, D_2, and D_3

the functions $e_k(x) := (2\pi)^{-1/2}\mathrm{e}^{\mathrm{i}kx}$ are orthonormal in $L^2(-\pi,\pi)$ and hence the Fourier series is just the expansion with respect to this orthogonal set. Hence we obtain

Theorem 2.18. *For every square integrable function $f \in L^2(-\pi,\pi)$, the Fourier coefficients $\hat{f}_k$ are square summable*

$$\sum_{k\in\mathbb{Z}} |\hat{f}_k|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2 dx \tag{2.48}$$

and the Fourier series converges to f in the sense of L^2. Moreover, this is a continuous bijection between $L^2(-\pi,\pi)$ and $\ell^2(\mathbb{Z})$.

Proof. To show this theorem it suffices to show that the functions e_k form a basis. This will follow from Theorem 2.20 below (see the discussion after this theorem). It will also follow as a special case of Theorem 3.11 below (see the examples after this theorem) as well as from the Stone–Weierstraß theorem — Problem 2.20. □

This gives a satisfactory answer in the Hilbert space $L^2(-\pi,\pi)$ but does not answer the question about pointwise or uniform convergence. The latter will be the case if the Fourier coefficients are summable. First of all we note that for integrable functions the Fourier coefficients will at least tend to zero.

Lemma 2.19 (Riemann–Lebesgue lemma)**.** *Suppose $f \in L^1(-\pi,\pi)$, then the Fourier coefficients $\hat{f}_k$ converge to zero as $|k| \to \infty$.*

Proof. By our previous theorem this holds for continuous functions. But the map $f \to \hat{f}$ is bounded from $C[-\pi,\pi] \subset L^1(-\pi,\pi)$ to $c_0(\mathbb{Z})$ (the sequences vanishing as $|k| \to \infty$) since $|\hat{f}_k| \le (2\pi)^{-1}\|f\|_1$ and there is a unique extension to all of $L^1(-\pi,\pi)$. □

It turns out that this result is best possible in general and we cannot say more without additional assumptions on f. For example, if f is periodic and differentiable, then integration by parts shows

$$\hat{f}_k = \frac{1}{2\pi \mathrm{i} k} \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}kx} f'(x) dx. \tag{2.49}$$

Then, since both k^{-1} and the Fourier coefficients of f' are square summable, we conclude that $\hat{f}$ is summable and hence the Fourier series converges uniformly. So we have a simple sufficient criterion for summability of the Fourier coefficients, but can it be improved? Of course continuity of f is a necessary condition but this alone will not even be enough for pointwise convergence as we will see in Example 4.3. Moreover, continuity will not tell us more about the decay of the Fourier coefficients than what we already know in the integrable case from the Riemann–Lebesgue lemma (see Example 4.4).

A few improvements are easy: First of all, piecewise continuously differentiable would be sufficient for this argument. Or, slightly more general, an absolutely continuous function whose derivative is square integrable would also do (cf. Lemma 11.52). However, even for an absolutely continuous function the Fourier coefficients might not be summable: For an absolutely continuous function f we have a derivative which is integrable (Theorem 11.51) and hence the above formula combined with the Riemann–Lebesgue lemma implies $\hat{f}_k = o(\frac{1}{k})$. But on the other hand we can choose a summable sequence c_k which does not obey this asymptotic requirement, say $c_k = \frac{1}{k}$ for $k = l^2$ and $c_k = 0$ else. Then

$$f(x) := \sum_{k\in\mathbb{Z}} c_k \mathrm{e}^{\mathrm{i}kx} = \sum_{l\in\mathbb{N}} \frac{1}{l^2} \mathrm{e}^{\mathrm{i}l^2 x} \tag{2.50}$$

is a function with summable Fourier coefficients $\hat{f}_k = c_k$ (by uniform convergence we can interchange summation and integration) but which is not absolutely continuous. There are further criteria for summability of the Fourier coefficients, but no simple necessary and sufficient one.

Note however, that the situation looks much brighter if one looks at mean values

$$\bar{S}_n(f)(x) := \frac{1}{n} \sum_{k=0}^{n-1} S_n(f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F_n(x-y) f(y) dy, \tag{2.51}$$

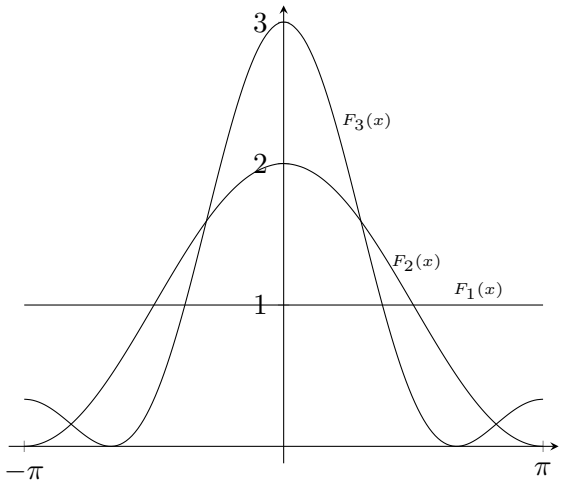


Figure 2.2. The Fejér kernels F_1, F_2, and F_3

where

$$F_n(x) = \frac{1}{n}\sum_{k=0}^{n-1} D_k(x) = \frac{1}{n}\left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^2 \tag{2.52}$$

is the **Fejér kernel**. To see the second form we use the closed form for the Dirichlet kernel to obtain

$$\begin{aligned} nF_n(x) &= \sum_{k=0}^{n-1} \frac{\sin((k+1/2)x)}{\sin(x/2)} = \frac{1}{\sin(x/2)} \operatorname{Im} \sum_{k=0}^{n-1} \mathrm{e}^{\mathrm{i}(k+1/2)x} \\ &= \frac{1}{\sin(x/2)} \operatorname{Im}\left(\mathrm{e}^{\mathrm{i}x/2}\frac{\mathrm{e}^{\mathrm{i}nx}-1}{\mathrm{e}^{\mathrm{i}x}-1}\right) = \frac{1-\cos(nx)}{2\sin(x/2)^2} = \left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^2. \end{aligned}$$

The main difference to the Dirichlet kernel is positivity: $F_n(x) \geq 0$. Of course the property $\int_{-\pi}^{\pi} F_n(x)dx = 1$ is inherited from the Dirichlet kernel.

Theorem 2.20 (Fejér)**.** *Suppose f is continuous and periodic with period 2π. Then $\bar{S}_n(f) \to f$ uniformly.*

Proof. Let us set $F_n = 0$ outside $[-\pi, \pi]$. Then $F_n(x) \leq \frac{1}{n\sin(\delta/2)^2}$ for $\delta \leq |x| \leq \pi$ implies that a straightforward adaption of Lemma 1.2 to the periodic case is applicable. □

In particular, this shows that the functions $\{e_k\}_{k\in\mathbb{Z}}$ are total in $C_{per}[-\pi, \pi]$ (continuous periodic functions) and hence also in $L^p(-\pi, \pi)$ for $1 \leq p < \infty$ (Problem 2.19).

Note that for a given continuous function f this result shows that if $S_n(f)(x)$ converges, then it must converge to $\bar{S}_n(f)(x) = f(x)$. We also

remark that one can extend this result (see Lemma 10.19) to show that for $f \in L^p(-\pi,\pi)$, $1 \le p < \infty$, one has $\bar{S}_n(f) \to f$ in the sense of L^p. As a consequence note that the Fourier coefficients uniquely determine f for integrable f (for square integrable f this follows from Theorem 2.18).

Finally, we look at pointwise convergence.

Theorem 2.21. *Suppose*

$$\frac{f(x)-f(x_0)}{x-x_0} \tag{2.53}$$

is integrable (e.g. f is Hölder continuous), then

$$\lim_{m,n\to\infty} \sum_{k=-m}^{n} \hat{f}(k) \mathrm{e}^{\mathrm{i}kx_0} = f(x_0). \tag{2.54}$$

Proof. Without loss of generality we can assume $x_0 = 0$ (by shifting $x \to x - x_0$ modulo 2π implying $\hat{f}_k \to \mathrm{e}^{-\mathrm{i}kx_0}\hat{f}_k$) and $f(x_0) = 0$ (by linearity since the claim is trivial for constant functions). Then by assumption

$$g(x) := \frac{f(x)}{\mathrm{e}^{\mathrm{i}x}-1}$$

is integrable and $f(x) = (\mathrm{e}^{\mathrm{i}x}-1)g(x)$ implies $\hat{f}_k = \hat{g}_{k-1} - \hat{g}_k$ and hence

$$\sum_{k=m}^{n} \hat{f}_k = \hat{g}_{-m-1} - \hat{g}_n.$$

Now the claim follows from the Riemann–Lebesgue lemma. □

If we look at symmetric partial sums $S_n(f)$ we can do even better.

Corollary 2.22 (Dirichlet–Dini criterion)**.** *Suppose there is some α such that*

$$\frac{f(x_0+x)+f(x_0-x)-2\alpha}{x}$$

is integrable. Then $S_n(f)(x_0) \to \alpha$.

Proof. Without loss of generality we can assume $x_0 = 0$. Now observe (since $D_n(-x) = D_n(x)$) $S_n(f)(0) = \alpha + S_n(g)(0)$, where $g(x) := \frac{1}{2}(f(x) + f(-x)) - \alpha$ and apply the previous result. □

Problem 2.17. *Compute the Fourier series of D_n and F_n.*

Problem 2.18. *Show $|D_n(x)| \le \min(2n+1, \frac{\pi}{|x|})$ and $F_n(x) \le \min(n, \frac{\pi^2}{nx^2})$.*

Problem 2.19. *Show that $C_{per}[-\pi,\pi]$ is dense in $L^p(-\pi,\pi)$ for $1 \le p < \infty$.*

Problem 2.20. *Show that the functions $e_k(x) := \frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathrm{i}kx}$, $k \in \mathbb{Z}$, form an orthonormal basis for $\mathfrak{H} = L^2(-\pi,\pi)$. (Hint: Start with $K = [-\pi,\pi]$ where $-\pi$ and π are identified and use the Stone–Weierstraß theorem.)*

Chapter 3

Compact operators

Typically, linear operators are much more difficult to analyze than matrices and many new phenomena appear which are not present in the finite dimensional case. So we have to be modest and slowly work our way up. A class of operators which still preserves some of the nice properties of matrices is the class of compact operators to be discussed in this chapter.

3.1. Compact operators

A linear operator $A : X \to Y$ defined between normed spaces X, Y is called **compact** if every sequence Af_n has a convergent subsequence whenever f_n is bounded. Equivalently (cf. Corollary B.20), A is compact if it maps bounded sets to relatively compact ones. The set of all compact operators is denoted by $\mathscr{C}(X,Y)$. If $X = Y$ we will just write $\mathscr{C}(X) := \mathscr{C}(X,X)$ as usual.

Example 3.1. Every linear map between finite dimensional spaces is compact by the Bolzano–Weierstraß theorem. Slightly more general, a bounded operator is compact if its range is finite dimensional. ⋄

The following elementary properties of compact operators are left as an exercise (Problem 3.1):

Theorem 3.1. *Let X, Y, and Z be normed spaces. Every compact linear operator is bounded, $\mathscr{C}(X,Y) \subseteq \mathscr{L}(X,Y)$. Linear combinations of compact operators are compact, that is, $\mathscr{C}(X,Y)$ is a subspace of $\mathscr{L}(X,Y)$. Moreover, the product of a bounded and a compact operator is again compact, that is, $A \in \mathscr{L}(X,Y)$, $B \in \mathscr{C}(Y,Z)$ or $A \in \mathscr{C}(X,Y)$, $B \in \mathscr{L}(Y,Z)$ implies $BA \in \mathscr{C}(X,Z)$.*

In particular, the set of compact operators $\mathscr{C}(X)$ is an ideal of the set of bounded operators. Moreover, if X is a Banach space this ideal is even closed:

Theorem 3.2. *Suppose X is a normed and Y a Banach space. Let $A_n \in \mathscr{C}(X,Y)$ be a convergent sequence of compact operators. Then the limit A is again compact.*

Proof. Let f_j^0 be a bounded sequence. Choose a subsequence f_j^1 such that $A_1 f_j^1$ converges. From f_j^1 choose another subsequence f_j^2 such that $A_2 f_j^2$ converges and so on. Since there might be nothing left from f_j^n as $n \to \infty$, we consider the diagonal sequence $f_j := f_j^j$. By construction, f_j is a subsequence of f_j^n for $j \geq n$ and hence $A_n f_j$ is Cauchy for every fixed n. Now

$$\begin{aligned}\|Af_j - Af_k\| &= \|(A - A_n)(f_j - f_k) + A_n(f_j - f_k)\| \\ &\leq \|A - A_n\|\|f_j - f_k\| + \|A_n f_j - A_n f_k\|\end{aligned}$$

shows that Af_j is Cauchy since the first term can be made arbitrary small by choosing n large and the second by the Cauchy property of $A_n f_j$. □

Example 3.2. Let $X := \ell^p(\mathbb{N})$ and consider the operator

$$(Qa)_j := q_j a_j$$

for some sequence $q = (q_j)_{j=1}^\infty \in c_0(\mathbb{N})$ converging to zero. Let Q_n be associated with $q_j^n = q_j$ for $j \leq n$ and $q_j^n = 0$ for $j > n$. Then the range of Q^n is finite dimensional and hence Q_n is compact. Moreover, by $\|Q_n - Q\| = \sup_{j>n} |q_j|$ we see $Q_n \to Q$ and thus Q is also compact by the previous theorem. ◇

Example 3.3. Let $X := C^1[0,1]$, $Y := C[0,1]$ (cf. Problem 1.31) then the embedding $X \hookrightarrow Y$ is compact. Indeed, a bounded sequence in X has both the functions and the derivatives uniformly bounded. Hence by the mean value theorem the functions are equicontinuous and hence there is a uniformly convergent subsequence by the Arzelà–Ascoli theorem (Theorem 1.14). Of course the same conclusion holds if we take $X := C^{0,\gamma}[0,1]$ to be Hölder continuous functions and if we replace $[0,1]$ by a compact metric space (cf. Theorem 1.22). ◇

If $A : X \to Y$ is a bounded operator there is a unique extension $\overline{A} : \overline{X} \to \overline{Y}$ to the completion by Theorem 1.17. Moreover, if $A \in \mathscr{C}(X,Y)$, then $A \in \mathscr{C}(X,\overline{Y})$ is immediate. That we also have $\overline{A} \in \mathscr{C}(\overline{X},\overline{Y})$ will follow from the next lemma. In particular, it suffices to verify compactness on a dense set.

Lemma 3.3. *Let X, Y be normed spaces and $A \in \mathscr{C}(X,Y)$. Let $\overline{X}$, $\overline{Y}$ be the completion of X, Y, respectively. Then $\overline{A} \in \mathscr{C}(\overline{X},\overline{Y})$, where $\overline{A}$ is the unique extension of A (cf. Theorem 1.17).*

Proof. Let $f_n \in \overline{X}$ be a given bounded sequence. We need to show that $\overline{A}f_n$ has a convergent subsequence. Pick $f_n^j \in X$ such that $\|f_n^j - f_n\| \leq \frac{1}{j}$ and by compactness of A we can assume that $Af_n^n \to g$. But then $\|\overline{A}f_n - g\| \leq \|A\|\|f_n - f_n^n\| + \|Af_n^n - g\|$ shows that $\overline{A}f_n \to g$. □

One of the most important examples of compact operators are integral operators. The proof will be based on the Arzelà–Ascoli theorem (Theorem 1.14).

Lemma 3.4. *Let $X := C([a,b])$ or $X := \mathcal{L}^2_{cont}(a,b)$. The integral operator $K : X \to X$ defined by*

$$(Kf)(x) := \int_a^b K(x,y)f(y)dy, \tag{3.1}$$

where $K(x,y) \in C([a,b]\times[a,b])$, is compact.

Proof. First of all note that $K(.,..)$ is continuous on $[a,b]\times[a,b]$ and hence uniformly continuous. In particular, for every $\varepsilon > 0$ we can find a $\delta > 0$ such that $|K(y,t) - K(x,t)| \leq \varepsilon$ for any $t \in [a,b]$ whenever $|y - x| \leq \delta$. Moreover, $\|K\|_\infty = \sup_{x,y\in[a,b]} |K(x,y)| < \infty$.

We begin with the case $X := \mathcal{L}^2_{cont}(a,b)$. Let $g := Kf$. Then

$$|g(x)| \leq \int_a^b |K(x,t)|\,|f(t)|dt \leq \|K\|_\infty \int_a^b |f(t)|dt \leq \|K\|_\infty \|1\|\,\|f\|,$$

where we have used Cauchy–Schwarz in the last step (note that $\|1\| = \sqrt{b-a}$). Similarly,

$$\begin{aligned} |g(x) - g(y)| &\leq \int_a^b |K(y,t) - K(x,t)|\,|f(t)|dt \\ &\leq \varepsilon \int_a^b |f(t)|dt \leq \varepsilon\|1\|\,\|f\|, \end{aligned}$$

whenever $|y - x| \leq \delta$. Hence, if $f_n(x)$ is a bounded sequence in $\mathcal{L}^2_{cont}(a,b)$, then $g_n := Kf_n$ is bounded and equicontinuous and hence has a uniformly convergent subsequence by the Arzelà–Ascoli theorem (Theorem 1.14). But a uniformly convergent sequence is also convergent in the norm induced by the scalar product. Therefore K is compact.

The case $X := C([a,b])$ follows by the same argument upon observing $\int_a^b |f(t)|dt \leq (b-a)\|f\|_\infty$. □

Compact operators share many similarities with (finite) matrices as we will see in the next section.

Problem* 3.1. *Show Theorem 3.1.*

Problem* 3.2. *Show that the adjoint of the integral operator K on $\mathcal{L}^2_{cont}(a,b)$ from Lemma 3.4 is the integral operator with kernel $K(y,x)^*$:*

$$(K^*f)(x) = \int_a^b K(y,x)^* f(y) dy.$$

(Hint: Fubini.)

Problem 3.3. *Show that the operator $\frac{d}{dx} : C^2[a,b] \to C[a,b]$ is compact. (Hint: Arzelà–Ascoli.)*

3.2. The spectral theorem for compact symmetric operators

Let $\mathfrak{H}$ be an inner product space. A linear operator $A : \mathfrak{D}(A) \subseteq \mathfrak{H} \to \mathfrak{H}$ is called **symmetric** if its domain is dense and if

$$\langle g, Af\rangle = \langle Ag, f\rangle \qquad f, g \in \mathfrak{D}(A). \tag{3.2}$$

If A is bounded (with $\mathfrak{D}(A) = \mathfrak{H}$), then A is symmetric precisely if $A = A^*$, that is, if A is **self-adjoint**. However, for unbounded operators there is a subtle but important difference between symmetry and self-adjointness.

A number $z \in \mathbb{C}$ is called **eigenvalue** of A if there is a nonzero vector $u \in \mathfrak{D}(A)$ such that

$$Au = zu. \tag{3.3}$$

The vector u is called a corresponding **eigenvector** in this case. The set of all eigenvectors corresponding to z is called the **eigenspace**

$$\operatorname{Ker}(A - z) \tag{3.4}$$

corresponding to z. Here we have used the shorthand notation $A - z$ for $A - z\mathbb{I}$. An eigenvalue is called (geometrically) **simple** if there is only one linearly independent eigenvector.

Example 3.4. Let $\mathfrak{H} := \ell^2(\mathbb{N})$ and consider the shift operators $(S^\pm a)_j := a_{j\pm1}$ (with $a_0 := 0$). Suppose $z \in \mathbb{C}$ is an eigenvalue, then the corresponding eigenvector u must satisfy $u_{j\pm1} = zu_j$. For S^- the special case $j = 1$ gives $0 = u_0 = zu_1$. So either $z = 0$ and $u = 0$ or $z \neq 0$ and again $u = 0$. Hence there are no eigenvalues. For S^+ we get $u_j = z^j u_1$ and this will give an element in $\ell^2(\mathbb{N})$ if and only if $|z| < 1$. Hence z with $|z| < 1$ is an eigenvalue. All these eigenvalues are simple. ⋄

Example 3.5. Let $\mathfrak{H} := \ell^2(\mathbb{N})$ and consider the multiplication operator $(Qa)_j := q_j a_j$ with a bounded sequence $q \in \ell^\infty(\mathbb{N})$. Suppose $z \in \mathbb{C}$ is an eigenvalue, then the corresponding eigenvector u must satisfy $(q_j - z)u_j = 0$.

Hence every value q_j is an eigenvalue with corresponding eigenvector $u = \delta^j$. If there is only one j with $z = q_j$ the eigenvalue is simple (otherwise the numbers of independent eigenvectors equals the number of times z appears in the sequence q). If z is different from all entries of the sequence then $u = 0$ and z is no eigenvalue. ⋄

Note that in the last example Q will be self-adjoint if and only if q is real-valued and hence if and only if all eigenvalues are real-valued. Moreover, the corresponding eigenfunctions are orthogonal. This has nothing to do with the simple structure of our operator and is in fact always true.

Theorem 3.5. *Let A be symmetric. Then all eigenvalues are real and eigenvectors corresponding to different eigenvalues are orthogonal.*

Proof. Suppose λ is an eigenvalue with corresponding normalized eigenvector u. Then $\lambda = \langle u, Au\rangle = \langle Au, u\rangle = \lambda^*$, which shows that λ is real. Furthermore, if $Au_j = \lambda_j u_j$, $j = 1, 2$, we have

$$(\lambda_1 - \lambda_2)\langle u_1, u_2\rangle = \langle Au_1, u_2\rangle - \langle u_1, Au_2\rangle = 0$$

finishing the proof. □

Note that while eigenvectors corresponding to the same eigenvalue λ will in general not automatically be orthogonal, we can of course replace each set of eigenvectors corresponding to λ by an set of orthonormal eigenvectors having the same linear span (e.g. using Gram–Schmidt orthogonalization).

Example 3.6. Let $\mathfrak{H} := \ell^2(\mathbb{N})$ and consider the Jacobi operator $J := \frac{1}{2}(S^+ + S^-)$:

$$(Jc)_j := \frac{1}{2}(c_{j+1} + c_{j-1})$$

with the convention $c_0 = 0$. Recall that $J^* = J$. If we look for an eigenvalue $Ju = zu$, we need to solve the corresponding recursion $u_{j+1} = 2zu_j - u_{j-1}$ starting from $u_0 = 0$ (our convention) and $u_1 = 1$ (normalization). Like an ordinary differential equation, a linear recursion relation with constant coefficients can be solved by an exponential ansatz $u_j = k^j$ which leads to the characteristic polynomial $k^2 = 2zk - 1$. This gives two linearly independent solutions and our requirements lead us to

$$u_j(z) = \frac{k^j - k^{-j}}{k - k^{-1}}, \qquad k = z - \sqrt{z^2 - 1}.$$

Note that $k^{-1} = z + \sqrt{z^2 - 1}$ and in the case $k = z = \pm 1$ the above expression has to be understood as its limit $u_j(\pm 1) = (\pm 1)^{j+1} j$. In fact, $U_j(z) := u_{j+1}(z)$ are polynomials of degree j known as **Chebyshev polynomials** of the second kind.

Now for $z \in \mathbb{R} \setminus [-1,1]$ we have $|k| < 1$ and u_j explodes exponentially. For $z \in [-1,1]$ we have $|k| = 1$ and hence we can write $k = \mathrm{e}^{\mathrm{i}\kappa}$ with $\kappa \in \mathbb{R}$. Thus $u_j = \frac{\sin(\kappa j)}{\sin(\kappa)}$ is oscillating. So for no value of $z \in \mathbb{R}$ our potential eigenvector u is square summable and thus J has no eigenvalues. $\diamond$

The previous example shows that in the infinite dimensional case symmetry is not enough to guarantee existence of even a single eigenvalue. In order to always get this, we will need an extra condition. In fact, we will see that compactness provides a suitable extra condition to obtain an orthonormal basis of eigenfunctions. The crucial step is to prove existence of one eigenvalue, the rest then follows as in the finite dimensional case.

Theorem 3.6. *Let $\mathfrak{H}$ be an inner product space. A symmetric compact operator A has an eigenvalue α_1 which satisfies $|\alpha_1| = \|A\|$.*

Proof. We set $\alpha := \|A\|$ and assume $\alpha \neq 0$ (i.e, $A \neq 0$) without loss of generality. Since

$$\|A\|^2 = \sup_{f:\|f\|=1} \|Af\|^2 = \sup_{f:\|f\|=1} \langle Af, Af\rangle = \sup_{f:\|f\|=1} \langle f, A^2 f\rangle$$

there exists a normalized sequence u_n such that

$$\lim_{n\to\infty} \langle u_n, A^2 u_n\rangle = \alpha^2.$$

Since A is compact, it is no restriction to assume that $A^2 u_n$ converges, say $\lim_{n\to\infty} A^2 u_n = \alpha^2 u$. Now

$$\begin{aligned}\|(A^2 - \alpha^2)u_n\|^2 &= \|A^2 u_n\|^2 - 2\alpha^2\langle u_n, A^2 u_n\rangle + \alpha^4 \\ &\leq 2\alpha^2(\alpha^2 - \langle u_n, A^2 u_n\rangle)\end{aligned}$$

(where we have used $\|A^2 u_n\| \leq \|A\|\|Au_n\| \leq \|A\|^2\|u_n\| = \alpha^2$) implies $\lim_{n\to\infty}(A^2 u_n - \alpha^2 u_n) = 0$ and hence $\lim_{n\to\infty} u_n = u$. In addition, u is a normalized eigenvector of A^2 since $(A^2 - \alpha^2)u = 0$. Factorizing this last equation according to $(A - \alpha)u = v$ and $(A + \alpha)v = 0$ shows that either $v \neq 0$ is an eigenvector corresponding to $-\alpha$ or $v = 0$ and hence $u \neq 0$ is an eigenvector corresponding to α. $\square$

Note that for a bounded operator A, there cannot be an eigenvalue with absolute value larger than $\|A\|$, that is, the set of eigenvalues is bounded by $\|A\|$ (Problem 3.4).

Now consider a symmetric compact operator A with eigenvalue α_1 (as above) and corresponding normalized eigenvector u_1. Setting

$$\mathfrak{H}_1 := \{u_1\}^\perp = \{f \in \mathfrak{H} | \langle u_1, f\rangle = 0\} \tag{3.5}$$

we can restrict A to $\mathfrak{H}_1$ since $f \in \mathfrak{H}_1$ implies

$$\langle u_1, Af\rangle = \langle Au_1, f\rangle = \alpha_1\langle u_1, f\rangle = 0 \tag{3.6}$$

and hence $Af \in \mathfrak{H}_1$. Denoting this restriction by A_1, it is not hard to see that A_1 is again a symmetric compact operator. Hence we can apply Theorem 3.6 iteratively to obtain a sequence of eigenvalues α_j with corresponding normalized eigenvectors u_j. Moreover, by construction, u_j is orthogonal to all u_k with $k < j$ and hence the eigenvectors $\{u_j\}$ form an orthonormal set. By construction we also have $|\alpha_j| = \|A_j\| \leq \|A_{j-1}\| = |\alpha_{j-1}|$. This procedure will not stop unless $\mathfrak{H}$ is finite dimensional. However, note that $\alpha_j = 0$ for $j \geq n$ might happen if $A_n = 0$.

Theorem 3.7 (Spectral theorem for compact symmetric operators; Hilbert)**.** *Suppose $\mathfrak{H}$ is an infinite dimensional Hilbert space and $A : \mathfrak{H} \to \mathfrak{H}$ is a compact symmetric operator. Then there exists a sequence of real eigenvalues α_j converging to 0. The corresponding normalized eigenvectors u_j form an orthonormal set and every $f \in \mathfrak{H}$ can be written as*

$$f = \sum_{j=1}^{\infty} \langle u_j, f\rangle u_j + h, \tag{3.7}$$

where h is in the kernel of A, that is, $Ah = 0$.

In particular, if 0 is not an eigenvalue, then the eigenvectors form an orthonormal basis (in addition, $\mathfrak{H}$ need not be complete in this case).

Proof. Existence of the eigenvalues α_j and the corresponding eigenvectors u_j has already been established. Since the sequence $|\alpha_j|$ is decreasing it has a limit $\varepsilon \geq 0$ and we have $|\alpha_j| \geq \varepsilon$. If this limit is nonzero, then $v_j = \alpha_j^{-1} u_j$ is a bounded sequence ($\|v_j\| \leq \frac{1}{\varepsilon}$) for which Av_j has no convergent subsequence since $\|Av_j - Av_k\|^2 = \|u_j - u_k\|^2 = 2$, a contradiction.

Next, setting

$$f_n := \sum_{j=1}^{n} \langle u_j, f\rangle u_j,$$

we have

$$\|A(f - f_n)\| \leq |\alpha_{n+1}| \|f - f_n\| \leq |\alpha_{n+1}| \|f\|$$

since $f - f_n \in \mathfrak{H}_n$ and $\|A_n\| = |\alpha_{n+1}|$. Letting $n \to \infty$ shows $A(f_\infty - f) = 0$ proving (3.7). Finally, note that without completeness f_∞ might not be well-defined unless $h = 0$. □

By applying A to (3.7) we obtain the following canonical form of compact symmetric operators.

Corollary 3.8. *Every compact symmetric operator A can be written as*

$$Af = \sum_{j=1}^{N} \alpha_j \langle u_j, f\rangle u_j, \tag{3.8}$$

where α_j are the nonzero eigenvalues with corresponding eigenvectors u_j from the previous theorem.

Remark: There are two cases where our procedure might fail to construct an orthonormal basis of eigenvectors. One case is where there is an infinite number of nonzero eigenvalues. In this case α_n never reaches 0 and all eigenvectors corresponding to 0 are missed. In the other case, 0 is reached, but there might not be a countable basis and hence again some of the eigenvectors corresponding to 0 are missed. In any case, by adding vectors from the kernel (which are automatically eigenvectors), one can always extend the eigenvectors u_j to an orthonormal basis of eigenvectors.

Corollary 3.9. *Every compact symmetric operator A has an associated orthonormal basis of eigenvectors $\{u_j\}_{j\in J}$. The corresponding unitary map $U : \mathfrak{H} \to \ell^2(J)$, $f \mapsto \{\langle u_j, f\rangle\}_{j\in J}$ diagonalizes A in the sense that UAU^{-1} is the operator which multiplies each basis vector $\delta^j = Uu_j$ by the corresponding eigenvalue α_j.*

Example 3.7. Let $a, b \in c_0(\mathbb{N})$ be real-valued sequences and consider the operator

$$(Jc)_j := a_j c_{j+1} + b_j c_j + a_{j-1} c_{j-1}.$$

If A, B denote the multiplication operators by the sequences a, b, respectively, then we already know that A and B are compact. Moreover, using the shift operators $S^{\pm}$ we can write

$$J = AS^{+} + B + S^{-}A,$$

which shows that J is self-adjoint since $A^* = A$, $B^* = B$, and $(S^{\pm})^* = S^{\mp}$. Hence we can conclude that J has a countable number of eigenvalues converging to zero and a corresponding orthonormal basis of eigenvectors. $\diamond$

In particular, in the new picture it is easy to define functions of our operator (thus extending the functional calculus from Problem 1.35). To this end set $\Sigma := \{\alpha_j\}_{j\in J}$ and denote by $B(K)$ the Banach algebra of bounded functions $F : K \to \mathbb{C}$ together with the sup norm.

Corollary 3.10 (Functional calculus)**.** *Let A be a compact symmetric operator with associated orthonormal basis of eigenvectors $\{u_j\}_{j\in J}$ and corresponding eigenvalues $\{\alpha_j\}_{j\in J}$. Suppose $F \in B(\Sigma)$, then*

$$F(A)f = \sum_{j\in J} F(\alpha_j)\langle u_j, f\rangle u_j \tag{3.9}$$

defines a continuous algebra homomorphism from the Banach algebra $B(\Sigma)$ to the algebra $\mathscr{L}(\mathfrak{H})$ with $1(A) = \mathbb{I}$ and $\mathbb{I}(A) = A$. Moreover $F(A)^* = F^*(A)$, where F^* is the function which takes complex conjugate values.

Proof. This is straightforward to check for multiplication operators in $\ell^2(J)$ and hence the result follows by the previous corollary. □

In many applications F will be given by a function on $\mathbb{R}$ (or at least on $[-\|A\|, \|A\|]$) and, since only the values $F(\alpha_j)$ are used, two functions which agree on all eigenvalues will give the same result.

As a brief application we will say a few words about general spectral theory for bounded operators $A \in \mathscr{L}(X)$ in a Banach space X. In the finite dimensional case, the spectrum is precisely the set of eigenvalues. In the infinite dimensional case one defines the **spectrum** as

$$\sigma(A) := \mathbb{C} \setminus \{z \in \mathbb{C} | \exists (A - z)^{-1} \in \mathscr{L}(X)\}. \tag{3.10}$$

It is important to emphasize that the inverse is required to exist as a bounded operator. Hence there are several ways in which this can fail: First of all, $A - z$ could not be injective. In this case z is an eigenvalue and thus all eigenvalues belong to the spectrum. Secondly, it could not be surjective. And finally, even if it is bijective it could be unbounded. However, it will follow form the open mapping theorem that this last case cannot happen for a bounded operator. The inverse of $A - z$ for $z \in \mathbb{C} \setminus \sigma(A)$ is known as the **resolvent** of A and plays a crucial role in spectral theory. Using Problem 1.34 one can show that the complement of the spectrum is open, and hence the spectrum is closed. Since we will discuss this in detail in Chapter 6, we will not pursue this here but only look at our special case of symmetric compact operators.

To compute the inverse of $A - z$ we will use the functional calculus and consider $F(\alpha) = \frac{1}{\alpha - z}$. Of course this function is unbounded on $\mathbb{R}$ but if z is neither an eigenvalue nor zero it is bounded on Σ and hence satisfies our requirements. Then

$$R_A(z) f := \sum_{j \in J} \frac{1}{\alpha_j - z} \langle u_j, f \rangle u_j \tag{3.11}$$

satisfies $(A - z) R_A(z) = R_A(z)(A - z) = \mathbb{I}$, that is, $R_A(z) = (A - z)^{-1} \in \mathscr{L}(\mathfrak{H})$. Of course, if z is an eigenvalue, then the above formula breaks down. However, in the infinite dimensional case it also breaks down if $z = 0$ even if 0 is not an eigenvalue! In this case the above definition will still give an operator which is the inverse of $A - z$, however, since the sequence α_j^{-1} is unbounded, so will be the corresponding multiplication operator in $\ell^2(J)$ and the sum in (3.11) will only converge if $\{\alpha_j^{-1} \langle u_j, f \rangle\}_{j \in J} \in \ell^2(J)$. So in the infinite dimensional case 0 is in the spectrum even if it is not an eigenvalue. In particular,

$$\sigma(A) = \overline{\{\alpha_j\}_{j \in J}}. \tag{3.12}$$

Moreover, if we use $\frac{1}{\alpha_j - z} = \frac{\alpha_j}{z(\alpha_j - z)} - \frac{1}{z}$ we can rewrite this as

$$R_A(z)f = \frac{1}{z}\left(\sum_{j=1}^{N} \frac{\alpha_j}{\alpha_j - z}\langle u_j, f\rangle u_j - f\right)$$

where it suffices to take the sum over all nonzero eigenvalues.

This is all we need and it remains to apply these results to Sturm–Liouville operators.

Problem 3.4. *Show that if $A \in \mathscr{L}(\mathfrak{H})$, then every eigenvalue α satisfies $|\alpha| \leq \|A\|$.*

Problem 3.5. *Find the eigenvalues and eigenfunctions of the integral operator $K \in \mathscr{L}(\mathcal{L}^2_{cont}(0,1))$ given by*

$$(Kf)(x) := \int_0^1 u(x)v(y)f(y)dy,$$

where $u, v \in C([0,1])$ are some given continuous functions.

Problem 3.6. *Find the eigenvalues and eigenfunctions of the integral operator $K \in \mathscr{L}(\mathcal{L}^2_{cont}(0,1))$ given by*

$$(Kf)(x) := 2\int_0^1 (2xy - x - y + 1)f(y)dy.$$

3.3. Applications to Sturm–Liouville operators

Now, after all this hard work, we can show that our Sturm–Liouville operator

$$L := -\frac{d^2}{dx^2} + q(x), \tag{3.13}$$

where q is continuous and real, defined on

$$\mathfrak{D}(L) := \{f \in C^2[0,1] | f(0) = f(1) = 0\} \subset \mathcal{L}^2_{cont}(0,1), \tag{3.14}$$

has an orthonormal basis of eigenfunctions.

The corresponding eigenvalue equation $Lu = zu$ explicitly reads

$$-u''(x) + q(x)u(x) = zu(x). \tag{3.15}$$

It is a second order homogeneous linear ordinary differential equations and hence has two linearly independent solutions. In particular, specifying two initial conditions, e.g. $u(0) = 0, u'(0) = 1$ determines the solution uniquely. Hence, if we require $u(0) = 0$, the solution is determined up to a multiple and consequently the additional requirement $u(1) = 0$ cannot be satisfied by a nontrivial solution in general. However, there might be some $z \in \mathbb{C}$ for which the solution corresponding to the initial conditions $u(0) = 0, u'(0) = 1$

happens to satisfy $u(1) = 0$ and these are precisely the eigenvalues we are looking for.

Note that the fact that $\mathcal{L}^2_{cont}(0,1)$ is not complete causes no problems since we can always replace it by its completion $\mathfrak{H} = L^2(0,1)$. A thorough investigation of this completion will be given later, at this point this is not essential.

We first verify that L is symmetric:

$$\begin{aligned}\langle f, Lg\rangle &= \int_0^1 f(x)^*(-g''(x) + q(x)g(x))dx \\ &= \int_0^1 f'(x)^* g'(x)dx + \int_0^1 f(x)^* q(x)g(x)dx \\ &= \int_0^1 -f''(x)^* g(x)dx + \int_0^1 f(x)^* q(x)g(x)dx \\ &= \langle Lf, g\rangle.\end{aligned} \tag{3.16}$$

Here we have used integration by parts twice (the boundary terms vanish due to our boundary conditions $f(0) = f(1) = 0$ and $g(0) = g(1) = 0$).

Of course we want to apply Theorem 3.7 and for this we would need to show that L is compact. But this task is bound to fail, since L is not even bounded (see Example 1.17)!

So here comes the trick: If L is unbounded its inverse L^{-1} might still be bounded. Moreover, L^{-1} might even be compact and this is the case here! Since L might not be injective (0 might be an eigenvalue), we consider $R_L(z) := (L - z)^{-1}$, $z \in \mathbb{C}$, which is also known as the **resolvent** of L.

In order to compute the resolvent, we need to solve the inhomogeneous equation $(L - z)f = g$. This can be done using the variation of constants formula from ordinary differential equations which determines the solution up to an arbitrary solution of the homogeneous equation. This homogeneous equation has to be chosen such that $f \in \mathfrak{D}(L)$, that is, such that $f(0) = f(1) = 0$.

Define

$$\begin{aligned}f(x) :=& \frac{u_+(z,x)}{W(z)}\Big(\int_0^x u_-(z,t)g(t)dt\Big) \\ &+ \frac{u_-(z,x)}{W(z)}\Big(\int_x^1 u_+(z,t)g(t)dt\Big),\end{aligned} \tag{3.17}$$

where $u_\pm(z,x)$ are the solutions of the homogeneous differential equation $-u''_\pm(z,x)+(q(x)-z)u_\pm(z,x) = 0$ satisfying the initial conditions $u_-(z,0) = 0$, $u'_-(z,0) = 1$ respectively $u_+(z,1) = 0$, $u'_+(z,1) = 1$ and

$$W(z) := W(u_+(z), u_-(z)) = u'_-(z,x)u_+(z,x) - u_-(z,x)u'_+(z,x) \tag{3.18}$$

is the Wronski determinant, which is independent of x (check this!).

Then clearly $f(0) = 0$ since $u_-(z,0) = 0$ and similarly $f(1) = 0$ since $u_+(z,1) = 0$. Furthermore, f is differentiable and a straightforward computation verifies

$$\begin{aligned} f'(x) =& \frac{u_+'(z,x)}{W(z)}\Big(\int_0^x u_-(z,t)g(t)dt\Big) \\ &+ \frac{u_-'(z,x)}{W(z)}\Big(\int_x^1 u_+(z,t)g(t)dt\Big). \end{aligned} \tag{3.19}$$

Thus we can differentiate once more giving

$$\begin{aligned} f''(x) =& \frac{u_+''(z,x)}{W(z)}\Big(\int_0^x u_-(z,t)g(t)dt\Big) \\ &+ \frac{u_-''(z,x)}{W(z)}\Big(\int_x^1 u_+(z,t)g(t)dt\Big) - g(x) \\ =& (q(x) - z)f(x) - g(x). \end{aligned} \tag{3.20}$$

In summary, f is in the domain of L and satisfies $(L - z)f = g$.

Note that z is an eigenvalue if and only if $W(z) = 0$. In fact, in this case $u_+(z,x)$ and $u_-(z,x)$ are linearly dependent and hence $u_+(z,x) = c\,u_-(z,x)$ with $c = u_+'(z,0)$. Evaluating this identity at $x = 0$ shows $u_+(z,0) = c\,u_-(z,0) = 0$ that $u_+(z,x)$ satisfies both boundary conditions and is thus an eigenfunction.

Introducing the **Green function**

$$G(z,x,t) := \frac{1}{W(u_+(z), u_-(z))} \begin{cases} u_+(z,x)u_-(z,t), & x \geq t, \\ u_+(z,t)u_-(z,x), & x \leq t, \end{cases} \tag{3.21}$$

we see that $(L - z)^{-1}$ is given by

$$(L - z)^{-1}g(x) = \int_0^1 G(z,x,t)g(t)dt. \tag{3.22}$$

Moreover, from $G(z,x,t) = G(z,t,x)$ it follows that $(L - z)^{-1}$ is symmetric for $z \in \mathbb{R}$ (Problem 3.7) and from Lemma 3.4 it follows that it is compact. Hence Theorem 3.7 applies to $(L - z)^{-1}$ once we show that we can find a real z which is not an eigenvalue.

Theorem 3.11. *The Sturm–Liouville operator L has a countable number of discrete and simple eigenvalues E_n which accumulate only at ∞. They are bounded from below and can hence be ordered as follows:*

$$\min_{x\in[0,1]} q(x) < E_0 < E_1 < \cdots . \tag{3.23}$$

The corresponding normalized eigenfunctions u_n form an orthonormal basis for $\mathcal{L}^2_{cont}(0,1)$, that is, every $f \in \mathcal{L}^2_{cont}(0,1)$ can be written as

$$f(x) = \sum_{n=0}^{\infty} \langle u_n, f\rangle u_n(x). \tag{3.24}$$

Moreover, for $f \in \mathfrak{D}(L)$ this series is absolutely uniformly convergent.

Proof. If E_j is an eigenvalue with corresponding normalized eigenfunction u_j we have

$$E_j = \langle u_j, Lu_j\rangle = \int_0^1 \big(|u_j'(x)|^2 + q(x)|u_j(x)|^2\big)dx < \min_{x\in[0,1]} q(x) \tag{3.25}$$

where we have used integration by parts as in (3.16). Note that equality could only occur if u_j is constant, which is incompatible with our boundary conditions. Hence the eigenvalues are bounded from below.

Now pick a value $\lambda \in \mathbb{R}$ such that $R_L(\lambda)$ exists ($\lambda < \min_{x\in[0,1]} q(x)$ say). By Lemma 3.4 $R_L(\lambda)$ is compact and by Lemma 3.3 this remains true if we replace $\mathcal{L}^2_{cont}(0,1)$ by its completion. By Theorem 3.7 there are eigenvalues α_n of $R_L(\lambda)$ with corresponding eigenfunctions u_n. Moreover, $R_L(\lambda)u_n = \alpha_n u_n$ is equivalent to $Lu_n = (\lambda + \frac{1}{\alpha_n})u_n$, which shows that $E_n = \lambda + \frac{1}{\alpha_n}$ are eigenvalues of L with corresponding eigenfunctions u_n. Now everything follows from Theorem 3.7 except that the eigenvalues are simple. To show this, observe that if u_n and v_n are two different eigenfunctions corresponding to E_n, then $u_n(0) = v_n(0) = 0$ implies $W(u_n, v_n) = 0$ and hence u_n and v_n are linearly dependent.

To show that (3.24) converges uniformly if $f \in \mathfrak{D}(L)$ we begin by writing $f = R_L(\lambda)g$, $g \in \mathcal{L}^2_{cont}(0,1)$, implying

$$\sum_{n=0}^{\infty} \langle u_n, f\rangle u_n(x) = \sum_{n=0}^{\infty} \langle R_L(\lambda)u_n, g\rangle u_n(x) = \sum_{n=0}^{\infty} \alpha_n \langle u_n, g\rangle u_n(x).$$

Moreover, the Cauchy–Schwarz inequality shows

$$\left|\sum_{j=m}^{n} |\alpha_j \langle u_j, g\rangle u_j(x)|\right|^2 \le \sum_{j=m}^{n} |\langle u_j, g\rangle|^2 \sum_{j=m}^{n} |\alpha_j u_j(x)|^2.$$

Now, by (2.18), $\sum_{j=0}^{\infty} |\langle u_j, g\rangle|^2 = \|g\|^2$ and hence the first term is part of a convergent series. Similarly, the second term can be estimated independent of x since

$$\alpha_n u_n(x) = R_L(\lambda)u_n(x) = \int_0^1 G(\lambda, x, t)u_n(t)dt = \langle u_n, G(\lambda, x, .)\rangle$$

implies

$$\sum_{j=m}^{n}|\alpha_j u_j(x)|^2 \le \sum_{j=0}^{\infty}|\langle u_j, G(\lambda,x,.)\rangle|^2 = \int_0^1 |G(\lambda,x,t)|^2 dt \le M(\lambda)^2,$$

where $M(\lambda) := \max_{x,t\in[0,1]} |G(\lambda,x,t)|$, again by (2.18). □

Moreover, it is even possible to weaken our assumptions for uniform convergence. To this end we consider the sequilinear form associated with L:

$$s_L(f,g) := \langle f, Lg\rangle = \int_0^1 \left(f'(x)^* g'(x) + q(x) f(x)^* g(x)\right) dx \tag{3.26}$$

for $f, g \in \mathfrak{D}(L)$, where we have used integration by parts as in (3.16). In fact, the above formula continues to hold for f in a slightly larger class of functions,

$$\mathfrak{Q}(L) := \{f \in C_p^1[0,1] | f(0) = f(1) = 0\} \supseteq \mathfrak{D}(L), \tag{3.27}$$

which we call the **form domain** of L. Here $C_p^1[a,b]$ denotes the set of piecewise continuously differentiable functions f in the sense that f is continuously differentiable except for a finite number of points at which it is continuous and the derivative has limits from the left and right. In fact, any class of functions for which the partial integration needed to obtain (3.26) can be justified would be good enough (e.g. the set of absolutely continuous functions to be discussed in Section 11.8).

Lemma 3.12. *For a regular Sturm–Liouville problem* (3.24) *converges absolutely uniformly provided* $f \in \mathfrak{Q}(L)$.

Proof. By replacing $L \to L - q_0$ for $q_0 < \min_{x\in[0,1]} q(x)$ we can assume $q(x) > 0$ without loss of generality. (This will shift the eigenvalues $E_n \to E_n - q_0$ and leave the eigenvectors unchanged.) In particular, we have $q_L(f) := s_L(f,f) > 0$ after this change. By (3.26) we also have $E_j = \langle u_j, Lu_j\rangle = q_L(u_j) > 0$.

Now let $f \in \mathfrak{Q}(L)$ and consider (3.24). Then, observing that $s_L(f,g)$ is a symmetric sesquilinear form (after our shift it is even a scalar product) as

well as $s_L(f,u_j) = E_j\langle f,u_j\rangle$ one obtains

$$\begin{aligned}0 \le & q_L\big(f - \sum_{j=m}^{n}\langle u_j,f\rangle u_j\big) = q_L(f) - \sum_{j=m}^{n}\langle u_j,f\rangle s_L(f,u_j)\\ & - \sum_{j=m}^{n}\langle u_j,f\rangle^* s_L(u_j,f) + \sum_{j,k=m}^{n}\langle u_j,f\rangle^*\langle u_k,f\rangle s_L(u_j,u_k)\\ = & q_L(f) - \sum_{j=m}^{n} E_j|\langle u_j,f\rangle|^2\end{aligned}$$

which implies

$$\sum_{j=m}^{n} E_j|\langle u_j,f\rangle|^2 \le q_L(f).$$

In particular, note that this estimate applies to $f(y) = G(\lambda,x,y)$. Now we can proceed as in the proof of the previous theorem (with $\lambda = 0$ and $\alpha_j = E_j^{-1}$)

$$\begin{aligned}\sum_{j=m}^{n}|\langle u_j,f\rangle u_j(x)| &= \sum_{j=m}^{n} E_j|\langle u_j,f\rangle\langle u_j,G(0,x,.)\rangle|\\ &\le \left(\sum_{j=m}^{n} E_j|\langle u_j,f\rangle|^2 \sum_{j=m}^{n} E_j|\langle u_j,G(0,x,.)\rangle|^2\right)^{1/2}\\ &\le \left(\sum_{j=m}^{n} E_j|\langle u_j,f\rangle|^2\right)^{1/2} q_L(G(0,x,.))^{1/2},\end{aligned}$$

where we have used the Cauchy–Schwarz inequality for the weighted scalar product $(f_j,g_j) \mapsto \sum_j f_j^* g_j E_j$. Finally note that $q_L(G(0,x,.))$ is continuous with respect to x and hence can be estimated by its maximum over $[0,1]$. This shows that the sum (3.24) is absolutely convergent, uniformly with respect to x. □

Another consequence of the computations in the previous proof is also worthwhile noting:

Corollary 3.13. *We have*

$$G(z,x,y) = \sum_{j=0}^{\infty}\frac{1}{E_j - z}u_j(x)u_j(y), \tag{3.28}$$

where the sum is uniformly convergent. Moreover, we have the following ***trace formula***

$$\int_0^1 G(z,x,x)dx = \sum_{j=0}^{\infty} \frac{1}{E_j - z}. \tag{3.29}$$

Proof. Using the conventions from the proof of the previous lemma we have $\langle u_j, G(0,x,.)\rangle = E_j^{-1} u_j(x)$ and since $G(0,x,.) \in \mathfrak{Q}(L)$ for fixed $x \in [a,b]$ we have

$$\sum_{j=0}^{\infty} \frac{1}{E_j} u_j(x) u_j(y) = G(0,x,y),$$

where the convergence is uniformly with respect to y (and x fixed). Moreover, for $x = y$ Dini's theorem (cf. Problem B.29) shows that the convergence is uniform with respect to $x = y$ and this also proves uniform convergence of our sum since

$$\sum_{j=0}^{n} \frac{1}{|E_j - z|} |u_j(x)u_j(y)| \le C(z) \left(\sum_{j=0}^{n} \frac{1}{E_j} u_j(x)^2 \right)^{1/2} \left(\sum_{j=0}^{n} \frac{1}{E_j} u_j(y)^2 \right)^{1/2},$$

where $C(z) := \sup_j \frac{E_j}{|E_j - z|}$.

Finally, the last claim follows upon computing the integral using (3.28) and observing $\|u_j\| = 1$. □

Example 3.8. Let us look at the Sturm–Liouville problem with $q = 0$. Then the underlying differential equation is

$$-u''(x) = z\, u(x)$$

whose solution is given by $u(x) = c_1 \sin(\sqrt{z}x) + c_2 \cos(\sqrt{z}x)$. The solution satisfying the boundary condition at the left endpoint is $u_-(z,x) = \sin(\sqrt{z}x)$ and it will be an eigenfunction if and only if $u_-(z,1) = \sin(\sqrt{z}) = 0$. Hence the corresponding eigenvalues and normalized eigenfunctions are

$$E_n = \pi^2 n^2, \quad u_n(x) = \sqrt{2}\sin(n\pi x), \qquad n \in \mathbb{N}.$$

Moreover, every function $f \in \mathcal{L}^2_{cont}(0,1)$ can be expanded into a **Fourier sine series**

$$f(x) = \sum_{n=1}^{\infty} f_n u_n(x), \qquad f_n := \int_0^1 u_n(x) f(x) dx,$$

which is convergent with respect to our scalar product. If $f \in C^1_p[0,1]$ with $f(0) = f(1) = 0$ the series will converge uniformly. For an application of the trace formula see Problem 3.10. ◇

Example 3.9. We could also look at the same equation as in the previous problem but with different boundary conditions

$$u'(0) = u'(1) = 0.$$

Then

$$E_n = \pi^2 n^2, \quad u_n(x) = \begin{cases} 1, & n = 0, \\ \sqrt{2}\cos(n\pi x), & n \in \mathbb{N}. \end{cases}$$

Moreover, every function $f \in \mathcal{L}^2_{cont}(0,1)$ can be expanded into a **Fourier cosine series**

$$f(x) = \sum_{n=1}^{\infty} f_n u_n(x), \qquad f_n := \int_0^1 u_n(x) f(x) dx,$$

which is convergent with respect to our scalar product. ⋄

Example 3.10. Combining the last two examples we see that every symmetric function on $[-1,1]$ can be expanded into a Fourier cosine series and every anti-symmetric function into a Fourier sine series. Moreover, since every function $f(x)$ can be written as the sum of a symmetric function $\frac{f(x)+f(-x)}{2}$ and an anti-symmetric function $\frac{f(x)-f(-x)}{2}$, it can be expanded into a Fourier series. Hence we recover Theorem 2.18. ⋄

Problem* 3.7. *Show that for our Sturm–Liouville operator $u_\pm(z,x)^* = u_\pm(z^*,x)$. Conclude $R_L(z)^* = R_L(z^*)$. (Hint: Problem 3.2.)*

Problem 3.8. *Show that the resolvent $R_A(z) = (A-z)^{-1}$ (provided it exists and is densely defined) of a symmetric operator A is again symmetric for $z \in \mathbb{R}$. (Hint: $g \in \mathfrak{D}(R_A(z))$ if and only if $g = (A-z)f$ for some $f \in \mathfrak{D}(A)$.)*

Problem 3.9. *Suppose $E_0 > 0$ and equip $\mathfrak{Q}(L)$ with the scalar product s_L. Show that*

$$f(x) = s_L(G(0,x,.), f).$$

In other words, point evaluations are continuous functionals associated with the vectors $G(0,x,.) \in \mathfrak{Q}(L)$. In this context, $G(0,x,y)$ is called a ***reproducing kernel****.*

Problem 3.10. *Show that*

$$\sum_{n=1}^{\infty} \frac{1}{n^2 - z} = \frac{1 - \pi\sqrt{z}\cot(\pi\sqrt{z})}{2z}, \qquad z \in \mathbb{C} \setminus \mathbb{N}.$$

In particular, for $z = 0$ this gives Euler's solution of the **Basel problem***:*

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

In fact, comparing the power series of both sides at $z = 0$ gives

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = \frac{(-1)^{k+1}(2\pi)^{2k} B_{2k}}{2(2k)!}, \qquad k \in \mathbb{N},$$

where B_k are the ***Bernoulli numbers*** *defined via $\frac{x}{\mathrm{e}^x - 1} = \sum_{k=0}^{\infty} \frac{B_k}{k!} z^k$. (Hint: Use the trace formula* (3.29)*.)*

Problem 3.11. *Consider the Sturm–Liouville problem on a compact interval $[a, b]$ with domain*

$$\mathfrak{D}(L) = \{f \in C^2[a, b] | f'(a) - \alpha f(a) = f'(b) - \beta f(b) = 0\}$$

for some real constants $\alpha, \beta \in \mathbb{R}$. Show that Theorem 3.11 continues to hold except for the lower bound on the eigenvalues.

3.4. Estimating eigenvalues

In general, there is no way of computing eigenvalues and their corresponding eigenfunctions explicitly. Hence it is important to be able to determine the eigenvalues at least approximately.

Let A be a symmetric operator which has a lowest eigenvalue α_1 (e.g., A is a Sturm–Liouville operator). Suppose we have a vector f which is an approximation for the eigenvector u_1 of this lowest eigenvalue α_1. Moreover, suppose we can write

$$A := \sum_{j=1}^{\infty} \alpha_j \langle u_j, . \rangle u_j, \qquad \mathfrak{D}(A) := \{f \in \mathfrak{H} | \sum_{j=1}^{\infty} |\alpha_j \langle u_j, f \rangle|^2 < \infty\}, \quad (3.30)$$

where $\{u_j\}_{j \in \mathbb{N}}$ is an orthonormal basis of eigenvectors. Since α_1 is supposed to be the lowest eigenvalue we have $\alpha_j \geq \alpha_1$ for all $j \in \mathbb{N}$.

Writing $f = \sum_j \gamma_j u_j$, $\gamma_j = \langle u_j, f \rangle$, one computes

$$\langle f, Af \rangle = \langle f, \sum_{j=1}^{\infty} \alpha_j \gamma_j u_j \rangle = \sum_{j=1}^{\infty} \alpha_j |\gamma_j|^2, \qquad f \in \mathfrak{D}(A), \quad (3.31)$$

and we clearly have

$$\alpha_1 \leq \frac{\langle f, Af \rangle}{\|f\|^2}, \qquad f \in \mathfrak{D}(A), \quad (3.32)$$

with equality for $f = u_1$. In particular, any f will provide an upper bound and if we add some free parameters to f, one can optimize them and obtain quite good upper bounds for the first eigenvalue. For example we could take some orthogonal basis, take a finite number of coefficients and optimize them. This is known as the **Rayleigh–Ritz method**.

Example 3.11. Consider the Sturm–Liouville operator L with potential $q(x) = x$ and Dirichlet boundary conditions $f(0) = f(1) = 0$ on the interval $[0, 1]$. Our starting point is the quadratic form

$$q_L(f) := \langle f, Lf\rangle = \int_0^1 \big(|f'(x)|^2 + q(x)|f(x)|^2\big)dx$$

which gives us the lower bound

$$\langle f, Lf\rangle \geq \min_{0\leq x\leq 1} q(x) = 0.$$

While the corresponding differential equation can in principle be solved in terms of Airy functions, there is no closed form for the eigenvalues.

First of all we can improve the above bound upon observing $0 \leq q(x) \leq 1$ which implies

$$\langle f, L_0 f\rangle \leq \langle f, Lf\rangle \leq \langle f, (L_0 + 1)f\rangle, \qquad f \in \mathfrak{D}(L) = \mathfrak{D}(L_0),$$

where L_0 is the Sturm–Liouville operator corresponding to $q(x) = 0$. Since the lowest eigenvalue of L_0 is π^2 we obtain

$$\pi^2 \leq E_1 \leq \pi^2 + 1$$

for the lowest eigenvalue E_1 of L.

Moreover, using the lowest eigenfunction $f_1(x) = \sqrt{2}\sin(\pi x)$ of L_0 one obtains the improved upper bound

$$E_1 \leq \langle f_1, Lf_1\rangle = \pi^2 + \frac{1}{2} \approx 10.3696.$$

Taking the second eigenfunction $f_2(x) = \sqrt{2}\sin(2\pi x)$ of L_0 we can make the ansatz $f(x) = (1 + \gamma^2)^{-1/2}(f_1(x) + \gamma f_2(x))$ which gives

$$\langle f, Lf\rangle = \pi^2 + \frac{1}{2} + \frac{\gamma}{1+\gamma^2}\big(3\pi^2\gamma - \frac{32}{9\pi^2}\big).$$

The right-hand side has a unique minimum at $\gamma = \frac{32}{27\pi^4+\sqrt{1024+729\pi^8}}$ giving the bound

$$E_1 \leq \frac{5}{2}\pi^2 + \frac{1}{2} - \frac{\sqrt{1024 + 729\pi^8}}{18\pi^2} \approx 10.3685$$

which coincides with the exact eigenvalue up to five digits. ⋄

But is there also something one can say about the next eigenvalues? Suppose we know the first eigenfunction u_1. Then we can restrict A to the orthogonal complement of u_1 and proceed as before: E_2 will be the minimum of $\langle f, Af\rangle$ over all f restricted to this subspace. If we restrict to the orthogonal complement of an approximating eigenfunction f_1, there will still be a component in the direction of u_1 left and hence the infimum of the expectations will be lower than E_2. Thus the optimal choice $f_1 = u_1$ will give the maximal value E_2.

Theorem 3.14 (Max-min)**.** *Let A be a symetric operator and let $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_N$ be eigenvalues of A with corresponding orthonormal eigenvectors $u_1, u_2, \ldots, u_N$. Suppose*

$$A = \sum_{j=1}^{N} \alpha_j \langle u_j, . \rangle u_j + \tilde{A} \tag{3.33}$$

with $\langle f, \tilde{A} f\rangle \ge \alpha_N \|f\|^2$ for all $f \in \mathfrak{D}(A)$ and $u_1, \ldots, u_N \in \operatorname{Ker}(\tilde{A})$. Then

$$\alpha_j = \sup_{f_1,\ldots,f_{j-1}} \inf_{f \in U(f_1,\ldots,f_{j-1})} \langle f, Af\rangle, \qquad 1 \le j \le N, \tag{3.34}$$

where

$$U(f_1, \ldots, f_j) := \{f \in \mathfrak{D}(A) | \, \|f\| = 1, \; f \in \operatorname{span}\{f_1, \ldots, f_j\}^{\perp}\}. \tag{3.35}$$

Proof. We have

$$\inf_{f \in U(f_1,\ldots,f_{j-1})} \langle f, Af\rangle \le \alpha_j.$$

In fact, set $f = \sum_{k=1}^{j} \gamma_k u_k$ and choose γ_k such that $f \in U(f_1, \ldots, f_{j-1})$. Then

$$\langle f, Af\rangle = \sum_{k=1}^{j} |\gamma_k|^2 \alpha_k \le \alpha_j$$

and the claim follows.

Conversely, let $\gamma_k = \langle u_k, f\rangle$ and write $f = \sum_{k=1}^{j} \gamma_k u_k + \tilde{f}$. Then

$$\inf_{f \in U(u_1,\ldots,u_{j-1})} \langle f, Af\rangle = \inf_{f \in U(u_1,\ldots,u_{j-1})} \left(\sum_{k=j}^{N} |\gamma_k|^2 \alpha_k + \langle \tilde{f}, \tilde{A}\tilde{f}\rangle \right) = \alpha_j. \quad \square$$

Of course if we are interested in the largest eigenvalues all we have to do is consider $-A$.

Note that this immediately gives an estimate for eigenvalues if we have a corresponding estimate for the operators. To this end we will write

$$A \le B \qquad \Leftrightarrow \qquad \langle f, Af\rangle \le \langle f, Bf\rangle, \quad f \in \mathfrak{D}(A) \cap \mathfrak{D}(B). \tag{3.36}$$

Corollary 3.15. *Suppose A and B are symmetric operators with corresponding eigenvalues α_j and β_j as in the previous theorem. If $A \le B$ and $\mathfrak{D}(B) \subseteq \mathfrak{D}(A)$ then $\alpha_j \le \beta_j$.*

Proof. By assumption we have $\langle f, Af\rangle \le \langle f, Bf\rangle$ for $f \in \mathfrak{D}(B)$ implying

$$\inf_{f \in U_A(f_1,\ldots,f_{j-1})} \langle f, Af\rangle \le \inf_{f \in U_B(f_1,\ldots,f_{j-1})} \langle f, Af\rangle \le \inf_{f \in U_B(f_1,\ldots,f_{j-1})} \langle f, Bf\rangle,$$

where we have indicated the dependence of U on the operator via a subscript. Taking the sup on both sides the claim follows. $\square$

Example 3.12. Let L be again our Sturm–Liouville operator and L_0 the corresponding operator with $q(x) = 0$. Set $q_- = \min_{0\le x\le 1} q(x)$ and $q_+ = \max_{0\le x\le 1} q(x)$. Then $L_0 + q_- \le L \le L_0 + q_+$ implies

$$\pi^2 n^2 + q_- \le E_n \le \pi^2 n^2 + q_+.$$

In particular, we have proven the famous **Weyl asymptotic**

$$E_n = \pi^2 n^2 + O(1)$$

for the eigenvalues. ⋄

There is also an alternative version which can be proven similar (Problem 3.12):

Theorem 3.16 (Min-max)**.** *Let A be as in the previous theorem. Then*

$$\alpha_j = \inf_{V_j\subset\mathfrak{D}(A),\dim(V_j)=j} \sup_{f\in V_j,\|f\|=1} \langle f, Af\rangle, \tag{3.37}$$

where the inf is taken over subspaces with the indicated properties.

Problem* 3.12. *Prove Theorem 3.16.*

Problem 3.13. *Suppose A, A_n are self-adjoint, bounded and $A_n \to A$. Then $\alpha_k(A_n) \to \alpha_k(A)$. (Hint: For B self-adjoint $\|B\| \le \varepsilon$ is equivalent to $-\varepsilon \le B \le \varepsilon$.)*

3.5. Singular value decomposition of compact operators

Our first aim is to find a generalization of Corollary 3.8 for general compact operators between Hilbert spaces. The key observation is that if $K \in \mathscr{C}(\mathfrak{H}_1, \mathfrak{H}_2)$ is compact, then $K^*K \in \mathscr{C}(\mathfrak{H}_1)$ is compact and symmetric and thus, by Corollary 3.8, there is a countable orthonormal set $\{u_j\} \subset \mathfrak{H}_1$ and nonzero real numbers $s_j^2 \neq 0$ such that

$$K^*Kf = \sum_j s_j^2 \langle u_j, f\rangle u_j. \tag{3.38}$$

Moreover, $\|Ku_j\|^2 = \langle u_j, K^*Ku_j\rangle = \langle u_j, s_j^2 u_j\rangle = s_j^2$ shows that we can set

$$s_j := \|Ku_j\| > 0. \tag{3.39}$$

The numbers $s_j = s_j(K)$ are called **singular values** of K. There are either finitely many singular values or they converge to zero.

Theorem 3.17 (Singular value decomposition of compact operators)**.** *Let $K \in \mathscr{C}(\mathfrak{H}_1, \mathfrak{H}_2)$ be compact and let s_j be the singular values of K and $\{u_j\} \subset \mathfrak{H}_1$ corresponding orthonormal eigenvectors of K^*K. Then*

$$K = \sum_j s_j \langle u_j, .\rangle v_j, \tag{3.40}$$

where $v_j = s_j^{-1} K u_j$. The norm of K is given by the largest singular value

$$\|K\| = \max_j s_j(K). \tag{3.41}$$

Moreover, the vectors $\{v_j\} \subset \mathfrak{H}_2$ are again orthonormal and satisfy $K^* v_j = s_j u_j$. In particular, v_j are eigenvectors of KK^* corresponding to the eigenvalues s_j^2.

Proof. For any $f \in \mathfrak{H}_1$ we can write

$$f = \sum_j \langle u_j, f\rangle u_j + f_\perp$$

with $f_\perp \in \operatorname{Ker}(K^*K) = \operatorname{Ker}(K)$ (Problem 3.14). Then

$$Kf = \sum_j \langle u_j, f\rangle K u_j = \sum_j s_j \langle u_j, f\rangle v_j$$

as required. Furthermore,

$$\langle v_j, v_k\rangle = (s_j s_k)^{-1} \langle Ku_j, Ku_k\rangle = (s_j s_k)^{-1} \langle K^*Ku_j, u_k\rangle = s_j s_k^{-1} \langle u_j, u_k\rangle$$

shows that $\{v_j\}$ are orthonormal. By definition $K^* v_j = s_j^{-1} K^* K u_j = s_j u_j$ which also shows $KK^* v_j = s_j K u_j = s_j^2 v_j$.

Finally, (3.41) follows using Bessel's inequality

$$\|Kf\|^2 = \|\sum_j s_j \langle u_j, f\rangle v_j\|^2 = \sum_j s_j^2 |\langle u_j, f\rangle|^2 \le \Big(\max_j s_j(K)^2\Big) \|f\|^2,$$

where equality holds for $f = u_{j_0}$ if $s_{j_0} = \max_j s_j(K)$. □

If $K \in \mathscr{C}(\mathfrak{H})$ is self-adjoint, then $u_j = \sigma_j v_j$, $\sigma_j^2 = 1$, are the eigenvectors of K and $\sigma_j s_j$ are the corresponding eigenvalues. In particular, for a self-adjoint operators the singular values are the absolute values of the nonzero eigenvalues.

The above theorem also gives rise to the **polar decomposition**

$$K = U|K| = |K^*|U, \tag{3.42}$$

where

$$|K| := \sqrt{K^*K} = \sum_j s_j \langle u_j, .\rangle u_j, \quad |K^*| = \sqrt{KK^*} = \sum_j s_j \langle v_j, .\rangle v_j \tag{3.43}$$

are self-adjoint (in fact nonnegative) and

$$U := \sum_j \langle u_j, .\rangle v_j \tag{3.44}$$

is an isometry from $\overline{\operatorname{Ran}(K^*)} = \overline{\operatorname{span}\{u_j\}}$ onto $\overline{\operatorname{Ran}(K)} = \overline{\operatorname{span}\{v_j\}}$.

From the max-min theorem (Theorem 3.14) we obtain:

Lemma 3.18. *Let $K \in \mathscr{C}(\mathfrak{H}_1, \mathfrak{H}_2)$ be compact; then*

$$s_j(K) = \min_{f_1,\dots,f_{j-1}} \sup_{f \in U(f_1,\dots,f_{j-1})} \|Kf\|, \tag{3.45}$$

where $U(f_1, \dots, f_j) := \{f \in \mathfrak{H}_1 | \, \|f\| = 1, \ f \in \operatorname{span}\{f_1, \dots, f_j\}^\perp\}$.

In particular, note

$$s_j(AK) \le \|A\| s_j(K), \qquad s_j(KA) \le \|A\| s_j(K) \tag{3.46}$$

whenever K is compact and A is bounded (the second estimate follows from the first by taking adjoints).

An operator $K \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is called a **finite rank operator** if its range is finite dimensional. The dimension

$$\operatorname{rank}(K) := \dim \operatorname{Ran}(K)$$

is called the **rank** of K. Since for a compact operator

$$\overline{\operatorname{Ran}(K)} = \overline{\operatorname{span}\{v_j\}} \tag{3.47}$$

we see that a compact operator is finite rank if and only if the sum in (3.40) is finite. Note that the finite rank operators form an ideal in $\mathscr{L}(\mathfrak{H})$ just as the compact operators do. Moreover, every finite rank operator is compact by the Heine–Borel theorem (Theorem B.22).

Now truncating the sum in the canonical form gives us a simple way to approximate compact operators by finite rank ones. Moreover, this is in fact the best approximation within the class of finite rank operators:

Lemma 3.19. *Let $K \in \mathscr{C}(\mathfrak{H}_1, \mathfrak{H}_2)$ be compact and let its singular values be ordered. Then*

$$s_j(K) = \min_{\operatorname{rank}(F)<j} \|K - F\|, \tag{3.48}$$

where the minimum is attained for

$$F_{j-1} := \sum_{k=1}^{j-1} s_k \langle u_k, . \rangle v_k. \tag{3.49}$$

In particular, the closure of the ideal of finite rank operators in $\mathscr{L}(\mathfrak{H})$ is the ideal of compact operators.

Proof. That there is equality for $F = F_{j-1}$ follows from (3.41). In general, the restriction of F to $\operatorname{span}\{u_1, \dots, u_j\}$ will have a nontrivial kernel. Let $f = \sum_{k=1}^{j} \alpha_j u_j$ be a normalized element of this kernel, then $\|(K-F)f\|^2 = \|Kf\|^2 = \sum_{k=1}^{j} |\alpha_k s_k|^2 \ge s_j^2$.

In particular, every compact operator can be approximated by finite rank ones and since the limit of compact operators is compact, we cannot get more than the compact operators. □

Two more consequences are worthwhile noting.

Corollary 3.20. *An operator $K \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is compact if and only if K^*K is.*

Proof. Just observe that K^*K compact is all that was used to show Theorem 3.17. □

Corollary 3.21. *An operator $K \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is compact (finite rank) if and only $K^* \in \mathscr{L}(\mathfrak{H}_2, \mathfrak{H}_1)$ is. In fact, $s_j(K) = s_j(K^*)$ and*

$$K^* = \sum_j s_j \langle v_j, .\rangle u_j. \tag{3.50}$$

Proof. First of all note that (3.50) follows from (3.40) since taking adjoints is continuous and $(\langle u_j, .\rangle v_j)^* = \langle v_j, .\rangle u_j$ (cf. Problem 2.7). The rest is straightforward. □

From this last lemma one easily gets a number of useful inequalities for the singular values:

Corollary 3.22. *Let K_1 and K_2 be compact and let $s_j(K_1)$ and $s_j(K_2)$ be ordered. Then*

(i) $s_{j+k-1}(K_1 + K_2) \leq s_j(K_1) + s_k(K_2)$,
(ii) $s_{j+k-1}(K_1 K_2) \leq s_j(K_1) s_k(K_2)$,
(iii) $|s_j(K_1) - s_j(K_2)| \leq \|K_1 - K_2\|$.

Proof. Let F_1 be of rank $j-1$ and F_2 of rank $k-1$ such that $\|K_1 - F_1\| = s_j(K_1)$ and $\|K_2 - F_2\| = s_k(K_2)$. Then $s_{j+k-1}(K_1 + K_2) \leq \|(K_1 + K_2) - (F_1 + F_2)\| = \|K_1 - F_1\| + \|K_2 - F_2\| = s_j(K_1) + s_k(K_2)$ since $F_1 + F_2$ is of rank at most $j + k - 2$.

Similarly $F = F_1(K_2 - F_2) + K_1 F_2$ is of rank at most $j + k - 2$ and hence $s_{j+k-1}(K_1 K_2) \leq \|K_1 K_2 - F\| = \|(K_1 - F_1)(K_2 - F_2)\| \leq \|K_1 - F_1\| \|K_2 - F_2\| = s_j(K_1) s_k(K_2)$.

Next, choosing $k = 1$ and replacing $K_2 \to K_2 - K_1$ in (i) gives $s_j(K_2) \leq s_j(K_1) + \|K_2 - K_1\|$. Reversing the roles gives $s_j(K_1) \leq s_j(K_2) + \|K_1 - K_2\|$ and proves (iii). □

Example 3.13. On might hope that item (i) from the previous corollary can be improved to $s_j(K_1 + K_2) \leq s_j(K_1) + s_j(K_2)$. However, this is not the case as the following example shows:

$$K_1 := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad K_2 := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then $1 = s_2(K_1 + K_2) \not\leq s_2(K_1) + s_2(K_2) = 0$. ◇

Problem* 3.14. *Show that* $\operatorname{Ker}(A^*A) = \operatorname{Ker}(A)$ *for any* $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$.

Problem 3.15. *Let* K *be multiplication by a sequence* $k \in c_0(\mathbb{N})$ *in the Hilbert space* $\ell^2(\mathbb{N})$. *What are the singular values of* K*?*

Problem 3.16. *Let* K *be multiplication by a sequence* $k \in c_0(\mathbb{N})$ *in the Hilbert space* $\ell^2(\mathbb{N})$ *and consider* $L = KS^-$. *What are the singular values of* L*? Does* L *have any eigenvalues?*

Problem 3.17. *Let* $K \in \mathscr{C}(\mathfrak{H}_1, \mathfrak{H}_2)$ *be compact and let its singular values be ordered. Let* $M \subseteq \mathfrak{H}_1$, $N \subseteq \mathfrak{H}_1$ *be subspaces whith corresponding orthogonal projections* P_M, P_N, *respectively. Then*

$$s_j(K) = \min_{\dim(M)<j} \|K - KP_M\| = \min_{\dim(N)<j} \|K - P_N K\|,$$

where the minimum is taken over all subspaces with the indicated dimension. Moreover, the minimum is attained for

$$M = \operatorname{span}\{u_k\}_{k=1}^{j-1}, \qquad N = \operatorname{span}\{v_k\}_{k=1}^{j-1}.$$

3.6. Hilbert–Schmidt and trace class operators

We can further subdivide the class of compact operators $\mathscr{C}(\mathfrak{H})$ according to the decay of their singular values. We define

$$\|K\|_p := \Big(\sum_j s_j(K)^p\Big)^{1/p} \tag{3.51}$$

plus corresponding spaces

$$\mathscr{J}_p(\mathfrak{H}) = \{K \in \mathscr{C}(\mathfrak{H}) | \|K\|_p < \infty\}, \tag{3.52}$$

which are known as **Schatten p-classes**. Even though our notation hints at the fact that $\|.\|_p$ is a norm, we will only prove this here for $p = 1, 2$ (the only nontrivial part is the triangle inequality). Note that by (3.41)

$$\|K\| \le \|K\|_p \tag{3.53}$$

and that by $s_j(K) = s_j(K^*)$ we have

$$\|K\|_p = \|K^*\|_p. \tag{3.54}$$

The two most important cases are $p = 1$ and $p = 2$: $\mathscr{J}_2(\mathfrak{H})$ is the space of **Hilbert–Schmidt operators** and $\mathscr{J}_1(\mathfrak{H})$ is the space of **trace class** operators.

Example 3.14. Any multiplication operator by a sequence from $\ell^p(\mathbb{N})$ is in the Schatten p-class of $\mathfrak{H} = \ell^2(\mathbb{N})$. ⋄

Example 3.15. By virtue of the Weyl asymptotics (see Example 3.12) the resolvent of our Sturm–Liouville operator is trace class. ⋄

Example 3.16. Let k be a periodic function which is square integrable over $[-\pi, \pi]$. Then the integral operator

$$(Kf)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} k(y-x) f(y) dy$$

has the eigenfunctions $u_j(x) = (2\pi)^{-1/2} \mathrm{e}^{-\mathrm{i}jx}$ with corresponding eigenvalues $\hat{k}_j$, $j \in \mathbb{Z}$, where $\hat{k}_j$ are the Fourier coefficients of k. Since $\{u_j\}_{j\in\mathbb{Z}}$ is an ONB we have found all eigenvalues. In particular, the Fourier transform maps K to the multiplication operator with the sequence of its eigenvalues $\hat{k}_j$. Hence the singular values are the absolute values of the nonzero eigenvalues and (3.40) reads

$$K = \sum_{j\in\mathbb{Z}} \hat{k}_j \langle u_j, .\rangle u_j.$$

Moreover, since the eigenvalues are in $\ell^2(\mathbb{Z})$ we see that K is a Hilbert–Schmidt operator. If k is continuous with summable Fourier coefficients (e.g. $k \in C^2_{per}[-\pi, \pi]$), then K is trace class. ⋄

We first prove an alternate definition for the Hilbert–Schmidt norm.

Lemma 3.23. *A bounded operator K is Hilbert–Schmidt if and only if*

$$\sum_{j\in J} \|Kw_j\|^2 < \infty \tag{3.55}$$

for some orthonormal basis and

$$\|K\|_2 = \Big(\sum_{j\in J} \|Kw_j\|^2 \Big)^{1/2}, \tag{3.56}$$

for every orthonormal basis in this case.

Proof. First of all note that (3.55) implies that K is compact. To see this, let P_n be the projection onto the space spanned by the first n elements of the orthonormal basis $\{w_j\}$. Then $K_n = KP_n$ is finite rank and converges to K since

$$\|(K - K_n)f\| = \|\sum_{j>n} c_j K w_j\| \le \sum_{j>n} |c_j| \|Kw_j\| \le \Big(\sum_{j>n} \|Kw_j\|^2 \Big)^{1/2} \|f\|,$$

where $f = \sum_j c_j w_j$.

The rest follows from (3.40) and

$$\begin{aligned} \sum_j \|Kw_j\|^2 &= \sum_{k,j} |\langle v_k, Kw_j\rangle|^2 = \sum_{k,j} |\langle K^* v_k, w_j\rangle|^2 = \sum_k \|K^* v_k\|^2 \\ &= \sum_k s_k(K)^2 = \|K\|_2^2. \end{aligned}$$

Here we have used $\overline{\operatorname{span}\{v_k\}} = \operatorname{Ker}(K^*)^\perp = \overline{\operatorname{Ran}(K)}$ in the first step. □

Corollary 3.24. *The Hilbert–Schmidt norm satisfies the triangle inequality and hence is indeed a norm.*

Proof. This follows from (3.56) upon using the triangle inequality for $\mathfrak{H}$ and for $\ell^2(J)$. □

Now we can show

Lemma 3.25. *The set of Hilbert–Schmidt operators forms an ideal in $\mathscr{L}(\mathfrak{H})$ and*

$$\|KA\|_2 \leq \|A\|\|K\|_2, \quad \textit{respectively,} \quad \|AK\|_2 \leq \|A\|\|K\|_2. \tag{3.57}$$

Proof. If K_1 and K_2 are Hilbert–Schmidt operators, then so is their sum since

$$\begin{aligned}\|K_1 + K_2\|_2 &= \Big(\sum_{j\in J} \|(K_1 + K_2)w_j\|^2\Big)^{1/2} \leq \Big(\sum_{j\in J} (\|K_1 w_j\| + \|K_2 w_j\|)^2\Big)^{1/2} \\ &\leq \|K_1\|_2 + \|K_2\|_2,\end{aligned}$$

where we have used the triangle inequality for $\ell^2(J)$.

Let K be Hilbert–Schmidt and A bounded. Then AK is compact and

$$\|AK\|_2^2 = \sum_j \|AKw_j\|^2 \leq \|A\|^2 \sum_j \|Kw_j\|^2 = \|A\|^2\|K\|_2^2.$$

For KA just consider adjoints. □

Example 3.17. Consider $\ell^2(\mathbb{N})$ and let K be some compact operator. Let $K_{jk} = \langle \delta^j, K\delta^k\rangle = (K\delta^j)_k$ be its matrix elements such that

$$(Ka)_j = \sum_{k=1}^{\infty} K_{jk} a_k.$$

Then, choosing $w_j = \delta^j$ in (3.56) we get

$$\|K\|_2 = \Big(\sum_{j=1}^{\infty} \|K\delta^j\|^2\Big)^{1/2} = \Big(\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} |K_{jk}|^2\Big)^{1/2}.$$

Hence K is Hilbert–Schmidt if and only if its matrix elements are in $\ell^2(\mathbb{N}\times\mathbb{N})$ and the Hilbert–Schmidt norm coincides with the $\ell^2(\mathbb{N}\times\mathbb{N})$ norm of the matrix elements. Especially in the finite dimensional case the Hilbert–Schmidt norm is also known as **Frobenius norm**.

Of course the same calculation shows that a bounded operator is Hilbert–Schmidt if and only if its matrix elements $\langle w_j, Kw_k\rangle$ with respect to some orthonormal basis $\{w_j\}_{j\in J}$ are in $\ell^2(J\times J)$ and the Hilbert–Schmidt norm coincides with the $\ell^2(J\times J)$ norm of the matrix elements. ⋄

Example 3.18. Let $I = [a, b]$ be a compact interval. Suppose $K : L^2(I) \to C(I)$ is continuous, then $K : L^2(I) \to L^2(I)$ is Hilbert–Schmidt with Hilbert–Schmidt norm $\|K\|_2 \leq \sqrt{b-a}M$, where $M := \|K\|_{L^2(I)\to C(I)}$.

To see this start by observing that point evaluations are continuous functionals on $C(I)$ and hence $f \mapsto (Kf)(x)$ is a continuous linear functional on $L^2(I)$ satisfying $|(Kf)(x)| \leq M\|f\|$. By the Riesz lemma there is some $K_x \in L^2(I)$ with $\|K_x\| \leq M$ such that

$$(Kf)(x) = \langle K_x, f\rangle$$

and hence for any orthonormal basis $\{w_j\}_{j\in\mathbb{N}}$ we have

$$\sum_{j\in\mathbb{N}} |(Kw_j)(x)|^2 = \sum_{j\in\mathbb{N}} |\langle K_x, w_j\rangle|^2 = \|K_x\|^2 \leq M^2.$$

But then

$$\begin{aligned}\sum_{j\in\mathbb{N}} \|Kw_j\|^2 = \sum_{j\in\mathbb{N}} \int_a^b |(Kw_j)(x)|^2 dx &= \int_a^b \Big(\sum_{j\in\mathbb{N}} |(Kw_j)(x)|^2\Big) dx \\ &\leq (b-a)M^2\end{aligned}$$

as claimed. ⋄

Since Hilbert–Schmidt operators turn out easy to identify (cf. also Section 10.5), it is important to relate $\mathcal{J}_1(\mathfrak{H})$ with $\mathcal{J}_2(\mathfrak{H})$:

Lemma 3.26. *An operator is trace class if and only if it can be written as the product of two Hilbert–Schmidt operators, $K = K_1K_2$, and in this case we have*

$$\|K\|_1 \leq \|K_1\|_2\|K_2\|_2. \tag{3.58}$$

In fact, K_1, K_2 can be chosen such that $\|K\|_1 = \|K_1\|_2\|K_2\|_2$.

Proof. Using (3.40) (where we can extend u_n and v_n to orthonormal bases if necessary) and Cauchy–Schwarz we have

$$\begin{aligned}\|K\|_1 = \sum_n \langle v_n, Ku_n\rangle = \sum_n |\langle K_1^* v_n, K_2 u_n\rangle| \\ \leq \Big(\sum_n \|K_1^* v_n\|^2 \sum_n \|K_2 u_n\|^2\Big)^{1/2} = \|K_1\|_2\|K_2\|_2\end{aligned}$$

and hence $K = K_1K_2$ is trace class if both K_1 and K_2 are Hilbert–Schmidt operators. To see the converse, let K be given by (3.40) and choose $K_1 = \sum_j \sqrt{s_j(K)}\langle u_j, .\rangle v_j$, respectively, $K_2 = \sum_j \sqrt{s_j(K)}\langle u_j, .\rangle u_j$. Note that in this case $\|K\|_1 = \|K_1\|_2^2 = \|K_2\|_2^2$. □

Now we can also explain the name trace class:

Lemma 3.27. *If K is trace class, then for every orthonormal basis $\{w_n\}$ the* ***trace***

$$\operatorname{tr}(K) = \sum_n \langle w_n, K w_n \rangle \tag{3.59}$$

is finite,

$$|\operatorname{tr}(K)| \le \|K\|_1, \tag{3.60}$$

and independent of the orthonormal basis.

Proof. If we write $K = K_1 K_2$ with K_1, K_2 Hilbert–Schmidt such that $\|K\|_1 = \|K_1\|_2 \|K_2\|_2$, then the Cauchy–Schwarz inequality implies $|\operatorname{tr}(K)| \le \|K_1^*\|_2 \|K_2\|_2 = \|K\|_1$. Moreover, if $\{\tilde w_n\}$ is another orthonormal basis, we have

$$\begin{aligned}\sum_n \langle w_n, K_1 K_2 w_n\rangle &= \sum_n \langle K_1^* w_n, K_2 w_n\rangle = \sum_{n,m} \langle K_1^* w_n, \tilde w_m\rangle\langle \tilde w_m, K_2 w_n\rangle \\ &= \sum_{m,n} \langle K_2^* \tilde w_m, w_n\rangle\langle w_n, K_1 \tilde w_m\rangle = \sum_m \langle K_2^* \tilde w_m, K_1 \tilde w_m\rangle \\ &= \sum_m \langle \tilde w_m, K_2 K_1 \tilde w_m\rangle.\end{aligned}$$

In the special case $w = \tilde w$ we see $\operatorname{tr}(K_1 K_2) = \operatorname{tr}(K_2 K_1)$ and the general case now shows that the trace is independent of the orthonormal basis. □

Clearly for self-adjoint trace class operators, the trace is the sum over all eigenvalues (counted with their multiplicity). To see this, one just has to choose the orthonormal basis to consist of eigenfunctions. This is even true for all trace class operators and is known as Lidskij trace theorem (see [**34**] for an easy to read introduction).

Example 3.19. We already mentioned that the resolvent of our Sturm–Liouville operator is trace class. Choosing a basis of eigenfunctions we see that the trace of the resolvent is the sum over its eigenvalues and combining this with our trace formula (3.29) gives

$$\operatorname{tr}(R_L(z)) = \sum_{j=0}^{\infty} \frac{1}{E_j - z} = \int_0^1 G(z, x, x)\, dx$$

for $z \in \mathbb{C}$ no eigenvalue. ◇

Example 3.20. For our integral operator K from Example 3.16 we have in the trace class case

$$\operatorname{tr}(K) = \sum_{j\in\mathbb{Z}} \hat k_j = k(0).$$

Note that this can again be interpreted as the integral over the diagonal $(2\pi)^{-1} k(x - x) = (2\pi)^{-1} k(0)$ of the kernel. ◇

We also note the following elementary properties of the trace:

Lemma 3.28. *Suppose K, K_1, K_2 are trace class and A is bounded.*

(i) *The trace is linear.*

(ii) $\operatorname{tr}(K^*) = \operatorname{tr}(K)^*$.

(iii) *If $K_1 \le K_2$, then $\operatorname{tr}(K_1) \le \operatorname{tr}(K_2)$.*

(iv) $\operatorname{tr}(AK) = \operatorname{tr}(KA)$.

Proof. (i) and (ii) are straightforward. (iii) follows from $K_1 \le K_2$ if and only if $\langle f, K_1 f\rangle \le \langle f, K_2 f\rangle$ for every $f \in \mathfrak{H}$. (iv) By Problem 2.12 and (i), it is no restriction to assume that A is unitary. Let $\{w_n\}$ be some ONB and note that $\{\tilde{w}_n = Aw_n\}$ is also an ONB. Then

$$\begin{aligned}\operatorname{tr}(AK) &= \sum_n \langle \tilde{w}_n, AK\tilde{w}_n\rangle = \sum_n \langle Aw_n, AKAw_n\rangle \\ &= \sum_n \langle w_n, KAw_n\rangle = \operatorname{tr}(KA)\end{aligned}$$

and the claim follows. □

We also mention a useful criterion for K to be trace class.

Lemma 3.29. *An operator K is trace class if and only if it can be written as*

$$K = \sum_j \langle f_j, .\rangle g_j \tag{3.61}$$

for some sequences f_j, g_j satisfying

$$\sum_j \|f_j\|\|g_j\| < \infty. \tag{3.62}$$

Moreover, in this case

$$\operatorname{tr}(K) = \sum_j \langle f_j, g_j\rangle \tag{3.63}$$

and

$$\|K\|_1 = \min \sum_j \|f_j\|\|g_j\|, \tag{3.64}$$

where the minimum is taken over all representations as in (3.61).

Proof. To see that a trace class operator (3.40) can be written in such a way choose $f_j = u_j$, $g_j = s_j v_j$. This also shows that the minimum in (3.64) is attained. Conversely note that the sum converges in the operator norm

and hence K is compact. Moreover, for every finite N we have

$$\sum_{k=1}^{N} s_k = \sum_{k=1}^{N} \langle v_k, K u_k\rangle = \sum_{k=1}^{N}\sum_{j} \langle v_k, g_j\rangle\langle f_j, u_k\rangle = \sum_{j}\sum_{k=1}^{N} \langle v_k, g_j\rangle\langle f_j, u_k\rangle$$
$$\leq \sum_{j} \left(\sum_{k=1}^{N} |\langle v_k, g_j\rangle|^2\right)^{1/2} \left(\sum_{k=1}^{N} |\langle f_j, u_k\rangle|^2\right)^{1/2} \leq \sum_{j} \|f_j\|\|g_j\|.$$

This also shows that the right-hand side in (3.64) cannot exceed $\|K\|_1$. To see the last claim we choose an ONB $\{w_k\}$ to compute the trace

$$\operatorname{tr}(K) = \sum_{k} \langle w_k, K w_k\rangle = \sum_{k}\sum_{j} \langle w_k, \langle f_j, w_k\rangle g_j\rangle = \sum_{j}\sum_{k} \langle\langle w_k, f_j\rangle w_k, g_j\rangle$$
$$= \sum_{j} \langle f_j, g_j\rangle. \qquad \square$$

An immediate consequence of (3.64) is:

Corollary 3.30. *The trace norm satisfies the triangle inequality and hence is indeed a norm.*

Finally, note that

$$\|K\|_2 = \big(\operatorname{tr}(K^*K)\big)^{1/2} \tag{3.65}$$

which shows that $\mathcal{J}_2(\mathfrak{H})$ is in fact a Hilbert space with scalar product given by

$$\langle K_1, K_2\rangle = \operatorname{tr}(K_1^* K_2). \tag{3.66}$$

Problem 3.18. *Let $\mathfrak{H} := \ell^2(\mathbb{N})$ and let A be multiplication by a sequence $a = (a_j)_{j=1}^{\infty}$. Show that A is Hilbert–Schmidt if and only if $a \in \ell^2(\mathbb{N})$. Furthermore, show that $\|A\|_2 = \|a\|$ in this case.*

Problem 3.19. *An operator of the form $K : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$, $f_n \mapsto \sum_{j\in\mathbb{N}} k_{n+j} f_j$ is called **Hankel operator**.*

- *Show that K is Hilbert–Schmidt if and only if $\sum_{j\in\mathbb{N}} j|k_{j+1}|^2 < \infty$ and this number equals $\|K\|_2$.*
- *Show that K is Hilbert–Schmidt with $\|K\|_2 \leq \|c\|_1$ if $|k_j| \leq c_j$, where c_j is decreasing and summable.*

(Hint: For the first item use summation by parts.)

Chapter 4

The main theorems about Banach spaces

Despite the many advantages of Hilbert spaces, there are also situations where a non-Hilbert space is better suited (in fact the choice of the *right* space is typically crucial for many problems). Hence we will devote our attention to Banach spaces next.

4.1. The Baire theorem and its consequences

Recall that the interior of a set is the largest open subset (that is, the union of all open subsets). A set is called **nowhere dense** if its closure has empty interior. The key to several important theorems about Banach spaces is the observation that a Banach space cannot be the countable union of nowhere dense sets.

Theorem 4.1 (Baire category theorem)**.** *Let X be a (nonempty) complete metric space. Then X cannot be the countable union of nowhere dense sets.*

Proof. Suppose $X = \bigcup_{n=1}^{\infty} X_n$. We can assume that the sets X_n are closed and none of them contains a ball; that is, $X \setminus X_n$ is open and nonempty for every n. We will construct a Cauchy sequence x_n which stays away from all X_n.

Since $X \setminus X_1$ is open and nonempty, there is a ball $B_{r_1}(x_1) \subseteq X \setminus X_1$. Reducing r_1 a little, we can even assume $\overline{B_{r_1}(x_1)} \subseteq X \setminus X_1$. Moreover, since X_2 cannot contain $B_{r_1}(x_1)$, there is some $x_2 \in B_{r_1}(x_1)$ that is not in X_2. Since $B_{r_1}(x_1) \cap (X \setminus X_2)$ is open, there is a closed ball $\overline{B_{r_2}(x_2)} \subseteq B_{r_1}(x_1) \cap (X \setminus X_2)$. Proceeding recursively, we obtain a sequence (here we

use the axion of choice) of balls such that

$$\overline{B_{r_n}(x_n)} \subseteq B_{r_{n-1}}(x_{n-1}) \cap (X \setminus X_n).$$

Now observe that in every step we can choose r_n as small as we please; hence without loss of generality $r_n \to 0$. Since by construction $x_n \in \overline{B_{r_N}(x_N)}$ for $n \geq N$, we conclude that x_n is Cauchy and converges to some point $x \in X$. But $x \in \overline{B_{r_n}(x_n)} \subseteq X \setminus X_n$ for every n, contradicting our assumption that the X_n cover X. □

In other words, if $X_n \subseteq X$ is a sequence of closed subsets which cover X, at least one X_n contains a ball of radius $\varepsilon > 0$.

Example 4.1. The set of rational numbers $\mathbb{Q}$ can be written as a countable union of its elements. This shows that the completeness assumption is crucial. ⋄

Remark: Sets which can be written as the countable union of nowhere dense sets are said to be of **first category** or **meager**. All other sets are **second category** or **fat**. Hence explaining the name category theorem.

Since a closed set is nowhere dense if and only if its complement is open and dense (cf. Problem B.6), there is a reformulation which is also worthwhile noting:

Corollary 4.2. *Let X be a complete metric space. Then any countable intersection of open dense sets is again dense.*

Proof. Let $\{O_n\}$ be a family of open dense sets whose intersection is not dense. Then this intersection must be missing some closed ball $\overline{B_\varepsilon}$. This ball will lie in $\bigcup_n X_n$, where $X_n := X \setminus O_n$ are closed and nowhere dense. Now note that $\tilde{X}_n := X_n \cap \overline{B_\varepsilon}$ are closed nowhere dense sets in $\overline{B_\varepsilon}$. But $\overline{B_\varepsilon}$ is a complete metric space, a contradiction. □

Countable intersections of open sets are in some sense the next general sets after open sets (cf. also Section 8.6) and are called G_δ sets (here G and δ stand for the German words *Gebiet* and *Durchschnitt*, respectively). The complement of a G_δ set is a countable union of closed sets also known as an F_σ set (here F and σ stand for the French words *fermé* and *somme*, respectively). The complement of a dense G_δ set will be a countable intersection of nowhere dense sets and hence by definition meager. Consequently properties which hold on a dense G_δ are considered *generic* in this context.

Example 4.2. The irrational numbers are a dense G_δ set in $\mathbb{R}$. To see this, let x_n be an enumeration of the rational numbers and consider the intersection of the open sets $O_n := \mathbb{R} \setminus \{x_n\}$. The rational numbers are hence an F_σ set. ⋄

Now we are ready for the first important consequence:

Theorem 4.3 (Banach–Steinhaus)**.** *Let X be a Banach space and Y some normed vector space. Let $\{A_\alpha\} \subseteq \mathscr{L}(X,Y)$ be a family of bounded operators. Then*

- *either $\{A_\alpha\}$ is uniformly bounded, $\|A_\alpha\| \le C$,*
- *or the set $\{x \in X | \sup_\alpha \|A_\alpha x\| = \infty\}$ is a dense G_δ.*

Proof. Consider the sets

$$O_n := \{x | \, \|A_\alpha x\| > n \text{ for some } \alpha\} = \bigcup_\alpha \{x | \, \|A_\alpha x\| > n\}, \quad n \in \mathbb{N}.$$

By continuity of A_α and the norm, each O_n is a union of open sets and hence open. Now either all of these sets are dense and hence their intersection

$$\bigcap_{n\in\mathbb{N}} O_n = \{x | \sup_\alpha \|A_\alpha x\| = \infty\}$$

is a dense G_δ by Corollary 4.2. Otherwise, $X \setminus \overline{O_n}$ is nonempty and open for one n and we can find a ball of positive radius $\overline{B_\varepsilon(x_0)} \subset X \setminus O_n$. Now observe

$$\|A_\alpha y\| = \|A_\alpha(y + x_0 - x_0)\| \le \|A_\alpha(y + x_0)\| + \|A_\alpha x_0\| \le 2n$$

for $\|y\| \le \varepsilon$. Setting $y = \varepsilon \frac{x}{\|x\|}$, we obtain

$$\|A_\alpha x\| \le \frac{2n}{\varepsilon}\|x\|$$

for every x. □

Note that there is also a variant of the Banach–Steinhaus theorem for pointwise limits of bounded operators which will be discussed in Lemma 4.32.

Hence there are two ways to use this theorem by excluding one of the two possible options. Showing that the pointwise bound holds on a sufficiently large set (e.g. a ball), thereby ruling out the second option, implies a uniform bound and is known as the **uniform boundedness principle**.

Corollary 4.4. *Let X be a Banach space and Y some normed vector space. Let $\{A_\alpha\} \subseteq \mathscr{L}(X,Y)$ be a family of bounded operators. Suppose $\|A_\alpha x\| \le C(x)$ is bounded for every fixed $x \in X$. Then $\{A_\alpha\}$ is uniformly bounded, $\|A_\alpha\| \le C$.*

Conversely, if there is no uniform bound, the pointwise bound must fail on a dense G_δ. This is illustrated in the following example.

Example 4.3. Consider the Fourier series (2.44) of a continuous periodic function $f \in C_{per}[-\pi,\pi] = \{f \in C[-\pi,\pi] | f(-\pi) = f(\pi)\}$. (Note that this is a closed subspace of $C[-\pi,\pi]$ and hence a Banach space — it is the kernel of the linear functional $\ell(f) = f(-\pi) - f(\pi)$.) We want to show that for every fixed $x \in [-\pi,\pi]$ there is a dense G_δ set of functions in $C_{per}[-\pi,\pi]$ for which the Fourier series will diverge at x (it will even be unbounded).

Without loss of generality we fix $x = 0$ as our point. Then the n'th partial sum gives rise to the linear functional

$$\ell_n(f) := S_n(f)(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(x)f(x)dx$$

and it suffices to show that the family $\{\ell_n\}_{n\in\mathbb{N}}$ is not uniformly bounded.

By Example 1.21 (adapted to our present periodic setting) we have

$$\|\ell_n\| = \frac{1}{2\pi}\|D_n\|_1.$$

Now we estimate

$$\begin{aligned}\|D_n\|_1 &= 2\int_0^{\pi} |D_n(x)|dx \geq 2\int_0^{\pi} \frac{|\sin((n+1/2)x)|}{x/2}dx \\ &= 4\int_0^{(n+1/2)\pi} |\sin(y)|\frac{dy}{y} \geq 4\sum_{k=1}^{n}\int_{(k-1)\pi}^{k\pi} |\sin(y)|\frac{dy}{k\pi} = \frac{8}{\pi}\sum_{k=1}^{n}\frac{1}{k}\end{aligned}$$

and note that the harmonic series diverges.

In fact, we can even do better. Let $G(x) \subset C_{per}[-\pi,\pi]$ be the dense G_δ of functions whose Fourier series diverges at x. Then, given countably many points $\{x_j\}_{j\in\mathbb{N}} \subset [-\pi,\pi]$, the set $G = \bigcap_{j\in\mathbb{N}} G(x_j)$ is still a dense G_δ by Corollary 4.2. Hence there is a dense G_δ of functions whose Fourier series diverges on a given countable set of points. ◇

Example 4.4. Recall that the Fourier coefficients of an absolutely continuous function satisfy the estimate

$$|\hat{f}_k| \leq \begin{cases} \|f\|_\infty, & k = 0, \\ \frac{\|f'\|_\infty}{|k|}, & k \neq 0. \end{cases}$$

This raises the question if a similar estimate can be true for continuous functions. More precisely, can we find a sequence $c_k > 0$ such that

$$|\hat{f}_k| \leq C_f c_k,$$

where C_f is some constant depending on f. If this were true, the linear functionals

$$\ell_k(f) := \frac{\hat{f}_k}{c_k}, \qquad k \in \mathbb{Z},$$

satisfy the assumptions of the uniform boundedness principle implying $\|\ell_k\| \leq C$. In other words, we must have an estimate of the type

$$|\hat{f}_k| \leq C\|f\|_\infty c_k$$

which implies $1 \leq C\, c_k$ upon choosing $f(x) = \mathrm{e}^{\mathrm{i}kx}$. Hence our assumption cannot hold for any sequence c_k converging to zero and there is no universal decay rate for the Fourier coefficients of continuous functions beyond the fact that they must converge to zero by the Riemann–Lebesgue lemma. ⋄

The next application is

Theorem 4.5 (Open mapping)**.** *Let $A \in \mathscr{L}(X, Y)$ be a continuous linear operator between Banach spaces. Then A is open (i.e., maps open sets to open sets) if and only if it is onto.*

Proof. Set $B_r^X := B_r^X(0)$ and similarly for $B_r^Y(0)$. By translating balls (using linearity of A), it suffices to prove that for every $\varepsilon > 0$ there is a $\delta > 0$ such that $B_\delta^Y \subseteq A(B_\varepsilon^X)$.

So let $\varepsilon > 0$ be given. Since A is surjective we have

$$Y = AX = A\bigcup_{n=1}^{\infty} nB_\varepsilon^X = \bigcup_{n=1}^{\infty} A(nB_\varepsilon^X) = \bigcup_{n=1}^{\infty} nA(B_\varepsilon^X)$$

and the Baire theorem implies that for some n, $n\overline{A(B_\varepsilon^X)}$ contains a ball. Since multiplication by n is a homeomorphism, the same must be true for $n = 1$, that is, $B_\delta^Y(y) \subset \overline{A(B_\varepsilon^X)}$. Consequently

$$B_\delta^Y \subseteq -y + \overline{A(B_\varepsilon^X)} \subset \overline{A(B_\varepsilon^X)} + \overline{A(B_\varepsilon^X)} \subseteq \overline{A(B_\varepsilon^X) + A(B_\varepsilon^X)} \subseteq \overline{A(B_{2\varepsilon}^X)}.$$

So it remains to get rid of the closure. To this end choose $\varepsilon_n > 0$ such that $\sum_{n=1}^\infty \varepsilon_n < \varepsilon$ and corresponding $\delta_n \to 0$ such that $B_{\delta_n}^Y \subset \overline{A(B_{\varepsilon_n}^X)}$. Now for $y \in B_{\delta_1}^Y \subset \overline{A(B_{\varepsilon_1}^X)}$ we have $x_1 \in B_{\varepsilon_1}^X$ such that Ax_1 is arbitrarily close to y, say $y - Ax_1 \in B_{\delta_2}^Y \subset \overline{A(B_{\varepsilon_2}^X)}$. Hence we can find $x_2 \in B_{\varepsilon_2}^X$ such that $(y - Ax_1) - Ax_2 \in B_{\delta_3}^Y \subset \overline{A(B_{\varepsilon_3}^X)}$ and proceeding like this a sequence $x_n \in B_{\varepsilon_n}^X$ such that

$$y - \sum_{k=1}^{n} Ax_k \in B_{\delta_{n+1}}^Y.$$

By construction the limit $x := \sum_{k=1}^\infty x_k$ exists and satisfies $x \in B_\varepsilon^X$ as well as $y = Ax \in A(B_\varepsilon^X)$. That is, $B_{\delta_1}^Y \subseteq A(B_\varepsilon^X)$ as desired.

Conversely, if A is open, then the image of the unit ball contains again some ball $B_\varepsilon^Y \subseteq A(B_1^X)$. Hence by scaling $B_{r\varepsilon}^Y \subseteq A(B_r^X)$ and letting $r \to \infty$ we see that A is onto: $Y = A(X)$. □

Example 4.5. However, note that, under the assumptions of the open mapping theorem, the image of a closed set might not be closed. For example, consider the bounded linear operator $A : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$, $x \mapsto (x_2, x_4, \dots)$ which is clearly surjective. Then the image of the closed set $U = \{x \in \ell^2(\mathbb{N}) | x_{2n} = x_{2n-1}/n\}$ is dense (it contains all sequences with finite support) but not all of $\ell^2(\mathbb{N})$ (e.g. $y_n = \frac{1}{n}$ is missing since this would imply $x_{2n} = 1$). ⋄

As an immediate consequence we get the inverse mapping theorem:

Theorem 4.6 (Inverse mapping)**.** *Let $A \in \mathscr{L}(X, Y)$ be a continuous linear bijection between Banach spaces. Then A^{-1} is continuous.*

Example 4.6. Consider the operator $(Aa)_{j=1}^n = (\frac{1}{j} a_j)_{j=1}^n$ in $\ell^2(\mathbb{N})$. Then its inverse $(A^{-1}a)_{j=1}^n = (j\, a_j)_{j=1}^n$ is unbounded (show this!). This is in agreement with our theorem since its range is dense (why?) but not all of $\ell^2(\mathbb{N})$: For example, $(b_j = \frac{1}{j})_{j=1}^\infty \notin \operatorname{Ran}(A)$ since $b = Aa$ gives the contradiction

$$\infty = \sum_{j=1}^{\infty} 1 = \sum_{j=1}^{\infty} |j b_j|^2 = \sum_{j=1}^{\infty} |a_j|^2 < \infty.$$

This should also be compared with Corollary 4.9 below. ⋄

Example 4.7. Consider the Fourier series (2.44) of an integrable function. Using the inverse function theorem we can show that not every sequence tending to 0 (which is a necessary condition according to the Riemann–Lebesgue lemma) arises as the Fourier coefficients of an integrable function:

By the elementary estimate

$$\|\hat{f}\|_\infty \le \frac{1}{2\pi} \|f\|_1$$

we see that that the mapping $F(f) := \hat{f}$ continuously maps $F : L^1(-\pi, \pi) \to c_0(\mathbb{Z})$ (the Banach space of sequences converging to 0). In fact, this estimate holds for continuous functions and hence there is a unique continuous extension of F to all of $L^1(-\pi, \pi)$ by Theorem 1.17. Moreover, it can be shown that F is injective (for $f \in L^2$ this follows from Theorem 2.18, the general case $f \in L^1$ will be established in Example 10.7). Now if F were onto, the inverse mapping theorem would show that the inverse is also continuous, that is, we would have an estimate $\|\hat{f}\|_\infty \ge C \|f\|_1$ for some $C > 0$. However, considering the Dirichlet kernel D_n we have $\|\hat{D}_n\|_\infty = 1$ but $\|D_n\|_1 \to \infty$ as shown in Example 4.3. ⋄

Another important consequence is the closed graph theorem. The **graph** of an operator A is just

$$\Gamma(A) := \{(x, Ax) | x \in \mathfrak{D}(A)\}. \tag{4.1}$$

If A is linear, the graph is a subspace of the Banach space $X \oplus Y$ (provided X and Y are Banach spaces), which is just the Cartesian product together with the norm

$$\|(x, y)\|_{X\oplus Y} := \|x\|_X + \|y\|_Y. \tag{4.2}$$

Note that $(x_n, y_n) \to (x, y)$ if and only if $x_n \to x$ and $y_n \to y$. We say that A has a closed graph if $\Gamma(A)$ is a closed subset of $X \oplus Y$.

Theorem 4.7 (Closed graph)**.** *Let $A : X \to Y$ be a linear map from a Banach space X to another Banach space Y. Then A is continuous if and only if its graph is closed.*

Proof. If $\Gamma(A)$ is closed, then it is again a Banach space. Now the projection $\pi_1(x, Ax) = x$ onto the first component is a continuous bijection onto X. So by the inverse mapping theorem its inverse π_1^{-1} is again continuous. Moreover, the projection $\pi_2(x, Ax) = Ax$ onto the second component is also continuous and consequently so is $A = \pi_2 \circ \pi_1^{-1}$. The converse is easy. □

Remark: The crucial fact here is that A is defined on *all* of X!

Operators whose graphs are closed are called **closed operators**. Warning: By Example 4.5 a closed operator will not map closed sets to closed sets in general. In particular, the concept of a closed operator should not be confused with the concept of a closed map in topology!

Being closed is the next option you have once an operator turns out to be unbounded. If A is closed, then $x_n \to x$ does not guarantee you that Ax_n converges (like continuity would), but it at least guarantees that if Ax_n converges, it converges to the right thing, namely Ax:

- A bounded (with $\mathfrak{D}(A) = X$): $x_n \to x$ implies $Ax_n \to Ax$.
- A closed (with $\mathfrak{D}(A) \subseteq X$): $x_n \to x$, $x_n \in \mathfrak{D}(A)$, and $Ax_n \to y$ implies $x \in \mathfrak{D}(A)$ and $y = Ax$.

If an operator is not closed, you can try to take the closure of its graph, to obtain a closed operator. If A is bounded this always works (which is just the content of Theorem 1.17). However, in general, the closure of the graph might not be the graph of an operator as we might pick up points $(x, y_1), (x, y_2) \in \overline{\Gamma(A)}$ with $y_1 \neq y_2$. Since $\overline{\Gamma(A)}$ is a subspace, we also have $(x, y_2) - (x, y_1) = (0, y_2 - y_1) \in \overline{\Gamma(A)}$ in this case and thus $\overline{\Gamma(A)}$ is the graph of some operator if and only if

$$\overline{\Gamma(A)} \cap \{(0, y) | y \in Y\} = \{(0, 0)\}. \tag{4.3}$$

If this is the case, A is called **closable** and the operator $\overline{A}$ associated with $\overline{\Gamma(A)}$ is called the **closure** of A.

In particular, A is closable if and only if $x_n \to 0$ and $Ax_n \to y$ implies $y = 0$. In this case

$$\mathfrak{D}(\overline{A}) = \{x \in X | \exists x_n \in \mathfrak{D}(A),\ y \in Y : x_n \to x \text{ and } Ax_n \to y\},$$
$$\overline{A}x = y. \tag{4.4}$$

For yet another way of defining the closure see Problem 4.12.

Example 4.8. Consider the operator A in $\ell^p(\mathbb{N})$ defined by $Aa_j := ja_j$ on $\mathfrak{D}(A) = \{a \in \ell^p(\mathbb{N}) | a_j \neq 0 \text{ for finitely many } j\}$.

(i). A is closable. In fact, if $a^n \to 0$ and $Aa^n \to b$ then we have $a_j^n \to 0$ and thus $ja_j^n \to 0 = b_j$ for any $j \in \mathbb{N}$.

(ii). The closure of A is given by

$$\mathfrak{D}(\overline{A}) = \begin{cases} \{a \in \ell^p(\mathbb{N}) | (ja_j)_{j=1}^\infty \in \ell^p(\mathbb{N})\}, & 1 \leq p < \infty, \\ \{a \in c_0(\mathbb{N}) | (ja_j)_{j=1}^\infty \in c_0(\mathbb{N})\}, & p = \infty, \end{cases}$$

and $\overline{A}a_j = ja_j$. In fact, if $a^n \to a$ and $Aa^n \to b$ then we have $a_j^n \to a_j$ and $ja_j^n \to b_j$ for any $j \in \mathbb{N}$ and thus $b_j = ja_j$ for any $j \in \mathbb{N}$. In particular, $(ja_j)_{j=1}^\infty = (b_j)_{j=1}^\infty \in \ell^p(\mathbb{N})$ ($c_0(\mathbb{N})$ if $p = \infty$). Conversely, suppose $(ja_j)_{j=1}^\infty \in \ell^p(\mathbb{N})$ ($c_0(\mathbb{N})$ if $p = \infty$) and consider

$$a_j^n := \begin{cases} a_j, & j \leq n, \\ 0, & j > n. \end{cases}$$

Then $a^n \to a$ and $Aa^n \to (ja_j)_{j=1}^\infty$.

(iii). Extending the basis vectors $\{\delta^n\}_{n\in\mathbb{N}}$ to a Hamel basis (Problem 1.6) and setting $Aa = 0$ for every other element from this Hamel basis we obtain a (still unbounded) operator which is everywhere defined. However, this extension cannot be closed! ⋄

Example 4.9. Here is a simple example of a nonclosable operator: Let $X := \ell^2(\mathbb{N})$ and consider $Ba := (\sum_{j=1}^\infty a_j)\delta^1$ defined on $\ell^1(\mathbb{N}) \subset \ell^2(\mathbb{N})$. Let $a_j^n := \frac{1}{n}$ for $1 \leq j \leq n$ and $a_j^n := 0$ for $j > n$. Then $\|a^n\|_2 = \frac{1}{\sqrt{n}}$ implying $a^n \to 0$ but $Ba^n = \delta^1 \not\to 0$. ⋄

Example 4.10. Another example are point evaluations in $L^2(0,1)$: Let $x_0 \in [0,1]$ and consider $\ell_{x_0} : \mathfrak{D}(\ell_{x_0}) \to \mathbb{C}$, $f \mapsto f(x_0)$ defined on $\mathfrak{D}(\ell_{x_0}) := C[0,1] \subseteq L^2(0,1)$. Then $f_n(x) := \max(0, 1 - n|x - x_0|)$ satisfies $f_n \to 0$ but $\ell_{x_0}(f_n) = 1$. In fact, a linear functional is closable if and only if it is bounded (Problem 4.8). ⋄

Lemma 4.8. *Suppose A is closable and $\overline{A}$ is injective. Then $\overline{A}^{-1} = \overline{A^{-1}}$.*

Proof. If we set

$$\Gamma^{-1} = \{(y, x) | (x, y) \in \Gamma\}$$

then $\Gamma(A^{-1}) = \Gamma^{-1}(A)$ and

$$\overline{\Gamma(A^{-1})} = \overline{\Gamma(A)^{-1}} = \overline{\Gamma(A)}^{-1} = \Gamma(\overline{A})^{-1} = \Gamma(\overline{A}^{-1}). \qquad \square$$

Note that A injective does not imply $\overline{A}$ injective in general.

Example 4.11. Let P_M be the projection in $\ell^2(\mathbb{N})$ on $M := \{b\}^\perp$, where $b := (2^{-j/2})_{j=1}^\infty$. Explicitly we have $P_M a = a - \langle b, a\rangle b$. Then P_M restricted to the space of sequences with finitely many nonzero terms is injective, but its closure is not. $\diamond$

As a consequence of the closed graph theorem we obtain:

Corollary 4.9. *Suppose $A : \mathfrak{D}(A) \subseteq X \to Y$ is closed and injective. Then A^{-1} defined on $\mathfrak{D}(A^{-1}) = \operatorname{Ran}(A)$ is closed. Moreover, in this case $\operatorname{Ran}(A)$ is closed if and only if A^{-1} is bounded.*

The question when $\operatorname{Ran}(A)$ is closed plays an important role when investigating solvability of the equation $Ax = y$ and the last part gives us a convenient criterion. Moreover, note that A^{-1} is bounded if and only if there is some $c > 0$ such that

$$\|Ax\| \geq c\|x\|, \qquad x \in \mathfrak{D}(A). \tag{4.5}$$

Indeed, this follows upon setting $x = A^{-1}y$ in the above inequality which also shows that $c = \|A^{-1}\|^{-1}$ is the best possible constant. Factoring out the kernel we even get a criterion for the general case:

Corollary 4.10. *Suppose $A : \mathfrak{D}(A) \subseteq X \to Y$ is closed. Then $\operatorname{Ran}(A)$ is closed if and only if*

$$\|Ax\| \geq c \operatorname{dist}(x, \operatorname{Ker}(A)), \qquad x \in \mathfrak{D}(A), \tag{4.6}$$

*for some $c > 0$. The sup over all possible c is known as the (reduced) **minimum modulus** of A.*

Proof. Consider the quotient space $\tilde{X} := X/\operatorname{Ker}(A)$ and the induced operator $\tilde{A} : \mathfrak{D}(\tilde{A}) \to Y$ where $\mathfrak{D}(\tilde{A}) = \mathfrak{D}(A)/\operatorname{Ker}(A) \subseteq \tilde{X}$. By construction $\tilde{A}[x] = 0$ iff $x \in \operatorname{Ker}(A)$ and hence $\tilde{A}$ is injective. To see that $\tilde{A}$ is closed we use $\tilde{\pi} : X \times Y \to \tilde{X} \times Y$, $(x, y) \mapsto ([x], y)$ which is bounded, surjective and hence open. Moreover, $\tilde{\pi}(\Gamma(A)) = \Gamma(\tilde{A})$. In fact, we even have $(x, y) \in \Gamma(A)$ iff $([x], y) \in \Gamma(\tilde{A})$ and thus $\tilde{\pi}(X \times Y \setminus \Gamma(A)) = \tilde{X} \times Y \setminus \Gamma(\tilde{A})$ implying that $Y \setminus \Gamma(\tilde{A})$ is open. Finally, observing $\operatorname{Ran}(A) = \operatorname{Ran}(\tilde{A})$ we have reduced it to the previous corollary. $\square$

There is also another criterion which does not involve the distance to the kernel.

Corollary 4.11. *Suppose $A : \mathfrak{D}(A) \subseteq X \to Y$ is closed. Then* $\operatorname{Ran}(A)$ *is closed if for some given $\varepsilon > 0$ and $0 \leq \delta < 1$ we can find for every $y \in \overline{\operatorname{Ran}(A)}$ a corresponding $x \in \mathfrak{D}(X)$ such that*

$$\varepsilon\|x\| + \|y - Ax\| \leq \delta\|y\|. \tag{4.7}$$

Conversely, if $\operatorname{Ran}(A)$ *is closed this can be done whenever $\varepsilon < c\delta$ with c from the previous corollary.*

Proof. If $\operatorname{Ran}(A)$ is closed and $\varepsilon < c\delta$ there is some $x \in \mathfrak{D}(A)$ with $y = Ax$ and $\|Ax\| \geq \frac{\varepsilon}{\delta}\|x\|$ after maybe adding an element from the kernel to x. This x satisfies $\varepsilon\|x\| + \|y - Ax\| = \varepsilon\|x\| \leq \delta\|y\|$ as required.

Conversely, fix $y \in \overline{\operatorname{Ran}(A)}$ and recursively choose a sequence x_n such that

$$\varepsilon\|x_n\| + \|(y - A\tilde{x}_{n-1}) - Ax_n\| \leq \delta\|y - A\tilde{x}_{n-1}\|, \quad \tilde{x}_n := \sum_{m \leq n} x_m.$$

In particular, $\|y - A\tilde{x}_n\| \leq \delta^n\|y\|$ as well as $\varepsilon\|x_n\| \leq \delta^n\|y\|$, which shows $\tilde{x}_n \to x$ and $A\tilde{x}_n \to y$. Hence $x \in \mathfrak{D}(A)$ and $y = Tx \in \operatorname{Ran}(A)$. □

The closed graph theorem tells us that closed linear operators can be defined on all of X if and only if they are bounded. So if we have an unbounded operator we cannot have both! That is, if we want our operator to be at least closed, we have to live with domains. This is the reason why in quantum mechanics most operators are defined on domains. In fact, there is another important property which does not allow unbounded operators to be defined on the entire space:

Theorem 4.12 (Hellinger–Toeplitz)**.** *Let $A : \mathfrak{H} \to \mathfrak{H}$ be a linear operator on some Hilbert space $\mathfrak{H}$. If A is symmetric, that is $\langle g, Af\rangle = \langle Ag, f\rangle$, $f, g \in \mathfrak{H}$, then A is bounded.*

Proof. It suffices to prove that A is closed. In fact, $f_n \to f$ and $Af_n \to g$ implies

$$\langle h, g\rangle = \lim_{n\to\infty} \langle h, Af_n\rangle = \lim_{n\to\infty} \langle Ah, f_n\rangle = \langle Ah, f\rangle = \langle h, Af\rangle$$

for every $h \in \mathfrak{H}$. Hence $Af = g$. □

Problem 4.1. *An infinite dimensional Banach space cannot have a countable Hamel basis (see Problem 1.6). (Hint: Apply Baire's theorem to $X_n := \operatorname{span}\{u_j\}_{j=1}^n$.)*

Problem 4.2. *Let $X := C[0, 1]$. Show that the set of functions which are nowhere differentiable contains a dense G_δ. (Hint: Consider $F_k := \{f \in X | \exists x \in [0, 1] : |f(x) - f(y)| \leq k|x - y|, \forall y \in [0, 1]\}$. Show that this set is closed and nowhere dense. For the first property Bolzano–Weierstraß*

might be useful, for the latter property show that the set of piecewise linear functions whose slopes are bounded below by some fixed number in absolute value are dense.)

Problem 4.3. *Let X be the space of sequences with finitely many nonzero terms together with the sup norm. Consider the family of operators $\{A_n\}_{n\in\mathbb{N}}$ given by $(A_n a)_j := j a_j$, $j \le n$ and $(A_n a)_j := 0$, $j > n$. Then this family is pointwise bounded but not uniformly bounded. Does this contradict the Banach–Steinhaus theorem?*

Problem 4.4. *Let X be a complete metric space without isolated points. Show that a dense G_δ set cannot be countable. (Hint: A single point is nowhere dense.)*

Problem 4.5. *Show that a bilinear map $B : X \times Y \to Z$ is bounded, $\|B(x,y)\| \le C\|x\|\|y\|$, if and only if it is separately continuous with respect to both arguments. (Hint: Uniform boundedness principle.)*

Problem 4.6. *Consider a Schauder basis as in* (1.31)*. Show that the coordinate functionals α_n are continuous. (Hint: Denote the set of all possible sequences of Schauder coefficients by $\mathcal{A}$ and equip it with the norm $\|\alpha\| := \sup_n \|\sum_{k=1}^n \alpha_k u_k\|$; note that $\mathcal{A}$ is precisely the set of sequences for which this norm is finite. By construction the operator $A : \mathcal{A} \to X$, $\alpha \mapsto \sum_k \alpha_k u_k$ has norm one. Now show that $\mathcal{A}$ is complete and apply the inverse mapping theorem.)*

Problem 4.7. *Show that a compact symmetric operator in an infinite-dimensional Hilbert space cannot be surjective.*

Problem* 4.8. *A linear functional defined on a dense subspace is closable if and only if it is bounded.*

Problem 4.9. *Show that if A is a closed and B a bounded operator, then $A + B$ is closed. Show that this in general fails if B is not bounded. (Here $A + B$ is defined on $\mathfrak{D}(A + B) = \mathfrak{D}(A) \cap \mathfrak{D}(B)$.)*

Problem 4.10. *Suppose $B \in \mathscr{L}(X,Y)$ is bounded and $A : \mathfrak{D}(A) \subseteq Y \to Z$ is closed. Then $AB : \mathfrak{D}(AB) \subseteq X \to Z$ is closed, where $\mathfrak{D}(AB) := \{x \in \mathfrak{D}(B) | Ax \in \mathfrak{D}(A)\}$.*

Problem 4.11. *Show that the differential operator $A = \frac{d}{dx}$ defined on $\mathfrak{D}(A) = C^1[0,1] \subset C[0,1]$ (sup norm) is a closed operator. (Compare the example in Section 1.6.)*

Problem* 4.12. *Consider a linear operator $A : \mathfrak{D}(A) \subseteq X \to Y$, where X and Y are Banach spaces. Define the* ***graph norm*** *associated with A by*

$$\|x\|_A := \|x\|_X + \|Ax\|_Y, \quad x \in \mathfrak{D}(A). \tag{4.8}$$

Show that $A : \mathfrak{D}(A) \to Y$ is bounded if we equip $\mathfrak{D}(A)$ with the graph norm. Show that the completion X_A of $(\mathfrak{D}(A), \|.\|_A)$ can be regarded as a subset of X if and only if A is closable. Show that in this case the completion can be identified with $\mathfrak{D}(\overline{A})$ and that the closure of A in X coincides with the extension from Theorem 1.17 of A in X_A. In particular, A is closed if and only if $(\mathfrak{D}(A), \|.\|_A)$ is complete.

Problem 4.13. *Let $X := \ell^2(\mathbb{N})$ and $(Aa)_j := j\, a_j$ with $\mathfrak{D}(A) := \{a \in \ell^2(\mathbb{N}) | (ja_j)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N})\}$ and $Ba := (\sum_{j\in\mathbb{N}} a_j)\delta^1$. Then we have seen that A is closed while B is not closable. Show that $A+B$, $\mathfrak{D}(A+B) = \mathfrak{D}(A) \cap \mathfrak{D}(B) = \mathfrak{D}(A)$ is closed.*

4.2. The Hahn–Banach theorem and its consequences

Let X be a Banach space. Recall that we have called the set of all bounded linear functionals the dual space X^* (which is again a Banach space by Theorem 1.18).

Example 4.12. Consider the Banach space $\ell^p(\mathbb{N})$, $1 \le p < \infty$. Taking the Kronecker deltas δ^n as a Schauder basis the n'th term x_n of a sequence $x \in \ell^p(\mathbb{N})$ can also be considered as the n'th coordinate of x with respect to this basis. Moreover, the map $l_n(x) = x_n$ is a bounded linear functional, that is, $l_n \in \ell^p(\mathbb{N})^*$, since $|l_n(x)| = |x_n| \le \|x\|_p$. It is a special case of the following more general example (in fact, we have $l_n = l_{\delta^n}$). Since the coordinates of a vector carry all the information this explains why understanding linear functionals if of key importance. ⋄

Example 4.13. Consider the Banach space $\ell^p(\mathbb{N})$, $1 \le p < \infty$. We have already seen that by Hölder's inequality (1.25) every $y \in \ell^q(\mathbb{N})$ gives rise to a bounded linear functional

$$l_y(x) := \sum_{n\in\mathbb{N}} y_n x_n \tag{4.9}$$

whose norm is $\|l_y\| = \|y\|_q$ (Problem 4.19). But can every element of $\ell^p(\mathbb{N})^*$ be written in this form?

Suppose $p := 1$ and choose $l \in \ell^1(\mathbb{N})^*$. Now define

$$y_n := l(\delta^n).$$

Then

$$|y_n| = |l(\delta^n)| \le \|l\|\, \|\delta^n\|_1 = \|l\|$$

shows $\|y\|_\infty \le \|l\|$, that is, $y \in \ell^\infty(\mathbb{N})$. By construction $l(x) = l_y(x)$ for every $x \in \operatorname{span}\{\delta^n\}$. By continuity of l it even holds for $x \in \overline{\operatorname{span}\{\delta^n\}} = \ell^1(\mathbb{N})$. Hence the map $y \mapsto l_y$ is an isomorphism, that is, $\ell^1(\mathbb{N})^* \cong \ell^\infty(\mathbb{N})$. A similar argument shows $\ell^p(\mathbb{N})^* \cong \ell^q(\mathbb{N})$, $1 \le p < \infty$ (Problem 4.20). One usually identifies $\ell^p(\mathbb{N})^*$ with $\ell^q(\mathbb{N})$ using this canonical isomorphism and

simply writes $\ell^p(\mathbb{N})^* = \ell^q(\mathbb{N})$. In the case $p = \infty$ this is not true, as we will see soon. ⋄

It turns out that many questions are easier to handle after applying a linear functional $\ell \in X^*$. For example, suppose $x(t)$ is a function $\mathbb{R} \to X$ (or $\mathbb{C} \to X$), then $\ell(x(t))$ is a function $\mathbb{R} \to \mathbb{C}$ (respectively $\mathbb{C} \to \mathbb{C}$) for any $\ell \in X^*$. So to investigate $\ell(x(t))$ we have all tools from real/complex analysis at our disposal. But how do we translate this information back to $x(t)$? Suppose we have $\ell(x(t)) = \ell(y(t))$ for all $\ell \in X^*$. Can we conclude $x(t) = y(t)$? The answer is yes and will follow from the Hahn–Banach theorem.

We first prove the real version from which the complex one then follows easily.

Theorem 4.13 (Hahn–Banach, real version)**.** *Let X be a real vector space and $\varphi : X \to \mathbb{R}$ a convex function (i.e., $\varphi(\lambda x+(1-\lambda)y) \le \lambda\varphi(x)+(1-\lambda)\varphi(y)$ for $\lambda \in (0,1)$).*

If ℓ is a linear functional defined on some subspace $Y \subset X$ which satisfies $\ell(y) \le \varphi(y)$, $y \in Y$, then there is an extension $\overline{\ell}$ to all of X satisfying $\overline{\ell}(x) \le \varphi(x)$, $x \in X$.

Proof. Let us first try to extend ℓ in just one direction: Take $x \notin Y$ and set $\tilde{Y} = \operatorname{span}\{x, Y\}$. If there is an extension $\tilde{\ell}$ to $\tilde{Y}$ it must clearly satisfy

$$\tilde{\ell}(y + \alpha x) = \ell(y) + \alpha\tilde{\ell}(x).$$

So all we need to do is to choose $\tilde{\ell}(x)$ such that $\tilde{\ell}(y + \alpha x) \le \varphi(y + \alpha x)$. But this is equivalent to

$$\sup_{\alpha>0, y\in Y} \frac{\varphi(y - \alpha x) - \ell(y)}{-\alpha} \le \tilde{\ell}(x) \le \inf_{\alpha>0, y\in Y} \frac{\varphi(y + \alpha x) - \ell(y)}{\alpha}$$

and is hence only possible if

$$\frac{\varphi(y_1 - \alpha_1 x) - \ell(y_1)}{-\alpha_1} \le \frac{\varphi(y_2 + \alpha_2 x) - \ell(y_2)}{\alpha_2}$$

for every $\alpha_1, \alpha_2 > 0$ and $y_1, y_2 \in Y$. Rearranging this last equations we see that we need to show

$$\alpha_2\ell(y_1) + \alpha_1\ell(y_2) \le \alpha_2\varphi(y_1 - \alpha_1 x) + \alpha_1\varphi(y_2 + \alpha_2 x).$$

Starting with the left-hand side we have

$$\begin{aligned}
\alpha_2\ell(y_1) + \alpha_1\ell(y_2) &= (\alpha_1 + \alpha_2)\ell\left(\lambda y_1 + (1-\lambda)y_2\right) \\
&\le (\alpha_1 + \alpha_2)\varphi\left(\lambda y_1 + (1-\lambda)y_2\right) \\
&= (\alpha_1 + \alpha_2)\varphi\left(\lambda(y_1 - \alpha_1 x) + (1-\lambda)(y_2 + \alpha_2 x)\right) \\
&\le \alpha_2\varphi(y_1 - \alpha_1 x) + \alpha_1\varphi(y_2 + \alpha_2 x),
\end{aligned}$$

where $\lambda = \frac{\alpha_2}{\alpha_1+\alpha_2}$. Hence one dimension works.

To finish the proof we appeal to Zorn's lemma (see Appendix A): Let E be the collection of all extensions $\tilde{\ell}$ satisfying $\tilde{\ell}(x) \leq \varphi(x)$. Then E can be partially ordered by inclusion (with respect to the domain) and every linear chain has an upper bound (defined on the union of all domains). Hence there is a maximal element $\overline{\ell}$ by Zorn's lemma. This element is defined on X, since if it were not, we could extend it as before contradicting maximality. □

Note that linearity gives us a corresponding lower bound $-\varphi(-x) \leq \overline{\ell}(x)$, $x \in X$, for free. In particular, if $\varphi(x) = \varphi(-x)$ then $|\overline{\ell}(x)| \leq \varphi(x)$.

Theorem 4.14 (Hahn–Banach, complex version)**.** *Let X be a complex vector space and $\varphi : X \to \mathbb{R}$ a convex function satisfying $\varphi(\alpha x) \leq \varphi(x)$ if $|\alpha| = 1$.*

If ℓ is a linear functional defined on some subspace $Y \subset X$ which satisfies $|\ell(y)| \leq \varphi(y)$, $y \in Y$, then there is an extension $\overline{\ell}$ to all of X satisfying $|\overline{\ell}(x)| \leq \varphi(x)$, $x \in X$.

Proof. Set $\ell_r = \mathrm{Re}(\ell)$ and observe

$$\ell(x) = \ell_r(x) - \mathrm{i}\ell_r(\mathrm{i}x).$$

By our previous theorem, there is a real linear extension $\overline{\ell}_r$ satisfying $\overline{\ell}_r(x) \leq \varphi(x)$. Now set $\overline{\ell}(x) = \overline{\ell}_r(x) - \mathrm{i}\overline{\ell}_r(\mathrm{i}x)$. Then $\overline{\ell}(x)$ is real linear and by $\overline{\ell}(\mathrm{i}x) = \overline{\ell}_r(\mathrm{i}x) + \mathrm{i}\overline{\ell}_r(x) = \mathrm{i}\overline{\ell}(x)$ also complex linear. To show $|\overline{\ell}(x)| \leq \varphi(x)$ we abbreviate $\alpha = \frac{\overline{\ell}(x)^*}{|\overline{\ell}(x)|}$ and use

$$|\overline{\ell}(x)| = \alpha\overline{\ell}(x) = \overline{\ell}(\alpha x) = \overline{\ell}_r(\alpha x) \leq \varphi(\alpha x) \leq \varphi(x),$$

which finishes the proof. □

Note that $\varphi(\alpha x) \leq \varphi(x)$, $|\alpha| = 1$ is in fact equivalent to $\varphi(\alpha x) = \varphi(x)$, $|\alpha| = 1$.

If ℓ is a bounded linear functional defined on some subspace, the choice $\varphi(x) = \|\ell\|\|x\|$ implies:

Corollary 4.15. *Let X be a normed space and let ℓ be a bounded linear functional defined on some subspace $Y \subseteq X$. Then there is an extension $\overline{\ell} \in X^*$ preserving the norm.*

Example 4.14. Note that in a Hilbert space this result is trivial: First of all there is a unique extension to $\overline{Y}$ by Theorem 1.17. Now set $\bar{\ell} = 0$ on $Y^\perp$. Moreover, any other extension is of the form $\bar{\ell} + \ell_1$, where ℓ_1 vanishes on Y. Then $\|\bar{\ell} + \ell_1\|^2 = \|\ell\|^2 + \|\ell_1\|^2$ and the norm will increase unless $\ell_1 = 0$. ⋄

Example 4.15. In a Banach space this extension will in general not be unique: Consider $X := \ell^1(\mathbb{N})$ and $\ell(x) := x_1$ on $Y := \operatorname{span}\{\delta^1\}$. Then by Example 4.13 any extension is of the form $\bar{\ell} = l_y$ with $y \in \ell^\infty(\mathbb{N})$ and $y_1 = 1$, $\|y\|_\infty \leq 1$. (Sometimes it still might be unique: Problems 4.14 and 4.15). $\diamond$

Moreover, we can now easily prove our anticipated result

Corollary 4.16. *Let X be a normed space and $x \in X$ fixed. Suppose $\ell(x) = 0$ for all ℓ in some total subset $Y \subseteq X^*$. Then $x = 0$.*

Proof. Clearly, if $\ell(x) = 0$ holds for all ℓ in some total subset, this holds for all $\ell \in X^*$. If $x \neq 0$ we can construct a bounded linear functional on $\operatorname{span}\{x\}$ by setting $\ell(\alpha x) = \alpha$ and extending it to X^* using the previous corollary. But this contradicts our assumption. $\square$

Example 4.16. Let us return to our example $\ell^\infty(\mathbb{N})$. Let $c(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$ be the subspace of convergent sequences. Set

$$l(x) = \lim_{n\to\infty} x_n, \qquad x \in c(\mathbb{N}), \tag{4.10}$$

then l is bounded since

$$|l(x)| = \lim_{n\to\infty} |x_n| \leq \|x\|_\infty. \tag{4.11}$$

Hence we can extend it to $\ell^\infty(\mathbb{N})$ by Corollary 4.15. Then $l(x)$ cannot be written as $l(x) = l_y(x)$ for some $y \in \ell^1(\mathbb{N})$ (as in (4.9)) since $y_n = l(\delta^n) = 0$ shows $y = 0$ and hence $\ell_y = 0$. The problem is that $\overline{\operatorname{span}\{\delta^n\}} = c_0(\mathbb{N}) \neq \ell^\infty(\mathbb{N})$, where $c_0(\mathbb{N})$ is the subspace of sequences converging to 0.

Moreover, there is also no other way to identify $\ell^\infty(\mathbb{N})^*$ with $\ell^1(\mathbb{N})$, since $\ell^1(\mathbb{N})$ is separable whereas $\ell^\infty(\mathbb{N})$ is not. This will follow from Lemma 4.21 (iii) below. $\diamond$

Another useful consequence is

Corollary 4.17. *Let $Y \subseteq X$ be a subspace of a normed vector space and let $x_0 \in X \setminus \overline{Y}$. Then there exists an $\ell \in X^*$ such that (i) $\ell(y) = 0$, $y \in Y$, (ii) $\ell(x_0) = \operatorname{dist}(x_0, Y)$, and (iii) $\|\ell\| = 1$.*

Proof. Replacing Y by $\overline{Y}$ we see that it is no restriction to assume that Y is closed. (Note that $x_0 \in X \setminus \overline{Y}$ if and only if $\operatorname{dist}(x_0, Y) > 0$.) Let $\tilde{Y} = \operatorname{span}\{x_0, Y\}$. Since every element of $\tilde{Y}$ can be uniquely written as $y + \alpha x_0$ we can define

$$\ell(y + \alpha x_0) = \alpha \operatorname{dist}(x_0, Y).$$

By construction ℓ is linear on $\tilde{Y}$ and satisfies (i) and (ii). Moreover, by $\operatorname{dist}(x_0, Y) \leq \|x_0 - \frac{-y}{\alpha}\|$ for every $y \in Y$ we have

$$|\ell(y + \alpha x_0)| = |\alpha| \operatorname{dist}(x_0, Y) \leq \|y + \alpha x_0\|, \quad y \in Y.$$

Hence $\|\ell\| \leq 1$ and there is an extension to X by Corollary 4.15. To see that the norm is in fact equal to one, take a sequence $y_n \in Y$ such that $\operatorname{dist}(x_0, Y) \geq (1 - \frac{1}{n})\|x_0 + y_n\|$. Then

$$|\ell(y_n + x_0)| = \operatorname{dist}(x_0, Y) \geq (1 - \frac{1}{n})\|y_n + x_0\|$$

establishing (iii). □

Two more straightforward consequences of the last corollary are also worthwhile noting:

Corollary 4.18. *Let $Y \subseteq X$ be a subspace of a normed vector space. Then $x \in \overline{Y}$ if and only if $\ell(x) = 0$ for every $\ell \in X^*$ which vanishes on Y.*

Corollary 4.19. *Let Y be a closed subspace and let $\{x_j\}_{j=1}^n$ be a linearly independent subset of X. If $Y \cap \operatorname{span}\{x_j\}_{j=1}^n = \{0\}$, then there exists a **biorthogonal system** $\{\ell_j\}_{j=1}^n \subset X^*$ such that $\ell_j(x_k) = 0$ for $j \neq k$, $\ell_j(x_j) = 1$ and $\ell(y) = 0$ for $y \in Y$.*

Proof. Fix j_0. Since $Y_{j_0} = Y \dotplus \operatorname{span}\{x_j\}_{1 \leq j \leq n; j \neq j_0}$ is closed (Corollary 1.20), $x_{j_0} \notin Y_{j_0}$ implies $\operatorname{dist}(x_{j_0}, Y_{j_0}) > 0$ and existence of ℓ_{j_0} follows from Corollary 4.17. □

If we take the **bidual** (or **double dual**) X^{**} of a normed space X, then the Hahn–Banach theorem tells us, that X can be identified with a subspace of X^{**}. In fact, consider the linear map $J : X \to X^{**}$ defined by $J(x)(\ell) = \ell(x)$ (i.e., $J(x)$ is evaluation at x). Then

Theorem 4.20. *Let X be a normed space. Then $J : X \to X^{**}$ is isometric (norm preserving).*

Proof. Fix $x_0 \in X$. By $|J(x_0)(\ell)| = |\ell(x_0)| \leq \|\ell\|_* \|x_0\|$ we have at least $\|J(x_0)\|_{**} \leq \|x_0\|$. Next, by Hahn–Banach there is a linear functional ℓ_0 with norm $\|\ell_0\|_* = 1$ such that $\ell_0(x_0) = \|x_0\|$. Hence $|J(x_0)(\ell_0)| = |\ell_0(x_0)| = \|x_0\|$ shows $\|J(x_0)\|_{**} = \|x_0\|$. □

Example 4.17. This gives another quick way of showing that a normed space has a completion: Take $\bar{X} = \overline{J(X)} \subseteq X^{**}$ and recall that a dual space is always complete (Theorem 1.18). ⋄

Thus $J : X \to X^{**}$ is an isometric embedding. In many cases we even have $J(X) = X^{**}$ and X is called **reflexive** in this case.

Example 4.18. The Banach spaces $\ell^p(\mathbb{N})$ with $1 < p < \infty$ are reflexive: Identify $\ell^p(\mathbb{N})^*$ with $\ell^q(\mathbb{N})$ (cf. Problem 4.20) and choose $z \in \ell^p(\mathbb{N})^{**}$. Then

there is some $x \in \ell^p(\mathbb{N})$ such that

$$z(y) = \sum_{j\in\mathbb{N}} y_j x_j, \qquad y \in \ell^q(\mathbb{N}) \cong \ell^p(\mathbb{N})^*.$$

But this implies $z(y) = y(x)$, that is, $z = J(x)$, and thus J is surjective. (Warning: It does not suffice to just argue $\ell^p(\mathbb{N})^{**} \cong \ell^q(\mathbb{N})^* \cong \ell^p(\mathbb{N})$.)

However, ℓ^1 is not reflexive since $\ell^1(\mathbb{N})^* \cong \ell^\infty(\mathbb{N})$ but $\ell^\infty(\mathbb{N})^* \ncong \ell^1(\mathbb{N})$ as noted earlier. Things get even a bit more explicit if we look at $c_0(\mathbb{N})$, where we can identify (cf. Problem 4.21) $c_0(\mathbb{N})^*$ with $\ell^1(\mathbb{N})$ and $c_0(\mathbb{N})^{**}$ with $\ell^\infty(\mathbb{N})$. Under this identification $J(c_0(\mathbb{N})) = c_0(\mathbb{N}) \subseteq \ell^\infty(\mathbb{N})$. ⋄

Example 4.19. By the same argument, every Hilbert space is reflexive. In fact, by the Riesz lemma we can identify $\mathfrak{H}^*$ with $\mathfrak{H}$ via the (conjugate linear) map $x \mapsto \langle x, .\rangle$. Taking $z \in \mathfrak{H}^{**}$ we have, again by the Riesz lemma, that $z(y) = \langle\langle x, .\rangle, \langle y, .\rangle\rangle_{\mathfrak{H}^*} = \langle x, y\rangle^* = \langle y, x\rangle = J(x)(y)$. ⋄

Lemma 4.21. *Let X be a Banach space.*

(i) *If X is reflexive, so is every closed subspace.*

(ii) *X is reflexive if and only if X^* is.*

(iii) *If X^* is separable, so is X.*

Proof. (i) Let Y be a closed subspace. Denote by $j : Y \hookrightarrow X$ the natural inclusion and define $j_{**} : Y^{**} \to X^{**}$ via $(j_{**}(y''))(\ell) = y''(\ell|_Y)$ for $y'' \in Y^{**}$ and $\ell \in X^*$. Note that j_{**} is isometric by Corollary 4.15. Then

$$\begin{array}{ccc} X & \xrightarrow{J_X} & X^{**} \\ j\uparrow & & \uparrow j_{**} \\ Y & \xrightarrow[J_Y]{} & Y^{**} \end{array}$$

commutes. In fact, we have $j_{**}(J_Y(y))(\ell) = J_Y(y)(\ell|_Y) = \ell(y) = J_X(y)(\ell)$. Moreover, since J_X is surjective, for every $y'' \in Y^{**}$ there is an $x \in X$ such that $j_{**}(y'') = J_X(x)$. Since $j_{**}(y'')(\ell) = y''(\ell|_Y)$ vanishes on all $\ell \in X^*$ which vanish on Y, so does $\ell(x) = J_X(x)(\ell) = j_{**}(y'')(\ell)$ and thus $x \in Y$ by Corollary 4.18. That is, $j_{**}(Y^{**}) = J_X(Y)$ and $J_Y = j \circ J_X \circ j_{**}^{-1}$ is surjective.

(ii) Suppose X is reflexive. Then the two maps

$$\begin{array}{rccc} (J_X)_* : & X^* & \to & X^{***} \\ & x' & \mapsto & x' \circ J_X^{-1} \end{array} \qquad \begin{array}{rccc} (J_X)^* : & X^{***} & \to & X^* \\ & x''' & \mapsto & x''' \circ J_X \end{array}$$

are inverse of each other. Moreover, fix $x'' \in X^{**}$ and let $x = J_X^{-1}(x'')$. Then $J_{X^*}(x')(x'') = x''(x') = J(x)(x') = x'(x) = x'(J_X^{-1}(x''))$, that is $J_{X^*} = (J_X)_*$ respectively $(J_{X^*})^{-1} = (J_X)^*$, which shows X^* reflexive if X reflexive. To see the converse, observe that X^* reflexive implies X^{**} reflexive and hence $J_X(X) \cong X$ is reflexive by (i).

(iii) Let $\{\ell_n\}_{n=1}^\infty$ be a dense set in X^*. Then we can choose $x_n \in X$ such that $\|x_n\| = 1$ and $\ell_n(x_n) \geq \|\ell_n\|/2$. We will show that $\{x_n\}_{n=1}^\infty$ is total in X. If it were not, we could find some $x \in X \setminus \overline{\operatorname{span}\{x_n\}_{n=1}^\infty}$ and hence there is a functional $\ell \in X^*$ as in Corollary 4.17. Choose a subsequence $\ell_{n_k} \to \ell$. Then

$$\|\ell - \ell_{n_k}\| \geq |(\ell - \ell_{n_k})(x_{n_k})| = |\ell_{n_k}(x_{n_k})| \geq \|\ell_{n_k}\|/2,$$

which implies $\ell_{n_k} \to 0$ and contradicts $\|\ell\| = 1$. □

If X is reflexive, then the converse of (iii) is also true (since $X \cong X^{**}$ separable implies X^* separable), but in general this fails as the example $\ell^1(\mathbb{N})^* \cong \ell^\infty(\mathbb{N})$ shows. In fact, this can be used to show that a separable space is not reflexive, by showing that its dual is not separable.

Example 4.20. The space $C(I)$ is not reflexive. To see this observe that the dual space contains point evaluations $\ell_{x_0}(f) := f(x_0)$, $x_0 \in I$. Moreover, for $x_0 \neq x_1$ we have $\|\ell_{x_0} - \ell_{x_1}\| = 2$ and hence $C(I)^*$ is not separable. You should appreciate the fact that it was not necessary to know the full dual space which is quite intricate (see Theorem 12.5). ⋄

Note that the product of two reflexive spaces is also reflexive. In fact, this even holds for countable products — Problem 4.23.

Problem 4.14. *Let $X := \mathbb{C}^3$ equipped with the norm $|(x, y, z)|_1 := |x| + |y| + |z|$ and $Y := \{(x, y, z)|x + y = 0,\ z = 0\}$. Find at least two extensions of $\ell(x, y, z) := x$ from Y to X which preserve the norm. What if we take $Y := \{(x, y, z)|x + y = 0\}$?*

Problem 4.15. *Show that the extension from Corollary 4.15 is unique if X^* is strictly convex. (Hint: Problem 1.12.)*

Problem* 4.16. *Let X be some normed space. Show that*

$$\|x\| = \sup_{\ell \in V,\, \|\ell\|=1} |\ell(x)|, \tag{4.12}$$

where $V \subset X^*$ is some dense subspace. Show that equality is attained if $V = X^*$.

Problem 4.17. *Let X be some normed space. By definition we have*

$$\|\ell\| = \sup_{x \in X, \|x\|=1} |\ell(x)|$$

*for every $\ell \in X^*$. One calls $\ell \in X^*$ **norm-attaining**, if the supremum is attained, that is, there is some $x \in X$ such that $\|\ell\| = |\ell(x)|$.*

Show that in a reflexive Banach space every linear functional is norm-attaining. Give an example of a linear functional which is not norm-attaining. For uniqueness see Problem 5.31. (Hint: For the first part apply the previous problem to X^*. For the second part consider Problem 4.21 below.)

Problem 4.18. *Let X, Y be some normed spaces and $A : \mathfrak{D}(A) \subseteq X \to Y$. Show*

$$\|A\| = \sup_{x\in X,\, \|x\|=1;\, \ell\in V,\, \|\ell\|=1} |\ell(Ax)|, \tag{4.13}$$

where $V \subset Y^*$ is a dense subspace.

Problem* 4.19. *Show that $\|l_y\| = \|y\|_q$, where $l_y \in \ell^p(\mathbb{N})^*$ as defined in (4.9). (Hint: Choose $x \in \ell^p$ such that $x_n y_n = |y_n|^q$.)*

Problem* 4.20. *Show that every $l \in \ell^p(\mathbb{N})^*$, $1 \le p < \infty$, can be written as*

$$l(x) = \sum_{n\in\mathbb{N}} y_n x_n$$

with some $y \in \ell^q(\mathbb{N})$. (Hint: To see $y \in \ell^q(\mathbb{N})$ consider x^N defined such that $x_n^N = |y_n|^q/y_n$ for $n \le N$ with $y_n \neq 0$ and $x_n^N = 0$ else. Now look at $|l(x^N)| \le \|l\|\|x^N\|_p$.)

Problem* 4.21. *Let $c_0(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$ be the subspace of sequences which converge to 0, and $c(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$ the subspace of convergent sequences.*

(i) *Show that $c_0(\mathbb{N})$, $c(\mathbb{N})$ are both Banach spaces and that $c(\mathbb{N}) = \operatorname{span}\{c_0(\mathbb{N}), e\}$, where $e = (1, 1, 1, \dots) \in c(\mathbb{N})$.*

(ii) *Show that every $l \in c_0(\mathbb{N})^*$ can be written as*

$$l(a) = \sum_{n\in\mathbb{N}} b_n a_n$$

with some $b \in \ell^1(\mathbb{N})$ which satisfies $\|b\|_1 = \|\ell\|$.

(iii) *Show that every $l \in c(\mathbb{N})^*$ can be written as*

$$l(a) = \sum_{n\in\mathbb{N}} b_n a_n + b_0 \lim_{n\to\infty} a_n$$

with some $b \in \ell^1(\mathbb{N})$ which satisfies $|b_0| + \|b\|_1 = \|\ell\|$.

Problem 4.22. *Let $u_n \in X$ be a Schauder basis and suppose the complex numbers c_n satisfy $|c_n| \le c\|u_n\|$. Is there a bounded linear functional $\ell \in X^*$ with $\ell(u_n) = c_n$? (Hint: Consider e.g. $X = \ell^2(\mathbb{Z})$.)*

Problem* 4.23. *Let $X := \bigtimes_{p,j\in\mathbb{N}} X_j$ be defined as in Problem 1.38 and let $\frac{1}{p} + \frac{1}{q} = 1$. Show that for $1 \le p < \infty$ we have $X^* \cong \bigtimes_{q,j\in\mathbb{N}} X_j^*$, where the identification is given by*

$$y(x) = \sum_{j\in\mathbb{N}} y_j(x_j), \qquad x = (x_j)_{j\in\mathbb{N}} \in \bigtimes_{p,j\in\mathbb{N}} X_j, \quad y = (y_j)_{j\in\mathbb{N}} \in \bigtimes_{q,j\in\mathbb{N}} X_j^*.$$

Moreover, if all X_j are reflexive, so is X.

Problem 4.24 (Banach limit)**.** *Let $\mathfrak{c}(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$ be the subspace of all bounded sequences for which the limit of the Cesàro means*

$$L(x) := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} x_k$$

exists. Note that $c(\mathbb{N}) \subseteq \mathfrak{c}(\mathbb{N})$ and $L(x) = \lim_{n\to\infty} x_n$ for $x \in c(\mathbb{N})$.

Show that L can be extended to all of $\ell^\infty(\mathbb{N})$ such that

(i) *L is linear,*

(ii) *$|L(x)| \le \|x\|_\infty$,*

(iii) *$L(Sx) = L(x)$ where $(Sx)_n = x_{n+1}$ is the shift operator,*

(iv) *$L(x) \ge 0$ when $x_n \ge 0$ for all n,*

(v) *$\liminf_n x_n \le L(x) \le \limsup x_n$ for all real-valued sequences.*

(Hint: Of course existence follows from Hahn–Banach and (i), (ii) will come for free. Also (iii) will be inherited from the construction. For (iv) note that the extension can assumed to be real-valued and investigate $L(e - x)$ for $x \ge 0$ with $\|x\|_\infty = 1$ where $e = (1, 1, 1, \dots)$. (v) then follows from (iv).)

Problem* 4.25. *Show that a finite dimensional subspace M of a Banach space X can be complemented. (Hint: Start with a basis $\{x_j\}$ for M and choose a corresponding dual basis $\{\ell_k\}$ with $\ell_k(x_j) = \delta_{j,k}$ which can be extended to X^*.)*

4.3. The adjoint operator

Given two normed spaces X and Y and a bounded operator $A \in \mathscr{L}(X, Y)$ we can define its **adjoint** $A' : Y^* \to X^*$ via $A'y' = y' \circ A$, $y' \in Y^*$. It is immediate that A' is linear and boundedness follows from

$$\begin{aligned}\|A'\| &= \sup_{y'\in Y^*:\, \|y'\|=1} \|A'y'\| = \sup_{y'\in Y^*:\, \|y'\|=1} \left(\sup_{x\in X:\, \|x\|=1} |(A'y')(x)| \right) \\ &= \sup_{y'\in Y^*:\, \|y'\|=1} \left(\sup_{x\in X:\, \|x\|=1} |y'(Ax)| \right) = \sup_{x\in X:\, \|x\|=1} \|Ax\| = \|A\|,\end{aligned}$$

where we have used Problem 4.16 to obtain the fourth equality. In summary,

Theorem 4.22. *Let $A \in \mathscr{L}(X, Y)$, then $A' \in \mathscr{L}(Y^*, X^*)$ with $\|A\| = \|A'\|$.*

Note that for $A, B \in \mathscr{L}(X, Y)$ and $\alpha, \beta \in \mathbb{C}$ we have

$$(\alpha A + \beta B)' = \alpha A' + \beta B' \tag{4.14}$$

and for $A \in \mathscr{L}(X, Y)$ and $B \in \mathscr{L}(Y, Z)$ we have

$$(BA)' = A'B' \tag{4.15}$$

which is immediate from the definition. Moreover, note that $(\mathbb{I}_X)' = \mathbb{I}_{X^*}$ which shows that if A is invertible then so is A' is with

$$(A^{-1})' = (A')^{-1}. \tag{4.16}$$

That A is invertible if A' is will follow from Theorem 4.26 below.

Example 4.21. Given a Hilbert space $\mathfrak{H}$ we have the conjugate linear isometry $C : \mathfrak{H} \to \mathfrak{H}^*$, $f \mapsto \langle f, \cdot \rangle$. Hence for given $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ we have $A'C_2 f = \langle f, A\,\cdot\rangle = \langle A^* f, \cdot\rangle$ which shows $A' = C_1 A^* C_2^{-1}$. ⋄

Example 4.22. Let $X := Y := \ell^p(\mathbb{N})$, $1 \le p < \infty$, such that $X^* = \ell^q(\mathbb{N})$, $\frac{1}{p} + \frac{1}{q} = 1$. Consider the right shift $R \in \mathscr{L}(\ell^p(\mathbb{N}))$ given by

$$Rx := (0, x_1, x_2, \dots).$$

Then for $y' \in \ell^q(\mathbb{N})$

$$y'(Rx) = \sum_{j=1}^{\infty} y'_j (Rx)_j = \sum_{j=2}^{\infty} y'_j x_{j-1} = \sum_{j=1}^{\infty} y'_{j+1} x_j$$

which shows $(R'y')_k = y'_{k+1}$ upon choosing $x = \delta^k$. Hence $R' = L$ is the left shift: $Ly := (y_2, y_3, \dots)$. Similarly, $L' = R$. ⋄

Example 4.23. Recall that an operator $K \in \mathscr{L}(X, Y)$ is called a **finite rank operator** if its range is finite dimensional. The dimension of its range $\operatorname{rank}(K) := \dim \operatorname{Ran}(K)$ is called the **rank** of K. Choosing a basis $\{y_j = Kx_j\}_{j=1}^n$ for $\operatorname{Ran}(K)$ and a corresponding dual basis $\{y'_j\}_{j=1}^n$ (cf. Problem 4.25), then $x'_j := K'y'_j$ is a dual basis for x_j and

$$Kx = \sum_{j=1}^{n} y'_j(Kx) y_j = \sum_{j=1}^{n} x'_j(x) y_j, \qquad K'y' = \sum_{j=1}^{n} y'(y_j) x'_j.$$

In particular, $\operatorname{rank}(K) = \operatorname{rank}(K')$. ⋄

Of course we can also consider the doubly adjoint operator A''. Then a simple computation

$$A''(J_X(x))(y') = J_X(x)(A'y') = (A'y')(x) = y'(Ax) = J_Y(Ax)(y') \tag{4.17}$$

shows that the following diagram commutes

$$\begin{array}{ccc} X & \xrightarrow{A} & Y \\ {\scriptstyle J_X}\downarrow & & \downarrow{\scriptstyle J_Y} \\ X^{**} & \xrightarrow[A'']{} & Y^{**} \end{array}$$

Consequently

$$A'' \restriction_{\operatorname{Ran}(J_X)} = J_Y A J_X^{-1}, \qquad A = J_Y^{-1} A'' J_X. \tag{4.18}$$

Hence, regarding X as a subspace $J_X(X) \subseteq X^{**}$ and Y as a subspace $J_Y(Y) \subseteq Y^{**}$, then A'' is an extension of A to X^{**} but with values in Y^{**}.

In particular, note that $B \in \mathscr{L}(Y^*, X^*)$ is the adjoint of some other operator $B = A'$ if and only if $B'(J_X(X)) = A''(J_X(X)) \subseteq J_Y(Y)$ (for the converse note that $A := J_Y^{-1} B' J_X$ will do the trick). This can be used to show that not every operator is an adjoint (Problem 4.26).

Theorem 4.23 (Schauder)**.** *Suppose X, Y are Banach spaces and $A \in \mathscr{L}(X, Y)$. Then A is compact if and only if A' is.*

Proof. If A is compact, then $A(B_1^X(0))$ is relatively compact and hence $K = \overline{A(B_1^X(0))}$ is a compact metric space. Let $y_n' \in Y^*$ be a bounded sequence and consider the family of functions $f_n := y_n'|_K \in C(K)$. Then this family is bounded and equicontinuous since

$$|f_n(y_1) - f_n(y_2)| \le \|y_n'\| \|y_1 - y_2\| \le C\|y_1 - y_2\|.$$

Hence the Arzelà–Ascoli theorem (Theorem B.39) implies existence of a uniformly converging subsequence f_{n_j}. For this subsequence we have

$$\|A'y_{n_j}' - A'y_{n_k}'\| \le \sup_{x \in B_1^X(0)} |y_{n_j}'(Ax) - y_{n_k}'(Ax)| = \|f_{n_j} - f_{n_k}\|_\infty$$

since $A(B_1^X(0)) \subseteq K$ is dense. Thus y_{n_j}' is the required subsequence and A' is compact.

To see the converse note that if A' is compact then so is A'' by the first part and hence also $A = J_Y^{-1} A'' J_X$. □

Finally we discuss the relation between solvability of $Ax = y$ and the corresponding adjoint equation $A'y' = x'$. To this end we need the analog of the orthogonal complement of a set. Given subsets $M \subseteq X$ and $N \subseteq X^*$ we define their **annihilator** as

$$\begin{aligned} M^\perp &:= \{\ell \in X^* | \ell(x) = 0 \, \forall x \in M\} = \{\ell \in X^* | M \subseteq \operatorname{Ker}(\ell)\} \\ &= \bigcap_{x \in M} \{\ell \in X^* | \ell(x) = 0\} = \bigcap_{x \in M} \{x\}^\perp, \\ N_\perp &:= \{x \in X | \ell(x) = 0 \, \forall \ell \in N\} = \bigcap_{\ell \in N} \operatorname{Ker}(\ell) = \bigcap_{\ell \in N} \{\ell\}_\perp. \end{aligned} \tag{4.19}$$

In particular, $\{\ell\}_\perp = \operatorname{Ker}(\ell)$ while $\{x\}^\perp = \operatorname{Ker}(J(x))$ (with $J : X \hookrightarrow X^{**}$ the canonical embedding). Note $\{0\}^\perp = X^*$ and $\{0\}_\perp = X$.

Example 4.24. In a Hilbert space the annihilator is simply the orthogonal complement. ◇

The following properties are immediate from the definition (by linearity and continuity):

- $M^\perp$ is a closed subspace of X^* and $M^\perp = (\overline{\operatorname{span}(M)})^\perp$.
- $N_\perp$ is a closed subspace of X and $N_\perp = (\overline{\operatorname{span}(N)})_\perp$.

Note also that

$$\begin{aligned} \overline{\operatorname{span}(M)} = X \quad &\Leftrightarrow \quad M^\perp = \{0\}, \\ \overline{\operatorname{span}(N)} = X^* \quad &\Rightarrow \quad N_\perp = \{0\} \end{aligned} \tag{4.20}$$

by Corollary 4.17 and Corollary 4.16, respectively. The converse of the last statement is wrong in general.

Example 4.25. Consider $X := \ell^1(\mathbb{N})$ and $N := \{\delta^n\}_{n\in\mathbb{N}} \subset \ell^\infty(\mathbb{N}) \cong X^*$. Then $\overline{\operatorname{span}(N)} = c_0(\mathbb{N})$ but $N_\perp = \{0\}$. ⋄

Lemma 4.24. *We have* $(M^\perp)_\perp = \overline{\operatorname{span}(M)}$ *and* $(N_\perp)^\perp \supseteq \overline{\operatorname{span}(N)}$.

Proof. By the preceding remarks we can assume M, N to be closed subspaces. The first part

$$(M^\perp)_\perp = \{x \in X | \ell(x) = 0 \, \forall \ell \in X^* \text{ with } M \subseteq \operatorname{Ker}(\ell)\} = \overline{\operatorname{span}(M)}$$

is Corollary 4.18 and for the second part one just has to spell out the definition:

$$(N_\perp)^\perp = \{\ell \in X^* | \bigcap_{\tilde\ell \in N} \operatorname{Ker}(\tilde\ell) \subseteq \operatorname{Ker}(\ell)\} \supseteq \overline{\operatorname{span}(N)}. \qquad \square$$

Note that we have equality in the preceding lemma if N is finite dimensional (Problem 4.32). Moreover, with a little more machinery one can show equality if X is reflexive (Problem 5.13). For non-reflexive spaces the inclusion can be strict as the previous example shows.

Warning: Some authors call a set $N \subseteq X^*$ total if $N_\perp = \{0\}$. By the preceding discussion this is equivalent to our definition if X is reflexive, but otherwise might differ.

Furthermore, we have the following analog of (2.28).

Lemma 4.25. *Suppose X, Y are normed spaces and $A \in \mathscr{L}(X,Y)$. Then* $\operatorname{Ran}(A')_\perp = \operatorname{Ker}(A)$ *and* $\operatorname{Ran}(A)^\perp = \operatorname{Ker}(A')$.

Proof. For the first claim observe: $x \in \operatorname{Ker}(A) \Leftrightarrow Ax = 0 \Leftrightarrow \ell(Ax) = 0$, $\forall \ell \in X^* \Leftrightarrow (A'\ell)(x) = 0$, $\forall \ell \in X^* \Leftrightarrow x \in \operatorname{Ran}(A')^\perp$.

For the second claim observe: $\ell \in \operatorname{Ker}(A') \Leftrightarrow A'\ell = 0 \Leftrightarrow (A'\ell)(x) = 0$, $\forall x \in X \Leftrightarrow \ell(Ax) = 0$, $\forall x \in X \Leftrightarrow \ell \in \operatorname{Ran}(A)^\perp$. $\square$

Taking annihilators in these formulas we obtain

$$\operatorname{Ker}(A')_\perp = (\operatorname{Ran}(A)^\perp)_\perp = \overline{\operatorname{Ran}(A)} \tag{4.21}$$

and

$$\operatorname{Ker}(A)^\perp = (\operatorname{Ran}(A')_\perp)^\perp \supseteq \overline{\operatorname{Ran}(A')} \tag{4.22}$$

which raises the question of equality in the latter.

Theorem 4.26 (Closed range)**.** *Suppose X, Y are Banach spaces and $A \in \mathscr{L}(X,Y)$. Then the following items are equivlaent:*

(i) $\operatorname{Ran}(A)$ *is closed.*

(ii) $\operatorname{Ker}(A)^\perp = \operatorname{Ran}(A')$.

(iii) $\operatorname{Ran}(A')$ *is closed.*

(iv) $\operatorname{Ker}(A')_\perp = \operatorname{Ran}(A)$.

Proof. (i) $\Leftrightarrow$ (vi): Immediate from (4.21).

(i) $\Rightarrow$ (ii): Note that if $\ell \in \operatorname{Ran}(A')$ then $\ell = A'(\tilde{\ell}) = \tilde{\ell} \circ A$ vanishes on $\operatorname{Ker}(A)$ and hence $\ell \in \operatorname{Ker}(A)^\perp$. Conversely, if $\ell \in \operatorname{Ker}(A)^\perp$ we can set $\tilde{\ell}(y) = \ell(\tilde{A}^{-1}y)$ for $y \in \operatorname{Ran}(A)$ and extend it to all of Y using Corollary 4.15. Here $\tilde{A} : X/\operatorname{Ker}(A) \to \operatorname{Ran}(A)$ is the induced map (cf. Problem 1.42) which has a bounded inverse by Theorem 4.6. By construction $\ell = A'(\tilde{\ell}) \in \operatorname{Ran}(A')$.

(ii) $\Rightarrow$ (iii): Clear since annihilators are closed.

(iii) $\Rightarrow$ (i): Let $Z = \overline{\operatorname{Ran}(A)}$ and let $\tilde{A} : X \to Z$ be the range restriction of A. Then $\tilde{A}'$ is injective (since $\operatorname{Ker}(\tilde{A}') = \operatorname{Ran}(\tilde{A})^\perp = \{0\}$) and has the same range $\operatorname{Ran}(\tilde{A}') = \operatorname{Ran}(A')$ (since every linear functional in Z^* can be extended to one in Y^* by Corollary 4.15). Hence we can assume $Z = Y$ and hence A' injective without loss of generality.

Suppose $\operatorname{Ran}(A)$ were not closed. Then, given $\varepsilon > 0$ and $0 \le \delta < 1$, by Corollary 4.11 there is some $y \in Y$ such that $\varepsilon\|x\| + \|y - Ax\| > \delta\|y\|$ for all $x \in X$. Hence there is a linear functional $\ell \in Y^*$ such that $\delta \le \|\ell\| \le 1$ and $\|A'\ell\| \le \varepsilon$. Indeed consider $X \oplus Y$ and use Corollary 4.17 to choose $\bar{\ell} \in (X \oplus Y)^*$ such that $\bar{\ell}$ vanishes on the closed set $V := \{(\varepsilon x, Ax) | x \in X\}$, $\|\bar{\ell}\| = 1$, and $\bar{\ell}(0,y) = \operatorname{dist}((0,y),V)$ (note that $(0,y) \notin V$ since $y \neq 0$). Then $\ell(.) = \bar{\ell}(0,.)$ is the functional we are looking for since $\operatorname{dist}((0,y),V) \ge \delta\|y\|$ and $(A'\ell)(x) = \bar{\ell}(0,Ax) = \bar{\ell}(-\varepsilon x, 0) = -\varepsilon\bar{\ell}(x,0)$. Now this allows us to choose ℓ_n with $\|\ell_n\| \to 1$ and $\|A'\ell_n\| \to 0$ such that Corollary 4.10 implies that $\operatorname{Ran}(A')$ is not closed. □

With the help of annihilators we can also describe the dual spaces of subspaces.

Theorem 4.27. *Let M be a closed subspace of a normed space X. Then there are canonical isometries*

$$(X/M)^* \cong M^\perp, \qquad M^* \cong X^*/M^\perp. \tag{4.23}$$

Proof. In the first case the isometry is given by $\ell \mapsto \ell \circ j$, where $j : X \to X/M$ is the quotient map. In the second case $x' + M^\perp \mapsto x'|_M$. The details are easy to check. □

Problem* 4.26. *Let $X := Y := c_0(\mathbb{N})$ and recall that $X^* \cong \ell^1(\mathbb{N})$ and $X^{**} \cong \ell^\infty(\mathbb{N})$. Consider the operator $A \in \mathscr{L}(\ell^1(\mathbb{N}))$ given by*

$$Ax := (\sum_{n\in\mathbb{N}} x_n, 0, \dots).$$

Show that

$$A'x' = (x'_1, x'_1, \dots).$$

Conclude that A is not the adjoint of an operator from $\mathscr{L}(c_0(\mathbb{N}))$.

Problem* 4.27. *Show*

$$\operatorname{Ker}(A') \cong \operatorname{Coker}(A)^*, \qquad \operatorname{Coker}(A') \cong \operatorname{Ker}(A)^*$$

for $A \in \mathscr{L}(X, Y)$ with $\operatorname{Ran}(A)$ closed.

Problem 4.28. *Let X_j be Banach spaces. A sequence of operators $A_j \in \mathscr{L}(X_j, X_{j+1})$*

$$X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \cdots X_n \xrightarrow{A_n} X_{n+1}$$

*is said to be **exact** if $\operatorname{Ran}(A_j) = \operatorname{Ker}(A_{j+1})$ for $1 \le j \le n$. Show that a sequence is exact if and only if the corresponding dual sequence*

$$X_1^* \xleftarrow{A_1'} X_2^* \xleftarrow{A_2'} X_3^* \cdots X_n^* \xleftarrow{A_n'} X_{n+1}^*$$

is exact.

Problem 4.29. *Suppose X is separable. Show that there exists a countable set $N \subset X^*$ with $N_\perp = \{0\}$.*

Problem 4.30. *Show that for $A \in \mathscr{L}(X, Y)$ we have*

$$\operatorname{rank}(A) = \operatorname{rank}(A').$$

Problem* 4.31 (Riesz lemma)**.** *Let X be a normed vector space and $Y \subset X$ some subspace. Show that if $\overline{Y} \neq X$, then for every $\varepsilon \in (0, 1)$ there exists an x_ε with $\|x_\varepsilon\| = 1$ and*

$$\inf_{y\in Y} \|x_\varepsilon - y\| \ge 1 - \varepsilon. \tag{4.24}$$

Note: In a Hilbert space the claim holds with $\varepsilon = 0$ for any normalized x in the orthogonal complement of Y and hence x_ε can be thought of a replacement of an orthogonal vector. (Hint: Choose a $y_\varepsilon \in Y$ which is close to x and look at $x - y_\varepsilon$.)

Problem* 4.32. *Suppose X is a vector space and ℓ, $\ell_1, \dots, \ell_n$ are linear functionals such that $\bigcap_{j=1}^n \operatorname{Ker}(\ell_j) \subseteq \operatorname{Ker}(\ell)$. Then $\ell = \sum_{j=1}^n \alpha_j \ell_j$ for some constants $\alpha_j \in \mathbb{C}$. (Hint: Find a dual basis $x_k \in X$ such that $\ell_j(x_k) = \delta_{j,k}$ and look at $x - \sum_{j=1}^n \ell_j(x) x_j$.)*

Problem 4.33. *Suppose M_1, M_2 are closed subspaces of X. Show*

$$M_1 \cap M_2 = (M_1^\perp + M_2^\perp)_\perp, \qquad M_1^\perp \cap M_2^\perp = (M_1 + M_2)^\perp$$

and

$$(M_1 \cap M_2)^\perp \supseteq \overline{(M_1^\perp + M_2^\perp)}, \qquad (M_1^\perp \cap M_2^\perp)_\perp = \overline{(M_1 + M_2)}.$$

Problem 4.34. *Let us write $\ell_n \overset{*}{\rightharpoonup} \ell$ provided the sequence converges pointwise, that is, $\ell_n(x) \to \ell(x)$ for all $x \in X$. Let $N \subseteq X^*$ and suppose $\ell_n \overset{*}{\rightharpoonup} \ell$ with $\ell_n \in N$. Show that $\ell \in (N_\perp)^\perp$.*

4.4. Weak convergence

In Section 4.2 we have seen that $\ell(x) = 0$ for all $\ell \in X^*$ implies $x = 0$. Now what about convergence? Does $\ell(x_n) \to \ell(x)$ for every $\ell \in X^*$ imply $x_n \to x$? In fact, in a finite dimensional space component-wise convergence is equivalent to convergence. Unfortunately in the infinite dimensional this is no longer true in general:

Example 4.26. Let u_n be an infinite orthonormal set in some Hilbert space. Then $\langle g, u_n\rangle \to 0$ for every g since these are just the expansion coefficients of g which are in $\ell^2(\mathbb{N})$ by Bessel's inequality. Since by the Riesz lemma (Theorem 2.10), every bounded linear functional is of this form, we have $\ell(u_n) \to 0$ for every bounded linear functional. (Clearly u_n does not converge to 0, since $\|u_n\| = 1$.) ⋄

If $\ell(x_n) \to \ell(x)$ for every $\ell \in X^*$ we say that x_n **converges weakly** to x and write

$$\operatorname*{w-lim}_{n\to\infty} x_n = x \quad \text{or} \quad x_n \rightharpoonup x. \tag{4.25}$$

Clearly, $x_n \to x$ implies $x_n \rightharpoonup x$ and hence this notion of convergence is indeed weaker. Moreover, the weak limit is unique, since $\ell(x_n) \to \ell(x)$ and $\ell(x_n) \to \ell(\tilde{x})$ imply $\ell(x - \tilde{x}) = 0$. A sequence x_n is called a **weak Cauchy sequence** if $\ell(x_n)$ is Cauchy (i.e. converges) for every $\ell \in X^*$.

Lemma 4.28. *Let X be a Banach space.*

(i) *$x_n \rightharpoonup x$, $y_n \rightharpoonup y$ and $\alpha_n \to \alpha$ implies $x_n + y_n \rightharpoonup x + y$ and $\alpha_n x_n \rightharpoonup \alpha x$.*

(ii) *$x_n \rightharpoonup x$ implies $\|x\| \le \liminf \|x_n\|$.*

(iii) *Every weak Cauchy sequence x_n is bounded: $\|x_n\| \le C$.*

(iv) *If X is reflexive, then every weak Cauchy sequence converges weakly.*

(v) *A sequence x_n is Cauchy if and only if $\ell(x_n)$ is Cauchy, uniformly for $\ell \in X^*$ with $\|\ell\| = 1$.*

Proof. (i) Follows from $\ell(\alpha_n x_n + y_n) = \alpha_n \ell(x_n) + \ell(y_n) \to \alpha\ell(x) + \ell(y)$. (ii) Choose $\ell \in X^*$ such that $\ell(x) = \|x\|$ (for the limit x) and $\|\ell\| = 1$. Then

$$\|x\| = \ell(x) = \liminf \ell(x_n) \le \liminf \|x_n\|.$$

(iii) For every ℓ we have that $|J(x_n)(\ell)| = |\ell(x_n)| \le C(\ell)$ is bounded. Hence by the uniform boundedness principle we have $\|x_n\| = \|J(x_n)\| \le C$.
(iv) If x_n is a weak Cauchy sequence, then $\ell(x_n)$ converges and we can define $j(\ell) = \lim \ell(x_n)$. By construction j is a linear functional on X^*. Moreover, by (iii) we have $|j(\ell)| \le \sup |\ell(x_n)| \le \|\ell\| \sup \|x_n\| \le C\|\ell\|$ which shows $j \in X^{**}$. Since X is reflexive, $j = J(x)$ for some $x \in X$ and by construction $\ell(x_n) \to J(x)(\ell) = \ell(x)$, that is, $x_n \rightharpoonup x$.
(v) This follows from

$$\|x_n - x_m\| = \sup_{\|\ell\|=1} |\ell(x_n - x_m)|$$

(cf. Problem 4.16). □

Item (ii) says that the norm is sequentially weakly lower semicontinuous (cf. Problem 8.20) while the previous example shows that it is not sequentially weakly continuous (this will in fact be true for any convex function as we will see later). However, bounded linear operators turn out to be sequentially weakly continuous (Problem 4.36).

Example 4.27. Consider $L^2(0,1)$ and recall (see Example 3.8) that

$$u_n(x) = \sqrt{2}\sin(n\pi x), \qquad n \in \mathbb{N},$$

form an ONB and hence $u_n \rightharpoonup 0$. However, $v_n = u_n^2 \rightharpoonup 1$. In fact, one easily computes

$$\langle u_m, v_n\rangle = \frac{\sqrt{2}((-1)^m - 1)}{m\pi} \frac{4k^2}{(m^2 + 4k^2)} \to \frac{\sqrt{2}((-1)^m - 1)}{m\pi} = \langle u_m, 1\rangle$$

and the claim follows from Problem 4.39 since $\|v_n\| = \sqrt{\frac{3}{2}}$. ⋄

Remark: One can equip X with the weakest topology for which all $\ell \in X^*$ remain continuous. This topology is called the **weak topology** and it is given by taking all finite intersections of inverse images of open sets as a base. By construction, a sequence will converge in the weak topology if and only if it converges weakly. By Corollary 4.17 the weak topology is Hausdorff, but it will not be metrizable in general. In particular, sequences do not suffice to describe this topology. Nevertheless we will stick with sequences for now and come back to this more general point of view in Section 5.3.

In a Hilbert space there is also a simple criterion for a weakly convergent sequence to converge in norm (see Theorem 5.19 for a generalization).

Lemma 4.29. *Let $\mathfrak{H}$ be a Hilbert space and let $f_n \rightharpoonup f$. Then $f_n \to f$ if and only if $\limsup \|f_n\| \leq \|f\|$.*

Proof. By (ii) of the previous lemma we have $\lim \|f_n\| = \|f\|$ and hence

$$\|f - f_n\|^2 = \|f\|^2 - 2\mathrm{Re}(\langle f, f_n\rangle) + \|f_n\|^2 \to 0.$$

The converse is straightforward. □

Now we come to the main reason why weakly convergent sequences are of interest: A typical approach for solving a given equation in a Banach space is as follows:

(i) Construct a (bounded) sequence x_n of approximating solutions (e.g. by solving the equation restricted to a finite dimensional subspace and increasing this subspace).

(ii) Use a compactness argument to extract a convergent subsequence.

(iii) Show that the limit solves the equation.

Our aim here is to provide some results for the step (ii). In a finite dimensional vector space the most important compactness criterion is boundedness (Heine–Borel theorem, Theorem B.22). In infinite dimensions this breaks down as we have seen in Theorem 1.11 However, if we are willing to treat convergence for weak convergence, the situation looks much brighter!

Theorem 4.30. *Let X be a reflexive Banach space. Then every bounded sequence has a weakly convergent subsequence.*

Proof. Let x_n be some bounded sequence and consider $Y = \overline{\mathrm{span}\{x_n\}}$. Then Y is reflexive by Lemma 4.21 (i). Moreover, by construction Y is separable and so is Y^* by the remark after Lemma 4.21.

Let ℓ_k be a dense set in Y^*. Then by the usual diagonal sequence argument we can find a subsequence x_{n_m} such that $\ell_k(x_{n_m})$ converges for every k. Denote this subsequence again by x_n for notational simplicity. Then,

$$\begin{aligned}|\ell(x_n) - \ell(x_m)| \leq& |\ell(x_n) - \ell_k(x_n)| + |\ell_k(x_n) - \ell_k(x_m)| \\ &+ |\ell_k(x_m) - \ell(x_m)| \\ \leq& 2C\|\ell - \ell_k\| + |\ell_k(x_n) - \ell_k(x_m)|\end{aligned}$$

shows that $\ell(x_n)$ converges for every $\ell \in \overline{\mathrm{span}\{\ell_k\}} = Y^*$. Thus there is a limit by Lemma 4.28 (iv). □

Note that this theorem breaks down if X is not reflexive.

Example 4.28. Consider the sequence of vectors δ^n (with $\delta^n_n = 1$ and $\delta^n_m = 0$, $n \neq m$) in $\ell^p(\mathbb{N})$, $1 \leq p < \infty$. Then $\delta^n \rightharpoonup 0$ for $1 < p < \infty$. In fact, since every $l \in \ell^p(\mathbb{N})^*$ is of the form $l = l_y$ for some $y \in \ell^q(\mathbb{N})$ we have $l_y(\delta^n) = y_n \to 0$.

If we consider the same sequence in $\ell^1(\mathbb{N})$ there is no weakly convergent subsequence. In fact, since $l_y(\delta^n) \to 0$ for every sequence $y \in \ell^\infty(\mathbb{N})$ with finitely many nonzero entries, the only possible weak limit is zero. On the other hand choosing the constant sequence $y = (1)_{j=1}^\infty$ we see $l_y(\delta^n) = 1 \not\to 0$, a contradiction. ⋄

Example 4.29. Let $X := L^1[-1,1]$. Every bounded integrable φ gives rise to a linear functional

$$\ell_\varphi(f) := \int f(x)\varphi(x)\,dx$$

in $L^1[-1,1]^*$. Take some nonnegative u_1 with compact support, $\|u_1\|_1 = 1$, and set $u_k(x) = k u_1(k\,x)$ (implying $\|u_k\|_1 = 1$). Then we have

$$\int u_k(x)\varphi(x)\,dx \to \varphi(0)$$

(see Problem 10.25) for every continuous φ. Furthermore, if $u_{k_j} \rightharpoonup u$ we conclude

$$\int u(x)\varphi(x)\,dx = \varphi(0).$$

In particular, choosing $\varphi_k(x) = \max(0, 1-k|x|)$ we infer from the dominated convergence theorem

$$1 = \int u(x)\varphi_k(x)\,dx \to \int u(x)\chi_{\{0\}}(x)\,dx = 0,$$

a contradiction.

In fact, u_k *converges* to the Dirac measure centered at 0, which is not in $L^1[-1,1]$. ⋄

Note that the above theorem also shows that in an infinite dimensional reflexive Banach space weak convergence is always weaker than strong convergence since otherwise every bounded sequence had a weakly, and thus by assumption also norm, convergent subsequence contradicting Theorem 1.11. In a non-reflexive space this situation can however occur.

Example 4.30. In $\ell^1(\mathbb{N})$ every weakly convergent sequence is in fact (norm) convergent (such Banach spaces are said to have the **Schur property**). First of all recall that $\ell^1(\mathbb{N})^* \cong \ell^\infty(\mathbb{N})$ and $a^n \rightharpoonup 0$ implies

$$l_b(a^n) = \sum_{k=1}^\infty b_k a^n_k \to 0, \qquad \forall b \in \ell^\infty(\mathbb{N}).$$

Now suppose we could find a sequence $a^n \rightharpoonup 0$ for which $\limsup_n \|a^n\|_1 \geq \varepsilon > 0$. After passing to a subsequence we can assume $\|a^n\|_1 \geq \varepsilon/2$ and after rescaling the vector even $\|a^n\|_1 = 1$. Now weak convergence $a^n \rightharpoonup 0$ implies $a^n_j = l_{\delta^j}(a^n) \to 0$ for every fixed $j \in \mathbb{N}$. Hence the main contribution to the norm of a^n must move towards ∞ and we can find a subsequence n_j and a corresponding increasing sequence of integers k_j such that $\sum_{k_j \leq k < k_{j+1}} |a^{n_j}_k| \geq \frac{2}{3}$. Now set

$$b_k = \operatorname{sign}(a^{n_j}_k), \qquad k_j \leq k < k_{j+1}.$$

Then

$$|l_b(a^{n_j})| \geq \sum_{k_j \leq k < k_{j+1}} |a^{n_j}_k| - \left| \sum_{1 \leq k < k_j;\, k_{j+1} \leq k} b_k a^{n_j}_k \right| \geq \frac{2}{3} - \frac{1}{3} = \frac{1}{3},$$

contradicting $a^{n_j} \rightharpoonup 0$. ⋄

It is also useful to observe that compact operators will turn weakly convergent into (norm) convergent sequences.

Theorem 4.31. *Let $A \in \mathscr{C}(X, Y)$ be compact. Then $x_n \rightharpoonup x$ implies $Ax_n \to Ax$. If X is reflexive the converse is also true.*

Proof. If $x_n \rightharpoonup x$ we have $\sup_n \|x_n\| \leq C$ by Lemma 4.28 (ii). Consequently Ax_n is bounded and we can pass to a subsequence such that $Ax_{n_k} \to y$. Moreover, by Problem 4.36 we even have $y = Ax$ and Lemma B.5 shows $Ax_n \to Ax$.

Conversely, if X is reflexive, then by Theorem 4.30 every bounded sequence x_n has a subsequence $x_{n_k} \rightharpoonup x$ and by assumption $Ax_{n_k} \to x$. Hence A is compact. □

Operators which map weakly convergent sequences to convergent sequences are also called **completely continuous**. However, be warned that some authors use completely continuous for compact operators. By the above theorem every compact operator is completely continuous and the converse also holds in reflexive spaces. However, the last example shows that the identity map in $\ell^1(\mathbb{N})$ is completely continuous but it is clearly not compact by Theorem 1.11.

Let me remark that similar concepts can be introduced for operators. This is of particular importance for the case of unbounded operators, where convergence in the operator norm makes no sense at all.

A sequence of operators A_n is said to **converge strongly** to A,

$$\operatorname*{s-lim}_{n\to\infty} A_n = A \quad :\Leftrightarrow \quad A_n x \to Ax \quad \forall x \in \mathfrak{D}(A) \subseteq \mathfrak{D}(A_n). \tag{4.26}$$

It is said to **converge weakly** to A,

$$\operatorname*{w-lim}_{n\to\infty} A_n = A \quad :\Leftrightarrow \quad A_n x \rightharpoonup Ax \quad \forall x \in \mathfrak{D}(A) \subseteq \mathfrak{D}(A_n). \tag{4.27}$$

Clearly norm convergence implies strong convergence and strong convergence implies weak convergence. If Y is finite dimensional strong and weak convergence will be the same and this is in particular the case for $Y = \mathbb{C}$.

Example 4.31. Consider the operator $S_n \in \mathscr{L}(\ell^2(\mathbb{N}))$ which shifts a sequence n places to the left, that is,

$$S_n\,(x_1, x_2, \dots) = (x_{n+1}, x_{n+2}, \dots) \tag{4.28}$$

and the operator $S_n^* \in \mathscr{L}(\ell^2(\mathbb{N}))$ which shifts a sequence n places to the right and fills up the first n places with zeros, that is,

$$S_n^*\,(x_1, x_2, \dots) = (\underbrace{0, \dots, 0}_{n \text{ places}}, x_1, x_2, \dots). \tag{4.29}$$

Then S_n converges to zero strongly but not in norm (since $\|S_n\| = 1$) and S_n^* converges weakly to zero (since $\langle x, S_n^* y\rangle = \langle S_n x, y\rangle$) but not strongly (since $\|S_n^* x\| = \|x\|$) . ⋄

Lemma 4.32. *Suppose $A_n, B_n \in \mathscr{L}(X, Y)$ are sequences of bounded operators.*

(i) $\operatorname*{s-lim}_{n\to\infty} A_n = A$, $\operatorname*{s-lim}_{n\to\infty} B_n = B$, *and* $\alpha_n \to \alpha$ *implies* $\operatorname*{s-lim}_{n\to\infty}(A_n + B_n) = A + B$ *and* $\operatorname*{s-lim}_{n\to\infty} \alpha_n A_n = \alpha A$.

(ii) $\operatorname*{s-lim}_{n\to\infty} A_n = A$ *implies* $\|A\| \le \liminf_{n\to\infty} \|A_n\|$.

(iii) *If $A_n x$ converges for all $x \in X$ then $\|A_n\| \le C$ and there is an operator $A \in \mathscr{L}(X, Y)$ such that* $\operatorname*{s-lim}_{n\to\infty} A_n = A$.

(iv) *If $A_n y$ converges for y in a total set and $\|A_n\| \le C$, then there is an operator $A \in \mathscr{L}(X, Y)$ such that* $\operatorname*{s-lim}_{n\to\infty} A_n = A$.

The same result holds if strong convergence is replaced by weak convergence.

Proof. (i) $\lim_{n\to\infty}(\alpha_n A_n + B_n)x = \lim_{n\to\infty}(\alpha_n A_n x + B_n x) = \alpha A x + B x$.
(ii) follows from

$$\|Ax\| = \lim_{n\to\infty} \|A_n x\| \le \liminf_{n\to\infty} \|A_n\|$$

for every $x \in X$ with $\|x\| = 1$.
(iii) by linearity of the limit, $Ax := \lim_{n\to\infty} A_n x$ is a linear operator. Moreover, since convergent sequences are bounded, $\|A_n x\| \le C(x)$, the uniform boundedness principle implies $\|A_n\| \le C$. Hence $\|Ax\| = \lim_{n\to\infty} \|A_n x\| \le C\|x\|$.
(iv) By taking linear combinations we can replace the total set by a dense one. Moreover, we can define a linear operator A on this dense set via

$Ay := \lim_{n\to\infty} A_n y$. By $\|A_n\| \le C$ we see $\|A\| \le C$ and there is a unique extension to all of X. Now just use

$$\begin{aligned}\|A_n x - Ax\| &\le \|A_n x - A_n y\| + \|A_n y - Ay\| + \|Ay - Ax\| \\ &\le 2C\|x - y\| + \|A_n y - Ay\|\end{aligned}$$

and choose y in the dense subspace such that $\|x-y\| \le \frac{\varepsilon}{4C}$ and n large such that $\|A_n y - Ay\| \le \frac{\varepsilon}{2}$.

The case of weak convergence is left as an exercise (Problem 4.16 might be useful). □

Item (iii) of this lemma is sometimes also known as Banach–Steinhaus theorem. For an application of this lemma see Lemma 10.19.

Example 4.32. Let X be a Banach space of functions $f : [-\pi, \pi] \to \mathbb{C}$ such the functions $\{e_k(x) := \mathrm{e}^{\mathrm{i}kx}\}_{k\in\mathbb{Z}}$ are total. E.g. $X = C_{per}[-\pi, \pi]$ or $X = L^p[-\pi, \pi]$ for $1 \le p < \infty$. Then the Fourier series (2.44) converges on a total set and hence it will converge on all of X if and only if $\|S_n\| \le C$. For example, if $X = C_{per}[-\pi, \pi]$ then

$$\|S_n\| = \sup_{\|f\|_\infty=1} \|S_n(f)\| = \sup_{\|f\|_\infty=1} |S_n(f)(0)| = \frac{1}{2\pi}\|D_n\|_1$$

which is unbounded as we have seen in Example 4.3. In fact, in this example we have even shown failure of pointwise convergence and hence this is nothing new. However, if we consider $X = L^1[-\pi, \pi]$ we have (recall the Fejér kernel which satisfies $\|F_n\|_1 = 1$ and use (2.52) together with $S_n(D_m) = D_{\min(m,n)}$)

$$\|S_n\| = \sup_{\|f\|_1=1} \|S_n(f)\| \ge \lim_{m\to\infty} \|S_n(F_m)\|_1 = \|D_n\|_1$$

and we get that the Fourier series does not converge for some L^1 function. ⋄

Lemma 4.33. *Suppose $A_n \in \mathscr{L}(Y, Z)$, $B_n \in \mathscr{L}(X, Y)$ are two sequences of bounded operators.*

(i) $\operatorname{s-lim}_{n\to\infty} A_n = A$ *and* $\operatorname{s-lim}_{n\to\infty} B_n = B$ *implies* $\operatorname{s-lim}_{n\to\infty} A_n B_n = AB$.

(ii) $\operatorname{w-lim}_{n\to\infty} A_n = A$ *and* $\operatorname{s-lim}_{n\to\infty} B_n = B$ *implies* $\operatorname{w-lim}_{n\to\infty} A_n B_n = AB$.

(iii) $\lim_{n\to\infty} A_n = A$ *and* $\operatorname{w-lim}_{n\to\infty} B_n = B$ *implies* $\operatorname{w-lim}_{n\to\infty} A_n B_n = AB$.

Proof. For the first case just observe

$$\|(A_n B_n - AB)x\| \le \|(A_n - A)Bx\| + \|A_n\|\|(B_n - B)x\| \to 0.$$

The remaining cases are similar and again left as an exercise. □

Example 4.33. Consider again the last example. Then

$$S_n^* S_n (x_1, x_2, \dots) = (\underbrace{0, \dots, 0}_{n \text{ places}}, x_{n+1}, x_{n+2}, \dots)$$

converges to 0 weakly (in fact even strongly) but

$$S_n S_n^* (x_1, x_2, \dots) = (x_1, x_2, \dots)$$

does not! Hence the order in the second claim is important. ⋄

For a sequence of linear functionals ℓ_n, strong convergence is also called **weak-$*$** convergence. That is, the weak-$*$ limit of ℓ_n is ℓ if $\ell_n(x) \to \ell(x)$ for all $x \in X$ and we will write

$$\operatorname*{w^*-lim}_{n\to\infty} \ell_n = \ell \quad \text{or} \quad \ell_n \xrightarrow{*} \ell \tag{4.30}$$

in this case. Note that this is not the same as weak convergence on X^* unless X is reflexive: ℓ is the weak limit of ℓ_n if

$$j(\ell_n) \to j(\ell) \qquad \forall j \in X^{**}, \tag{4.31}$$

whereas for the weak-$*$ limit this is only required for $j \in J(X) \subseteq X^{**}$ (recall $J(x)(\ell) = \ell(x)$).

Example 4.34. In a Hilbert space weak-$*$ convergence of the linear functionals $\langle x_n, . \rangle$ is the same as weak convergence of the vectors x_n. ⋄

Example 4.35. Consider $X = c_0(\mathbb{N})$, $X^* \cong \ell^1(\mathbb{N})$, and $X^{**} \cong \ell^\infty(\mathbb{N})$ with J corresponding to the inclusion $c_0(\mathbb{N}) \hookrightarrow \ell^\infty(\mathbb{N})$. Then weak convergence on X^* implies

$$l_b(a^n - a) = \sum_{k=1}^{\infty} b_k (a_k^n - a_k) \to 0$$

for all $b \in \ell^\infty(\mathbb{N})$ and weak-* convergence implies that this holds for all $b \in c_0(\mathbb{N})$. Whereas we already have seen that weak convergence is equivalent to norm convergence, it is not hard to see that weak-* convergence is equivalent to the fact that the sequence is bounded and each component converges (cf. Problem 4.40). ⋄

With this notation it is also possible to slightly generalize Theorem 4.30 (Problem 4.41):

Lemma 4.34 (Helly)**.** *Suppose X is a separable Banach space. Then every bounded sequence $\ell_n \in X^*$ has a weak-$*$ convergent subsequence.*

Example 4.36. Let us return to the example after Theorem 4.30. Consider the Banach space of continuous functions $X = C[-1, 1]$. Using $\ell_f(\varphi) = \int \varphi f \, dx$ we can regard $L^1[-1, 1]$ as a subspace of X^*. Then the Dirac measure centered at 0 is also in X^* and it is the weak-$*$ limit of the sequence u_k. ⋄

Problem 4.35. *Suppose $\ell_n \to \ell$ in X^* and $x_n \rightharpoonup x$ in X. Then $\ell_n(x_n) \to \ell(x)$. Similarly, suppose* s-lim $\ell_n \to \ell$ *and $x_n \to x$. Then $\ell_n(x_n) \to \ell(x)$. Does this still hold if* s-lim $\ell_n \to \ell$ *and $x_n \rightharpoonup x$?*

Problem* 4.36. *Show that $x_n \rightharpoonup x$ implies $Ax_n \rightharpoonup Ax$ for $A \in \mathscr{L}(X,Y)$. Conversely, show that if $x_n \to 0$ implies $Ax_n \rightharpoonup 0$ then $A \in \mathscr{L}(X,Y)$.*

Problem 4.37. *Let $X := X_1 \oplus X_2$ show that $(x_{1,n}, x_{2,n}) \rightharpoonup (x_1, x_2)$ if and only if $x_{j,n} \rightharpoonup x_j$ for $j = 1, 2$.*

Problem 4.38. *Suppose $A_n, A \in \mathscr{L}(X,Y)$. Show that* s-lim $A_n = A$ *and* lim $x_n = x$ *implies* lim $A_n x_n = Ax$.

Problem* 4.39. *Show that if $\{\ell_j\} \subseteq X^*$ is some total set, then $x_n \rightharpoonup x$ if and only if x_n is bounded and $\ell_j(x_n) \to \ell_j(x)$ for all j. Show that this is wrong without the boundedness assumption (Hint: Take e.g. $X = \ell^2(\mathbb{N})$).*

Problem* 4.40. *Show that if $\{x_j\} \subseteq X$ is some total set, then $\ell_n \overset{*}{\rightharpoonup} \ell$ if and only if $\ell_n \in X^*$ is bounded and $\ell_n(x_j) \to \ell(x_j)$ for all j.*

Problem* 4.41. *Prove Lemma 4.34.*

4.5. Applications to minimizing nonlinear functionals

Finally, let me discuss a simple application of the above ideas to the **calculus of variations**. Many problems lead to finding the minimum of a given function. For example, many physical problems can be described by an energy functional and one seeks a solution which minimizes this energy. So we have a Banach space X (typically some function space) and a functional $F : M \subseteq X \to \mathbb{R}$ (of course this functional will in general be nonlinear). If M is compact and F is continuous, then we can proceed as in the finite-dimensional case to show that there is a minimizer: Start with a sequence x_n such that $F(x_n) \to \inf_M F$. By compactness we can assume that $x_n \to x_0$ after passing to a subsequence and by continuity $F(x_n) \to F(x_0) = \inf_M F$. Now in the infinite dimensional case we will use weak convergence to get compactness and hence we will also need weak (sequential) continuity of F. However, since there are more weakly than strongly convergent subsequences, weak (sequential) continuity is in fact a stronger property than just continuity!

Example 4.37. By Lemma 4.28 (ii) the norm is weakly sequentially lower semicontinuous but it is in general not weakly sequentially continuous as any infinite orthonormal set in a Hilbert space converges weakly to 0. However, note that this problem does not occur for linear maps. This is an immediate consequence of the very definition of weak convergence (Problem 4.36). ⋄

Hence weak continuity might be too much to hope for in concrete applications. In this respect note that, for our argument to work lower semicontinuity (cf. Problem 8.20) will already be sufficient:

Theorem 4.35 (Variational principle)**.** *Let X be a reflexive Banach space and let $F : M \subseteq X \to (-\infty, \infty]$. Suppose M is nonempty, weakly sequentially closed and that either F is* ***weakly coercive****, that is $F(x) \to \infty$ whenever $\|x\| \to \infty$, or that M is bounded. Then, if F is weakly sequentially lower semicontinuous, there exists some $x_0 \in M$ with $F(x_0) = \inf_M F$.*

Proof. Without loss of generality we can assume $F(x) < \infty$ for some $x \in M$. As above we start with a sequence $x_n \in M$ such that $F(x_n) \to \inf_M F$. If $M = X$ then the fact that F is coercive implies that x_n is bounded. Otherwise, it is bounded since we assumed M to be bounded. Hence we can pass to a subsequence such that $x_n \rightharpoonup x_0$ with $x_0 \in M$ since M is assumed sequentially closed. Now since F is weakly sequentially lower semicontinuous we finally get $\inf_M F = \lim_{n\to\infty} F(x_n) = \liminf_{n\to\infty} F(x_n) \geq F(x_0)$. □

Of course in a metric space the definition of closedness in terms of sequences agrees with the corresponding topological definition. In the present situation sequentially weakly closed implies (sequentially) closed and the converse holds at least for convex sets.

Lemma 4.36. *Suppose $M \subseteq X$ is convex. Then M is closed if and only if it is sequentially weakly closed.*

Proof. Suppose M is closed and let x be in the weak sequential closure of M, that is, there is a sequence $x_n \rightharpoonup x$. If $x \notin M$, then by Corollary 5.4 we can find a linear functional ℓ which separates $\{x\}$ and M. But this contradicts $\ell(x) = d < c \leq \ell(x_n) \to \ell(x)$. □

Similarly, the same is true with lower semicontinuity. In fact, a slightly weaker assumption suffices. Let X be a vector space and $M \subseteq X$ a convex subset. A function $F : M \to \overline{\mathbb{R}}$ is called **quasiconvex** if

$$F(\lambda x + (1-\lambda)y) \leq \max\{F(x), F(y)\}, \qquad \lambda \in (0,1), \quad x, y \in M. \tag{4.32}$$

It is called **strictly quasiconvex** if the inequality is strict for $x \neq y$. By $\lambda F(x) + (1-\lambda)F(y) \leq \max\{F(x), F(y)\}$ every (strictly) convex function is (strictly) quasiconvex. The converse is not true as the following example shows.

Example 4.38. Every (strictly) monotone function on $\mathbb{R}$ is (strictly) quasiconvex. Moreover, the same is true for symmetric functions which are (strictly) monotone on $[0, \infty)$. Hence the function $F(x) = \sqrt{|x|}$ is strictly quasiconvex. But it is clearly not convex on $M = \mathbb{R}$. ⋄

Now we are ready for the next

Lemma 4.37. *Suppose $M \subseteq X$ is a closed convex set and suppose $F : M \to \overline{\mathbb{R}}$ is quasiconvex. Then F is weakly sequentially lower semicontinuous if and only if it is (sequentially) lower semicontinuous.*

Proof. Suppose F is lower semicontinuous. If it were not weakly sequentially lower semicontinuous we could find a sequence $x_n \rightharpoonup x_0$ with $F(x_n) \to a < F(x_0)$. But then $x_n \in F^{-1}((-\infty, a])$ for n sufficiently large implying $x_0 \in F^{-1}((-\infty, a])$ as this set is convex (Problem 4.43) and closed (Problem 8.20). But this gives the contradiction $a < F(x_0) \leq a$. □

Corollary 4.38. *Let X be a reflexive Banach space and let M be a nonempty closed convex subset. If $F : M \subseteq X \to \overline{\mathbb{R}}$ is quasiconvex, lower semicontinuous, and, if M is unbounded, weakly coercive, then there exists some $x_0 \in M$ with $F(x_0) = \inf_M F$. If F is strictly quasiconvex then x_0 is unique.*

Proof. It remains to show uniqueness. Let x_0 and x_1 be two different minima. Then $F(\lambda x_0 + (1-\lambda)x_1) < \max\{F(x_0), F(x_1)\} = \inf_M F$, a contradiction. □

Example 4.39. Let X be a reflexive Banach space. Suppose $M \subseteq X$ is a nonempty closed convex set. Then for every $x \in X$ there is a point $x_0 \in M$ with minimal distance, $\|x - x_0\| = \operatorname{dist}(x, M)$. Indeed, $F(z) = \operatorname{dist}(x, z)$ is convex, continuous and, if M is unbounded weakly coercive. Hence the claim follows from Corollary 4.38. Note that the assumption that X is reflexive is crucial (Problem 4.42). Moreover, we also get that x_0 is unique if X is strictly convex (see Problem 1.12). ⋄

Example 4.40. Let $\mathfrak{H}$ be a Hilbert space and $\ell \in \mathfrak{H}^*$ a linear functional. We will give a variational proof of the Riesz lemma (Theorem 2.10). Since we already need to know that Hilbert spaces are reflexive, it should not be taken too serious. To this end consider

$$F(x) = \frac{1}{2}\|x\|^2 - \operatorname{Re}\big(\ell(x)\big), \qquad x \in \mathfrak{H}.$$

Then F is convex, continuous, and weakly coercive. Hence there is some $x_0 \in \mathfrak{H}$ with $F(x_0) = \inf_{x\in\mathfrak{H}} F(x)$. Moreover, for fixed $x \in \mathfrak{H}$,

$$\mathbb{R} \to \mathbb{R}, \quad \varepsilon \mapsto F(x_0 + \varepsilon x) = F(x_0) + \varepsilon \operatorname{Re}\big(\langle x_0, x\rangle - \ell(x)\big) + \frac{\varepsilon^2}{2}\|x\|^2$$

is a smooth map which has a minimum at $\varepsilon = 0$. Hence its derivative at $\varepsilon = 0$ must vanish: $\operatorname{Re}\big(\langle x_0, x\rangle - \ell(x)\big) = 0$ for all $x \in \mathfrak{H}$. Replacing $x \to -\mathrm{i}x$ we also get $\operatorname{Im}\big(\langle x_0, x\rangle - \ell(x)\big) = 0$ and hence $\ell(x) = \langle x_0, x\rangle$. ⋄

Example 4.41. Let $\mathfrak{H}$ be a Hilbert space and let us consider the problem of finding the lowest eigenvalue of a positive operator $A \geq 0$. Of course

this is bound to fail since the eigenvalues could accumulate at 0 without 0 being an eigenvalue (e.g. the multiplication operator with the sequence $\frac{1}{n}$ in $\ell^2(\mathbb{N})$). Nevertheless it is instructive to see how things can go wrong (and it underlines the importance of our various assumptions).

To this end consider its quadratic form $q_A(f) = \langle f, Af\rangle$. Then, since $q_A^{1/2}$ is a seminorm (Problem 1.21) and taking squares is convex, q_A is convex. If we consider it on $M = \bar{B}_1(0)$ we get existence of a minimum from Theorem 4.35. However this minimum is just $q_A(0) = 0$ which is not very interesting. In order to obtain a minimal eigenvalue we would need to take $M = S_1 = \{f| \, \|f\| = 1\}$, however, this set is not weakly closed (its weak closure is $\bar{B}_1(0)$ as we will see in the next section). In fact, as pointed out before, the minimum is in general not attained on M in this case.

Note that our problem with the trivial minimum at 0 would also disappear if we would search for a maximum instead. However, our lemma above only guarantees us weak sequential lower semicontinuity but not weak sequential upper semicontinuity. In fact, note that not even the norm (the quadratic form of the identity) is weakly sequentially upper continuous (cf. Lemma 4.28 (ii) versus Lemma 4.29). If we make the additional assumption that A is compact, then q_A is weakly sequentially continuous as can be seen from Theorem 4.31. Hence for compact operators the maximum is attained at some vector f_0. Of course we will have $\|f_0\| = 1$ but is it an eigenvalue? To see this we resort to a small ruse: Consider the real function

$$\phi(t) = \frac{q_A(f_0 + tf)}{\|f_0 + tf\|^2} = \frac{\alpha_0 + 2t\mathrm{Re}\langle f, Af_0\rangle + t^2 q_A(f)}{1 + 2t\mathrm{Re}\langle f, f_0\rangle + t^2\|f\|^2}, \quad \alpha_0 = q_A(f_0),$$

which has a maximum at $t = 0$ for any $f \in \mathfrak{H}$. Hence we must have $\phi'(0) = 2\mathrm{Re}\langle f, (A - \alpha_0)f_0\rangle = 0$ for all $f \in \mathfrak{H}$. Replacing $f \to \mathrm{i}f$ we get $2\mathrm{Im}\langle f, (A - \alpha_0)f_0\rangle = 0$ and hence $\langle f, (A - \alpha_0)f_0\rangle = 0$ for all f, that is $Af_0 = \alpha_0 f$. So we have recovered Theorem 3.6. ⋄

Problem 4.42. *Consider $X = C[0,1]$ and $M = \{f| \int_0^1 f(x)dx = 1,\ f(0) = 0\}$. Show that M is closed and convex. Show that $d(0, M) = 1$ but there is no minimizer. If we replace the boundary condition by $f(0) = 1$ there is a unique minimizer and for $f(0) = 2$ there are infinitely many minimizers.*

Problem* 4.43. *Show that $F : M \to \overline{\mathbb{R}}$ is quasiconvex if and only if the sublevel sets $F^{-1}((-\infty, a])$ are convex for every $a \in \mathbb{R}$.*

Chapter 5

Further topics on Banach spaces

5.1. The geometric Hahn–Banach theorem

The Hahn–Banach theorem is about constructing (continuous) linear functionals with given properties. In our original version this was done by showing that a functional defined on a subspace can be extended to the entire space. In this section we will establish a geometric version which establishes existence of functionals separating given convex sets. The key ingredient will be an association between convex sets and convex functions such that we can apply our original version of the Hahn–Banach theorem.

Let X be a vector space. For every subset $U \subset X$ we define its **Minkowski functional** (or **gauge**)

$$p_U(x) = \inf\{t > 0 | x \in t\,U\}. \tag{5.1}$$

Here $t\,U = \{tx | x \in U\}$. Note that $0 \in U$ implies $p_U(0) = 0$ and $p_U(x)$ will be finite for all x when U is **absorbing**, that is, for every $x \in X$ there is some r such that $x \in \alpha U$ for every $|\alpha| \geq r$. Note that every absorbing set contains 0 and every neighborhood of 0 in a Banach space is absorbing.

Example 5.1. Let X be a Banach space and $U = B_1(0)$, then $p_U(x) = \|x\|$. If $X = \mathbb{R}^2$ and $U = (-1,1) \times \mathbb{R}$ then $p_U(x) = |x_1|$. If $X = \mathbb{R}^2$ and $U = (-1,1) \times \{0\}$ then $p_U(x) = |x_1|$ if $x_2 = 0$ and $p_U(x) = \infty$ else. $\diamond$

We will only need minimal requirements and it will suffice if X is a **topological vector space**, that is, a vector space which carries a topology such that both vector addition $X \times X \to X$ and scalar multiplication $\mathbb{C} \times X \to X$ are continuous mappings. Of course every normed vector space is

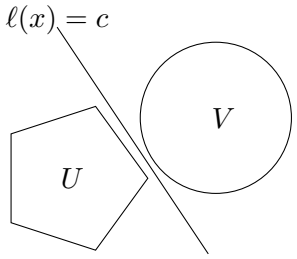


Figure 5.1. Separation of convex sets via a hyperplane

a topological vector space with the usual topology generated by open balls. As in the case of normed linear spaces, X^* will denote the vector space of all continuous linear functionals on X.

Lemma 5.1. *Let X be a vector space and U a convex subset containing 0. Then*

$$p_U(x+y) \le p_U(x) + p_U(y), \qquad p_U(\lambda\, x) = \lambda p_U(x), \qquad \lambda \ge 0. \tag{5.2}$$

Moreover, $\{x|p_U(x) < 1\} \subseteq U \subseteq \{x|p_U(x) \le 1\}$. If, in addition, X is a topological vector space and U is open, then $U = \{x|p_U(x) < 1\}$.

Proof. The homogeneity condition $p(\lambda\, x) = \lambda p(x)$ for $\lambda > 0$ is straightforward. To see the sublinearity Let $t, s > 0$ with $x \in t\, U$ and $y \in s\, U$, then

$$\frac{t}{t+s}\frac{x}{t} + \frac{s}{t+s}\frac{y}{s} = \frac{x+y}{t+s}$$

is in U by convexity. Moreover, $p_U(x+y) \le s+t$ and taking the infimum over all t and s we find $p_U(x+y) \le p_U(x) + p_U(y)$.

Suppose $p_U(x) < 1$, then $t^{-1}x \in U$ for some $t < 1$ and thus $x \in U$ by convexity. Similarly, if $x \in U$ then $t^{-1}x \in U$ for $t \ge 1$ by convexity and thus $p_U(x) \le 1$. Finally, let U be open and $x \in U$, then $(1+\varepsilon)x \in U$ for some $\varepsilon > 0$ and thus $p(x) \le (1+\varepsilon)^{-1}$. □

Note that (5.2) implies convexity

$$p_U(\lambda x + (1-\lambda)y) \le \lambda p_U(x) + (1-\lambda)p_U(y), \qquad \lambda \in [0,1]. \tag{5.3}$$

Theorem 5.2 (geometric Hahn–Banach, real version)**.** *Let U, V be disjoint nonempty convex subsets of a real topological vector space X and let U be open. Then there is a linear functional $\ell \in X^*$ and some $c \in \mathbb{R}$ such that*

$$\ell(x) < c \le \ell(y), \qquad x \in U,\, y \in V. \tag{5.4}$$

If V is also open, then the second inequality is also strict.

Proof. Choose $x_0 \in U$ and $y_0 \in V$, then

$$W = (U - x_0) - (V - y_0) = \{(x - x_0) - (y - y_0) | x \in U, y \in V\}$$

is open (since U is), convex (since U and V are) and contains 0. Moreover, since U and V are disjoint we have $z_0 = y_0 - x_0 \notin W$. By the previous lemma, the associated Minkowski functional p_W is convex and by the Hahn–Banach theorem there is a linear functional satisfying

$$\ell(tz_0) = t, \qquad |\ell(x)| \le p_W(x).$$

Note that since $z_0 \notin W$ we have $p_W(z_0) \ge 1$. Moreover, $W = \{x | p_W(x) < 1\} \subseteq \{x | |\ell(x)| < 1\}$ which shows that ℓ is continuous at 0 by scaling. By translations ℓ is continuous everywhere.

Finally we again use $p_W(z) < 1$ for $z \in W$ implying

$$\ell(x) - \ell(y) + 1 = \ell(x - y + z_0) \le p_W(x - y + z_0) < 1$$

and hence $\ell(x) < \ell(y)$ for $x \in U$ and $y \in V$. Therefore $\ell(U)$ and $\ell(V)$ are disjoint convex subsets of $\mathbb{R}$. Finally, let us suppose that there is some x_1 for which $\ell(x_1) = \sup \ell(U)$. Then, by continuity of the map $t \mapsto x_1 + tz_0$ there is some $\varepsilon > 0$ such that $x_1 + \varepsilon z_0 \in U$. But this gives a contradiction $\ell(x_1) + \varepsilon = \ell(x_1 + \varepsilon z_0) \le \ell(x_1)$. Thus the claim holds with $c = \sup \ell(U)$. If V is also open an analogous argument shows $\inf \ell(V) < \ell(y)$ for all $y \in V$. □

Of course there is also a complex version.

Theorem 5.3 (geometric Hahn–Banach, complex version)**.** *Let U, V be disjoint nonempty convex subsets of a topological vector space X and let U be open. Then there is a linear functional $\ell \in X^*$ and some $c \in \mathbb{R}$ such that*

$$\operatorname{Re}(\ell(x)) < c \le \operatorname{Re}(\ell(y)), \qquad x \in U,\ y \in V. \tag{5.5}$$

If V is also open, then the second inequality is also strict.

Proof. Consider X as a real Banach space. Then there is a continuous real-linear functional $\ell_r : X \to \mathbb{R}$ by the real version of the geometric Hahn–Banach theorem. Then $\ell(x) = \ell_r(x) - \mathrm{i}\ell_r(\mathrm{i}x)$ is the functional we are looking for (check this). □

Example 5.2. The assumption that one set is open is crucial as the following example shows. Let $X = c_0(\mathbb{N})$, $U = \{a \in c_0(\mathbb{N}) | \exists N : a_N > 0 \text{ and } a_n = 0, n > N\}$ and $V = \{0\}$. Note that U is convex but not open and that $U \cap V = \emptyset$. Suppose we could find a linear functional ℓ as in the geometric Hahn–Banach theorem (of course we can choose $\alpha = \ell(0) = 0$ in this case). Then by Problem 4.21 there is some $b_j \in \ell^\infty(\mathbb{N})$ such that $\ell(a) = \sum_{j=1}^\infty b_j a_j$. Moreover, we must have $b_j = \ell(\delta^j) < 0$. But then $a = (b_2, -b_1, 0, \dots) \in U$ and $\ell(a) = 0 \not< 0$. ◇

Note that two disjoint closed convex sets can be separated strictly if one of them is compact. However, this will require that every point has a neighborhood base of convex open sets. Such topological vector spaces are called **locally convex spaces** and they will be discussed further in Section 5.4. For now we just remark that every normed vector space is locally convex since balls are convex.

Corollary 5.4. *Let U, V be disjoint nonempty closed convex subsets of a locally convex space X and let U be compact. Then there is a linear functional $\ell \in X^*$ and some $c, d \in \mathbb{R}$ such that*

$$\operatorname{Re}(\ell(x)) \le d < c \le \operatorname{Re}(\ell(y)), \qquad x \in U,\ y \in V. \tag{5.6}$$

Proof. Since V is closed, for every $x \in U$ there is a convex open neighborhood N_x of 0 such that $x + N_x$ does not intersect V. By compactness of U there are $x_1, \dots, x_n$ such that the corresponding neighborhoods $x_j + \frac{1}{2}N_{x_j}$ cover U. Set $N = \bigcap_{j=1}^n N_{x_j}$ which is a convex open neighborhood of 0. Then

$$\tilde{U} = U + \frac{1}{2}N \subseteq \bigcup_{j=1}^n (x_j + \frac{1}{2}N_{x_j}) + \frac{1}{2}N \subseteq \bigcup_{j=1}^n (x_j + \frac{1}{2}N_{x_j} + \frac{1}{2}N_{x_j}) = \bigcup_{j=1}^n (x_j + N_{x_j})$$

is a convex open set which is disjoint from V. Hence by the previous theorem we can find some ℓ such that $\operatorname{Re}(\ell(x)) < c \le \operatorname{Re}(\ell(y))$ for all $x \in \tilde{U}$ and $y \in V$. Moreover, since $\ell(U)$ is a compact interval $[e, d]$ the claim follows. □

Note that if U and V are **absolutely convex**, that is, $\alpha U + \beta U \subseteq U$ for $|\alpha| + |\beta| \le 1$, then we can write the previous condition equivalently as

$$|\ell(x)| \le d < c \le |\ell(y)|, \qquad x \in U,\ y \in V, \tag{5.7}$$

since $x \in U$ implies $\theta x \in U$ for $\theta = \operatorname{sign}(\ell(x))$ and thus $|\ell(x)| = \theta\ell(x) = \ell(\theta x) = \operatorname{Re}(\ell(\theta x))$.

From the last corollary we can also obtain versions of Corollaries 4.17 and 4.15 for locally convex vector spaces.

Corollary 5.5. *Let $Y \subseteq X$ be a subspace of a locally convex space and let $x_0 \in X \setminus \overline{Y}$. Then there exists an $\ell \in X^*$ such that (i) $\ell(y) = 0$, $y \in Y$ and (ii) $\ell(x_0) = 1$.*

Proof. Consider ℓ from Corollary 5.4 applied to $U = \{x_0\}$ and $V = \overline{Y}$. Now observe that $\ell(Y)$ must be a subspace of $\mathbb{C}$ and hence $\ell(Y) = \{0\}$ implying $\operatorname{Re}(\ell(x_0)) < 0$. Finally $\ell(x_0)^{-1}\ell$ is the required functional. □

Corollary 5.6. *Let $Y \subseteq X$ be a subspace of a locally convex space and let $\ell : Y \to \mathbb{C}$ be a continuous linear functional. Then there exists a continuous extension $\bar{\ell} \in X^*$.*

Proof. Without loss of generality we can assume that ℓ is nonzero such that we can find $x_0 \in y$ with $\ell(x_0) = 1$. Since Y has the subset topology $x_0 \notin Y_0 := \overline{\mathrm{Ker}(\ell)}$, where the closure is taken in X. Now Corollary 5.5 gives a functional $\bar{\ell}$ with $\bar{\ell}(x_0) = 1$ and $Y_0 \subseteq \mathrm{Ker}(\bar{\ell})$. Moreover,

$$\bar{\ell}(x) - \ell(x) = \bar{\ell}(x) - \ell(x)\bar{\ell}(x_0) = \bar{\ell}(x - \ell(x)x_0) = 0, \qquad x \in Y,$$

since $x - \ell(x)x_0 \in \mathrm{Ker}(\ell)$. □

Problem* 5.1. *Let X be a topological vector space. Show that $U + V$ is open if one of the sets is open.*

Problem 5.2. *Show that Corollary 5.4 fails even in $\mathbb{R}^2$ unless one set is compact.*

Problem 5.3. *Let X be a topological vector space and $M \subseteq X$, $N \subseteq X^*$. Then the corresponding* ***polar, prepolar sets*** *are*

$$M^\circ = \{\ell \in X^* || \ell(x)| \leq 1 \, \forall x \in M\}, \quad N_\circ = \{x \in X || \ell(x)| \leq 1 \, \forall \ell \in N\},$$

respectively. Show

(i) *M° is closed and absolutely convex.*

(ii) *$M_1 \subseteq M_2$ implies $M_2^\circ \subseteq M_1^\circ$.*

(iii) *For $\alpha \neq 0$ we have $(\alpha M)^\circ = |\alpha|^{-1} M^\circ$.*

(iv) *If M is a subspace we have $M^\circ = M^\perp$.*

The same claims hold for prepolar sets.

Problem 5.4 (Bipolar theorem)**.** *Let X be a locally convex space and suppose $M \subseteq X$ is absolutely convex we have $\alpha M + \beta M \subseteq M$. Show $(M^\circ)_\circ = \overline{M}$. (Hint: Use Corollary 5.4 to show that for every $y \notin \overline{M}$ there is some $\ell \in X^*$ with $\mathrm{Re}(\ell(x)) \leq 1 < \ell(y)$, $x \in \overline{M}$.)*

5.2. Convex sets and the Krein–Milman theorem

Let X be a locally convex vector space. Since the intersection of arbitrary convex sets is again convex we can define the convex hull of a set U as the smallest convex set containing U, that is, the intersection of all convex sets containing U. It is straightforward to show (Problem 5.5) that the convex hull is given by

$$\mathrm{conv}(U) := \{\sum_{j=1}^{n} \lambda_j x_j | n \in \mathbb{N}, \, x_j \in U, \, \sum_{j=1}^{n} \lambda_j = 1, \, \lambda_j \geq 0\}. \tag{5.8}$$

A line segment is convex and can be generated as the convex hull of its endpoints. Similarly, a full triangle is convex and can be generated as the convex hull of its vertices. However, if we look at a ball, then we need its

entire boundary to recover it as the convex hull. So how can we characterize those points which determine a convex sets via the convex hull?

Let K be a set and $M \subseteq K$ a nonempty subset. Then M is called an **extremal subset** of K if no point of M can be written as a convex combination of two points unless both are in M: For given $x, y \in K$ and $\lambda \in (0,1)$ we have that

$$\lambda x + (1-\lambda)y \in M \quad \Rightarrow x, y \in M. \tag{5.9}$$

If $M = \{x\}$ is extremal, then x is called an **extremal point** of K. Hence an extremal point cannot be written as a convex combination of two other points from K.

Note that we did not require K to be convex. If K is convex, then M is extremal if and only if $K \setminus M$ is convex. Note that the nonempty intersection of extremal sets is extremal. Moreover, if $L \subseteq M$ is extremal and $M \subseteq K$ is extremal, then $L \subseteq K$ is extremal as well (Problem 5.6).

Example 5.3. Consider $\mathbb{R}^2$ with the norms $\|.\|_p$. Then the extremal points of the closed unit ball (cf. Figure 1.1) are the boundary points for $1 < p < \infty$ and the vertices for $p = 1, \infty$. In any case the boundary is an extremal set. Slightly more general, in a strictly convex space, (ii) of Problem 1.12 says that the extremal points of the unit ball are precisely its boundary points. ⋄

Example 5.4. Consider $\mathbb{R}^3$ and let $C = \{(x_1, x_2, 0) \in \mathbb{R}^3 | x_1^2 + x_2^2 = 1\}$. Take two more points $x_\pm = (0, 0, \pm 1)$ and consider the convex hull K of $M = C \cup \{x_+, x_-\}$. Then M is extremal in K and, moreover, every point from M is an extremal point. However, if we change the two extra points to be $x_\pm = (1, 0, \pm 1)$, then the point $(1, 0, 0)$ is no longer extremal. Hence the extremal points are now $M \setminus \{(1, 0, 0)\}$. Note in particular that the set of extremal points is not closed in this case. ⋄

Extremal sets arise naturally when minimizing linear functionals.

Lemma 5.7. *Suppose $K \subseteq X$ and $\ell \in X^*$. If*

$$K_\ell := \{x \in K | \ell(x) = \inf_{y \in K} \mathrm{Re}(\ell(y))\}$$

is nonempty (e.g. if K is compact), then it is extremal in K. If K is closed and convex, then K_ℓ is closed and convex.

Proof. Set $m = \inf_{y \in K} \mathrm{Re}(\ell(y))$. Let $x, y \in K$, $\lambda \in (0,1)$ and suppose $\lambda x + (1-\lambda)y \in K_\ell$. Then

$$m = \mathrm{Re}(\ell(\lambda x + (1-\lambda)y)) = \lambda \mathrm{Re}(\ell(x)) + (1-\lambda)\mathrm{Re}(\ell(y)) \geq \lambda m + (1-\lambda)m = m$$

with strict inequality if $\mathrm{Re}(\ell(x)) > m$ or $\mathrm{Re}(\ell(y)) > m$. Hence we must have $x, y \in K_\ell$. Finally by linearity K_ℓ is convex and by continuity it is closed. □

If K is a closed convex set, then nonempty subsets of the type K_ℓ are called **faces** of K and $H_\ell := \{x \in X | \ell(x) = \inf_{y\in K} \operatorname{Re}(\ell(y))\}$ is called a **support hyperplane** of K.

Conversely, if K is convex with nonempty interior, then every point x on the boundary has a supporting hyperplane (observe that the interior is convex and apply the geometric Hahn–Banach theorem with $U = K^\circ$ and $V = \{x\}$).

Next we want to look into existence of extremal points.

Example 5.5. Note that an interior point can never be extremal as it can be written as convex combination of some neighboring points. In particular, an open convex set will not have any extremal points (e.g. X, which is also closed, has no extremal points). Conversely, if K is closed and convex, then the boundary is extremal since $K \setminus \partial K = K^\circ$ is convex (Problem 5.7). ⋄

Example 5.6. Suppose X is a strictly convex Banach space. Then every nonempty compact subset K has an extremal point. Indeed, let $x \in K$ be such that $\|x\| = \sup_{y\in K} \|y\|$, then x is extremal: If $x = \lambda y + (1-\lambda)z$ then $\|x\| \le \lambda\|y\| + (1-\lambda)\|z\| \le \|x\|$ shows that we have equality in the triangle inequality and hence $x = y = z$ by Problem 1.12 (i). ⋄

Example 5.7. In a not strictly convex space the situation is quite different. For example, consider the closed unit ball in $\ell^\infty(\mathbb{N})$. Let $a \in \ell^\infty(\mathbb{N})$. If there is some index j such that $\lambda := |a_j| < 1$ then $a = \frac{1}{2}b + \frac{1}{2}c$ where $b = a + \varepsilon\delta^j$ and $c = a - \varepsilon\delta^j$ with $\varepsilon \le 1 - |a_j|$. Hence the only possible extremal points are those with $|a_j| = 1$ for all $j \in \mathbb{N}$. If we have such an a, then if $a = \lambda b + (1-\lambda)c$ we must have $1 = |\lambda b_n + (1-\lambda)c_n| \le \lambda|b_n| + (1-\lambda)|c_n| \le 1$ and hence $a_n = b_n = c_n$ by strict convexity of the absolute value. Hence all such sequences are extremal.

However, if we consider $c_0(\mathbb{N})$ the same argument shows that the closed unit ball contains no extremal points. In particular, the following lemma implies that there is no locally convex topology for which the closed unit ball in $c_0(\mathbb{N})$ is compact. Together with the Banach–Alaoglu theorem (Theorem 5.10) this will show that $c_0(\mathbb{N})$ is not the dual of any Banach space. ⋄

Lemma 5.8 (Krein–Milman)**.** *Let X be a locally convex space. Suppose $K \subseteq X$ is compact and nonempty. Then it contains at least one extremal point.*

Proof. We want to apply Zorn's lemma. To this end consider the family

$$\mathcal{M} = \{M \subseteq K | \text{compact and extremal in } K\}$$

with the partial order given by reversed inclusion. Since $K \in \mathcal{M}$ this family is nonempty. Moreover, given a linear chain $\mathcal{C} \subset \mathcal{M}$ we consider $M := \bigcap \mathcal{C}$. Then $M \subseteq K$ is nonempty by the finite intersection property and since it

is closed also compact. Moreover, as the nonempty intersection of extremal sets it is also extremal. Hence $M \in \mathcal{M}$ and thus $\mathcal{M}$ has a maximal element. Denote this maximal element by M.

We will show that M contains precisely one point (which is then extremal by construction). Indeed, suppose $x, y \in M$. If $x \neq y$ we can, by Corollary 5.4, choose a linear functional $\ell \in X^*$ with $\operatorname{Re}(\ell(x)) \neq \operatorname{Re}(\ell(y))$. Then by Lemma 5.7 $M_\ell \subset M$ is extremal in M and hence also in K. But by $\operatorname{Re}(\ell(x)) \neq \operatorname{Re}(\ell(y))$ it cannot contain both x and y contradicting maximality of M. □

Finally, we want to recover a convex set as the convex hull of its extremal points. In our infinite dimensional setting an additional closure will be necessary in general.

Since the intersection of arbitrary closed convex sets is again closed and convex we can define the closed convex hull of a set U as the smallest closed convex set containing U, that is, the intersection of all closed convex sets containing U. Since the closure of a convex set is again convex (Problem 5.7) the closed convex hull is simply the closure of the convex hull.

Theorem 5.9 (Krein–Milman)**.** *Let X be a locally convex space. Suppose $K \subseteq X$ is convex and compact. Then it is the closed convex hull of its extremal points.*

Proof. Let E be the extremal points and $M := \overline{\operatorname{conv}(E)} \subseteq K$ be its closed convex hull. Suppose $x \in K \setminus M$ and use Corollary 5.4 to choose a linear functional $\ell \in X^*$ with

$$\min_{y \in M} \operatorname{Re}(\ell(y)) > \operatorname{Re}(\ell(x)) \geq \min_{y \in K} \operatorname{Re}(\ell(y)).$$

Now consider K_ℓ from Lemma 5.7 which is nonempty and hence contains an extremal point $y \in E$. But $y \notin M$, a contradiction. □

While in the finite dimensional case the closure is not necessary (Problem 5.8), it is important in general as the following example shows.

Example 5.8. Consider the closed unit ball in $\ell^1(\mathbb{N})$. Then the extremal points are $\{\mathrm{e}^{\mathrm{i}\theta}\delta^n | n \in \mathbb{N}, \theta \in \mathbb{R}\}$. Indeed, suppose $\|a\|_1 = 1$ with $\lambda := |a_j| \in (0,1)$ for some $j \in \mathbb{N}$. Then $a = \lambda b + (1-\lambda)c$ where $b := \lambda^{-1} a_j \delta^j$ and $c := (1-\lambda)^{-1}(a - a_j\delta^j)$. Hence the only possible extremal points are of the form $\mathrm{e}^{\mathrm{i}\theta}\delta^n$. Moreover, if $\mathrm{e}^{\mathrm{i}\theta}\delta^n = \lambda b + (1-\lambda)c$ we must have $1 = |\lambda b_n + (1-\lambda)c_n| \leq \lambda|b_n| + (1-\lambda)|c_n| \leq 1$ and hence $a_n = b_n = c_n$ by strict convexity of the absolute value. Thus the convex hull of the extremal points are the sequences from the unit ball which have finitely many terms nonzero. While the closed unit ball is not compact in the norm topology it will be in

the weak-$*$ topology by the Banach–Alaoglu theorem (Theorem 5.10). To this end note that $\ell^1(\mathbb{N}) \cong c_0(\mathbb{N})^*$. ⋄

Also note that in the infinite dimensional case the extremal points can be dense.

Example 5.9. Let $X = C([0,1],\mathbb{R})$ and consider the convex set $K = \{f \in C^1([0,1],\mathbb{R}) | f(0) = 0,\ \|f'\|_\infty \le 1\}$. Note that the functions $f_\pm(x) = \pm x$ are extremal. For example, assume

$$x = \lambda f(x) + (1-\lambda) g(x)$$

then

$$1 = \lambda f'(x) + (1-\lambda) g'(x)$$

which implies $f'(x) = g'(x) = 1$ and hence $f(x) = g(x) = x$.

To see that there are no other extremal functions, suppose $|f'(x)| \le 1-\varepsilon$ on some interval I. Choose a nontrivial continuous function g which is 0 outside I and has integral 0 over I and $\|g\|_\infty \le \varepsilon$. Let $G = \int_0^x g(t)dt$. Then $f = \frac{1}{2}(f+G) + \frac{1}{2}(f-G)$ and hence f is not extremal. Thus $f_\pm$ are the only extremal points and their (closed) convex is given by $f_\lambda(x) = \lambda x$ for $\lambda \in [-1,1]$.

Of course the problem is that K is not closed. Hence we consider the Lipschitz continuous functions $\bar{K} := \{f \in C^{0,1}([0,1],\mathbb{R}) | f(0) = 0,\ [f]_1 \le 1\}$ (this is in fact the closure of K, but this is a bit tricky to see and we won't need this here). By the Arzelà–Ascoli theorem (Theorem 1.14) $\bar{K}$ is relatively compact and since the Lipschitz estimate clearly is preserved under uniform limits it is even compact.

Now note that piecewise linear functions with $f'(x) \in \{\pm 1\}$ away from the kinks are extremal in $\bar{K}$. Moreover, these functions are dense: Split $[0,1]$ into n pieces of equal length using $x_j = \frac{j}{n}$. Set $f_n(x_0) = 0$ and $f_n(x) = f_n(x_j) \pm (x - x_j)$ for $x \in [x_j, x_{j+1}]$ where the sign is chosen such that $|f(x_{j+1}) - f_n(x_{j+1})|$ gets minimal. Then $\|f - f_n\|_\infty \le \frac{1}{n}$. ⋄

Problem* 5.5. *Show that the convex hull is given by* (5.8).

Problem* 5.6. *Show that the nonempty intersection of extremal sets is extremal. Show that if $L \subseteq M$ is extremal and $M \subseteq K$ is extremal, then $L \subseteq K$ is extremal as well.*

Problem 5.7. *Let X be a topological vector space. Show that the closure and the interior of a convex set is convex. (Hint: One way of showing the first claim is to consider the continuous map $f : X \times X \to X$ given by $(x,y) \mapsto \lambda x + (1-\lambda)y$ and use Problem B.13.)*

Problem 5.8 (Carathéodory)**.** *Show that for a compact convex set $K \subseteq \mathbb{R}^n$ every point can be written as convex combination of $n+1$ extremal points.*

(Hint: Induction on n. Without loss assume that 0 is an extremal point. If K is contained in an $n-1$ dimensional subspace we are done. Otherwise K has an open interior. Now for a given point the line through this point and 0 intersects the boundary where we have a corresponding face.)

5.3. Weak topologies

In Section 4.4 we have defined weak convergence for sequences and this raises the question about a natural topology associated with this convergence. To this end we define the **weak topology** on X as the weakest topology for which all $\ell \in X^*$ remain continuous. Recall that a base for this topology is given by sets of the form

$$x + \bigcap_{j=1}^{n} |\ell_j|^{-1}\big([0,\varepsilon_j)\big) = \{\tilde{x} \in X \,|\, |\ell_j(x-\tilde{x})| < \varepsilon_j,\ 1 \le j \le n\},$$
$$x \in X,\ \ell_j \in X^*,\ \varepsilon_j > 0. \tag{5.10}$$

In particular, it is straightforward to check that a sequence converges with respect to this topology if and only if it converges weakly. Since the linear functionals separate points (cf. Corollary 4.16) the weak topology is Hausdorff.

Note that, if X^* is separable, given a total set $\{\ell_n\}_{n\in\mathbb{N}} \subset X^*$ of (w.l.o.g.) normalized linear functionals

$$d(x,\tilde{x}) = \sum_{n=1}^{\infty} \frac{1}{2^n} |\ell_n(x-\tilde{x})| \tag{5.11}$$

defines a metric on the unit ball $B_1(0) \subset X$ which can be shown to generate the weak topology (Problem 5.11).

Similarly, we define the **weak-$*$ topology** on X^* as the weakest topology for which all $j \in J(X) \subseteq X^{**}$ remain continuous. In particular, the weak-$*$ topology is weaker than the weak topology on X^* and both are equal if X is reflexive. Since different linear functionals must differ at least at one point the weak-$*$ topology is also Hausdorff. A base for the weak-$*$ topology is given by sets of the form

$$\ell + \bigcap_{j=1}^{n} |J(x_j)|^{-1}\big([0,\varepsilon_j)\big) = \{\tilde{\ell} \in X^* \,|\, |(\ell-\tilde{\ell})(x_j)| < \varepsilon_j,\ 1 \le j \le n\},$$
$$\ell \in X^*,\ x_j \in X,\ \varepsilon_j > 0. \tag{5.12}$$

Note that, if X is separable, given a total set $\{x_n\}_{n\in\mathbb{N}} \subset X$ of (w.l.o.g.) normalized vectors

$$d(\ell,\tilde{\ell}) = \sum_{n=1}^{\infty} \frac{1}{2^n} |(\ell-\tilde{\ell})(x_n)| \tag{5.13}$$

defines a metric on the unit ball $B_1^*(0) \subset X^*$ which can be shown to generate the weak-$*$ topology (Problem 5.11). Hence Lemma 4.34 could also be stated as $\bar{B}_1^*(0) \subset X^*$ being weak-$*$ compact. This is in fact true without assuming X to be separable and is known as Banach–Alaoglu theorem.

Theorem 5.10 (Banach–Alaoglu)**.** *Let X be a Banach space. Then $\bar{B}_1^*(0) \subset X^*$ is compact in the weak-$*$ topology.*

Proof. Abbreviate $B = \bar{B}_1^X(0)$, $B^* = \bar{B}_1^{X^*}(0)$, and $B_x = \bar{B}_{\|x\|}^{\mathbb{C}}(0)$. Consider the (injective) map $\Phi : X^* \to \mathbb{C}^X$ given by $|\Phi(\ell)(x)| = \ell(x)$ and identify X^* with $\Phi(X^*)$. Then the weak-$*$ topology on X^* coincides with the relative topology on $\Phi(X^*) \subseteq \mathbb{C}^X$ (recall that the product topology on $\mathbb{C}^X$ is the weakest topology which makes all point evaluations continuous). Moreover, $\Phi(\ell) \le \|\ell\|\|x\|$ implies $\Phi(B^*) \subset \bigtimes_{x\in X} B_x$ where the last product is compact by Tychonoff's theorem. Hence it suffices to show that $\Phi(B^*)$ is closed. To this end let $l \in \overline{\Phi(B^*)}$. We need to show that l is linear and bounded. Fix $x_1, x_2 \in X$, $\alpha \in \mathbb{C}$, and consider the open neighborhood

$$U(l) = \Big\{ h \in \bigtimes_{x\in B} B_x \Big| \begin{array}{l} |h(x_1+x_2) - l(x_1+\alpha x_2)| < \varepsilon, \\ |h(x_1) - l(x_1)| < \varepsilon,\ |\alpha||h(x_2) - l(x_2)| < \varepsilon \end{array} \Big\}$$

of l. Since $U(l) \cap \Phi(X^*)$ is nonempty we can choose an element h from this intersection to show $|l(x_1+\alpha x_2) - l(x_1) - \alpha l(x_2)| < 3\varepsilon$. Since $\varepsilon > 0$ is arbitrary we conclude $l(x_1+\alpha x_2) = l(x_1) - \alpha l(x_2)$. Moreover, $|l(x_1)| \le |h(x_1)| + \varepsilon \le \|x_1\| + \varepsilon$ shows $\|l\| \le 1$ and thus $l \in \Phi(B^*)$. □

Since the weak topology is weaker than the norm topology every weakly closed set is also (norm) closed. Moreover, the weak closure of a set will in general be larger than the norm closure. However, for convex sets both will coincide. In fact, we have the following characterization in terms of closed (affine) **half-spaces**, that is, sets of the form $\{x \in X | \mathrm{Re}(\ell(x)) \le \alpha\}$ for some $\ell \in X^*$ and some $\alpha \in \mathbb{R}$.

Theorem 5.11 (Mazur)**.** *The weak as well as the norm closure of a convex set K is the intersection of all half-spaces containing K. In particular, a convex set $K \subseteq X$ is weakly closed if and only if it is closed.*

Proof. Since the intersection of closed-half spaces is (weakly) closed, it suffices to show that for every x not in the (weak) closure there is a closed half-plane not containing x. Moreover, if x is not in the weak closure it is also not in the norm closure (the norm closure is contained in the weak closure) and by Theorem 5.3 with $U = B_{\mathrm{dist}(x,K)}(x)$ and $V = K$ there is a functional $\ell \in X^*$ such that $K \subseteq \mathrm{Re}(\ell)^{-1}([c,\infty))$ and $x \notin \mathrm{Re}(\ell)^{-1}([c,\infty))$. □

Example 5.10. Suppose X is infinite dimensional. The weak closure $\overline{S}^w$ of $S = \{x \in X | \, \|x\| = 1\}$ is the closed unit ball $\bar{B}_1(0)$. Indeed, since

$\bar{B}_1(0)$ is convex the previous lemma shows $\overline{S}^w \subseteq \bar{B}_1(0)$. Conversely, if $x \in B_1(0)$ is not in the weak closure, then there must be an open neighborhood $x + \bigcap_{j=1}^n |\ell_j|^{-1}([0,\varepsilon))$ not contained in the weak closure. Since X is infinite dimensional we can find a nonzero element $x_0 \in \bigcap_{j=1}^n \operatorname{Ker}(\ell_j)$ such that the affine line $x + tx_0$ is in this neighborhood and hence also avoids $\overline{S}^w$. But this is impossible since by the intermediate value theorem there is some $t_0 > 0$ such that $\|x + t_0 x_0\| = 1$. Hence $\bar{B}_1(0) \subseteq \overline{S}^w$. ⋄

Note that this example also shows that in an infinite dimensional space the weak and norm topologies are always different! In a finite dimensional space both topologies of course agree.

Corollary 5.12 (Mazur lemma)**.** *Suppose $x_k \rightharpoonup x$, then there are convex combinations $y_k = \sum_{j=1}^{n_k} \lambda_{k,j} x_j$ (with $\sum_{j=1}^{n_k} \lambda_{k,j} = 1$ and $\lambda_{k,j} \geq 0$) such that $y_k \to x$.*

Proof. Let $K = \{\sum_{j=1}^n \lambda_j x_j | n \in \mathbb{N}, \sum_{j=1}^n \lambda_j = 1, \lambda_j \geq 0\}$ be the convex hull of the points $\{x_n\}$. Then by the previous result $x \in \overline{K}$. □

Example 5.11. Let $\mathfrak{H}$ be a Hilbert space and $\{\varphi_j\}$ some infinite ONS. Then we already know $\varphi_j \rightharpoonup 0$. Moreover, the convex combination $\psi_j = \frac{1}{j}\sum_{k=1}^j \varphi_k \to 0$ since $\|\psi_j\| = j^{-1/2}$. ⋄

Finally, we note two more important results. For the first note that since X^{**} is the dual of X^* it has a corresponding weak-$*$ topology and by the Banach–Alaoglu theorem $\bar{B}_1^{**}(0)$ is weak-$*$ compact and hence weak-$*$ closed.

Theorem 5.13 (Goldstine)**.** *The image of the closed unit ball $\bar{B}_1(0)$ under the canonical embedding J into the closed unit ball $\bar{B}_1^{**}(0)$ is weak-$*$ dense.*

Proof. Let $j \in \bar{B}_1^{**}(0)$ be given. Since sets of the form $j + \bigcap_{k=1}^n |\ell_k|^{-1}([0,\varepsilon))$ provide a neighborhood base (where we can assume the $\ell_k \in X^*$ to be linearly independent without loss of generality) it suffices to find some $x \in \bar{B}_{1+\varepsilon}(0)$ with $\ell_k(x) = j(\ell_k)$ for $1 \leq k \leq n$ since then $(1+\varepsilon)^{-1}J(x)$ will be in the above neighborhood. Without the requirement $\|x\| \leq 1 + \varepsilon$ this follows from surjectivity of the map $F : X \to \mathbb{C}^n$, $x \mapsto (\ell_1(x), \dots, \ell_n(x))$. Moreover, given one such x the same is true for every element from $x + Y$, where $Y = \bigcap_k \operatorname{Ker}(\ell_k)$. So if $(x + Y) \cap \bar{B}_{1+\varepsilon}(0)$ were empty, we would have $\operatorname{dist}(x, Y) \geq 1 + \varepsilon$ and by Corollary 4.17 we could find some normalized $\ell \in X^*$ which vanishes on Y and satisfies $\ell(x) \geq 1 + \varepsilon$. But by Problem 4.32 we have $\ell \in \operatorname{span}(\ell_1, \dots, \ell_n)$ implying

$$1 + \varepsilon \leq \ell(x) = j(\ell) \leq \|j\|\|\ell\| \leq 1$$

a contradiction. □

Example 5.12. Consider $X = c_0(\mathbb{N})$, $X^* \cong \ell^1(\mathbb{N})$, and $X^{**} \cong \ell^\infty(\mathbb{N})$ with J corresponding to the inclusion $c_0(\mathbb{N}) \hookrightarrow \ell^\infty(\mathbb{N})$. Then we can consider the linear functionals $\ell_j(x) = x_j$ which are total in X^* and a sequence in X^{**} will be weak-$*$ convergent if and only if it is bounded and converges when composed with any of the ℓ_j (in other words, when the sequence converges componentwise — cf. Problem 4.40). So for example, cutting off a sequence in $\bar{B}_1^{**}(0)$ after n terms (setting the remaining terms equal to 0) we get a sequence from $\bar{B}_1(0) \hookrightarrow \bar{B}_1^{**}(0)$ which is weak-$*$ convergent (but of course not norm convergent). $\diamond$

Theorem 5.14. *A Banach space X is reflexive if and only if the closed unit ball $\bar{B}_1(0)$ is weakly compact.*

Proof. If X is reflexive that this result follows from the Banach–Alaoglu theorem since in this case $J(\bar{B}_1(0)) = \bar{B}_1^{**}(0)$ and the weak-$*$ topology agrees with the weak topology on X^{**}.

Conversely, suppose $\bar{B}_1(0)$ is weakly compact. Since the weak topology on $J(X)$ is the relative topology of the weak-$*$ topology on X^{**} we conclude that $J(\bar{B}_1(0))$ is compact (and thus closed) in the weak-$*$ topology on X^{**}. But now Glodstine's theorem implies $J(\bar{B}_1(0)) = \bar{B}_1^{**}(0)$ and hence X is reflexive. $\square$

Problem 5.9. *Show that a weakly sequentially compact set is bounded.*

Problem 5.10. *Show that a convex set $K \subseteq X$ is weakly closed if and only if it is weakly sequentially closed.*

Problem 5.11. *Show that* (5.11) *generates the weak topology on $B_1(0) \subset X$. Show that* (5.13) *generates the weak topology on $B_1^*(0) \subset X^*$.*

Problem 5.12. *Show that $A : \mathfrak{D}(A) \subseteq X \to Y$ is closed if and only if for $x_n \in \mathfrak{D}(A)$ with $x_n \rightharpoonup x$ and $Ax_n \rightharpoonup y$ we have $x \in \mathfrak{D}(A)$ and $Ax = y$.*

Problem* 5.13. *Show that the annihilator $M^\perp$ of a set $M \subseteq X$ is weak-$*$ closed. Moreover show that $(N_\perp)^\perp = \overline{\operatorname{span}(N)}^{\text{weak-}*}$. In particular, $(N_\perp)^\perp = \overline{\operatorname{span}(N)}$ if X is reflexive. (Hint: The first part and hence one inclusion of the second part are straightforward. For the other inclusion use Corollary 4.19.)*

Problem 5.14. *Suppose $K \subseteq X$ is convex and x is a boundary point of K. Then there is a supporting hyperplane at x. That is, there is some $\ell \in X^*$ such that $\ell(x) = 0$ and K is contained in the closed half-plane $\{y | \operatorname{Re}(\ell(y - x)) \leq 0\}$.*

5.4. Beyond Banach spaces: Locally convex spaces

We have already seen that it is often important to weaken the notion of convergence (i.e., to weaken the underlying topology) to get a larger class of converging sequences. It turns out that all cases considered so far fit within a general framework which we want to discuss in this section. We start with an alternate definition of a locally convex vector space which we already briefly encountered in Corollary 5.4 (equivalence of both definitions will be established below).

A vector space X together with a topology is called a **locally convex vector space** if there exists a family of seminorms $\{q_\alpha\}_{\alpha\in A}$ which generates the topology in the sense that the topology is the weakest topology for which the family of functions $\{q_\alpha(.-x)\}_{\alpha\in A,x\in X}$ is continuous. Hence the topology is generated by sets of the form $x+q_\alpha^{-1}(I)$, where $I\subseteq[0,\infty)$ is open (in the relative topology). Moreover, sets of the form

$$x+\bigcap_{j=1}^{n} q_{\alpha_j}^{-1}([0,\varepsilon_j)) \tag{5.14}$$

are a neighborhood base at x and hence it is straightforward to check that a locally convex vector space is a topological vector space, that is, both vector addition and scalar multiplication are continuous. For example, if $z=x+y$ then the preimage of the open neighborhood $z+\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j))$ contains the open neighborhood $(x+\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j/2)), y+\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j/2)))$ by virtue of the triangle inequality. Similarly, if $z=\gamma x$ then the preimage of the open neighborhood $z+\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j))$ contains the open neighborhood $(B_\varepsilon(\gamma), x+\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\frac{\varepsilon_j}{2(|\gamma|+\varepsilon)}))$ with $\varepsilon<\frac{\varepsilon_j}{2q_{\alpha_j}(x)}$.

Moreover, note that a sequence x_n will converge to x in this topology if and only if $q_\alpha(x_n-x)\to 0$ for all α.

Example 5.13. Of course every Banach space equipped with the norm topology is a locally convex vector space if we choose the single seminorm $q(x)=\|x\|$. ⋄

Example 5.14. A Banach space X equipped with the weak topology is a locally convex vector space. In this case we have used the continuous linear functionals $\ell\in X^*$ to generate the topology. However, note that the corresponding seminorms $q_\ell(x):=|\ell(x)|$ generate the same topology since $x+q_\ell^{-1}([0,\varepsilon))=\ell^{-1}(B_\varepsilon(x))$ in this case. The same is true for X^* equipped with the weak or the weak-$*$ topology. ⋄

Example 5.15. The bounded linear operators $\mathscr{L}(X,Y)$ together with the seminorms $q_x(A):=\|Ax\|$ for all $x\in X$ (strong convergence) or the seminorms $q_{\ell,x}(A):=|\ell(Ax)|$ for all $x\in X$, $\ell\in Y^*$ (weak convergence) are locally convex vector spaces. ⋄

Example 5.16. The continuous functions $C(I)$ together with the pointwise topology generated by the seminorms $q_x(f) := |f(x)|$ for all $x \in I$ is a locally convex vector space. ⋄

In all these examples we have one additional property which is often required as part of the definition: The seminorms are called **separated** if for every $x \in X$ there is a seminorm with $q_\alpha(x) \neq 0$. In this case the corresponding locally convex space is Hausdorff, since for $x \neq y$ the neighborhoods $U(x) = x + q_\alpha^{-1}([0, \varepsilon))$ and $U(y) = y + q_\alpha^{-1}([0, \varepsilon))$ will be disjoint for $\varepsilon = \frac{1}{2} q_\alpha(x - y) > 0$ (the converse is also true; Problem 5.21).

It turns out crucial to understand when a seminorm is continuous.

Lemma 5.15. *Let X be a locally convex vector space with corresponding family of seminorms $\{q_\alpha\}_{\alpha \in A}$. Then a seminorm q is continuous if and only if there are seminorms q_{α_j} and constants $c_j > 0$, $1 \leq j \leq n$, such that $q(x) \leq \sum_{j=1}^n c_j q_{\alpha_j}(x)$.*

Proof. If q is continuous, then $q^{-1}(B_1(0))$ contains an open neighborhood of 0 of the form $\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0, \varepsilon_j))$ and choosing $c_j = \max_{1 \leq j \leq n} \varepsilon_j^{-1}$ we obtain that $\sum_{j=1}^n c_j q_{\alpha_j}(x) < 1$ implies $q(x) < 1$ and the claim follows from Problem 5.16. Conversely note that if $q(x) = r$ then $q^{-1}(B_\varepsilon(r))$ contains the set $U(x) = x + \bigcap_{j=1}^n q_{\alpha_j}^{-1}([0, \varepsilon_j))$ provided $\sum_{j=1}^n c_j \varepsilon_j \leq \varepsilon$ since $|q(y) - q(x)| \leq q(y - x) \leq \sum_{j=1}^n c_j q_{\alpha_j}(x - y) < \varepsilon$ for $y \in U(x)$. □

Example 5.17. The weak topology on an infinite dimensional space cannot be generated by a norm. Indeed, let q be a continuous seminorm and $q_{\alpha_j} = |\ell_{\alpha_j}|$ as in the lemma. Then $\bigcap_{j=1}^n \operatorname{Ker}(\ell_{\alpha_j})$ has codimension at most n and hence contains some $x \neq 0$ implying that $q(x) \leq \sum_{j=1}^n c_j q_{\alpha_j}(x) = 0$. Thus q is no norm. Similarly, the other examples cannot be generated by a norm except in finite dimensional cases. ⋄

Moreover, note that the topology is translation invariant in the sense that $U(x)$ is a neighborhood of x if and only if $U(x) - x = \{y - x | y \in U(x)\}$ is a neighborhood of 0. Hence we can restrict our attention to neighborhoods of 0 (this is of course true for any topological vector space). Hence if X and Y are topological vector spaces, then a linear map $A : X \to Y$ will be continuous if and only if it is continuous at 0. Moreover, if Y is a locally convex space with respect to some seminorms p_β, then A will be continuous if and only if $p_\beta \circ A$ is continuous for every β (Lemma B.11). Finally, since $p_\beta \circ A$ is a seminorm, the previous lemma implies:

Corollary 5.16. *Let $(X, \{q_\alpha\})$ and $(Y, \{p_\beta\})$ be locally convex vector spaces. Then a linear map $A : X \to Y$ is continuous if and only if for every β there are some seminorms q_{α_j} and constants $c_j > 0$, $1 \leq j \leq n$, such that $p_\beta(Ax) \leq \sum_{j=1}^n c_j q_{\alpha_j}(x)$.*

It will shorten notation when sums of the type $\sum_{j=1}^n c_j q_{\alpha_j}(x)$, which appeared in the last two results, can be replaced by a single expression $c\, q_\alpha$. This can be done if the family of seminorms $\{q_\alpha\}_{\alpha\in A}$ is **directed**, that is, for given $\alpha, \beta \in A$ there is a $\gamma \in A$ such that $q_\alpha(x) + q_\beta(x) \le C q_\gamma(x)$ for some $C > 0$. Moreover, if $\mathcal{F}(A)$ is the set of all finite subsets of A, then $\{\tilde{q}_F = \sum_{\alpha\in F} q_\alpha\}_{F\in\mathcal{F}(A)}$ is a directed family which generates the same topology (since every $\tilde{q}_F$ is continuous with respect to the original family we do not get any new open sets).

While the family of seminorms is in most cases more convenient to work with, it is important to observe that different families can give rise to the same topology and it is only the topology which matters for us. In fact, it is possible to characterize locally convex vector spaces as topological vector spaces which have a neighborhood basis at 0 of absolutely convex sets. Here a set U is called **absolutely convex**, if for $|\alpha|+|\beta| \le 1$ we have $\alpha U + \beta U \subseteq U$. Since the sets $q_\alpha^{-1}([0,\varepsilon))$ are absolutely convex we always have such a basis in our case. To see the converse note that such a neighborhood U of 0 is also absorbing (Problem 5.15) und hence the corresponding Minkowski functional (5.1) is a seminorm (Problem 5.20). By construction, these seminorms generate the topology since if $U_0 = \bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j)) \subseteq U$ we have for the corresponding Minkowski functionals $p_U(x) \le p_{U_0}(x) \le \varepsilon^{-1} \sum_{j=1}^n q_{\alpha_j}(x)$, where $\varepsilon = \min \varepsilon_j$. With a little more work (Problem 5.19), one can even show that it suffices to assume to have a neighborhood basis at 0 of convex open sets.

Given a topological vector space X we can define its dual space X^* as the set of all continuous linear functionals. However, while it can happen in general that the dual space is empty, X^* will always be nontrivial for a locally convex space since the Hahn–Banach theorem can be used to construct linear functionals (using a continuous seminorm for φ in Theorem 4.14) and also the geometric Hahn–Banach theorem (Theorem 5.3) holds (see also its corollaries). In this respect note that for every continuous linear functional ℓ in a topological vector space $|\ell|^{-1}([0,\varepsilon))$ is an absolutely convex open neighborhoods of 0 and hence existence of such sets is necessary for the existence of nontrivial continuous functionals. As a natural topology on X^* we could use the weak-$*$ topology defined to be the weakest topology generated by the family of all point evaluations $q_x(\ell) = |\ell(x)|$ for all $x \in X$. Since different linear functionals must differ at least at one point the weak-$*$ topology is Hausdorff. Given a continuous linear operator $A : X \to Y$ between locally convex spaces we can define its adjoint $A' : Y^* \to X^*$ as before,

$$(A'y^*)(x) := y^*(Ax). \tag{5.15}$$

A brief calculation

$$q_x(A'y^*) = |(A'y^*)(x)| = |y^*(Ax)| = q_{Ax}(y^*) \tag{5.16}$$

verifies that A' is continuous in the weak-$*$ topology by virtue of Corollary 5.16.

The remaining theorems we have established for Banach spaces were consequences of the Baire theorem (which requires a complete metric space) and this leads us to the question when a locally convex space is a metric space. From our above analysis we see that a locally convex vector space will be first countable if and only if countably many seminorms suffice to determine the topology. In this case X turns out to be metrizable.

Theorem 5.17. *A locally convex Hausdorff space is metrizable if and only if it is first countable. In this case there is a countable family of separated seminorms $\{q_n\}_{n\in\mathbb{N}}$ generating the topology and a metric is given by*

$$d(x,y) := \max_{n\in\mathbb{N}} \frac{1}{2^n} \frac{q_n(x-y)}{1+q_n(x-y)}. \tag{5.17}$$

Proof. If X is first countable there is a countable neighborhood base at 0 and hence also a countable neighborhood base of absolutely convex sets. The Minkowski functionals corresponding to the latter base are seminorms of the required type.

Now in this case it is straightforward to check that (5.17) defines a metric (see also Problem B.3). Moreover, the balls $B_r^m(x) = \bigcap_{n:2^{-n}>r}\{y|q_n(y-x) < \frac{r}{2^{-n}-r}\}$ are clearly open and convex (note that the intersection is finite). Conversely, for every set of the form (5.14) we can choose $\varepsilon = \min\{2^{-\alpha_j}\frac{\varepsilon_j}{1+\varepsilon_j}|1 \le j \le n\}$ such that $B_\varepsilon(x)$ will be contained in this set. Hence both topologies are equivalent (cf. Lemma B.2). □

In general, a locally convex vector space X which has a separated countable family of seminorms is called a **Fréchet space** if it is complete with respect to the metric (5.17). Note that the metric (5.17) is **translation invariant**

$$d(f,g) = d(f-h, g-h). \tag{5.18}$$

Example 5.18. The continuous functions $C(\mathbb{R})$ together with local uniform convergence are a Fréchet space. A countable family of seminorms is for example

$$\|f\|_j = \sup_{|x|\le j} |f(x)|, \qquad j \in \mathbb{N}. \tag{5.19}$$

Then $f_k \to f$ if and only if $\|f_k - f\|_j \to 0$ for all $j \in \mathbb{N}$ and it follows that $C(\mathbb{R})$ is complete. ◇

Example 5.19. The space $C^\infty(\mathbb{R}^m)$ together with the seminorms

$$\|f\|_{j,k} = \sum_{|\alpha|\le k} \sup_{|x|\le j} |\partial_\alpha f(x)|, \qquad j \in \mathbb{N},\ k \in \mathbb{N}_0, \tag{5.20}$$

is a Fréchet space.

Note that $\partial_\alpha : C^\infty(\mathbb{R}^m) \to C^\infty(\mathbb{R}^m)$ is continuous. Indeed by Corollary 5.16 it suffices to observe that $\|\partial_\alpha f\|_{j,k} \le \|f\|_{j,k+|\alpha|}$. ⋄

Example 5.20. The **Schwartz space**

$$\mathcal{S}(\mathbb{R}^m) = \{f \in C^\infty(\mathbb{R}^m) | \sup_x |x^\alpha(\partial_\beta f)(x)| < \infty,\ \forall \alpha, \beta \in \mathbb{N}_0^m\} \tag{5.21}$$

together with the seminorms

$$q_{\alpha,\beta}(f) = \|x^\alpha(\partial_\beta f)(x)\|_\infty, \qquad \alpha, \beta \in \mathbb{N}_0^m. \tag{5.22}$$

To see completeness note that a Cauchy sequence f_n is in particular a Cauchy sequence in $C^\infty(\mathbb{R}^m)$. Hence there is a limit $f \in C^\infty(\mathbb{R}^m)$ such that all derivatives converge uniformly. Moreover, since Cauchy sequences are bounded $\|x^\alpha(\partial_\beta f_n)(x)\|_\infty \le C_{\alpha,\beta}$ we obtain $\|x^\alpha(\partial_\beta f)(x)\|_\infty \le C_{\alpha,\beta}$ and thus $f \in \mathcal{S}(\mathbb{R}^m)$.

Again $\partial_\gamma : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$ is continuous since $q_{\alpha,\beta}(\partial_\gamma f) \le q_{\alpha,\beta+\gamma}(f)$ and so is $x^\gamma : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$ since $q_{\alpha,\beta}(x^\gamma f) \le \sum_{\eta\le\beta} \binom{\beta}{\eta} \frac{\gamma!}{(\gamma-\eta)!} q_{\alpha+\gamma-\eta,\beta-\eta}(f)$.

The dual space $\mathcal{S}^*(\mathbb{R}^m)$ is known as the space of **tempered distributions**. ⋄

Example 5.21. The space of all entire functions $f(z)$ (i.e. functions which are holomorphic on all of $\mathbb{C}$) together with the seminorms $\|f\|_j = \sup_{|z|\le j} |f(z)|$, $j \in \mathbb{N}$, is a Fréchet space. Completeness follows from the Weierstraß convergence theorem which states that a limit of holomorphic functions which is uniform on every compact subset is again holomorphic. ⋄

Example 5.22. In all of the previous examples the topology cannot be generated by a norm. For example, if q is a norm for $C(\mathbb{R})$, then Lemma 5.15 that there is some index j such that $q(f) \le C\|f\|_j$. Now choose a nonzero function which vanishes on $[-j, j]$ to get a contradiction. ⋄

There is another useful criterion when the topology can be described by a single norm. To this end we call a set $B \subseteq X$ bounded if $\sup_{x\in B} q_\alpha(x) < \infty$ for every α. By Corollary 5.16 this will then be true for any continuous seminorm on X.

Theorem 5.18 (Kolmogorov)**.** *A locally convex vector space can be generated from a single seminorm if and only if it contains a bounded open set.*

Proof. In a Banach space every open ball is bounded and hence only the converse direction is nontrivial. So let U be a bounded open set. By shifting

and decreasing U if necessary we can assume U to be an absolutely convex open neighborhood of 0 and consider the associated Minkowski functional $q = p_U$. Then since $U = \{x|q(x) < 1\}$ and $\sup_{x\in U} q_\alpha(x) = C_\alpha < \infty$ we infer $q_\alpha(x) \le C_\alpha q(x)$ (Problem 5.16) and thus the single seminorm q generates the topology. □

Finally, we mention that, since the Baire category theorem holds for arbitrary complete metric spaces, the open mapping theorem (Theorem 4.5), the inverse mapping theorem (Theorem 4.6) and the closed graph (Theorem 4.7) hold for Fréchet spaces without modifications. In fact, they are formulated such that it suffices to replace Banach by Fréchet in these theorems as well as their proofs (concerning the proof of Theorem 4.5 take into account Problems 5.15 and 5.22).

Problem* 5.15. *In a topological vector space every neighborhood U of 0 is absorbing.*

Problem* 5.16. *Let p, q be two seminorms. Then $p(x) \le Cq(x)$ if and only if $q(x) < 1$ implies $p(x) < C$.*

Problem 5.17. *Let X be a vector space. We call a set U **balanced** if $\alpha U \subseteq U$ for every $|\alpha| \le 1$. Show that a set is balanced and convex if and only if it is absolutely convex.*

Problem* 5.18. *The intersection of arbitrary absolutely convex/balanced sets is again absolutely convex/balanced convex. Hence we can define the absolutely convex/balanced hull of a set U as the smallest absolutely convex/balanced set containing U, that is, the intersection of all absolutely convex/balanced sets containing U. Show that the absolutely convex hull is given by*

$$\operatorname{ahull}(U) := \{\sum_{j=1}^{n} \lambda_j x_j | n \in \mathbb{N},\ x_j \in U,\ \sum_{j=1}^{n} |\lambda_j| \le 1\}$$

and the balanced hull by

$$\operatorname{bhull}(U) := \{\alpha x | x \in U,\ |\alpha| \le 1\}.$$

Show that $\operatorname{ahull}(U) = \operatorname{conv}(\operatorname{bhull}(U))$.

Problem* 5.19. *In a topological vector space every convex open neighborhood U of zero contains an absolutely convex open neighborhood of zero. (Hint: By continuity of the scalar multiplication U contains a set of the form $B_\varepsilon^{\mathbb{C}}(0) \cdot V$, where V is an open neighborhood of zero.)*

Problem* 5.20. *Let X be a vector space. Show that the Minkowski functional of a balanced, convex, absorbing set is a seminorm.*

Problem* 5.21. *If a locally convex space is Hausdorff then any corresponding family of seminorms is separated.*

Problem* 5.22. *Suppose X is a complete vector space with a translation invariant metric d. Show that $\sum_{j=1}^{\infty} d(0, x_j) < \infty$ implies that*

$$\sum_{j=1}^{\infty} x_j = \lim_{n\to\infty} \sum_{j=1}^{n} x_j$$

exists and

$$d(0, \sum_{j=1}^{\infty} x_j) \leq \sum_{j=1}^{\infty} d(0, x_j)$$

in this case (compare also Problem 1.4).

Problem 5.23. *Instead of* (5.17) *one frequently uses*

$$\tilde{d}(x, y) := \sum_{n\in\mathbb{N}} \frac{1}{2^n} \frac{q_n(x-y)}{1+q_n(x-y)}.$$

Show that this metric generates the same topology.

Consider the Fréchet space $C(\mathbb{R})$ with $q_n(f) = \sup_{[-n,n]} |f|$. Show that the metric balls with respect to $\tilde{d}$ are not *convex.*

Problem 5.24. *Suppose X is a metric vector space. Then balls are convex if and only if the metric is quasiconvex:*

$$d(\lambda x + (1-\lambda)y, z) \leq \max\{d(x, z), d(y, z)\}, \qquad \lambda \in (0, 1).$$

(See also Problem 4.43.)

Problem 5.25. *Consider $\ell^p(\mathbb{N})$ for $p \in (0, 1)$ — compare Problem 1.14. Show that $\|.\|_p$ is not convex. Show that every convex open set is unbounded. Conclude that it is not a locally convex vector space. (Hint: Consider $B_R(0)$. Then for $r < R$ all vectors which have one entry equal to r and all other entries zero are in this ball. By taking convex combinations all vectors which have n entries equal to r/n are in the convex hull. The quasinorm of such a vector is $n^{1/p-1}r$.)*

Problem 5.26. *Show that $C_c^\infty(\mathbb{R}^m)$ is dense in $\mathcal{S}(\mathbb{R}^m)$.*

Problem 5.27. *Let X be a topological vector space and M a closed subspace. Show that the quotient space X/M is again a topological vector space and that $\pi : X \to X/M$ is linear, continuous, and open. Show that points in X/M are closed.*

5.5. Uniformly convex spaces

In a Banach space X, the unit ball is convex by the triangle inequality. Moreover, X is called **strictly convex** if the unit ball is a strictly convex set, that is, if for any two points on the unit sphere their average is inside the unit ball. See Problem 1.12 for some equivalent definitions. This is illustrated in Figure 1.1 which shows that in $\mathbb{R}^2$ this is only true for $1 < p < \infty$.

Example 5.23. By Problem 1.12 it follows that $\ell^p(\mathbb{N})$ is strictly convex for $1 < p < \infty$ but not for $p = 1, \infty$. ⋄

A more qualitative notion is to require that if two unit vectors x, y satisfy $\|x-y\| \geq \varepsilon$ for some $\varepsilon > 0$, then there is some $\delta > 0$ such that $\|\frac{x+y}{2}\| \leq 1-\delta$. In this case one calls X **uniformly convex** and

$$\delta(\varepsilon) := \inf \left\{ 1 - \|\tfrac{x+y}{2}\| \,\middle|\, \|x\| = \|y\| = 1,\ \|x-y\| \geq \varepsilon \right\}, \quad 0 \leq \varepsilon \leq 2, \tag{5.23}$$

is called the modulus of convexity. Of course every uniformly convex space is strictly convex. In finite dimensions the converse is also true (Problem 5.30).

Note that δ is nondecreasing and

$$\|\tfrac{x+y}{2}\| = \|x - \tfrac{x-y}{2}\| \geq 1 - \frac{\varepsilon}{2}$$

shows $0 \leq \delta(\varepsilon) \leq \frac{\varepsilon}{2}$. Moreover, $\delta(2) = 1$ implies X strictly convex. In fact in this case $1 = \delta(2) \leq 1 - \|\frac{x+y}{2}\| \leq 1$ for $2 \leq \|x-y\| \leq 2$. That is, $x = -y$ whenever $\|x-y\| = 2 = \|x\| + \|y\|$.

Example 5.24. Every Hilbert space is uniformly convex with modulus of convexity $\delta(\varepsilon) = 1 - \sqrt{1 - \frac{\varepsilon^2}{4}}$ (Problem 5.28). ⋄

Example 5.25. Consider $C[0,1]$ with the norm

$$\|x\| := \|x\|_\infty + \|x\|_2 = \max_{t\in[0,1]} |x(t)| + \left(\int_0^1 |x(t)|^2 dt \right)^{-1}.$$

Note that by $\|x\|_2 \leq \|x\|_\infty$ this norm is equivalent to the usual one: $\|x\|_\infty \leq \|x\| \leq 2\|x\|_\infty$. While with the usual norm $\|.\|_\infty$ this space is not strictly convex, it is with the new one. To see this we use (i) from Problem 1.12. Then if $\|x+y\| = \|x\| + \|y\|$ we must have both $\|x+y\|_\infty = \|x\|_\infty + \|y\|_\infty$ and $\|x+y\|_2 = \|x\|_2 + \|y\|_2$. Hence strict convexity of $\|.\|_2$ implies strict convexity of $\|.\|$.

Note however, that $\|.\|$ is not uniformly convex. In fact, since by the Milman–Pettis theorem below, every uniformly convex space is reflexive, there cannot be an equivalent norm on $C[0,1]$ which is uniformly convex (cf. Example 4.20). ⋄

Example 5.26. It can be shown that $\ell^p(\mathbb{N})$ is uniformly convex for $1 < p < \infty$ (see Theorem 10.11). ⋄

Equivalently, uniform convexity implies that if the average of two unit vectors is close to the boundary, then they must be close to each other. Specifically, if $\|x\| = \|y\| = 1$ and $\|\frac{x+y}{2}\| > 1 - \delta(\varepsilon)$ then $\|x - y\| < \varepsilon$. The following result (which generalizes Lemma 4.29) uses this observation:

Theorem 5.19 (Radon–Riesz theorem)**.** *Let X be a uniformly convex Banach space and let $x_n \rightharpoonup x$. Then $x_n \to x$ if and only if* $\limsup \|x_n\| \leq \|x\|$.

Proof. If $x = 0$ there is nothing to prove. Hence we can assume $x_n \neq 0$ for all n and consider $y_n := \frac{x_n}{\|x_n\|}$. Then $y_n \rightharpoonup y := \frac{x}{\|x\|}$ and it suffices to show $y_n \to y$. Next choose a linear functional $\ell \in X^*$ with $\|\ell\| = 1$ and $\ell(y) = 1$. Then

$$\ell\big(\frac{y_n + y}{2}\big) \leq \Big\|\frac{y_n + y}{2}\Big\| \leq 1$$

and letting $n \to \infty$ shows $\|\frac{y_n+y}{2}\| \to 1$. Finally uniform convexity shows $y_n \to y$. □

For the proof of the next result we need to following equivalent condition.

Lemma 5.20. *Let X be a Banach space. Then*

$$\delta(\varepsilon) = \inf\Big\{1 - \|\tfrac{x+y}{2}\| \Big| \ \|x\| \leq 1, \|y\| \leq 1, \ \|x - y\| \geq \varepsilon\Big\} \tag{5.24}$$

for $0 \leq \varepsilon \leq 2$.

Proof. It suffices to show that for given x and y which are not both on the unit sphere there is a better pair in the real subspace spanned by these vectors. By scaling we could get a better pair if both were strictly inside the unit ball and hence we can assume at least one vector to have norm one, say $\|x\| = 1$. Moreover, consider

$$u(t) := \frac{\cos(t)x + \sin(t)y}{\|\cos(t)x + \sin(t)y\|}, \qquad v(t) := u(t) + (y - x).$$

Then $\|v(0)\| = \|y\| < 1$. Moreover, let $t_0 \in (\frac{\pi}{2}, \frac{3\pi}{4})$ be the value such that the line from x to $u(t_0)$ passes through y. Then, by convexity we must have $\|v(t_0)\| > 1$ and by the intermediate value theorem there is some $0 < t_1 < t_0$ with $\|v(t_1)\| = 1$. Let $u := u(t_1)$, $v := v(t_1)$. The line through u and x is not parallel to the line through 0 and $x + y$ and hence there are $\alpha, \lambda \geq 0$ such that

$$\frac{\alpha}{2}(x + y) = \lambda u + (1 - \lambda)x.$$

Moreover, since the line from x to u is above the line from x to y (since $t_1 < t_0$) we have $\alpha \geq 1$. Rearranging this equation we get

$$\frac{\alpha}{2}(u + v) = (\alpha + \lambda)u + (1 - \alpha - \lambda)x.$$

Now, by convexity of the norm, if $\lambda \leq 1$ we have $\lambda + \alpha > 1$ and thus $\|\lambda u + (1-\lambda)x\| \leq 1 < \|(\alpha+\lambda)u + (1-\alpha-\lambda)x\|$. Similarly, if $\lambda > 1$ we have $\|\lambda u + (1-\lambda)x\| < \|(\alpha+\lambda)u + (1-\alpha-\lambda)x\|$ again by convexity of the norm. Hence $\|\frac{1}{2}(x+y)\| \leq \|\frac{1}{2}(u+v)\|$ and u, v is a better pair. □

Now we can prove:

Theorem 5.21 (Milman–Pettis)**.** *A uniformly convex Banach space is reflexive.*

Proof. Pick some $x'' \in X^{**}$ with $\|x''\| = 1$. It suffices to find some $x \in \bar{B}_1(0)$ with $\|x'' - J(x)\| \leq \varepsilon$. So fix $\varepsilon > 0$ and $\delta := \delta(\varepsilon)$, where $\delta(\varepsilon)$ is the modulus of convexity. Then $\|x''\| = 1$ implies that we can find some $\ell \in X^*$ with $\|\ell\| = 1$ and $|x''(\ell)| > 1 - \frac{\delta}{2}$. Consider the weak-$*$ neighborhood

$$U := \{y'' \in X^{**} \,|\, |(y'' - x'')(\ell)| < \tfrac{\delta}{2}\}$$

of x''. By Goldstine's theorem (Theorem 5.13) there is some $x \in \bar{B}_1(0)$ with $J(x) \in U$ and this is the x we are looking for. In fact, suppose this were not the case. Then the set $V := X^{**} \setminus \bar{B}^{**}_{\varepsilon}(J(x))$ is another weak-$*$ neighborhood of x'' (since $\bar{B}^{**}_{\varepsilon}(J(x))$ is weak-$*$ compact by the Banach-Alaoglu theorem) and appealing again to Goldstine's theorem there is some $y \in \bar{B}_1(0)$ with $J(y) \in U \cap V$. Since $x, y \in U$ we obtain

$$1 - \tfrac{\delta}{2} < |x''(\ell)| \leq |\ell(\tfrac{x+y}{2})| + \tfrac{\delta}{2} \quad \Rightarrow \quad 1 - \delta < |\ell(\tfrac{x+y}{2})| \leq \|\tfrac{x+y}{2}\|,$$

a contradiction to uniform convexity since $\|x - y\| \geq \varepsilon$. □

Problem* 5.28. *Show that a Hilbert space is uniformly convex. (Hint: Use the parallelogram law.)*

Problem 5.29. *A Banach space X is uniformly convex if and only if $\|x_n\| = \|y_n\| = 1$ and $\|\frac{x_n+y_n}{2}\| \to 1$ implies $\|x_n - y_n\| \to 0$.*

Problem* 5.30. *Show that a finite dimensional space is uniformly convex if and only if it is strictly convex.*

Problem 5.31. *Let X be strictly convex. Show that every nonzero linear functional attains its norm for at most one unit vector (cf. Problem 4.17).*

Chapter 6

Bounded linear operators

We have started out our study by looking at eigenvalue problems which, from a historic view point, were one of the key problems driving the development of functional analysis. In Chapter 3 we have investigated compact operators in Hilbert space and we have seen that they allow a treatment similar to what is known from matrices. However, more sophisticated problems will lead to operators whose spectra consist of more than just eigenvalues. Hence we want to go one step further and look at spectral theory for bounded operators. Here one of the driving forces was the development of quantum mechanics (there even the boundedness assumption is too much — but first things first). A crucial role is played by the algebraic structure, namely recall from Section 1.6 that the bounded linear operators on X form a Banach space which has a (non-commutative) multiplication given by composition. In order to emphasize that it is only this algebraic structure which matters, we will develop the theory from this abstract point of view. While the reader should always remember that bounded operators on a Hilbert space is what we have in mind as the prime application, examples will apply these ideas also to other cases thereby justifying the abstract approach.

To begin with, the operators could be on a Banach space (note that even if X is a Hilbert space, $\mathscr{L}(X)$ will only be a Banach space) but eventually again self-adjointness will be needed. Hence we will need the additional operation of taking adjoints.

6.1. Banach algebras

A Banach space X together with a multiplication satisfying

$$(x+y)z = xz + yz, \quad x(y+z) = xy + xz, \quad x, y, z \in X, \tag{6.1}$$

and

$$(xy)z = x(yz), \qquad \alpha\,(xy) = (\alpha x)y = x\,(\alpha y), \quad \alpha \in \mathbb{C}, \tag{6.2}$$

and

$$\|xy\| \le \|x\|\|y\|. \tag{6.3}$$

is called a **Banach algebra**. In particular, note that (6.3) ensures that multiplication is continuous (Problem 6.1). In fact, one can show that (separate) continuity of multiplication implies existence of an equivalent norm satisfying (6.3) (Problem 6.2).

An element $e \in X$ satisfying

$$ex = xe = x, \qquad \forall x \in X \tag{6.4}$$

is called **identity** (show that e is unique) and we will assume $\|e\| = 1$ in this case (by Problem 6.2 this can be done without loss of generality).

Example 6.1. The continuous functions $C(I)$ over some compact interval form a commutative Banach algebra with identity 1. ⋄

Example 6.2. The differentiable functions $C^n(I)$ over some compact interval do not form a commutative Banach algebra since (6.3) fails for $n \ge 1$. However, the equivalent norm

$$\|f\|_{\infty,n} := \sum_{k=0}^{n} \frac{\|f^{(k)}\|_\infty}{k!}$$

remedies this problem. ⋄

Example 6.3. The bounded linear operators $\mathscr{L}(X)$ form a Banach algebra with identity $\mathbb{I}$. ⋄

Example 6.4. The bounded sequences $\ell^\infty(\mathbb{N})$ together with the componentwise product form a commutative Banach algebra with identity 1. ⋄

Example 6.5. The space of all periodic continuous functions which have an absolutely convergent Fourier series $\mathcal{A}$ together with the norm

$$\|f\|_{\mathcal{A}} := \sum_{k\in\mathbb{Z}} |\hat{f}_k|$$

and the usual product is known as the **Wiener algebra**. Of course as a Banach space it is isomorphic to $\ell^1(\mathbb{Z})$ via the Fourier transform. To see

that it is a Banach algebra note that

$$f(x)g(x) = \sum_{k\in\mathbb{Z}} \hat{f}_k e^{ikx} \sum_{j\in\mathbb{Z}} \hat{g}_j e^{ijx} = \sum_{k,j\in\mathbb{Z}} \hat{f}_k \hat{g}_j e^{i(k+j)x}$$
$$= \sum_{k\in\mathbb{Z}} \Big(\sum_{j\in\mathbb{Z}} \hat{f}_j \hat{g}_{k-j} \Big) e^{ikx}.$$

Moreover, interchanging the order of summation

$$\|fg\|_{\mathcal{A}} = \sum_{k\in\mathbb{Z}} \Big| \sum_{j\in\mathbb{Z}} \hat{f}_j \hat{g}_{k-j} \Big| \le \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} |\hat{f}_j| |\hat{g}_{k-j}| = \|f\|_{\mathcal{A}} \|g\|_{\mathcal{A}}$$

shows that $\mathcal{A}$ is a Banach algebra. The identity is of course given by $e(x) \equiv 1$. Moreover, note that $\mathcal{A} \subseteq C_{per}[-\pi, \pi]$ and $\|f\|_\infty \le \|f\|_{\mathcal{A}}$. ⋄

Example 6.6. The space $L^1(\mathbb{R}^n)$ together with the convolution

$$(g * f)(x) := \int_{\mathbb{R}^n} g(x-y) f(y) dy = \int_{\mathbb{R}^n} g(y) f(x-y) dy \tag{6.5}$$

is a commutative Banach algebra (Problem 6.10) without identity. ⋄

A Banach algebra with identity is also known as **unital** and we will assume X to be a Banach algebra with identity e throughout the rest of this section. Note that an identity can always be added if needed (Problem 6.3).

An element $x \in X$ is called **invertible** if there is some $y \in X$ such that

$$xy = yx = e. \tag{6.6}$$

In this case y is called the inverse of x and is denoted by x^{-1}. It is straightforward to show that the inverse is unique (if one exists at all) and that

$$(xy)^{-1} = y^{-1}x^{-1}, \qquad (x^{-1})^{-1} = x. \tag{6.7}$$

In particular, the set of invertible elements $\mathcal{G}(X)$ forms a group under multiplication.

Example 6.7. If $X = \mathscr{L}(\mathbb{C}^n)$ is the set of n by n matrices, then $\mathcal{G}(X) = \mathrm{GL}(n)$ is the general linear group. ⋄

Example 6.8. Let $X = \mathscr{L}(\ell^p(\mathbb{N}))$ and recall the shift operators $S^\pm$ defined via $(S^\pm a)_j = a_{j\pm1}$ with the convention that $a_0 = 0$. Then $S^+S^- = \mathbb{I}$ but $S^-S^+ \neq \mathbb{I}$. Moreover, note that S^+S^- is invertible while S^-S^+ is not. So you really need to check both $xy = e$ and $yx = e$ in general. ⋄

If x is invertible then the same will be true for any nearby elements. This will be a consequence from the following straightforward generalization of the geometric series to our abstract setting.

Lemma 6.1. *Let X be a Banach algebra with identity e. Suppose $\|x\| < 1$. Then $e - x$ is invertible and*

$$(e-x)^{-1} = \sum_{n=0}^{\infty} x^n. \tag{6.8}$$

Proof. Since $\|x\| < 1$ the series converges and

$$(e-x)\sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} x^n - \sum_{n=1}^{\infty} x^n = e$$

respectively

$$\left(\sum_{n=0}^{\infty} x^n\right)(e-x) = \sum_{n=0}^{\infty} x^n - \sum_{n=1}^{\infty} x^n = e. \qquad \square$$

Corollary 6.2. *Suppose x is invertible and $\|x^{-1}y\| < 1$ or $\|yx^{-1}\| < 1$. Then $(x-y)$ is invertible as well and*

$$(x-y)^{-1} = \sum_{n=0}^{\infty} (x^{-1}y)^n x^{-1} \quad \text{or} \quad (x-y)^{-1} = \sum_{n=0}^{\infty} x^{-1}(yx^{-1})^n. \tag{6.9}$$

In particular, both conditions are satisfied if $\|y\| < \|x^{-1}\|^{-1}$ and the set of invertible elements $\mathcal{G}(X)$ is open and taking the inverse is continuous:

$$\|(x-y)^{-1} - x^{-1}\| \le \frac{\|y\|\|x^{-1}\|^2}{1 - \|x^{-1}y\|}. \tag{6.10}$$

Proof. Just observe $x - y = x(e - x^{-1}y) = (e - yx^{-1})x$. $\square$

The **resolvent set** is defined as

$$\rho(x) := \{\alpha \in \mathbb{C} | \exists (x-\alpha)^{-1}\} \subseteq \mathbb{C}, \tag{6.11}$$

where we have used the shorthand notation $x - \alpha = x - \alpha e$. Its complement is called the **spectrum**

$$\sigma(x) := \mathbb{C} \setminus \rho(x). \tag{6.12}$$

It is important to observe that the inverse has to exist as an element of X. That is, if the elements of X are bounded linear operators, it does not suffice that $x - \alpha$ is injective, as it might not be surjective. If it is bijective, boundedness of the inverse will come for free from the inverse mapping theorem.

Example 6.9. If $X := \mathscr{L}(\mathbb{C}^n)$ is the space of n by n matrices, then the spectrum is just the set of eigenvalues. More general, if X are the bounded linear operators on an infinite-dimensional Hilbert or Banach space, then every eigenvalue will be in the spectrum but the converse is not true in general as an injective operator might not be surjective. In fact, this already

can happen for compact operators where 0 could be in the spectrum without being an eigenvalue. ⋄

Example 6.10. If $X := C(I)$, then the spectrum of a function $x \in C(I)$ is just its range, $\sigma(x) = x(I)$. Indeed, if $\alpha \notin \operatorname{Ran}(x)$ then $t \mapsto (x(t) - \alpha)^{-1}$ is the inverse of $x-\alpha$ (note that $\operatorname{Ran}(x)$ is compact). Conversely, if $\alpha \in \operatorname{Ran}(x)$ and y were an inverse, then $y(t_0)(x(t_0) - \alpha) = 1$ gives a contradiction for any $t_0 \in I$ with $f(t_0) = \alpha$. ⋄

Example 6.11. If $X = \mathcal{A}$ is the Wiener algebra, then, as in the previous example, every function which vanishes at some point cannot be inverted. If it does not vanish anywhere, it can be inverted and the inverse will be a continuous function. But will it again have a convergent Fourier series, that is, will it be in the Wiener Algebra? The affirmative answer of this question is a famous theorem of Wiener, which will be given later in Theorem 6.24. ⋄

The map $\alpha \mapsto (x - \alpha)^{-1}$ is called the **resolvent** of $x \in X$. If $\alpha_0 \in \rho(x)$ we can choose $x \to x - \alpha_0$ and $y \to \alpha - \alpha_0$ in (6.9) which implies

$$(x-\alpha)^{-1} = \sum_{n=0}^{\infty} (\alpha-\alpha_0)^n (x-\alpha_0)^{-n-1}, \quad |\alpha-\alpha_0| < \|(x-\alpha_0)^{-1}\|^{-1}. \tag{6.13}$$

In particular, since the radius of convergence cannot exceed the distance to the spectrum (since everything within the radius of convergent must belong to the resolvent set), we see that the norm of the resolvent must diverge

$$\|(x - \alpha)^{-1}\| \geq \frac{1}{\operatorname{dist}(\alpha, \sigma(x))} \tag{6.14}$$

as α approaches the spectrum. Moreover, this shows that $(x - \alpha)^{-1}$ has a convergent power series with coefficients in X around every point $\alpha_0 \in \rho(x)$. As in the case of coefficients in $\mathbb{C}$, such functions will be called **analytic**. In particular, $\ell((x - \alpha)^{-1})$ is a complex-valued analytic function for every $\ell \in X^*$ and we can apply well-known results from complex analysis:

Theorem 6.3. *For every $x \in X$, the spectrum $\sigma(x)$ is compact, nonempty and satisfies*

$$\sigma(x) \subseteq \{\alpha \,|\, |\alpha| \leq \|x\|\}. \tag{6.15}$$

Proof. Equation (6.13) already shows that $\rho(x)$ is open. Hence $\sigma(x)$ is closed. Moreover, $x - \alpha = -\alpha(e - \frac{1}{\alpha}x)$ together with Lemma 6.1 shows

$$(x - \alpha)^{-1} = -\frac{1}{\alpha} \sum_{n=0}^{\infty} \left(\frac{x}{\alpha}\right)^n, \quad |\alpha| > \|x\|,$$

which implies $\sigma(x) \subseteq \{\alpha| \, |\alpha| \le \|x\|\}$ is bounded and thus compact. Moreover, taking norms shows

$$\|(x-\alpha)^{-1}\| \le \frac{1}{|\alpha|} \sum_{n=0}^{\infty} \frac{\|x\|^n}{|\alpha|^n} = \frac{1}{|\alpha| - \|x\|}, \quad |\alpha| > \|x\|,$$

which implies $(x-\alpha)^{-1} \to 0$ as $\alpha \to \infty$. In particular, if $\sigma(x)$ is empty, then $\ell((x-\alpha)^{-1})$ is an entire analytic function which vanishes at infinity. By Liouville's theorem we must have $\ell((x-\alpha)^{-1}) = 0$ for all $\ell \in X^*$ in this case, and so $(x-\alpha)^{-1} = 0$, which is impossible. □

Example 6.12. The spectrum of the matrix

$$A := \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -c_0 & -c_1 & \cdots & \cdots & -c_{n-1} \end{pmatrix}$$

is given by the zeros of the polynomial (show this)

$$\det(z\mathbb{I} - A) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0.$$

Hence the fact that $\sigma(A)$ is nonempty implies the **fundamental theorem of algebra**, that every non-constant polynomial has at least one zero. ⋄

As another simple consequence we obtain:

Theorem 6.4 (Gelfand–Mazur)**.** *Suppose X is a Banach algebra in which every element except 0 is invertible. Then X is isomorphic to $\mathbb{C}$.*

Proof. Pick $x \in X$ and $\alpha \in \sigma(x)$. Then $x - \alpha$ is not invertible and hence $x - \alpha = 0$, that is $x = \alpha$. Thus every element is a multiple of the identity. □

Now we look at functions of x. Given a polynomial $p(\alpha) = \sum_{j=0}^{n} p_j \alpha^j$ we of course set

$$p(x) := \sum_{j=0}^{n} p_j x^j. \tag{6.16}$$

In fact, we could easily extend this definition to arbitrary convergent power series whose radius of convergence is larger than $\|x\|$ (cf. Problem 1.35). While this will give a nice functional calculus sufficient for many applications our aim is the spectral theorem which will allow us to handle arbitrary continuous functions. Since continuous functions can be approximated by polynomials by the Weierstraß theorem, polynomials will be sufficient for now. Moreover, the following result will be one of the two key ingredients for the proof of the spectral theorem.

Theorem 6.5 (Spectral mapping)**.** *For every polynomial p and $x \in X$ we have*

$$\sigma(p(x)) = p(\sigma(x)), \tag{6.17}$$

where $p(\sigma(x)) := \{p(\alpha)|\alpha \in \sigma(x)\}$.

Proof. Let $\alpha \in \sigma(x)$ and observe

$$p(x) - p(\alpha) = (x - \alpha)q(x).$$

But since $(x - \alpha)$ is not invertible, the same is true for $(x - \alpha)q(x) = q(x)(x - \alpha)$ by Problem 6.6 and hence $p(\alpha) \in p(\sigma(x))$.

Conversely, let $\beta \in \sigma(p(x))$. Then

$$p(x) - \beta = a(x - \lambda_1) \cdots (x - \lambda_n)$$

and at least one $\lambda_j \in \sigma(x)$ since otherwise the right-hand side would be invertible. But then $\beta = p(\lambda_j) \in p(\sigma(x))$. □

The second key ingredient for the proof of the spectral theorem is the **spectral radius**

$$r(x) := \sup_{\alpha \in \sigma(x)} |\alpha| \tag{6.18}$$

of x. Note that by (6.15) we have

$$r(x) \leq \|x\|. \tag{6.19}$$

As our next theorem shows, it is related to the radius of convergence of the **Neumann series** for the resolvent

$$(x - \alpha)^{-1} = -\frac{1}{\alpha} \sum_{n=0}^{\infty} \left(\frac{x}{\alpha}\right)^n \tag{6.20}$$

encountered in the proof of Theorem 6.3 (which is just the Laurent expansion around infinity).

Theorem 6.6 (Beurling–Gelfand)**.** *The spectral radius satisfies*

$$r(x) = \inf_{n \in \mathbb{N}} \|x^n\|^{1/n} = \lim_{n \to \infty} \|x^n\|^{1/n}. \tag{6.21}$$

Proof. By spectral mapping we have $r(x)^n = r(x^n) \leq \|x^n\|$ and hence

$$r(x) \leq \inf \|x^n\|^{1/n}.$$

Conversely, fix $\ell \in X^*$, and consider

$$\ell((x - \alpha)^{-1}) = -\frac{1}{\alpha} \sum_{n=0}^{\infty} \frac{1}{\alpha^n} \ell(x^n). \tag{6.22}$$

Then $\ell((x - \alpha)^{-1})$ is analytic in $|\alpha| > r(x)$ and hence (6.22) converges absolutely for $|\alpha| > r(x)$ by Cauchy's integral formula for derivatives. Hence

for fixed α with $|\alpha| > r(x)$, $\ell(x^n/\alpha^n)$ converges to zero for every $\ell \in X^*$. Since every weakly convergent sequence is bounded we have

$$\frac{\|x^n\|}{|\alpha|^n} \leq C(\alpha)$$

and thus

$$\limsup_{n\to\infty} \|x^n\|^{1/n} \leq \limsup_{n\to\infty} C(\alpha)^{1/n}|\alpha| = |\alpha|.$$

Since this holds for every $|\alpha| > r(x)$ we have

$$r(x) \leq \inf \|x^n\|^{1/n} \leq \liminf_{n\to\infty} \|x^n\|^{1/n} \leq \limsup_{n\to\infty} \|x^n\|^{1/n} \leq r(x),$$

which finishes the proof. □

Note that it might be tempting to conjecture that the sequence $\|x^n\|^{1/n}$ is monotone, however this is false in general – see Problem 6.7. To end this section let us look at some examples illustrating these ideas.

Example 6.13. In $X := C(I)$ we have $\sigma(x) = x(I)$ and hence $r(x) = \|x\|_\infty$ for all x. ◇

Example 6.14. If $X := \mathscr{L}(\mathbb{C}^2)$ and $x := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ such that $x^2 = 0$ and consequently $r(x) = 0$. This is not surprising, since x has the only eigenvalue 0. In particular, the spectral radius can be strictly smaller then the norm (note that $\|x\| = 1$ in our example). The same is true for any nilpotent matrix. In general x will be called **nilpotent** if $x^n = 0$ for some $n \in \mathbb{N}$ and any nilpotent element will satisfy $r(x) = 0$. ◇

Example 6.15. Consider the linear **Volterra integral operator**

$$K(x)(t) := \int_0^t k(t,s)x(s)ds, \qquad x \in C([0,1]), \tag{6.23}$$

then, using induction, it is not hard to verify (Problem 6.9)

$$|K^n(x)(t)| \leq \frac{\|k\|_\infty^n t^n}{n!}\|x\|_\infty. \tag{6.24}$$

Consequently

$$\|K^n x\|_\infty \leq \frac{\|k\|_\infty^n}{n!}\|x\|_\infty,$$

that is $\|K^n\| \leq \frac{\|k\|_\infty^n}{n!}$, which shows

$$r(K) \leq \lim_{n\to\infty} \frac{\|k\|_\infty}{(n!)^{1/n}} = 0.$$

Hence $r(K) = 0$ and for every $\lambda \in \mathbb{C}$ and every $y \in C(I)$ the equation

$$x - \lambda K\, x = y \tag{6.25}$$

has a unique solution given by

$$x = (\mathbb{I} - \lambda K)^{-1} y = \sum_{n=0}^{\infty} \lambda^n K^n y. \tag{6.26}$$

Note that $\sigma(K) = \{0\}$ but 0 is in general not an eigenvalue (consider e.g. $k(t,s) = 1$). Elements of a Banach algebra with $r(x) = 0$ are called **quasinilpotent**. ⋄

In the last two examples we have seen a strict inequality in (6.19). If we regard $r(x)$ as a *spectral norm* for x, then the spectral norm does not control the *algebraic* norm in such a situation. On the other hand, if we had equality for some x, and moreover, this were also true for any polynomial $p(x)$, then spectral mapping would imply that the spectral norm $\sup_{\alpha\in\sigma(x)} |p(\alpha)|$ equals the algebraic norm $\|p(x)\|$ and convergence on one side would imply convergence on the other side. So by taking limits we could get an isometric identification of elements of the form $f(x)$ with functions $f \in C(\sigma(x))$. But this is nothing but the content of the spectral theorem and self-adjointness will be the property which will make all this work.

Problem* 6.1. *Show that the multiplication in a Banach algebra X is continuous: $x_n \to x$ and $y_n \to y$ imply $x_n y_n \to xy$.*

Problem* 6.2. *Suppose that X satisfies all requirements for a Banach algebra except that* (6.3) *is replaced by*

$$\|xy\| \le C\|x\|\|y\|, \qquad C > 0.$$

Of course one can rescale the norm to reduce it to the case $C = 1$. However, this might have undesirable side effects in case there is a unit. Show that if X has a unit e, then $\|e\| \ge C^{-1}$ and there is an equivalent norm $\|.\|_0$ which satisfies (6.3) *and $\|e\|_0 = 1$.*

Finally, note that for this construction to work it suffices to assume that multiplication is separately continuos by Problem 4.5.

(Hint: Identify $x \in X$ with the operator $L_x : X \to X$, $y \mapsto xy$ in $\mathscr{L}(X)$. For the last part use the uniform boundedness principle.)

Problem* 6.3 (Unitization)**.** *Show that if X is a Banach algebra then $\mathbb{C} \oplus X$ is a unital Banach algebra, where we set $\|(\alpha, x)\| = |\alpha| + \|x\|$ and $(\alpha, x)(\beta, y) = (\alpha\beta, \alpha y + \beta x + xy)$.*

Problem 6.4. *Show $\sigma(x^{-1}) = \sigma(x)^{-1}$ if x is invertible.*

Problem* 6.5. *Suppose x has both a right inverse y (i.e., $xy = e$) and a left inverse z (i.e., $zx = e$). Show that $y = z = x^{-1}$.*

Problem* 6.6. *Suppose xy and yx are both invertible, then so are x and y:*

$$y^{-1} = (xy)^{-1}x = x(yx)^{-1}, \qquad x^{-1} = (yx)^{-1}y = y(xy)^{-1}.$$

(Hint: Previous problem.)

Problem* 6.7. *Let $X := \mathscr{L}(\mathbb{C}^2)$ and compute $\|x^n\|^{1/n}$ for $x := \left(\begin{smallmatrix} 0 & \alpha \\ \beta & 0 \end{smallmatrix}\right)$. Conclude that this sequence is not monotone in general.*

Problem 6.8. *Let $X := \ell^\infty(\mathbb{N})$. Show $\sigma(x) = \overline{\{x_n\}_{n\in\mathbb{N}}}$. Also show that $r(x) = \|x\|$ for all $x \in X$.*

Problem* 6.9. *Show* (6.24).

Problem 6.10. *Show that $L^1(\mathbb{R}^n)$ with convolution as multiplication is a commutative Banach algebra without identity (Hint: Lemma 10.18).*

Problem 6.11. *Show the* ***first resolvent identity***

$$\begin{aligned}(x-\alpha)^{-1} - (x-\beta)^{-1} &= (\alpha-\beta)(x-\alpha)^{-1}(x-\beta)^{-1} \\ &= (\alpha-\beta)(x-\beta)^{-1}(x-\alpha)^{-1},\end{aligned} \tag{6.27}$$

for $\alpha, \beta \in \rho(x)$.

Problem 6.12. *Show $\sigma(xy) \setminus \{0\} = \sigma(yx) \setminus \{0\}$. (Hint: Find a relation between $(xy-\alpha)^{-1}$ and $(yx-\alpha)^{-1}$.)*

6.2. The C^* algebra of operators and the spectral theorem

We begin by recalling that if $\mathfrak{H}$ is some Hilbert space, then for every $A \in \mathscr{L}(\mathfrak{H})$ we can define its adjoint $A^* \in \mathscr{L}(\mathfrak{H})$. Hence the Banach algebra $\mathscr{L}(\mathfrak{H})$ has an additional operation in this case which will also give us self-adjointness, a property which has already turned out crucial for the spectral theorem in the case of compact operators. Even though this is not immediately evident, in some sense this additional structure adds the convenient geometric properties of Hilbert spaces to the picture.

A Banach algebra X together with an **involution** satisfying

$$(x+y)^* = x^* + y^*, \quad (\alpha x)^* = \alpha^* x^*, \quad x^{**} = x, \quad (xy)^* = y^* x^*, \tag{6.28}$$

and

$$\|x\|^2 = \|x^* x\| \tag{6.29}$$

is called a C^* **algebra**. Any subalgebra (we do not require a subalgebra to contain the identity) which is also closed under involution, is called a $*$-subalgebra.

The condition (6.29) might look a bit artificial at this point. Maybe a requirement like $\|x^*\| = \|x\|$ might seem more natural. In fact, at this point the only justification is that it holds for our guiding example $\mathscr{L}(\mathfrak{H})$

(cf. Lemma 2.14). Furthermore, it is important to emphasize that (6.29) is a rather strong condition as it implies that the norm is already uniquely determined by the algebraic structure. More precisely, Lemma 6.7 below implies that the norm of x can be computed from the spectral radius of x^*x via $\|x\| = r(x^*x)^{1/2}$. So while there might be several norms which turn X into a Banach algebra, there is at most one which will give a C^* algebra.

Note that (6.29) implies $\|x\|^2 = \|x^*x\| \le \|x\|\|x^*\|$ and hence $\|x\| \le \|x^*\|$. By $x^{**} = x$ this also implies $\|x^*\| \le \|x^{**}\| = \|x\|$ and hence

$$\|x\| = \|x^*\|, \qquad \|x\|^2 = \|x^*x\| = \|xx^*\|. \tag{6.30}$$

Example 6.16. The continuous functions $C(I)$ together with complex conjugation form a commutative C^* algebra. ⋄

Example 6.17. The Banach algebra $\mathscr{L}(\mathfrak{H})$ is a C^* algebra by Lemma 2.14. The compact operators $\mathscr{C}(\mathfrak{H})$ are a $*$-subalgebra. ⋄

Example 6.18. The bounded sequences $\ell^\infty(\mathbb{N})$ together with complex conjugation form a commutative C^* algebra. The set $c_0(\mathbb{N})$ of sequences converging to 0 are a $*$-subalgebra. ⋄

If X has an identity e, we clearly have $e^* = e$, $\|e\| = 1$, $(x^{-1})^* = (x^*)^{-1}$ (show this), and

$$\sigma(x^*) = \sigma(x)^*. \tag{6.31}$$

We will always assume that we have an identity and we note that it is always possible to add an identity (Problem 6.13).

If X is a C^* algebra, then $x \in X$ is called **normal** if $x^*x = xx^*$, **self-adjoint** if $x^* = x$, and **unitary** if $x^* = x^{-1}$. Moreover, x is called **positive** if $x = y^2$ for some $y = y^* \in X$. Clearly both self-adjoint and unitary elements are normal and positive elements are self-adjoint. If x is normal, then so is any polynomial $p(x)$ (it will be self-adjoint if x is and p is real-valued).

As already pointed out in the previous section, it is crucial to identify elements for which the spectral radius equals the norm. The key ingredient will be (6.29) which implies $\|x^2\| = \|x\|^2$ if x is self-adjoint. For unitary elements we have $\|x\| = \sqrt{\|x^*x\|} = \sqrt{\|e\|} = 1$. Moreover, for normal elements we get

Lemma 6.7. *If $x \in X$ is normal, then $\|x^2\| = \|x\|^2$ and $r(x) = \|x\|$.*

Proof. Using (6.29) three times we have

$$\|x^2\| = \|(x^2)^*(x^2)\|^{1/2} = \|(x^*x)^*(x^*x)\|^{1/2} = \|x^*x\| = \|x\|^2$$

and hence $r(x) = \lim_{k\to\infty} \|x^{2^k}\|^{1/2^k} = \|x\|$. □

The next result generalizes the fact that self-adjoint operators have only real eigenvalues.

Lemma 6.8. *If x is self-adjoint, then $\sigma(x) \subseteq \mathbb{R}$. If x is positive, then $\sigma(x) \subseteq [0, \infty)$.*

Proof. Suppose $\alpha + \mathrm{i}\beta \in \sigma(x)$, $\lambda \in \mathbb{R}$. Then $\alpha + \mathrm{i}(\beta + \lambda) \in \sigma(x + \mathrm{i}\lambda)$ and

$$\alpha^2 + (\beta + \lambda)^2 \leq \|x + \mathrm{i}\lambda\|^2 = \|(x + \mathrm{i}\lambda)(x - \mathrm{i}\lambda)\| = \|x^2 + \lambda^2\| \leq \|x\|^2 + \lambda^2.$$

Hence $\alpha^2 + \beta^2 + 2\beta\lambda \leq \|x\|^2$ which gives a contradiction if we let $|\lambda| \to \infty$ unless $\beta = 0$.

The second claim follows from the first using spectral mapping (Theorem 6.5). □

Example 6.19. If $X := \mathscr{L}(\mathbb{C}^2)$ and $x := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ then $\sigma(x) = \{0\}$. Hence the converse of the above lemma is not true in general. ⋄

Given $x \in X$ we can consider the C^* algebra $C^*(x)$ (with identity) generated by x (i.e., the smallest closed $*$-subalgebra containing e and x). If x is normal we explicitly have

$$C^*(x) = \overline{\{p(x, x^*) | p : \mathbb{C}^2 \to \mathbb{C} \text{ polynomial}\}}, \qquad xx^* = x^*x, \tag{6.32}$$

and, in particular, $C^*(x)$ is commutative (Problem 6.14). In the self-adjoint case this simplifies to

$$C^*(x) := \overline{\{p(x) | p : \mathbb{C} \to \mathbb{C} \text{ polynomial}\}}, \qquad x = x^*. \tag{6.33}$$

Moreover, in this case $C^*(x)$ is isomorphic to $C(\sigma(x))$ (the continuous functions on the spectrum):

Theorem 6.9 (Spectral theorem)**.** *If X is a C^* algebra and $x \in X$ is self-adjoint, then there is an isometric isomorphism $\Phi : C(\sigma(x)) \to C^*(x)$ such that $f(t) = t$ maps to $\Phi(t) = x$ and $f(t) = 1$ maps to $\Phi(1) = e$.*

Moreover, for every $f \in C(\sigma(x))$ we have

$$\sigma(f(x)) = f(\sigma(x)), \tag{6.34}$$

where $f(x) = \Phi(f)$.

Proof. First of all, Φ is well defined for polynomials p and given by $\Phi(p) = p(x)$. Moreover, since $p(x)$ is normal spectral mapping implies

$$\|p(x)\| = r(p(x)) = \sup_{\alpha \in \sigma(p(x))} |\alpha| = \sup_{\alpha \in \sigma(x)} |p(\alpha)| = \|p\|_\infty$$

for every polynomial p. Hence Φ is isometric. Next we use that the polynomials are dense in $C(\sigma(x))$. In fact, to see this one can either consider a compact interval I containing $\sigma(x)$ and use the Tietze extension theorem (Theorem B.29 to extend f to I and then approximate the extension using polynomials (Theorem 1.3) or use the Stone–Weierstraß theorem (Theorem B.41). Thus Φ uniquely extends to a map on all of $C(\sigma(x))$ by

Theorem 1.17. By continuity of the norm this extension is again isometric. Similarly, we have $\Phi(f\,g) = \Phi(f)\Phi(g)$ and $\Phi(f)^* = \Phi(f^*)$ since both relations hold for polynomials.

To show $\sigma(f(x)) = f(\sigma(x))$ fix some $\alpha \in \mathbb{C}$. If $\alpha \notin f(\sigma(x))$, then $g(t) = \frac{1}{f(t)-\alpha} \in C(\sigma(x))$ and $\Phi(g) = (f(x) - \alpha)^{-1} \in X$ shows $\alpha \notin \sigma(f(x))$. Conversely, if $\alpha \notin \sigma(f(x))$ then $g = \Phi^{-1}((f(x)-\alpha)^{-1}) = \frac{1}{f-\alpha}$ is continuous, which shows $\alpha \notin f(\sigma(x))$. □

In particular, this last theorem tells us that we have a functional calculus for self-adjoint operators, that is, if $A \in \mathscr{L}(\mathfrak{H})$ is self-adjoint, then $f(A)$ is well defined for every $f \in C(\sigma(A))$. Specifically, we can compute $f(A)$ by choosing a sequence of polynomials p_n which converge to f uniformly on $\sigma(A)$, then we have $p_n(A) \to f(A)$ in the operator norm. In particular, if f is given by a power series, then $f(A)$ defined via Φ coincides with $f(A)$ defined via its power series (cf. Problem 1.35).

Problem* 6.13 (Unitization)**.** *Show that if X is a non-unital C^* algebra then $\mathbb{C}\oplus X$ is a unital C^* algebra, where we set $\|(\alpha, x)\| := \sup\{\|\alpha y + xy\| | y \in X, \|y\| \le 1\}$, $(\alpha, x)(\beta, y) = (\alpha\beta, \alpha y + \beta x + xy)$ and $(\alpha, x)^* = (\alpha^*, x^*)$. (Hint: It might be helpful to identify $x \in X$ with the operator $L_x : X \to X$, $y \mapsto xy$ in $\mathscr{L}(X)$. Moreover, note $\|L_x\| = \|x\|$.)*

Problem* 6.14. *Let X be a C^* algebra and Y a $*$-subalgebra. Show that if Y is commutative, then so is $\overline{Y}$.*

Problem 6.15. *Show that the map Φ from the spectral theorem is positivity preserving, that is, $f \ge 0$ if and only if $\Phi(f)$ is positive.*

Problem 6.16. *Let x be self-adjoint. Show that the following are equivalent:*

(i) $\sigma(x) \subseteq [0, \infty)$.
(ii) *x is positive.*
(iii) $\|\lambda - x\| \le \lambda$ *for all* $\lambda \ge \|x\|$.
(iv) $\|\lambda - x\| \le \lambda$ *for one* $\lambda \ge \|x\|$.

Problem 6.17. *Let $A \in \mathscr{L}(\mathfrak{H})$. Show that A is normal if and only if*

$$\|Au\| = \|A^*u\|, \quad \forall u \in \mathfrak{H}.$$

In particular, $\operatorname{Ker}(A) = \operatorname{Ker}(A^*)$. *(Hint: Problem 1.19.)*

Problem 6.18. *Show that the* ***Cayley transform*** *of a self-adjoint element x,*

$$y = (x - \mathrm{i})(x + \mathrm{i})^{-1}$$

is unitary. Show that $1 \notin \sigma(y)$ and

$$x = \mathrm{i}(1 + y)(1 - y)^{-1}.$$

Problem 6.19. *Show if x is unitary then $\sigma(x) \subseteq \{\alpha \in \mathbb{C} | |\alpha| = 1\}$.*

Problem 6.20. *Suppose x is self-adjoint. Show that*

$$\|(x-\alpha)^{-1}\| = \frac{1}{\operatorname{dist}(\alpha, \sigma(x))}.$$

6.3. Spectral theory for bounded operators

So far we have developed spectral theory on an algebraic level based on the fact that bounded operators form a Banach algebra. In this section we want to take a more operator centered view and consider bounded linear operators $\mathscr{L}(X)$, where X is some Banach space. Now we can make a finer subdivision of the spectrum based on why our operator fails to have a bounded inverse. Since in the bijective case boundedness of the inverse comes for free from the inverse mapping theorem (Theorem 4.6), there are basically two things which can go wrong: Either our map is not injective or it is not surjective. Moreover, in the latter case one can also ask how far it is away from being surjective, that is, if the range is dense or not. Accordingly one defines the **point spectrum**

$$\sigma_p(A) := \{\alpha \in \sigma(A) | \operatorname{Ker}(A-\alpha) \neq \{0\}\} \tag{6.35}$$

as the set of all eigenvalues, the **continuous spectrum**

$$\sigma_c(A) := \{\alpha \in \sigma(A) \setminus \sigma_p(A) | \overline{\operatorname{Ran}(A-\alpha)} = X\} \tag{6.36}$$

and finally the **residual spectrum**

$$\sigma_r(A) := \{\alpha \in \sigma(A) \setminus \sigma_p(A) | \overline{\operatorname{Ran}(A-\alpha)} \neq X\}. \tag{6.37}$$

Clearly we have

$$\sigma(A) = \sigma_p(A) \uplus \sigma_c(A) \uplus \sigma_r(A). \tag{6.38}$$

Note that in a Hilbert space $\sigma_x(A^*) = \sigma_x(A')^*$ for $x \in \{p, c, r\}$.

Example 6.20. Suppose $\mathfrak{H}$ is a Hilbert space and $A = A^*$ is self-adjoint. Then by (2.28), $\sigma_r(A) = \emptyset$. ⋄

Example 6.21. Suppose $X := \ell^p(\mathbb{N})$ and L is the left shift. Then $\sigma(L) = \bar{B}_1(0)$. Indeed, a simple calculation shows that $\operatorname{Ker}(L-\alpha) = \operatorname{span}\{(\alpha^j)_{j\in\mathbb{N}}\}$ for $|\alpha| < 1$ if $1 \leq p < \infty$ and for $|\alpha| \leq 1$ if $p = \infty$. Hence $\sigma_p(L) = B_1(0)$ for $1 \leq p < \infty$ and $\sigma_p(L) = \bar{B}_1(0)$ if $p = \infty$. In particular, since the spectrum is closed and $\|L\| = 1$ we have $\sigma(L) = \bar{B}_1(0)$. Moreover, for $y \in \ell_c(\mathbb{N})$ we set $x_j := -\sum_{k=j}^{\infty} \alpha^{j-k-1} y_k$ such that $(L-\alpha)x = y$. In particular, $\ell_c(\mathbb{N}) \subset \operatorname{Ran}(L-\alpha)$ and hence $\operatorname{Ran}(S-\alpha)$ is dense for $1 \leq p < \infty$. Thus $\sigma_c(L) = \partial B_1(0)$ for $1 \leq p < \infty$. Consequently, $\sigma_r(L) = \emptyset$. ⋄

Since A is invertible if and only if A' is by Theorem 4.26 we obtain:

Lemma 6.10. *Suppose $A \in \mathscr{L}(X)$. Then*

$$\sigma(A) = \sigma(A'). \tag{6.39}$$

Moreover,

$$\begin{aligned} &\sigma_p(A') \subseteq \sigma_p(A) \uplus \sigma_r(A), && \sigma_p(A) \subseteq \sigma_p(A') \uplus \sigma_r(A'), \\ &\sigma_r(A') \subseteq \sigma_p(A) \uplus \sigma_c(A), && \sigma_r(A) \subseteq \sigma_p(A'), \\ &\sigma_c(A') \subseteq \sigma_c(A), && \sigma_c(A) \subseteq \sigma_r(A') \uplus \sigma_c(A'). \end{aligned} \tag{6.40}$$

If in addition, X is reflexive we have $\sigma_r(A') \subseteq \sigma_p(A)$ as well as $\sigma_c(A') = \sigma_c(A)$.

Proof. This follows from Lemma 4.25 and (4.20). In the reflexive case use $A \cong A''$. □

Example 6.22. Consider L' from the previous example, which is just the right shift in $\ell^q(\mathbb{N})$ if $1 \le p < \infty$. Then $\sigma(L') = \sigma(L) = \bar{B}_1(0)$. Moreover, it is easy to see that $\sigma_p(L') = \emptyset$. Thus in the reflexive case $1 < p < \infty$ we have $\sigma_c(L') = \sigma_c(L) = \partial B_1(0)$ as well as $\sigma_r(L') = \sigma(L') \setminus \sigma_c(L') = B_1(0)$. Otherwise, if $p = 1$, we only get $B_1(0) \subseteq \sigma_r(L')$ and $\sigma_c(L') \subseteq \sigma_c(L) = \partial B_1(0)$. Hence it remains to investigate $\operatorname{Ran}(L' - \alpha)$ for $|\alpha| = 1$: If we have $(L' - \alpha)x = y$ with some $y \in \ell^p(\mathbb{N})$ we must have $x_j := -\alpha^{-j-1} \sum_{k=1}^{j} \alpha^k y_k$. Thus $y = (\alpha^n)_{n\in\mathbb{N}}$ is clearly not in $\operatorname{Ran}(L' - \alpha)$. Moreover, if $\|y - \tilde{y}\|_\infty \le \varepsilon$ we have $|\tilde{x}_j| = |\sum_{k=1}^{j} \alpha^k \tilde{y}_k| \ge (1 - \varepsilon)j$ and hence $\tilde{y} \notin \operatorname{Ran}(L' - \alpha)$, which shows that the range is not dense and hence $\sigma_r(L') = \bar{B}_1(0)$, $\sigma_c(L') = \emptyset$. ⋄

Moreover, for compact operators the spectrum is particularly simple (cf. also Theorem 3.7). We start with the following observation:

Lemma 6.11. *Suppose that $K \in \mathscr{C}(X)$ and $\alpha \in \mathbb{C} \setminus \{0\}$. Then* $\operatorname{Ker}(K - \alpha)$ *is finite dimensional and the range* $\operatorname{Ran}(K - \alpha)$ *is closed.*

Proof. For $\alpha \neq 0$ we can consider $\mathbb{I} - \alpha^{-1}K$ and assume $\alpha = 1$ without loss of generality. First of all note that K restricted to $\operatorname{Ker}(\mathbb{I} - K)$ is the identity and since the identity is compact the corresponding space must be finite dimensional by Theorem 1.11. In particular, it is complemented (Problem 4.25), that is, there exists a closed subspace $X_0 \subseteq X$ such that $X = \operatorname{Ker}(\mathbb{I} - K) \dotplus X_0$.

To see that $\operatorname{Ran}(\mathbb{I} - K)$ is closed we consider $\mathbb{I} + K$ restricted to X_0 which is injective and has the same range. Hence if $\operatorname{Ran}(\mathbb{I} - K)$ were not closed, Corollary 4.10 would imply that there is a sequence $x_n \in X_0$ with $\|x_n\| = 1$ and $x_n - Kx_n \to 0$. By compactness of K we can pass to a subsequence such that $Kx_n \to y$ implying $x_n \to y \in X_0$ and hence $y \in \operatorname{Ker}(\mathbb{I} - K)$ contradicting $y \in X_0$ with $\|y\| = 1$. □

Next, we want to have a closer look at eigenvalues. Note that eigenvectors corresponding to different eigenvalues are always linearly independent (Problem 6.22). In Theorem 3.7 we have seen that for a symmetric compact operator in a Hilbert space we can choose an orthonormal basis of eigenfunctions. Without the symmetry assumption we know that even in the finite dimensional case we can in general no longer find a basis of eigenfunctions and that the Jordan canonical form is the best one can do. There the generalized eigenspaces $\operatorname{Ker}((A-\alpha)^k)$ play an important role. In this respect one looks at the following ascending and descending chains of subspaces associated to $A \in \mathscr{L}(X)$ (where we have assumed $\alpha = 0$ without loss of generality):

$$\{0\} \subseteq \operatorname{Ker}(A) \subseteq \operatorname{Ker}(A^2) \subseteq \operatorname{Ker}(A^3) \subseteq \cdots \tag{6.41}$$

and

$$X \supseteq \operatorname{Ran}(A) \supseteq \operatorname{Ran}(A^2) \supseteq \operatorname{Ran}(A^3) \supseteq \cdots \tag{6.42}$$

We will say that the kernel chain **stabilizes** at n if $\operatorname{Ker}(A^{n+1}) = \operatorname{Ker}(A^n)$. In this case the number n is also called the **ascent** of A. Substituting $x = Ay$ in the equivalence $A^n x = 0 \Leftrightarrow A^{n+1}x = 0$ gives $A^{n+1}y = 0 \Leftrightarrow A^{n+2}y = 0$ and hence by induction we have $\operatorname{Ker}(A^{n+k}) = \operatorname{Ker}(A^n)$ for all $k \in \mathbb{N}_0$ in this case. Similarly, will say that the range chain **stabilizes** at m if $\operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ and call m the **decent** of A. Again, if $x = A^{m+2}y \in \operatorname{Ran}(A^{m+2})$ we can write $A^{m+1}y = A^m z$ for some z which shows $x = A^{m+1}z \in \operatorname{Ran}(A^{m+1})$ and thus $\operatorname{Ran}(A^{m+k}) = \operatorname{Ran}(A^m)$ for all $k \in \mathbb{N}_0$ in this case. While in a finite dimensional space both chains eventually have to stabilize, there is no reason why the same should happen in an infinite dimensional space.

Example 6.23. For the left shift operator L we have $\operatorname{Ran}(L^n) = \ell^p(\mathbb{N})$ for all $n \in \mathbb{N}$ while the kernel chain does not stabilize as $\operatorname{Ker}(L^n) = \{a \in \ell^p(\mathbb{N}) | a_j = 0,\ j > n\}$. Similarly, for the right shift operator R we have $\operatorname{Ker}(R^n) = \{0\}$ while the range chain does not stabilize as $\operatorname{Ran}(R^n) = \{a \in \ell^p(\mathbb{N}) | a_j = 0,\ 1 \le j \le n\}$. ⋄

Lemma 6.12. *Suppose $A : X \to X$ is a linear operator.*

(i) *The kernel chain stabilizes at n if $\operatorname{Ran}(A^n) \cap \operatorname{Ker}(A) = \{0\}$. Conversely, if the kernel chain stabilizes at n, then $\operatorname{Ran}(A^n) \cap \operatorname{Ker}(A^n) = \{0\}$.*

(ii) *The range chain stabilizes at m if $\operatorname{Ker}(A^m) + \operatorname{Ran}(A) = X$. Conversely, if the range chain stabilizes at m, then $\operatorname{Ker}(A^m) + \operatorname{Ran}(A^m) = X$.*

(iii) *If both chains stabilize, then $m = n$ and $\operatorname{Ker}(A^m) \dot{+} \operatorname{Ran}(A^m) = X$.*

Proof. (i). If $\operatorname{Ran}(A^n) \cap \operatorname{Ker}(A) = \{0\}$ then $x \in \operatorname{Ker}(A^{n+1})$ implies $A^n x \in \operatorname{Ran}(A^n) \cap \operatorname{Ker}(A) = \{0\}$ and the kernel chain stabilizes at n. Conversely,

let $x \in \operatorname{Ran}(A^n) \cap \operatorname{Ker}(A^n)$, then $x = A^n y$ and $A^n x = A^{2n} y = 0$ implying $y \in \operatorname{Ker}(A^{2n}) = \operatorname{Ker}(A^n)$, that is, $x = A^n y = 0$.

(ii). If $\operatorname{Ker}(A^m) + \operatorname{Ran}(A) = X$, then for any $x = z + Ay$ we have $A^m x = A^{m+1} y$ and hence $\operatorname{Ran}(A^m) = \operatorname{Ran}(A^{m+1})$. Conversely, if the range chain stabilizes at m, then $A^m x = A^{2m} y$ and $x = A^m y + (x - A^m y)$.

(iii). Suppose $\operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ but $\operatorname{Ker}(A^m) \subsetneq \operatorname{Ker}(A^{m+1})$. Let $x \in \operatorname{Ker}(A^{m+1}) \setminus \operatorname{Ker}(A^m)$ and observe that by $0 \neq A^m x = A^{m+1} y$ there is an $x \in \operatorname{Ker}(A^{m+2}) \setminus \operatorname{Ker}(A^{m+1})$. Iterating this argument would shows that the kernel chain does not stabilize contradiction our assumption. Hence $n \leq m$.

Conversely, suppose $\operatorname{Ker}(A^{n+1}) = \operatorname{Ker}(A^n)$ and $\operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ for $m \geq n$. Then

$$A^m x = A^{m+1} y \;\Rightarrow\; x - Ay \in \operatorname{Ker}(A^m) = \operatorname{Ker}(A^n) \;\Rightarrow\; A^n x = A^{n+1} y$$

shows $\operatorname{Ran}(A^{n+1}) = \operatorname{Ran}(A^n)$, that is, $m \leq n$. The rest follows by combining (i) and (ii). □

Note that in case (iii) A is bijective when restricted to $\operatorname{Ran}(A^n) \to \operatorname{Ran}(A^n)$ and nilpotent when restricted to $\operatorname{Ker}(A^n) \to \operatorname{Ker}(A^n)$.

Example 6.24. In a finite dimensional space we are of course always in case (iii). ⋄

Example 6.25. Let $A \in \mathscr{L}(\mathfrak{H})$ be a self-adjoint operator in a Hilbert space. Then by (2.28) the kernel chain always stabilizes at $n = 1$ and the range chain stabilizes at $n = 1$ if $\operatorname{Ran}(A)$ is closed. ⋄

Now we can apply this to our situation.

Lemma 6.13. *Suppose that $K \in \mathscr{C}(X)$ and $\alpha \in \mathbb{C} \setminus \{0\}$. Then there is some $n = n(\alpha) \in \mathbb{N}$ such that $\operatorname{Ker}(K-\alpha)^n = \operatorname{Ker}(K-\alpha)^{n+k}$ and $\operatorname{Ran}(K-\alpha)^n = \operatorname{Ran}(K-\alpha)^{n+k}$ for every $k \geq 0$ and*

$$X = \operatorname{Ker}(K-\alpha)^n \dotplus \operatorname{Ran}(K-\alpha)^n. \tag{6.43}$$

Moreover, the space $\operatorname{Ker}(K-\alpha)^m$ is finite dimensional and the space $\operatorname{Ran}(K-\alpha)^m$ is closed for every $m \in \mathbb{N}$.

Proof. Since $\alpha \neq 0$ we can consider $\mathbb{I} - \alpha^{-1}K$ and assume $\alpha = 1$ without loss of generality. Moreover, since $(\mathbb{I}-K)^n - \mathbb{I} \in \mathscr{C}(X)$ we see that $\operatorname{Ker}(\mathbb{I}-K)^n$ is finite dimensional and $\operatorname{Ran}(\mathbb{I}-K)^n$ is closed for every $n \in \mathbb{N}$. Next suppose the kernel chain does not stabilize. Abbreviate $K_n := \operatorname{Ker}(\mathbb{I}-K)^n$. Then, by Problem 4.31, we can choose $x_n \in K_{n+1} \setminus K_n$ such that $\|x_n\| = 1$ and $\operatorname{dist}(x_n, K_n) \geq \frac{1}{2}$. But since $(\mathbb{I}-K)x_n \in K_n$ and $Kx_n \in K_{n+1}$, we see that

$$\|Kx_n - Kx_m\| = \|x_n - (\mathbb{I}-K)x_n - Ax_m\| \geq \operatorname{dist}(x_n, K_n) \geq \frac{1}{2}$$

for $n > m$ and hence the bounded sequence Kx_n has no convergent subsequence, a contradiction. Consequently the kernel sequence for K' also stabilizes, thus by Problem 4.27 $\operatorname{Coker}(\mathbb{I} - K)^* \cong \operatorname{Ker}(\mathbb{I} - K')$ the sequence of cokernels stabilizes, which finally implies that the range sequence stabilizes. The rest follows from the previous lemma. □

For an eigenvalue $\alpha \in \mathbb{C}$ the dimension $\dim(\operatorname{Ker}(A - \alpha))$ is called the **geometric multiplicity** of α and if the kernel chain stabilizes at n, then n is called the **index** of α and $\dim(\operatorname{Ker}(A - \alpha)^n)$ is called the **algebraic multiplicity** of α. Otherwise, if the kernel chain does not stabilize, both the index and the algebraic multiplicity are set to infinity. The **order** of a generalized eigenvector u corresponding to an eigenvalue α is the smallest n such that $(A - \alpha)^n u = 0$.

Example 6.26. Consider $X := \ell^p(\mathbb{N})$ and $K \in \mathscr{C}(X)$ given by $(Ka)_n = \frac{1}{n} a_{n+1}$. Then $Ka = \alpha\, a$ implies $a_n = \alpha^n (n-1)! a_1$ for $\alpha \neq 0$ (and hence $a_1 = 0$) and $a = a_1 \delta^1$ for $\alpha = 0$. Hence $\sigma(K) = \{0\}$ with 0 being an eigenvalue of geometric multiplicity one. Since $K^n a = 0$ implies $a_j = 0$ for $1 \leq j \leq n$ we see that its index as well as its algebraic multiplicity is ∞. Moreover, δ^n is a generalized eigenvalue of order n. ⋄

Theorem 6.14 (Spectral theorem for compact operators; Riesz)**.** *Suppose that $K \in \mathscr{C}(X)$. Then every $\alpha \in \sigma(K) \setminus \{0\}$ is an eigenvalue of finite algebraic multiplicity and X can de decomposed into invariant closed subspaces according to* (6.43)*, where n is the index of α. Furthermore, there are at most countably many eigenvalues which can only accumulate at 0. If X is infinite dimensional, we have $0 \in \sigma(K)$. In this case either $0 \in \sigma_p(K)$ with $\dim(\operatorname{Ker}(K)) = \infty$ or $\operatorname{Ran}(K)$ is not closed.*

Proof. That every eigenvalue $\alpha \neq 0$ has finite algebraic multiplicity follows from the previous two lemmas. Moreover if $\operatorname{Ker}(A - \alpha) = \{0\}$, then the kernel chain stabilizes as $n = 1$ and hence $\operatorname{Ran}(A - \alpha) = X$, that is $\alpha \notin \sigma(K)$.

Let α_n be a sequence of different eigenvalues with $|\alpha_n| \geq \varepsilon$. Let x_n be corresponding normalized eigenvectors and let $X_n := \operatorname{span}\{x_j\}_{j=1}^n$. The sequence of spaces X_n is increasing and by Problem 4.31 we can choose normalized vectors $\tilde{x}_n \in X_n$ such that $\operatorname{dist}(\tilde{x}_n, X_{n-1}) \geq \frac{1}{2}$. Now $\|K\tilde{x}_n - K\tilde{x}_m\| = \|\alpha_n \tilde{x}_n + ((K - \alpha_n)\tilde{x}_n - K\tilde{x}_m)\| \geq \frac{|\alpha_n|}{2} \geq \frac{\varepsilon}{2}$ for $m < n$ and hence there is no convergent subsequence, a contradiction. Moreover, if $0 \in \rho(K)$ then K^{-1} is bounded and hence $\mathbb{I} = K^{-1}K$ is compact, implying that X is finite dimensional.

Finally, if $\operatorname{Ran}(K)$ is closed we can consider the bijective operator $\tilde{K} : X/\operatorname{Ker}(A) \to \operatorname{Ran}(A)$ (cf. Problem 1.42) which is again compact. Hence $\mathbb{I}_{X/\operatorname{Ker}(A)} = \tilde{K}^{-1}\tilde{K}$ is compact and thus $X/\operatorname{Ker}(A)$ is finite dimensional. □

Example 6.27. Note that in contradistinction to the symmetric case, there might be no eigenvalues at all, as the Volterra integral operator from Example 6.15 shows. ⋄

As an immediate consequence we get the famous Fredholm alternative:

Theorem 6.15 (Fredholm alternative)**.** *Suppose that $K \in \mathscr{C}(X)$ and $\alpha \in \mathbb{C} \setminus \{0\}$. Then, either the inhomogeneous equation*

$$(K - \alpha)x = y \tag{6.44}$$

has a unique solution for every $y \in X$ or the corresponding homogeneous equation

$$(K - \alpha)x = 0 \tag{6.45}$$

has a nontrivial solution.

In particular, this applies to the case where K is a compact integral operator (cf. Lemma 3.4), which was the case originally studied by Fredholm.

Problem 6.21. *Discuss the spectrum of the right shift R on $\ell^1(\mathbb{N})$. Show $\sigma(R) = \sigma_r(R) = \bar{B}_1(0)$ and $\sigma_p(R) = \sigma_c(R) = \emptyset$.*

Problem* 6.22. *Suppose $A \in \mathscr{L}(X)$. Show that generalized eigenvectors corresponding to different eigenvalues or with different order are linearly independent.*

Problem 6.23. *Suppose $\mathfrak{H}$ is a Hilbert space and $A \in \mathscr{L}(\mathfrak{H})$ is normal. Then $\sigma_p(A) = \sigma_p(A^*)^*$, $\sigma_c(A) = \sigma_c(A^*)^*$, and $\sigma_r(A) = \sigma_r(A^*) = \emptyset$. (Hint: Problem 6.17.)*

Problem 6.24. *Suppose $A_j \in \mathscr{L}(X_j)$, $j = 1, 2$. Then $A_1 \oplus A_2 \in \mathscr{L}(X_1 \oplus X_2)$ and $\sigma(A_1 \oplus A_2) = \sigma(A_1) \cup \sigma(A_2)$.*

Problem 6.25. *Let $A : X \to Y$, $B : Y \to Z$. Show $\dim(\operatorname{Ker}(BA)) \leq \dim(\operatorname{Ker}(A)) + \dim(\operatorname{Ker}(B))$ and hence $\dim(\operatorname{Ker}(A^n)) \leq n \dim(\operatorname{Ker}(A))$ if $A : X \to X$.*

6.4. Spectral measures

The purpose of this section is to derive another formulation of the spectral theorem which is important in quantum mechanics. This reformulation requires familiarity with measure theory and can be skipped as the results will not be needed in the sequel.

Using the Riesz representation theorem we get a formulation in terms of spectral measures:

Theorem 6.16. *Let $\mathfrak{H}$ be a Hilbert space, and let $A \in \mathscr{L}(\mathfrak{H})$ be self-adjoint. For every $u, v \in \mathfrak{H}$ there is a corresponding complex Borel measure $\mu_{u,v}$ supported on $\sigma(A)$ (the **spectral measure**) such that*

$$\langle u, f(A)v\rangle = \int_{\sigma(A)} f(t)d\mu_{u,v}(t), \qquad f \in C(\sigma(A)). \tag{6.46}$$

We have

$$\mu_{u,v_1+v_2} = \mu_{u,v_1} + \mu_{u,v_2}, \quad \mu_{u,\alpha v} = \alpha\mu_{u,v}, \quad \mu_{v,u} = \mu_{u,v}^* \tag{6.47}$$

and $|\mu_{u,v}|(\sigma(A)) \le \|u\|\|v\|$. Furthermore, $\mu_u = \mu_{u,u}$ is a positive Borel measure with $\mu_u(\sigma(A)) = \|u\|^2$.

Proof. Consider the continuous functions on $I = [-\|A\|, \|A\|]$ and note that every $f \in C(I)$ gives rise to some $f \in C(\sigma(A))$ by restricting its domain. Clearly $\ell_{u,v}(f) = \langle u, f(A)v\rangle$ is a bounded linear functional and the existence of a corresponding measure $\mu_{u,v}$ with $|\mu_{u,v}|(I) = \|\ell_{u,v}\| \le \|u\|\|v\|$ follows from Theorem 12.5. Since $\ell_{u,v}(f)$ depends only on the value of f on $\sigma(A) \subseteq I$, $\mu_{u,v}$ is supported on $\sigma(A)$.

Moreover, if $f \ge 0$ we have $\ell_u(f) = \langle u, f(A)u\rangle = \langle f(A)^{1/2}u, f(A)^{1/2}u\rangle = \|f(A)^{1/2}u\|^2 \ge 0$ and hence ℓ_u is positive and the corresponding measure μ_u is positive. The rest follows from the properties of the scalar product. □

It is often convenient to regard $\mu_{u,v}$ as a complex measure on $\mathbb{R}$ by using $\mu_{u,v}(\Omega) = \mu_{u,v}(\Omega \cap \sigma(A))$. If we do this, we can also consider f as a function on $\mathbb{R}$. However, note that $f(A)$ depends only on the values of f on $\sigma(A)$! Moreover, it suffices to consider μ_u since using the polarization identity (1.55) we have

$$\mu_{u,v}(\Omega) = \frac{1}{4}(\mu_{u+v}(\Omega) - \mu_{u-v}(\Omega) + \mathrm{i}\mu_{u-\mathrm{i}v}(\Omega) - \mathrm{i}\mu_{u+\mathrm{i}v}(\Omega)). \tag{6.48}$$

Now the last theorem can be used to define $f(A)$ for every bounded measurable function $f \in B(\sigma(A))$ via Lemma 2.12 and extend the functional calculus from continuous to measurable functions:

Theorem 6.17 (Spectral theorem)**.** *If $\mathfrak{H}$ is a Hilbert space and $A \in \mathscr{L}(\mathfrak{H})$ is self-adjoint, then there is an homomorphism $\Phi : B(\sigma(A)) \to \mathscr{L}(\mathfrak{H})$ given by*

$$\langle u, f(A)v\rangle = \int_{\sigma(A)} f(t)d\mu_{u,v}(t), \qquad f \in B(\sigma(A)). \tag{6.49}$$

Moreover, if $f_n(t) \to f(t)$ pointwise and $\sup_n \|f_n\|_\infty$ is bounded, then $f_n(A)u \to f(A)u$ for every $u \in \mathfrak{H}$.

Proof. The map Φ is a well-defined linear operator by Lemma 2.12 since we have

$$\Big| \int_{\sigma(A)} f(t) d\mu_{u,v}(t) \Big| \leq \|f\|_\infty |\mu_{u,v}|(\sigma(A)) \leq \|f\|_\infty \|u\| \|v\|$$

and (6.47). Next, observe that $\Phi(f)^* = \Phi(f^*)$ and $\Phi(fg) = \Phi(f)\Phi(g)$ holds at least for continuous functions. To obtain it for arbitrary bounded functions, choose a (bounded) sequence f_n converging to f in $L^2(\sigma(A), d\mu_u)$ and observe

$$\|(f_n(A) - f(A))u\|^2 = \int |f_n(t) - f(t)|^2 d\mu_u(t)$$

(use $\|h(A)u\|^2 = \langle h(A)u, h(A)u\rangle = \langle u, h(A)^* h(A)u\rangle$). Thus $f_n(A)u \to f(A)u$ and for bounded g we also have that $(gf_n)(A)u \to (gf)(A)u$ and $g(A)f_n(A)u \to g(A)f(A)u$. This establishes the case where f is bounded and g is continuous. Similarly, approximating g removes the continuity requirement from g.

The last claim follows since $f_n \to f$ in L^2 by dominated convergence in this case. □

Our final aim is to generalize Corollary 3.8 to bounded self-adjoint operators. Since the spectrum of an arbitrary self-adjoint might contain more than just eigenvalues we need to replace the sum by an integral. To this end we recall the family of Borel sets $\mathfrak{B}(\mathbb{R})$ and begin by defining the **spectral projections**

$$P_A(\Omega) = \chi_\Omega(A), \qquad \Omega \in \mathfrak{B}(\mathbb{R}), \tag{6.50}$$

such that

$$\mu_{u,v}(\Omega) = \langle u, P_A(\Omega) v\rangle. \tag{6.51}$$

By $\chi_\Omega^2 = \chi_\Omega$ and $\chi_\Omega^* = \chi_\Omega$ they are **orthogonal projections**, that is $P^2 = P$ and $P^* = P$. Recall that any orthogonal projection P decomposes $\mathfrak{H}$ into an orthogonal sum

$$\mathfrak{H} = \operatorname{Ker}(P) \oplus \operatorname{Ran}(P), \tag{6.52}$$

where $\operatorname{Ker}(P) = (\mathbb{I} - P)\mathfrak{H}$, $\operatorname{Ran}(P) = P\mathfrak{H}$.

In addition, the spectral projections satisfy

$$P_A(\mathbb{R}) = \mathbb{I}, \quad P_A(\bigcupdot_{n=1}^{\infty} \Omega_n)u = \sum_{n=1}^{\infty} P_A(\Omega_n)u, \quad \Omega_n \cap \Omega_m = \emptyset,\ m \neq n, \tag{6.53}$$

for every $u \in \mathfrak{H}$. Here the dot inside the union just emphasizes that the sets are mutually disjoint. Such a family of projections is called a **projection-valued measure**. Indeed the first claim follows since $\chi_{\mathbb{R}} = 1$ and by $\chi_{\Omega_1 \cup \Omega_2} = \chi_{\Omega_1} + \chi_{\Omega_2}$ if $\Omega_1 \cap \Omega_2 = \emptyset$ the second claim follows at least for finite unions. The case of countable unions follows from the last part of the

previous theorem since $\sum_{n=1}^{N} \chi_{\Omega_n} = \chi_{\bigcup_{n=1}^{N} \Omega_n} \to \chi_{\bigcup_{n=1}^{\infty} \Omega_n}$ pointwise (note that the limit will not be uniform unless the Ω_n are eventually empty and hence there is no chance that this series will converge in the operator norm). Moreover, since all spectral measures are supported on $\sigma(A)$ the same is true for P_A in the sense that

$$P_A(\sigma(A)) = \mathbb{I}. \tag{6.54}$$

I also remark that in this connection the corresponding distribution function

$$P_A(t) := P_A((-\infty, t]) \tag{6.55}$$

is called a **resolution of the identity**.

Using our projection-valued measure we can define an operator-valued integral as follows: For every simple function $f = \sum_{j=1}^{n} \alpha_j \chi_{\Omega_j}$ (where $\Omega_j = f^{-1}(\alpha_j)$), we set

$$\int_{\mathbb{R}} f(t) dP_A(t) := \sum_{j=1}^{n} \alpha_j P_A(\Omega_j). \tag{6.56}$$

By (6.51) we conclude that this definition agrees with $f(A)$ from Theorem 6.17:

$$\int_{\mathbb{R}} f(t) dP_A(t) = f(A). \tag{6.57}$$

Extending this integral to functions from $B(\sigma(A))$ by approximating such functions with simple functions we get an alternative way of defining $f(A)$ for such functions. This can in fact be done by just using the definition of a projection-valued measure and hence there is a one-to-one correspondence between projection-valued measures (with bounded support) and (bounded) self-adjoint operators such that

$$A = \int t \, dP_A(t). \tag{6.58}$$

If $P_A(\{\alpha\}) \neq 0$, then α is an eigenvalue and $\operatorname{Ran}(P_A(\{\alpha\}))$ is the corresponding eigenspace (Problem 6.27). The fact that eigenspaces to different eigenvalues are orthogonal now generalizes to

Lemma 6.18. *Suppose $\Omega_1 \cap \Omega_2 = \emptyset$. Then*

$$\operatorname{Ran}(P_A(\Omega_1)) \perp \operatorname{Ran}(P_A(\Omega_2)). \tag{6.59}$$

Proof. Clearly $\chi_{\Omega_1} \chi_{\Omega_2} = \chi_{\Omega_1 \cap \Omega_2}$ and hence

$$P_A(\Omega_1) P_A(\Omega_2) = P_A(\Omega_1 \cap \Omega_2).$$

Now if $\Omega_1 \cap \Omega_2 = \emptyset$, then

$$\langle P_A(\Omega_1) u, P_A(\Omega_2) v \rangle = \langle u, P_A(\Omega_1) P_A(\Omega_2) v \rangle = \langle u, P_A(\emptyset) v \rangle = 0,$$

which shows that the ranges are orthogonal to each other. □

Example 6.28. Let $A \in \mathscr{L}(\mathbb{C}^n)$ be some symmetric matrix and let $\alpha_1, \dots, \alpha_m$ be its (distinct) eigenvalues. Then

$$A = \sum_{j=1}^{m} \alpha_j P_A(\{\alpha_j\}),$$

where $P_A(\{\alpha_j\})$ is the projection onto the eigenspace $\operatorname{Ker}(A - \alpha_j)$ corresponding to the eigenvalue α_j by Problem 6.27. In fact, using that P_A is supported on the spectrum, $P_A(\sigma(A)) = \mathbb{I}$, we see

$$P_A(\Omega) = P_A(\sigma(A))P_A(\Omega) = P_A(\sigma(A) \cap \Omega) = \sum_{\alpha_j \in \Omega} P_A(\{\alpha_j\}).$$

Hence using that any $f \in B(\sigma(A))$ is given as a simple function $f = \sum_{j=1}^{m} f(\alpha_j)\chi_{\{\alpha_j\}}$ we obtain

$$f(A) = \int f(t) dP_A(t) = \sum_{j=1}^{m} f(\alpha_j) P_A(\{\alpha_j\}).$$

In particular, for $f(t) = t$ we recover the above representation for A. ⋄

Example 6.29. Let $A \in \mathscr{L}(\mathbb{C}^n)$ be self-adjoint and let α be an eigenvalue. Let $P = P_A(\{\alpha\})$ be the projection onto the corresponding eigenspace and consider the restriction $\tilde{A} = A\big|_{\tilde{\mathfrak{H}}}$ onto the orthogonal complement of this eigenspace $\tilde{\mathfrak{H}} = (1 - P)\mathfrak{H}$. Then by Lemma 6.18 we have $\mu_{u,v}(\{\alpha\}) = 0$ for $u, v \in \tilde{\mathfrak{H}}$. Hence the integral in (6.49) does not see the point α in the sense that

$$\langle u, f(A)v\rangle = \int_{\sigma(A)} f(t) d\mu_{u,v}(t) = \int_{\sigma(A)\setminus\{\alpha\}} f(t) d\mu_{u,v}(t), \qquad u, v \in \tilde{\mathfrak{H}}.$$

Hence Φ extends to a homomorphism $\tilde{\Phi} : B(\sigma(A) \setminus \{\alpha\}) \to \mathscr{L}(\tilde{\mathfrak{H}})$. In particular, if α is an isolated eigenvalue, that is $(\alpha - \varepsilon, \alpha + \varepsilon) \cap \sigma(A) = \{\alpha\}$ for $\varepsilon > 0$ sufficiently small, we have $(. - \alpha)^{-1} \in B(\sigma(A) \setminus \{\alpha\})$ and hence $\alpha \in \rho(\tilde{A})$. ⋄

Problem 6.26. *Suppose A is self-adjoint. Let α be an eigenvalue and u a corresponding normalized eigenvector. Show $\int f(t) d\mu_u(t) = f(\alpha)$, that is, μ_u is the Dirac delta measure (with mass one) centered at α.*

Problem* 6.27. *Suppose A is self-adjoint. Show*

$$\operatorname{Ran}(P_A(\{\alpha\})) = \operatorname{Ker}(A - \alpha).$$

(Hint: Start by verifying $\operatorname{Ran}(P_A(\{\alpha\})) \subseteq \operatorname{Ker}(A - \alpha)$. To see the converse, let $u \in \operatorname{Ker}(A - \alpha)$ and use the previous example.)

6.5. The Gelfand representation theorem

In this section we look at an alternative approach to the spectral theorem by trying to find a canonical representation for a Banach algebra. The idea is as follows: Given the Banach algebra $C[a,b]$ we have a one-to-one correspondence between points $x_0 \in [a,b]$ and point evaluations $m_{x_0}(f) = f(x_0)$. These point evaluations are linear functionals which at the same time preserve multiplication. In fact, we will see that these are the only (nontrivial) multiplicative functionals and hence we also have a one-to-one correspondence between points in $[a,b]$ and multiplicative functionals. Now $m_{x_0}(f) = f(x_0)$ says that the action of a multiplicative functional on a function is the same as the action of the function on a point. So for a general algebra X we can try to go the other way: Consider the multiplicative functionals m als points and the elements $x \in X$ as functions acting on these points (the value of this function being $m(x)$). This gives a map, the Gelfand representation, from X into an algebra of functions.

A nonzero algebra homeomorphism $m : X \to \mathbb{C}$ will be called a **multiplicative linear functional** or **character**:

$$m(xy) = m(x)m(y), \qquad m(e) = 1. \tag{6.60}$$

Note that the last equation comes for free from multiplicativity since m is nontrivial. Moreover, there is no need to require that m is continuous as this will also follow automatically (cf. Lemma 6.20 below).

As we will see, they are closely related to **ideals**, that is linear subspaces I of X for which $a \in I$, $x \in X$ implies $ax \in I$ and $xa \in I$. An ideal is called **proper** if it is not equal to X and it is called **maximal** if it is not contained in any other proper ideal.

Example 6.30. Let $X := C([a,b])$ be the continuous functions over some compact interval. Then for fixed $x_0 \in [a,b]$, the linear functional $m_{x_0}(f) := f(x_0)$ is multiplicative. Moreover, its kernel $\operatorname{Ker}(m_{x_0}) = \{f \in C([a,b]) | f(x_0) = 0\}$ is a maximal ideal (we will prove this in more generality below). ⋄

Example 6.31. Let X be a Banach space. Then the compact operators are a closed ideal in $\mathscr{L}(X)$ (cf. Theorem 3.1). ⋄

We first collect a few elementary properties of ideals.

Lemma 6.19. *Let X be a unital Banach algebra.*

(i) *A proper ideal can never contain an invertible element.*

(ii) *If X is commutative every non-invertible element is contained in a proper ideal.*

(iii) *The closure of a (proper) ideal is again a (proper) ideal.*

(iv) *Maximal ideals are closed.*

(v) *Every proper ideal is contained in a maximal ideal.*

Proof. (i). If $x \in I$ is invertible then $y = x(x^{-1}y) \in I$ shows $I = X$. (ii). Consider the ideal $xX = \{x\, y | y \in X\}$. Then $xX = X$ if and only if there is some $y \in X$ with $xy = e$, that is , $y = x^{-1}$. (iii) and (iv). That the closure of an ideal is again an ideal follows from continuity of the product. Indeed, for $a \in \overline{I}$ choose a sequence $a_n \in I$ converging to a. Then $xa_n \in I \to xa \in \overline{I}$ as well as $a_n x \in I \to ax \in \overline{I}$. Moreover, note that by Lemma 6.1 all elements in the ball $B_1(e)$ are invertible and hence every proper ideal must be contained in the closed set $X \backslash B_1(e)$. So the closure of a proper ideal is proper and any maximal ideal must be closed. (v). To see that every ideal I is contained in a maximal ideal consider the family of proper ideals containing I ordered by inclusion. Then, since any union of a chain of proper ideals is again a proper ideal (that the union is again an ideal is straightforward, to see that it is proper note that it does not contain $B_1(e)$). Consequently Zorn's lemma implies existence of a maximal element. □

Note that if I is a closed ideal, then the quotient space X/I (cf. Lemma 1.19) is again a Banach algebra if we define

$$[x][y] = [xy]. \tag{6.61}$$

Indeed $(x + I)(y + I) = xy + I$ and hence the multiplication is well-defined and inherits the distributive and associative laws from X. Also $[e]$ is an identity. Finally,

$$\begin{aligned} \|[xy]\| &= \inf_{a\in I} \|xy + a\| = \inf_{b,c\in I} \|(x + b)(y + c)\| \le \inf_{b\in I} \|x + b\| \inf_{c\in I} \|y + c\| \\ &= \|[x]\|\|[y]\|. \end{aligned} \tag{6.62}$$

In particular, the projection map $\pi : X \to X/I$ is a Banach algebra homomorphism.

Example 6.32. Consider the Banach algebra $\mathscr{L}(X)$ together with the ideal of compact operators $\mathscr{C}(X)$. Then the Banach algebra $\mathscr{L}(X)/\mathscr{C}(X)$ is known as the **Calkin algebra**. Atkinson's theorem (Theorem 6.34) says that the invertible elements in the Calkin algebra are precisely the images of the Fredholm operators. ◇

Lemma 6.20. *Let X be a unital Banach algebra and m a character. Then* $\operatorname{Ker}(m)$ *is a maximal ideal and m is continuous with $\|m\| = m(e) = 1$.*

Proof. It is straightforward to check that $\operatorname{Ker}(m)$ is an ideal. Moreover, every x can be written as

$$x = m(x)e + y, \qquad y \in \operatorname{Ker}(m).$$

Let I be an ideal containing $\operatorname{Ker}(m)$. If there is some $x \in I \setminus \operatorname{Ker}(m)$ then $m(x)^{-1}x = e + m(x)^{-1}y \in I$ and hence also $e = (e + m(x)^{-1}y) -$

$m(x)^{-1}y \in I$. Thus $\operatorname{Ker}(m)$ is maximal. Since maximal ideals are closed by the previous lemma, we conclude that m is continuous by Problem 1.36. Clearly $\|m\| \geq m(e) = 1$. Conversely, suppose we can find some $x \in X$ with $\|x\| < 1$ and $m(x) = 1$. Consequently $\|x^n\| \leq \|x\|^n \to 0$ contradicting $m(x^n) = m(x)^n = 1$. □

In a commutative algebra the other direction is also true.

Lemma 6.21. *In a commutative unital Banach algebra the characters and maximal ideals are in one-to-one correspondence.*

Proof. We have already seen that for a character m there is corresponding maximal ideal $\operatorname{Ker}(m)$. Conversely, let I be a maximal ideal and consider the projection $\pi : X \to X/I$. We first claim that every nontrivial element in X/I is invertible. To this end suppose $[x_0] \neq [0]$ were not invertible. Then $J = [x_0]X/I$ is a proper ideal (if it would contain the identity, X/I would contain an inverse of $[x_0]$). Moreover, $I' = \{y \in X | [y] \in J\}$ is a proper ideal of X (since $e \in I'$ would imply $[e] \in J$) which contains I (since $[x] = [0] \in J$ for $x \in I$) but is strictly larger as $x_0 \in I' \setminus I$. This contradicts maximality and hence by the Gelfand–Mazur theorem (Theorem 6.4), every element of X/I is of the form $\alpha[e]$. If $h : X/I \to \mathbb{C}$, $h(\alpha[e]) \mapsto \alpha$ is the corresponding algebra isomorphism, then $m = h \circ \pi$ is a character with $\operatorname{Ker}(m) = I$. □

Now we continue with the following observation: For fixed $x \in X$ we get a map $X^* \to \mathbb{C}$, $\ell \mapsto \ell(x)$. Moreover, if we equip X^* with the weak-$*$ topology then this map will be continuous (by the very definition of the weak-$*$ topology). So we have a map $X \to C(X^*)$ and restricting this map to the set of all characters $\mathcal{M} \subseteq X^*$ (equipped with the relative topology of the weak-$*$ topology) it is known as the **Gelfand transform**:

$$\Gamma : X \to C(\mathcal{M}), \quad x \mapsto \hat{x}(m) = m(x). \tag{6.63}$$

Theorem 6.22 (Gelfand representation theorem)**.** *Let X be a unital Banach algebra. Then the set of all characters $\mathcal{M} \subseteq X^*$ is a compact Hausdorff space with respect to the weak-$*$ topology and the Gelfand transform is a continuous algebra homomorphism with $\hat{e} = 1$.*

Moreover, $\hat{x}(\mathcal{M}) \subseteq \sigma(x)$ and hence $\|\hat{x}\|_\infty \leq r(x) \leq \|x\|$ where $r(x)$ is the spectral radius of x. If X is commutative then $\hat{x}(\mathcal{M}) = \sigma(x)$ and hence $\|\hat{x}\|_\infty = r(x)$.

Proof. As pointed out before, for fixed $x, y \in X$ the map $X^* \to \mathbb{C}^3$, $\ell \mapsto (\ell(x), \ell(y), \ell(xy))$ is continuous and so is the map $X^* \to \mathbb{C}$, $\ell \mapsto \ell(x)\ell(y) - \ell(xy)$ as a composition of continuous maps. Hence the kernel of this map $M_{x,y} = \{\ell \in X^* | \ell(x)\ell(y) = \ell(xy)\}$ is weak-$*$ closed and so is $\mathcal{M} = M_0 \cap \bigcap_{x,y \in X} M_{x,y}$ where $M_0 = \{\ell \in X^* | \ell(e) = 1\}$. So $\mathcal{M}$ is a weak-$*$ closed subset

of the unit ball in X^* and the first claim follows form the Banach–Alaoglu theorem (Theorem 5.10).

Next $(x+y)^\wedge(m) = m(x+y) = m(x)+m(y) = \hat{x}(m)+\hat{y}(m)$, $(xy)^\wedge(m) = m(xy) = m(x)m(y) = \hat{x}(m)\hat{y}(m)$, and $\hat{e}(m) = m(e) = 1$ shows that the Gelfand transform is an algebra homomorphism.

Moreover, if $m(x) = \alpha$ then $x - \alpha \in \operatorname{Ker}(m)$ implying that $x - \alpha$ is not invertible (as maximal ideals cannot contain invertible elements), that is $\alpha \in \sigma(x)$. Conversely, if X is commutative and $\alpha \in \sigma(x)$, then $x - \alpha$ is not invertible and hence contained in some maximal ideal, which in turn is the kernel of some character m. Whence $m(x - \alpha) = 0$, that is $m(x) = \alpha$ for some m. □

Of course this raises the question about injectivity or surjectivity of the Gelfand transform. Clearly

$$x \in \operatorname{Ker}(\Gamma) \quad \Leftrightarrow \quad x \in \bigcap_{m\in\mathcal{M}} \operatorname{Ker}(m) \tag{6.64}$$

and it can only be injective if X is commutative. In this case

$$x \in \operatorname{Ker}(\Gamma) \quad \Leftrightarrow \quad x \in \operatorname{Rad}(X) := \bigcap_{I \text{ maximal ideal}} I, \tag{6.65}$$

where $\operatorname{Rad}(X)$ is known as the **Jacobson radical** of X and a Banach algebra is called **semi-simple** if the Jacobson radical is zero. So to put this result to use one needs to understand the set of characters, or equivalently, the set of maximal ideals. Two examples where this can be done are given below. The first one is not very surprising.

Example 6.33. If we start with a compact Hausdorff space K and consider $C(K)$ we get nothing new. Indeed, first of all notice that the map $K \to \mathcal{M}$, $x_0 \mapsto m_{x_0}$ which assigns each x_0 the corresponding point evaluation $m_{x_0}(f) = f(x_0)$ is injective and continuous. Hence, by compactness of K, it will be a homeomorphism once we establish surjectivity (Corollary B.17). To this end we will show that all maximal ideals are of the form $I = \operatorname{Ker}(m_{x_0})$ for some $x_0 \in K$. So let I be an ideal and suppose there is no point where all functions vanish. Then for every $x \in K$ there is a ball $B_{r(x)}(x)$ and a function $f_x \in C(K)$ such that $|f_x(y)| \geq 1$ for $y \in B_{r(x)}(x)$. By compactness finitely many of these balls will cover K. Now consider $f = \sum_j f_{x_j}^* f_{x_j} \in I$. Then $f \geq 1$ and hence f is invertible, that is $I = C(K)$. Thus maximal ideals are of the form $I_{x_0} = \{f \in C(K) | f(x_0) = 0\}$ which are precisely the kernels of the characters $m_{x_0}(f) = f(x_0)$. Thus $\mathcal{M} \cong K$ as well as $\hat{f} \cong f$. ⋄

Example 6.34. Consider the Wiener algebra $\mathcal{A}$ of all periodic continuous functions which have an absolutely convergent Fourier series. As in the previous example it suffices to show that all maximal ideals are of the form

$I_{x_0} = \{f \in \mathcal{A} | f(x_0) = 0\}$. To see this set $e_k(x) = \mathrm{e}^{\mathrm{i}kx}$ and note $\|e_k\|_{\mathcal{A}} = 1$. Hence for every character $m(e_k) = m(e_1)^k$ and $|m(e_k)| \leq 1$. Since the last claim holds for both positive and negative k, we conclude $|m(e_k)| = 1$ and thus there is some $x_0 \in [-\pi, \pi]$ with $m(e_k) = \mathrm{e}^{\mathrm{i}kx_0}$. Consequently $m(f) = f(x_0)$ and point evaluations are the only characters. Equivalently, every maximal ideal is of the form $\mathrm{Ker}(m_{x_0}) = I_{x_0}$.

So, as in the previous example, $\mathcal{M} \cong [-\pi, \pi]$ (with $-\pi$ and π identified) as well hat $\hat{f} \cong f$. Moreover, the Gelfand transform is injective but not surjective since there are continuous functions whose Fourier series are not absolutely convergent. Incidentally this also shows that the Wiener algebra is *no* C^* algebra (despite the fact that we have a natural conjugation which satisfies $\|f^*\|_{\mathcal{A}} = \|f\|_{\mathcal{A}}$ — this again underlines the special role of (6.29)) as the Gelfand–Naimark theorem below will show that the Gelfand transform is bijective for commutative C^* algebras. ⋄

Since $0 \notin \sigma(x)$ implies that x is invertible the Gelfand representation theorem also contains a useful criterion for invertibility.

Corollary 6.23. *In a commutative unital Banach algebra an element x is invertible if and only if $m(x) \neq 0$ for all characters m.*

And applying this to the last example we get the following famous theorem of Wiener:

Theorem 6.24 (Wiener)**.** *Suppose $f \in C_{per}[-\pi, \pi]$ has an absolutely convergent Fourier series and does not vanish on $[-\pi, \pi]$. Then the function $\frac{1}{f}$ also has an absolutely convergent Fourier series.*

If we turn to a commutative C^* algebra the situation further simplifies. First of all note that characters respect the additional operation automatically.

Lemma 6.25. *If X is a unital C^* algebra, then every character satisfies $m(x^*) = m(x)^*$. In particular, the Gelfand transform is a continuous $*$-algebra homomorphism with $\hat{e} = 1$ in this case.*

Proof. If x is self-adjoint then $\sigma(x) \subseteq \mathbb{R}$ (Lemma 6.8) and hence $m(x) \in \mathbb{R}$ by the Gelfand representation theorem. Now for general x we can write $x = a + \mathrm{i}b$ with $a = \frac{x+x^*}{2}$ and $b = \frac{x-x^*}{2\mathrm{i}}$ self-adjoint implying

$$m(x^*) = m(a - \mathrm{i}b) = m(a) - \mathrm{i}m(b) = (m(a) + \mathrm{i}m(b))^* = m(x)^*.$$

Consequently the Gelfand transform preserves the involution: $(x^*)^\wedge(m) = m(x^*) = m(x)^* = \hat{x}^*(m)$. □

Theorem 6.26 (Gelfand–Naimark)**.** *Suppose X is a unital commutative C^* algebra. Then the Gelfand transform is an isometric isomorphism between C^* algebras.*

Proof. Since in a commutative C^* algebra every element is normal, Lemma 6.7 implies that the Gelfand transform is isometric. Moreover, by the previous lemma the image of X under the Gelfand transform is a closed $*$-subalgebra which contains $\hat{e} \equiv 1$ and separates points (if $\hat{x}(m_1) = \hat{x}(m_2)$ for all $x \in X$ we have $m_1 = m_2$). Hence it must be all of $C(\mathcal{M})$ by the Stone–Weierstraß theorem (Theorem B.41). □

The first moral from this theorem is that from an abstract point of view there is only one commutative C^* algebra, namely $C(K)$ with K some compact Hausdorff space. Moreover, it also very much reassembles the spectral theorem and in fact, we can derive the spectral theorem by applying it to $C^*(x)$, the C^* algebra generated by x (cf. (6.32)). This will even give us the more general version for normal elements. As a preparation we show that it makes no difference whether we compute the spectrum in X or in $C^*(x)$.

Lemma 6.27 (Spectral permanence)**.** *Let X be a C^* algebra and $Y \subseteq X$ a closed $*$-subalgebra containing the identity. Then $\sigma(y) = \sigma_Y(y)$ for every $y \in Y$, where $\sigma_Y(y)$ denotes the spectrum computed in Y.*

Proof. Clearly we have $\sigma(y) \subseteq \sigma_Y(y)$ and it remains to establish the reverse inclusion. If $(y-\alpha)$ has an inverse in X, then the same is true for $(y-\alpha)^*(y-\alpha)$ and $(y-\alpha)(y-\alpha)^*$. But the last two operators are self-adjoint and hence have real spectrum in Y. Thus $((y-\alpha)^*(y-\alpha) + \frac{\mathrm{i}}{n})^{-1} \in Y$ and letting $n \to \infty$ shows $((y-\alpha)^*(y-\alpha))^{-1} \in Y$ since taking the inverse is continuous and Y is closed. Similarly $((y-\alpha)(y-\alpha)^*)^{-1} \in Y$ and whence $(y-\alpha)^{-1} \in Y$ by Problem 6.6. □

Now we can show

Theorem 6.28 (Spectral theorem)**.** *If X is a C^* algebra and x is normal, then there is an isometric isomorphism $\Phi : C(\sigma(x)) \to C^*(x)$ such that $f(t) = t$ maps to $\Phi(t) = x$ and $f(t) = 1$ maps to $\Phi(1) = e$.*

Moreover, for every $f \in C(\sigma(x))$ we have

$$\sigma(f(x)) = f(\sigma(x)), \tag{6.66}$$

where $f(x) = \Phi(f)$.

Proof. Given a normal element $x \in X$ we want to apply the Gelfand–Naimark theorem in $C^*(x)$. By our lemma we have $\sigma(x) = \sigma_{C^*(x)}(x)$. We first show that we can identify $\mathcal{M}$ with $\sigma(x)$. By the Gelfand representation theorem (applied in $C^*(x)$), $\hat{x} : \mathcal{M} \to \sigma(x)$ is continuous and surjective. Moreover, if for given $m_1, m_2 \in \mathcal{M}$ we have $\hat{x}(m_1) = m_1(x) = m_2(x) = \hat{x}(m_2)$ then

$$m_1(p(x, x^*)) = p(m_1(x), m_1(x)^*) = p(m_2(x), m_2(x)^*) = m_2(p(x, x^*))$$

for any polynomial $p : \mathbb{C}^2 \to \mathbb{C}$ and hence $m_1(y) = m_2(y)$ for every $y \in C^*(x)$ implying $m_1 = m_2$. Thus $\hat{x}$ is injective and hence a homeomorphism as $\mathcal{M}$ is compact. Thus we have an isometric isomorphism

$$\Psi : C(\sigma(x)) \to C(\mathcal{M}), \qquad f \mapsto f \circ \hat{x},$$

and the isometric isomorphism we are looking for is $\Phi = \Gamma^{-1} \circ \Psi$. Finally, $\sigma(f(x)) = \sigma_{C^*(x)}(\Phi(f)) = \sigma_{C(\sigma(x))}(f) = f(\sigma(x))$. □

Example 6.35. Let X be a C^* algebra and $x \in X$ normal. By the spectral theorem $C^*(x)$ is isomorphic to $C(\sigma(x))$. Hence every $y \in C^*(x)$ can be written as $y = f(x)$ for some $f \in C(\sigma(x))$ and every character is of the form $m(y) = m(f(x)) = f(\alpha)$ for some $\alpha \in \sigma(x)$. ⋄

Problem 6.28. *Show that $C^1[a, b]$ is a Banach algebra. What are the characters? Is it semi-simple?*

Problem 6.29. *Consider the subalgebra of the Wiener algebra consisting of all functions whose negative Fourier coefficients vanish. What are the characters? (Hint: Observe that these functions can be identified with holomorphic functions inside the unit disc with summable Taylor coefficients via $f(z) = \sum_{k=0}^{\infty} \hat{f}_k z^k$ known as the **Hardy space** H^1 of the disc.)*

6.6. Fredholm operators

In this section we want to investigate solvability of the equation

$$Ax = y \tag{6.67}$$

for $A \in \mathscr{L}(X, Y)$ given $y \in Y$. Clearly there exists a solution if $y \in \operatorname{Ran}(A)$ and this solution is unique if $\operatorname{Ker}(A) = \{0\}$. Hence these subspaces play a crucial role. Moreover, if the underlying Banach spaces are finite dimensional, the kernel has a complement $X = \operatorname{Ker}(A) \dot{+} X_0$ and after factoring out the kernel this complement is isomorphic to the range of A. As a consequence, the dimensions of these spaces are connected by the famous **rank-nullity theorem**

$$\dim \operatorname{Ker}(A) + \dim \operatorname{Ran}(A) = \dim X \tag{6.68}$$

from linear algebra. In our infinite dimensional setting (apart from the technical difficulties that the kernel might not be complemented and the range might not be closed) this formula does not contain much information, but if we rewrite it in terms of the index,

$$\operatorname{ind}(A) := \dim \operatorname{Ker}(A) - \dim \operatorname{Coker}(A) = \dim(X) - \dim(Y), \tag{6.69}$$

at least the left-hand side will be finite if we assume both $\operatorname{Ker}(A)$ and $\operatorname{Coker}(A)$ to be finite dimensional. One of the most useful consequences

of the rank-nullity theorem is that in the case $X = Y$ the index will vanish and hence uniqueness of solutions for $Ax = y$ will automatically give you existence for free (and vice versa). Indeed, for equations of the form $x + Kx = y$ with K compact originally studied by Fredholm this is still true by the famous Fredholm alternative (Theorem 6.15). It took a while until Noether found an example of singular integral equations which have a nonzero index and started investigating the general case.

We first note that in this case $\operatorname{Ran}(A)$ will be automatically closed.

Lemma 6.29. *Suppose $A \in \mathscr{L}(X, Y)$ with finite dimensional cokernel. Then $\operatorname{Ran}(A)$ is closed.*

Proof. First of all note that the induced map $\tilde{A} : X/\operatorname{Ker}(A) \to Y$ is injective (Problem 1.42). Moreover, the assumption that the cokernel is finite says that there is a finite subspace $Y_0 \subset Y$ such that $Y = Y_0 \dotplus \operatorname{Ran}(A)$. Then

$$\hat{A} : X/\operatorname{Ker}(A) \oplus Y_0 \to Y, \qquad \hat{A}(x, y) = \tilde{A}x + y$$

is bijective and hence a homeomorphism by Theorem 4.6. Since X is a closed subspace of $X \oplus Y_0$ we see that $\operatorname{Ran}(A) = \hat{A}(X)$ is closed in Y. □

Hence we call an operator $A \in \mathscr{L}(X, Y)$ a **Fredholm operator** (also **Noether operator**) if both its kernel and cokernel are finite dimensional. In this case we define its **index** as

$$\operatorname{ind}(A) := \dim \operatorname{Ker}(A) - \dim \operatorname{Coker}(A). \tag{6.70}$$

The set of Fredholm operators will be denoted by $\Phi(X, Y)$ and the set of Fredholm operators with index zero will be denoted by $\Phi_0(X, Y)$. An immediate consequence of Theorem 4.26 is:

Theorem 6.30 (Riesz)**.** *A bounded operator A is Fredholm if and only if A' is and in this case*

$$\operatorname{ind}(A') = -\operatorname{ind}(A). \tag{6.71}$$

Note that by Problem 4.27 we have

$$\operatorname{ind}(A) = \dim \operatorname{Ker}(A) - \dim \operatorname{Ker}(A') = \dim \operatorname{Coker}(A') - \dim \operatorname{Coker}(A) \tag{6.72}$$

since for a finite dimensional space the dual space has the same dimension.

Example 6.36. The right shift operator S in $X = Y = \ell^p(\mathbb{N})$, $1 \le p < 1$ is Fredholm. In fact, recall that S' is the right shift and $\operatorname{Ker}(S) = \{0\}$, $\operatorname{Ker}(S') = \operatorname{span}\{\delta^1\}$. In particular, $\operatorname{ind}(S) = 1$ and $\operatorname{ind}(S') = -1$. ⋄

In the case of Hilbert spaces $\operatorname{Ran}(A)$ closed implies $\mathfrak{H} = \operatorname{Ran}(A) \oplus \operatorname{Ran}(A)^\perp$ and thus $\operatorname{Coker}(A) \cong \operatorname{Ran}(A)^\perp$. Hence an operator is Fredholm if

$\operatorname{Ran}(A)$ is closed and $\operatorname{Ker}(A)$ and $\operatorname{Ran}(A)^\perp$ are both finite dimensional. In this case

$$\operatorname{ind}(A) = \dim \operatorname{Ker}(A) - \dim \operatorname{Ran}(A)^\perp \tag{6.73}$$

and $\operatorname{ind}(A^*) = -\operatorname{ind}(A)$ as is immediate from (2.28).

Example 6.37. Suppose $\mathfrak{H}$ is a Hilbert space and $A = A^*$ is a self-adjoint Fredholm operator, then (2.28) shows that $\operatorname{ind}(A) = 0$. In particular, a self-adjoint operator is Fredholm if $\dim \operatorname{Ker}(A) < \infty$ and $\operatorname{Ran}(A)$ is closed. For example, according to Example 6.29, $A - \lambda$ is Fredholm if λ is an eigenvalue of finite multiplicity (in fact, inspecting this example shows that the converse is also true).

It is however important to notice that $\operatorname{Ran}(A)^\perp$ finite dimensional does *not* imply $\operatorname{Ran}(A)$ closed! For example consider $(Ax)_n = \frac{1}{n} x_n$ in $\ell^2(\mathbb{N})$ whose range is dense but not closed. ⋄

Another useful formula concerns the product of two Fredholm operators. For its proof it will be convenient to use the notion of an exact sequence: Let X_j be Banach spaces. A sequence of operators $A_j \in \mathscr{L}(X_j, X_{j+1})$

$$X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \cdots X_n \xrightarrow{A_n} X_{n+1}$$

is said to be **exact** if $\operatorname{Ran}(A_j) = \operatorname{Ker}(A_{j+1})$ for $0 \le j < n$. We will also need the following two easily checked facts: If X_{j-1} and X_{j+1} are finite dimensional, so is X_j (Problem 6.30) and if the sequence of finite dimensional spaces starts with $X_0 = \{0\}$ and ends with $X_{n+1} = \{0\}$, then the alternating sum over the dimensions vanishes (Problem 6.31).

Lemma 6.31 (Atkinson)**.** *Suppose $A \in \mathscr{L}(X, Y)$, $B \in \mathscr{L}(Y, Z)$. If two of the operators A, B, AB are Fredholm, so is the third and we have*

$$\operatorname{ind}(AB) = \operatorname{ind}(A) + \operatorname{ind}(B). \tag{6.74}$$

Proof. It is straightforward to check that the sequence

$$\begin{aligned} 0 \longrightarrow \operatorname{Ker}(A) \longrightarrow \operatorname{Ker}(BA) \xrightarrow{A} \operatorname{Ker}(B) \longrightarrow \operatorname{Coker}(A) \\ \xrightarrow{B} \operatorname{Coker}(BA) \longrightarrow \operatorname{Coker}(B) \longrightarrow 0 \end{aligned}$$

is exact. Here the maps which are not explicitly stated are canonical inclusions/projections. Hence by Problem 6.30, if two operators are Fredholm, so is the third. Moreover, the formula for the index follows from Problem 6.31. □

Next we want to look a bit further into the structure of Fredholm operators. First of all, since $\operatorname{Ker}(A)$ is finite dimensional it is complemented (Problem 4.25), that is, there exists a closed subspace $X_0 \subseteq X$ such that $X = \operatorname{Ker}(A) \dotplus X_0$ and a corresponding projection $P \in \mathscr{L}(X)$

with $\operatorname{Ran}(P) = \operatorname{Ker}(A)$. Similarly, $\operatorname{Ran}(A)$ is complemented (Problem 1.43) and there exists a closed subspace $Y_0 \subseteq Y$ such that $Y = Y_0 \dotplus \operatorname{Ran}(A)$ and a corresponding projection $Q \in \mathscr{L}(Y)$ with $\operatorname{Ran}(Q) = Y_0$. With respect to the decomposition $\operatorname{Ker}(A) \oplus X_0 \to Y_0 \oplus \operatorname{Ran}(A)$ our Fredholm operator is given by

$$A = \begin{pmatrix} 0 & 0 \\ 0 & A_0 \end{pmatrix}, \tag{6.75}$$

where A_0 is the restriction of A to $X_0 \to \operatorname{Ran}(A)$. By construction A_0 is bijective and hence a homeomorphism (Theorem 4.6). Defining

$$B := \begin{pmatrix} 0 & 0 \\ 0 & A_0^{-1} \end{pmatrix} \tag{6.76}$$

we get

$$AB = \mathbb{I} - Q, \qquad BA = \mathbb{I} - P \tag{6.77}$$

and hence A is invertible up to finite rank operators. Now we are ready for showing that the index is stable under small perturbations.

Theorem 6.32 (Dieudonné)**.** *The set of Fredholm operators $\Phi(X, Y)$ is open in $\mathscr{L}(X, Y)$ and the index is locally constant.*

Proof. Let $C \in \mathscr{L}(X, Y)$ and write it as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

with respect to the above splitting. Then if $\|C_{22}\| < \|A_0^{-1}\|^{-1}$ we have that $A_0 + C_{22}$ is still invertible (Problem 1.34). Now introduce

$$D_1 = \begin{pmatrix} \mathbb{I} & -C_{12}(A_0 + C_{22})^{-1} \\ 0 & \mathbb{I} \end{pmatrix}, \qquad D_2 = \begin{pmatrix} \mathbb{I} & 0 \\ -(A_0 + C_{22})^{-1} C_{21} & \mathbb{I} \end{pmatrix}$$

and observe

$$D_1(A + C)D_2 = \begin{pmatrix} C_{11} - C_{12}(A_0 + C_{22})^{-1} C_{21} & 0 \\ 0 & A_0 + C_{22} \end{pmatrix}.$$

Since D_1, D_2 are homeomorphisms we see that $A + C$ is Fredholm since the right-hand side obviously is. Moreover, $\operatorname{ind}(A + C) = \operatorname{ind}(C_{11} - C_{12}(A_0 + C_{22})^{-1} C_{21}) = \dim(\operatorname{Ker}(A)) - \dim(Y_0) = \operatorname{ind}(A)$ since the second operator is between finite dimensional spaces and the index can be evaluated using (6.69). □

Since the index is locally constant, it is constant on every connected component of $\Phi(X, Y)$ which often is useful for computing the index. The next result identifies an important class of Fredholm operators and uses this observation for computing the index.

Theorem 6.33 (Riesz)**.** *For every $K \in \mathscr{C}(X)$ we have $\mathbb{I} - K \in \Phi_0(X)$.*

Proof. That $\mathbb{I}-K$ is Fredholm follows from Lemma 6.11 since K' is compact as well and $\operatorname{Coker}(\mathbb{I}-K)^* \cong \operatorname{Ker}(\mathbb{I}-K')$ by Problem 4.27. Furthermore, the index is constant along $[0,1] \to \Phi(X)$, $\alpha \mapsto \mathbb{I}-\alpha K$ and hence $\operatorname{ind}(\mathbb{I}-K) = \operatorname{ind}(\mathbb{I}) = 0$. □

Next we show that an operator is Fredholm if and only if it has a left/right inverse up to compact operators.

Theorem 6.34 (Atkinson)**.** *An operator $A \in \mathscr{L}(X,Y)$ is Fredholm if and only if there exist $B_1, B_2 \in \mathscr{L}(Y,X)$ such that $B_1A-\mathbb{I} \in \mathscr{C}(X)$ and $AB_2-\mathbb{I} \in \mathscr{C}(Y)$.*

Proof. If A is Fredholm we have already given an operator B in (6.76) such that $BA-\mathbb{I}$ and $AB-\mathbb{I}$ are finite rank. Conversely, according to Theorem 6.33 B_1A and AB_2 are Fredholm. Since $\operatorname{Ker}(A) \subseteq \operatorname{Ker}(B_2A)$ and $\operatorname{Ran}(AB_2) \subseteq \operatorname{Ran}(A)$ this shows that A is Fredholm. □

Operators B_1 and B_2 as in the previous theorem are also known as a left and right **parametrix**, respectively. As a consequence we can now strengthen Theorem 6.33:

Corollary 6.35 (Yood)**.** *For every $A \in \Phi(X,Y)$ and $K \in \mathscr{C}(X,Y)$ we have $A+K \in \Phi(X,Y)$ with $\operatorname{ind}(A+K) = \operatorname{ind}(A)$.*

Proof. Using (6.76) we see that $B(A+K)-\mathbb{I} = -P+BK \in \mathscr{C}(Y)$ and $(A+K)B-\mathbb{I} = -Q+KB \in \mathscr{C}(X)$ and hence $A+K$ is Fredholm. In fact, $A+\alpha K$ is Fredholm for $\alpha \in [0,1]$ and hence $\operatorname{ind}(A+K) = \operatorname{ind}(A)$ by continuity of the index. □

Fredholm operators are also used to split the spectrum. For $A \in \mathscr{L}(X)$ one defines the **essential spectrum**

$$\sigma_{ess}(A) := \{\alpha \in \mathbb{C} | A-\alpha \notin \Phi_0(X)\} \subseteq \sigma(A) \tag{6.78}$$

and the **Fredholm spectrum**

$$\sigma_\Phi(A) := \{\alpha \in \mathbb{C} | A-\alpha \notin \Phi(X)\} \subseteq \sigma_{ess}(A). \tag{6.79}$$

By Dieudonné's theorem both $\sigma_{ess}(A)$ and $\sigma_\Phi(A)$ are closed. Warning: These definitions are not universally accepted and several variants can be found in the literature.

Example 6.38. let X be infinite dimensional and $K \in \mathscr{C}(X)$. Then $\sigma_{ess}(K) = \sigma_\Phi(K) = \{0\}$. ◇

By Corollary 6.35 both the Fredholm spectrum and the essential spectrum are invariant under compact perturbations:

Theorem 6.36 (Weyl)**.** *Let $A \in \mathscr{L}(X)$, then*

$$\sigma_\Phi(A+K) = \sigma_\Phi(A), \quad \sigma_{ess}(A+K) = \sigma_{ess}(A), \quad K \in \mathscr{C}(X). \tag{6.80}$$

Note that if X is a Hilbert space and A is self-adjoint, then $\sigma(A) \subseteq \mathbb{R}$ and hence $\operatorname{ind}(A-\alpha) = -\operatorname{ind}((A-\alpha)^*) = \operatorname{ind}(A-\alpha^*)$ shows that the index is always zero for $\alpha \notin \sigma_\Phi(A)$. Thus $\sigma_{ess}(A) = \sigma_\Phi(A)$ for self-adjoint operators.

Moreover, if $\alpha \in \sigma_{ess}(A)$, using the notation from (6.75) for $A-\alpha$, we can find an invertible operator $K_0 : \operatorname{Ker}(A) \to Y_0$ and extend it to a finite rank operator $K : X \to X$ by setting it equal to 0 on X_0. Then $(A+K)-\alpha$ has a bounded inverse implying $\alpha \in \rho(A+K)$. Thus the essential spectrum is precisely the part which is stable under compact perturbations

$$\sigma_{ess}(A) = \bigcap_{K \in \mathscr{C}(X)} \sigma(A+K). \tag{6.81}$$

The complement is called the **discrete spectrum**

$$\sigma_d(A) := \sigma(A) \setminus \sigma_{ess}(A) = \{\alpha \in \sigma(A) | A-\alpha \in \Phi_0(X)\}. \tag{6.82}$$

Clearly points in the discrete spectrum are eigenvalues with finite geometric multiplicity. Moreover, if the algebraic multiplicity of an eigenvalue in $\sigma_d(A)$ is finite, then by Problem 6.33 and Lemma 6.12 there is an n such that $X = \operatorname{Ker}((A-\alpha)^n) \dot{+} \operatorname{Ran}((A-\alpha)^n)$. These spaces are invariant and $\operatorname{Ker}((A-\alpha)^n)$ is still finite dimensional (since $(A-\alpha)^n$ is also Fredholm). With respect to this decomposition A has a simple block structure with the first block $A_0 : \operatorname{Ker}((A-\alpha)^n) \to \operatorname{Ker}((A-\alpha)^n)$ such that $A_0 - \alpha$ is nilpotent and the second block $A_1 : \operatorname{Ran}((A-\alpha)^n) \to \operatorname{Ran}((A-\alpha)^n)$ such that $\alpha \in \rho(A_0)$. Hence for sufficiently small $\varepsilon > 0$ we will have $\alpha+\varepsilon \in \rho(A_0)$ and $\alpha+\varepsilon \in \rho(A_1)$ implying $\alpha+\varepsilon \in \rho(A)$ such that α is an isolated point of the spectrum. This happens for example if A is self-adjoint (in which case $n=1$). However, in the general case, σ_d could contain much more than just isolated eigenvalues with finite algebraic multiplicity as the following example shows.

Example 6.39. Let $X = \ell^2(\mathbb{N}) \oplus \ell^2(\mathbb{N})$ with $A = L \oplus R$, where L, R are the left, right shift, respectively. Explicitly, $A(x,y) = ((x_2, x_3, \dots), (0, y_1, y_2, \dots))$. Then $\sigma(A) = \sigma(L) \cup \sigma(R) = \bar{B}_1(0)$ and $\sigma_p(A) = \sigma_p(L) \cup \sigma_p(R) = B_1(0)$. Moreover, note that $A \in \Phi_0$ and that 0 is an eigenvalue of infinite algebraic multiplicity.

Now consider the rank-one operator $K(x,y) := ((0,0,\dots),(x_1,0,\dots))$ such that $(A+K)(x,y) = ((x_2,x_3,\dots),(x_1,y_1,y_2,\dots))$. Then $A+K$ is unitary (note that this is essentially a two-sided shift) and hence $\sigma(A+K) \subseteq \partial B_1(0)$. Thus $\sigma_d(A) = B_1(0)$ and $\sigma_{ess}(A) = \partial B_1(0)$. ⋄

Problem* 6.30. *Suppose $X \xrightarrow{A} Y \xrightarrow{B} Z$ is exact. Show that if X and Z are finite dimensional, so is Y.*

Problem* 6.31. *Let X_j be finite dimensional vector spaces and suppose*

$$0 \longrightarrow X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \cdots X_{n-1} \xrightarrow{A_{n-1}} X_n \longrightarrow 0$$

is exact. Show that

$$\sum_{j=1}^{n} (-1)^j \dim(X_j) = 0.$$

(Hint: Rank-nullity theorem.)

Problem 6.32. *Let $A \in \Phi(X,Y)$ with a corresponding parametrix $B_1, B_2 \in \mathscr{L}(Y,X)$. Set $K_1 := \mathbb{I} - B_1 A \in \mathscr{C}(X)$, $K_2 := \mathbb{I} - AB_2 \in \mathscr{C}(Y)$ and show*

$$B_1 - B = B_1 Q - K_1 B \in \mathscr{C}(Y,X), \quad B_2 - B = P B_2 - B K_2 \in \mathscr{C}(Y,X).$$

Hence a parametrix is unique up to compact operators. Moreover, $B_1, B_2 \in \Phi(Y,X)$.

Problem* 6.33. *Suppose $A \in \Phi(X)$. If the kernel chain stabilizes then $\operatorname{ind}(A) \leq 0$. If the range chain stabilizes then $\operatorname{ind}(A) \geq 0$. So if $A \in \Phi_0(X)$, then the kernel chain stabilizes if and only if the range chain stabilizes.*

Problem 6.34. *The definition of a Fredholm operator can be literally extended to unbounded operators, however, in this case A is additionally assumed to be densely defined and closed. Let us denote A regarded as an operator from $\mathfrak{D}(A)$ equipped with the graph norm $\|.\|_A$ to X by $\hat{A}$. Recall that $\hat{A}$ is bounded (cf. Problem 4.12). Show that a densely defined closed operator A is Fredholm if and only if $\hat{A}$ is.*

Chapter 7

Operator semigroups

In this chapter we want to look at ordinary linear differential equations in Banach spaces. As a preparation we briefly discuss a few relevant facts about differentiation and integration for Banach space valued functions.

7.1. Analysis for Banach space valued functions

Let X be a Banach space. Let $I \subseteq \mathbb{R}$ be some interval and denote by $C(I, X)$ the set of continuous functions from I to X. Given $t \in I$ we call $f : I \to X$ differentiable at t if the limit

$$\dot{f}(t) := \lim_{\varepsilon \to 0} \frac{f(t+\varepsilon) - f(t)}{\varepsilon} \tag{7.1}$$

exists. The set of functions $f : I \to X$ which are differentiable at all $t \in I$ and for which $\dot{f} \in C(I, X)$ is denoted by $C^1(I, X)$. Clearly $C^1(I, X) \subset C(I, X)$. As usual we set $C^{k+1}(I, X) := \{f \in C^1(I, x) | \dot{f} \in C^k(I, X)\}$. Note that if $U \in \mathscr{L}(X, Y)$ and $f \in C^k(I, X)$, then $Uf \in C^k(I, Y)$ and $\frac{d}{dt} Uf = U\dot{f}$.

The following version of the mean value theorem will be crucial.

Theorem 7.1 (Mean value theorem)**.** *Suppose $f(t) \in C^1(I, X)$. Then*

$$\|f(t) - f(s)\| \leq M|t - s|, \qquad M := \sup_{\tau \in [s,t]} \|\dot{f}(\tau)\|, \tag{7.2}$$

for $s \leq t \in I$.

Proof. Fix $\tilde{M} > M$ and consider $d(\tau) := \|f(\tau) - f(s)\| - \tilde{M}(\tau - s)$ for $\tau \in [s, t]$. Suppose τ_0 is the largest τ for which $d(\tau) \leq 0$ holds. So there is

a sequence $\varepsilon_n \downarrow 0$ such that for sufficiently large n

$$\begin{aligned} 0 < d(\tau_0 + \varepsilon_n) &= \|f(\tau_0 + \varepsilon_n) - f(\tau_0) + f(\tau_0) - f(s)\| - \tilde{M}(\tau_0 + \varepsilon_n - s) \\ &\le \|f(\tau_0 + \varepsilon_n) - f(\tau_0)\| - \tilde{M}\varepsilon_n = \|\dot{f}(\tau)\varepsilon_n + o(\varepsilon_n)\| - \tilde{M}\varepsilon_n \\ &\le (M - \tilde{M})\varepsilon_n + o(\varepsilon_n) < 0. \end{aligned}$$

This contradicts our assumption. □

In particular,

Corollary 7.2. *For $f \in C^1(I, X)$ we have $\dot{f} = 0$ if and only if f is constant.*

Next we turn to integration. Let $I := [a, b]$ be compact. A function $f : I \to X$ is called a **step function** provided there are numbers

$$t_0 = a < t_1 < t_2 < \cdots < t_{n-1} < t_n = b \tag{7.3}$$

such that $f(t)$ is constant on each of the open intervals (t_{i-1}, t_i). The set of all step functions $S(I, X)$ forms a linear space and can be equipped with the sup norm. The corresponding Banach space obtained after completion is called the set of **regulated functions** $R(I, X)$. In other words, a regulated function is the uniform limit of a step function.

Observe that $C(I, X) \subset R(I, X)$. In fact, consider the functions $f_n := \sum_{j=0}^{n-1} f(t_j)\chi_{[t_j, t_{j+1})} \in S(I, X)$, where $t_j := a + j\frac{b-a}{n}$ and χ is the characteristic function. Since $f \in C(I, X)$ is uniformly continuous, we infer that f_n converges uniformly to f.

For a step function $f \in S(I, X)$ we can define a linear map $\int : S(I, X) \to X$ by

$$\int_I f(t)dt := \sum_{i=1}^{n} x_i(t_i - t_{i-1}), \tag{7.4}$$

where x_i is the value of f on (t_{i-1}, t_i). This map satisfies

$$\left\| \int_I f(t)dt \right\| \le (b - a)\|f\|_\infty \tag{7.5}$$

and hence it can be extended uniquely to a bounded linear map $\int : R(I, X) \to X$ with the same norm $(b - a)$ by Theorem 1.17. Of course if $X = \mathbb{C}$ this coincides with the usual Riemann integral. We even have

$$\left\| \int_I f(t)dt \right\| \le \int_I \|f(t)\|dt. \tag{7.6}$$

In fact, by the triangle inequality this holds for step functions and thus extends to all regulated functions by continuity.

In addition, if $A \in \mathscr{L}(X,Y)$, then $f \in R(I,X)$ implies $Af \in R(I,Y)$ and

$$A \int_I f(t)dt = \int_I A\, f(t)dt. \tag{7.7}$$

Again this holds for step functions and thus extends to all regulated functions by continuity. In particular, if $\ell \in X^*$ is a continuous linear functional, then

$$\ell\left(\int_I f(t)dt\right) = \int_I \ell\big(f(t)\big)dt, \qquad f \in R(I,X). \tag{7.8}$$

Moreover, we will use the usual conventions $\int_{t_1}^{t_2} f(s)ds := \int_I \chi_{(t_1,t_2)}(s)f(s)ds$ and $\int_{t_2}^{t_1} f(s)ds := -\int_{t_1}^{t_2} f(s)ds$. Note that we could replace (t_1,t_2) by any other interval with the same endpoints (why?) and hence $\int_{t_1}^{t_3} f(s)ds = \int_{t_1}^{t_2} f(s)ds + \int_{t_2}^{t_3} f(s)ds$.

Theorem 7.3 (fundamental theorem of calculus)**.** *If $f \in C(I,X)$, then $F(t) := \int_a^t f(s)ds \in C^1(I,X)$ and $\dot F(t) = f(t)$. Conversely, for any $F \in C^1(I,X)$ we have*

$$F(t) = F(a) + \int_a^t \dot F(s)ds. \tag{7.9}$$

Proof. The first part can be seen from

$$\begin{aligned}\|\int_a^{t+\varepsilon} f(s)ds - \int_a^t f(s)ds - f(t)\varepsilon\| &= \|\int_t^{t+\varepsilon} (f(s)-f(t))ds\| \\ &\le |\varepsilon| \sup_{s\in[t,t+\varepsilon]} \|f(s)-f(t)\|.\end{aligned}$$

The second follows from the first part which implies $\frac{d}{dt}\big(F(t)-\int_a^t \dot F(s)ds\big) = 0$. Hence this difference is constant and equals its value at $t=a$. □

Problem* 7.1 (Product rule)**.** *Let X be a Banach algebra. Show that if $f,g \in C^1(I,X)$ then $fg \in C^1(I,X)$ and $\frac{d}{dt}fg = \dot f g + f\dot g$.*

Problem* 7.2. *Let $f \in R(I,X)$ and $\tilde I := I + t_0$. then $f(t-t_0) \in R(\tilde I, X)$ and*

$$\int_I f(t)dt = \int_{\tilde I} f(t-t_0)dt.$$

Problem* 7.3. *Let $A : \mathfrak{D}(A) \subseteq X \to X$ be a closed operator. Show that (7.7) holds for $f \in C(I,X)$ with $\operatorname{Ran}(f) \subseteq \mathfrak{D}(A)$ and $Af \in C(I,X)$.*

7.2. Uniformly continuous operator groups

Now we are ready to apply these ideas to the abstract Cauchy problem

$$\dot u = Au, \qquad u(0) = u_0 \tag{7.10}$$

in some Banach space X. Here A is some linear operator and we will assume that $A \in \mathscr{L}(X)$ to begin with. Note that in the simplest case $X = \mathbb{R}^n$ this is simply a linear first order system with constant coefficient matrix A. In this case the solution is given by

$$u(t) = T(t)u_0, \tag{7.11}$$

where

$$T(t) := \exp(tA) := \sum_{j=0}^{\infty} \frac{t^j}{j!} A^j \tag{7.12}$$

is the exponential of tA. It is not difficult to see that this also gives the solution in our Banach space setting.

Theorem 7.4. *Let $A \in \mathscr{L}(X)$. Then the series in* (7.12) *converges and defines a* ***uniformly continuous operator group****:*

(i) *The map $t \mapsto T(t)$ is continuous, $T \in C(\mathbb{R}, \mathscr{L}(X))$.*

(ii) *$T(0) = \mathbb{I}$ and $T(t+s) = T(t)T(s)$ for all $t, s \in \mathbb{R}$.*

Moreover, $T \in C^1(\mathbb{R}, \mathscr{L}(X))$ is the unique solution of $\dot{T}(t) = AT(t)$ with $T(0) = \mathbb{I}$ and it commutes with A, $AT(t) = T(t)A$.

Proof. Set

$$T_n(t) := \sum_{j=0}^{n} \frac{t^j}{j!} A^j.$$

Then (for $m \le n$)

$$\|T_n(t) - T_m(t)\| = \left\| \sum_{j=m+1}^{n} \frac{t^j}{j!} A^j \right\| \le \sum_{j=m+1}^{n} \frac{|t|^j}{j!} \|A\|^j \le \frac{|t|^{m+1}}{(m+1)!} \|A\|^{m+1} \mathrm{e}^{|t|\|A\|}.$$

In particular,

$$\|T(t)\| \le \mathrm{e}^{|t|\,\|A\|}$$

and $AT(t) = \lim_{n\to\infty} AT_n(t) = \lim_{n\to\infty} T_n(t)A = T(t)A$. Furthermore we have $\dot{T}_{n+1} = AT_n$ and thus

$$T_{n+1}(t) = \mathbb{I} + \int_0^t AT_n(s)ds.$$

Taking limits shows

$$T(t) = \mathbb{I} + \int_0^t AT(s)ds$$

or equivalently $T(t) \in C^1(\mathbb{R}, \mathscr{L}(X))$ and $\dot{T}(t) = AT(t)$, $T(0) = \mathbb{I}$.

Suppose $S(t)$ is another solution, $\dot{S} = AS$, $S(0) = \mathbb{I}$. Then, by the product rule (Problem 7.1), $\frac{d}{dt} T(-t)S(t) = T(-t)AS(t) - AT(-t)S(t) = 0$ implying $T(-t)S(t) = T(0)S(0) = \mathbb{I}$. In the special case $T = S$ this shows $T(-t) = T^{-1}(t)$ and in the general case it hence proves uniqueness $S = T$.

Finally, $T(t+s)$ and $T(t)T(s)$ both satisfy our differential equation and coincide at $t=0$. Hence they coincide for all t by uniqueness. □

Clearly A is uniquely determined by $T(t)$ via $A=\dot{T}(0)$. Moreover, from this we also easily get uniqueness for our original Cauchy problem. We will in fact be slightly more general and consider the inhomogeneous problem

$$\dot{u} = Au + g, \qquad u(0) = u_0, \tag{7.13}$$

where $g \in C(I,X)$. A solution necessarily satisfies

$$\frac{d}{dt}T(-t)u(t) = -AT(-t)u(t) + T(-t)\dot{u}(t) = T(-t)g(t)$$

and integrating this equation (fundamental theorem of calculus) shows the **Duhamel formula**

$$u(t) = T(t)\left(u_0 + \int_0^t T(-s)g(s)ds\right) = T(t)u_0 + \int_0^t T(t-s)g(s)ds. \tag{7.14}$$

Using Problem 7.4 it is straightforward to verify that this is indeed a solution for any given $g \in C(I,X)$.

Lemma 7.5. *Let $A \in \mathscr{L}(X)$ and $g \in C(I,X)$. Then* (7.13) *has a unique solution given by* (7.14).

Example 7.1. For example look at the discrete linear wave equation

$$\ddot{q}_n(t) = k\big(q_{n+1}(t) - 2q_n(t) + q_{n-1}(t)\big), \qquad n \in \mathbb{Z}.$$

Factorizing this equation according to

$$\dot{q}_n(t) = p_n(t), \qquad \dot{p}_n(t) = k\big(q_{n+1}(t) - 2q_n(t) + q_{n-1}(t)\big),$$

we can write this as a first order system

$$\frac{d}{dt}\begin{pmatrix} q_n \\ p_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ k\,A_0 & 0 \end{pmatrix}\begin{pmatrix} q_n \\ p_n \end{pmatrix}$$

with the Jacobi operator $A_0 q_n = q_{n+1} - 2q_n + q_{n-1}$. Since A_0 is a bounded operator on $X = \ell^p(\mathbb{Z})$ we obtain a well-defined uniformly continuous operator group in $\ell^p(\mathbb{Z}) \oplus \ell^p(\mathbb{Z})$. ◇

Problem* 7.4 (Product rule)**.** *Suppose $f \in C^1(I,X)$ and $T \in C^1(I,\mathscr{L}(X))$. Show that $Tf \in C^1(I,X)$ and $\frac{d}{dt}Tf = \dot{T}f + T\dot{f}$.*

7.3. Strongly continuous semigroups

In the previous section we have found a quite complete solution of the abstract Cauchy problem (7.13) in the case when A is bounded. However, since differential operators are typically unbounded, this assumption is too strong for applications to partial differential equations. Since it is unclear what the conditions on A should be, we will go the other way and impose

conditions on T. First of all, even rather simple equations like the heat equation are only solvable for positive times and hence we will only assume that the solutions give rise to a semigroup. Moreover, continuity in the operator topology is too much to ask for (in fact, it is equivalent to boundedness of A — Problem 7.5) and hence we go for the next best option, namely strong continuity. In this sense, our problem is still well-posed.

A **strongly continuous operator semigroup** (also C_0-semigoup) is a family of operators $T(t) \in \mathscr{L}(X)$, $t \geq 0$, such that

(i) $T(t)g \in C([0,\infty), X)$ for every $g \in X$ (strong continuity) and
(ii) $T(0) = \mathbb{I}$, $T(t+s) = T(t)T(s)$ for every $t, s \geq 0$ (semigroup property).

If item (ii) holds for all $t, s \in \mathbb{R}$ it is called a **strongly continuous operator group**.

We first note that $\|T(t)\|$ is uniformly bounded on compact time intervals.

Lemma 7.6. *Let $T(t)$ be a C_0-semigroup. Then there are constants $M \geq 1$, $\omega \geq 0$ such that*

$$\|T(t)\| \leq M\mathrm{e}^{\omega t}, \qquad t \geq 0. \tag{7.15}$$

In case of a C_0-group we have $\|T(t)\| \leq M\mathrm{e}^{\omega|t|}$, $t \in \mathbb{R}$.

Proof. Since $\|T(.)g\| \in C[0,1]$ for every $g \in X$ we have $\sup_{t\in[0,1]} \|T(t)g\| \leq M_g$. Hence by the uniform boundedness principle $\sup_{t\in[0,1]} \|T(t)\| \leq M$ for some $M \geq 1$. Setting $\omega = \log(M)$ the claim follows by induction using the semigroup property. For the group case apply the semigroup case to both $T(t)$ and $S(t) := T(-t)$. □

Inspired by the previous section we define the **generator** A of a strongly continuous semigroup as the linear operator

$$Af := \lim_{t\downarrow 0} \frac{1}{t}\big(T(t)f - f\big), \tag{7.16}$$

where the domain $\mathfrak{D}(A)$ is precisely the set of all $f \in X$ for which the above limit exists. Moreover, a C_0-semigroup is the solution of the abstract Cauchy problem associated with its generator A:

Lemma 7.7. *Let $T(t)$ be a C_0-semigroup with generator A. If $f \in \mathfrak{D}(A)$ then $T(t)f \in \mathfrak{D}(A)$ and $AT(t)f = T(t)Af$. Moreover, suppose $g \in X$ with $u(t) = T(t)g \in \mathfrak{D}(A)$ for $t > 0$. Then $u(t) \in C^1((0,\infty), X) \cap C([0,\infty), X)$ and $u(t)$ is the unique solution of the abstract Cauchy problem*

$$\dot{u}(t) = Au(t), \qquad u(0) = g. \tag{7.17}$$

This is, for example, the case if $g \in \mathfrak{D}(A)$ in which case we even have $u(t) \in C^1([0,\infty), X)$.

Similarly, if $T(t)$ is a C_0-group and $g \in \mathfrak{D}(A)$, then $u(t) := T(t)g \in C^1(\mathbb{R}, X)$ is the unique solution of (7.17) for all $t \in \mathbb{R}$.

Proof. Let $f \in \mathfrak{D}(A)$ and $t > 0$ (respectively $t \in \mathbb{R}$ for a group), then

$$\lim_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\big(u(t+\varepsilon) - u(t)\big) = \lim_{\varepsilon\downarrow 0} T(t)\frac{1}{\varepsilon}\big(T(\varepsilon)f - f\big) = T(t)Af.$$

This shows the first part. To show that $u(t)$ is differentiable it remains to compute

$$\begin{aligned}\lim_{\varepsilon\downarrow 0} \frac{1}{-\varepsilon}\big(u(t-\varepsilon) - u(t)\big) &= \lim_{\varepsilon\downarrow 0} T(t-\varepsilon)\frac{1}{\varepsilon}\big(T(\varepsilon)f - f\big)\\ &= \lim_{\varepsilon\downarrow 0} T(t-\varepsilon)\big(Af + o(1)\big) = T(t)Af\end{aligned}$$

since $\|T(t)\|$ is bounded on compact t intervals. Hence $u(t) \in C^1([0,\infty), X)$ (respectively $u(t) \in C^1(\mathbb{R}, X)$ for a group) solves (7.17). In the general case $f = T(t_0)g \in \mathfrak{D}(A)$ and $u(t) = T(t)g = T(t-t_0)f$ solves our differential equation for every $t > t_0$. Since $t_0 > 0$ is arbitrary it follows that $u(t)$ solves (7.17) by the first part. To see that it is the only solution, let $v(t)$ be a solution corresponding to the initial condition $v(0) = 0$. For $s \le t$ we have

$$\begin{aligned}\frac{d}{ds}T(t-s)v(s) =& \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\big(T(t-s-\varepsilon)v(s+\varepsilon) - T(t-s)v(s)\big)\\ =& \lim_{\varepsilon\to 0} T(t-s-\varepsilon)\frac{1}{\varepsilon}\big(v(s+\varepsilon) - v(s)\big)\\ &- T(t-s)\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\big(T(\varepsilon)v(s) - v(s)\big)\\ =& T(t-s)Av(s) - T(t-s)Av(s) = 0.\end{aligned}$$

Whence, $v(t) = T(t-t)v(t) = T(t-s)v(s) = T(t)v(0) = 0$. □

Before turning to some examples, we establish a useful criterion for a semigroup to be strongly continuous.

Lemma 7.8. *A (semi)group of bounded operators is strongly continuous if and only if $\limsup_{\varepsilon\downarrow 0}\|T(\varepsilon)g\| < \infty$ for every $g \in X$ and $\lim_{\varepsilon\downarrow 0} T(\varepsilon)f = f$ for f in a dense subset.*

Proof. We first show that $\limsup_{\varepsilon\downarrow 0}\|T(\varepsilon)g\| < \infty$ for every $g \in X$ implies that $T(t)$ is bounded in a small interval $[0,\delta]$. Otherwise there would exists a sequence $\varepsilon_n \downarrow 0$ with $\|T(\varepsilon_n)\| \to \infty$. Hence $\|T(\varepsilon_n)g\| \to \infty$ for some g by the uniform boundedness principle, a contradiction. Thus there exists some M such that $\sup_{t\in[0,\delta]}\|T(t)\| \le M$. Setting $\omega = \frac{\log(M)}{\delta}$ we even obtain

(7.15). Moreover, boundedness of $T(t)$ shows that that $\lim_{\varepsilon\downarrow 0} T(\varepsilon)f = f$ for all $f \in X$ by a simple approximation argument (Lemma 4.32 (iv)).

In case of a group this also shows $\|T(-t)\| \le \|T(\delta - t)\|\|T(-\delta)\| \le M\|T(-\delta)\|$ for $0 \le t \le \delta$. Choosing $\tilde{M} = \max(M, M\|T(-\delta)\|)$ we conclude $\|T(t)\| \le \tilde{M}\exp(\tilde{\omega}|t|)$.

Finally, right continuity is immediate from the semigroup property: $\lim_{\varepsilon\downarrow 0} T(t+\varepsilon)g = T(\varepsilon)T(t)g = T(t)g$. Left continuity follows from $\|T(t-\varepsilon)g - T(t)g\| = \|T(t-\varepsilon)(T(\varepsilon)g - g)\| \le \|T(t-\varepsilon)\|\|T(\varepsilon)g - g\|$. □

Example 7.2. Let $X := C_0(\mathbb{R})$ be the continuous functions vanishing as $|x| \to \infty$. Then it is straightforward to check that

$$(T(t)f)(x) := f(x+t)$$

defines a group of continuous operators on X. Since shifting a function does not alter its supremum we have $\|T(t)f\|_\infty = \|f\|_\infty$ and hence $\|T(t)\| = 1$. Moreover, strong continuity is immediate for uniformly continuous functions. Since every function with compact support is uniformly continuous and since such functions are dense, we get that T is strongly continuous. Moreover, for $f \in \mathfrak{D}(A)$ we have

$$\lim_{\varepsilon\to 0} \frac{f(t+\varepsilon) - f(t)}{\varepsilon} = (Af)(t)$$

uniformly. In particular, $f \in C^1(\mathbb{R})$ with $f, f' \in C_0(\mathbb{R})$. Conversely, for $f \in C^1(\mathbb{R})$ with $f, f' \in C_0(\mathbb{R})$ we have

$$\frac{f(t+\varepsilon) - f(t) - \varepsilon f'(t)}{\varepsilon} = \frac{1}{\varepsilon}\int_0^\varepsilon \big(f'(t+s) - f'(t)\big)ds \le \sup_{0\le s\le\varepsilon} \|T(s)f' - f'\|_\infty$$

which converges to zero as $\varepsilon \downarrow 0$ by strong continuity of T. Whence

$$A = \frac{d}{dx}, \qquad \mathfrak{D}(A) = \{f \in C^1(\mathbb{R}) \cap C_0(\mathbb{R}) | f' \in C_0(\mathbb{R})\}.$$

It is not hard to see that T is not uniformly continuous or, equivalently, that A is not bounded (cf. Problem 7.5).

Note that this group is not strongly continuous when considered on $X := C_b(\mathbb{R})$. Indeed for $f(x) = \cos(x^2)$ we can choose $x_n = \sqrt{2\pi n}$ and $t_n = \sqrt{2\pi}(\sqrt{n+\frac{1}{4}} - \sqrt{n}) = \frac{1}{4}\sqrt{\frac{\pi}{2n}} + O(n^{-3/2})$ such that $\|T(t_n)f - f\|_\infty \ge |f(x_n + t_n) - f(x_n)| = 1$. ⋄

Next consider

$$u(t) = T(t)g, \qquad v(t) := \int_0^t u(s)ds, \qquad g \in X. \tag{7.18}$$

Then $v \in C^1([0,\infty), X)$ with $\dot v(t) = u(t)$ and (Problem 7.2)

$$\lim_{\varepsilon\downarrow 0} \frac{1}{\varepsilon}\big(T(\varepsilon)v(t) - v(t)\big) = \lim_{\varepsilon\downarrow 0} \Big(-\frac{1}{\varepsilon} v(\varepsilon) + \frac{1}{\varepsilon}\big(v(t+\varepsilon) - v(t)\big)\Big) = -g + u(t). \tag{7.19}$$

Consequently $v(t) \in \mathfrak{D}(A)$ and $Av(t) = -g + u(t)$ implying that $u(t)$ solves the following integral version of our abstract Cauchy problem

$$u(t) = g + A\int_0^t u(s)ds. \tag{7.20}$$

Note that while in the case of a bounded generator both versions are equivalent, this will not be the case in general. So while $u(t) = T(t)g$ always solves the integral version, it will only solve the differential version if $u(t) \in \mathfrak{D}(A)$ for $t > 0$ (which is clearly also necessary for the differential version to make sense).

Two further consequences of these considerations are also worth while noticing:

Corollary 7.9. *Let $T(t)$ be a C_0-semigroup with generator A. Then A is a densely defined and closed operator.*

Proof. Since $v(t) \in \mathfrak{D}(A)$ and $\lim_{t\downarrow 0} \frac{1}{t} v(t) = g$ for arbitrary g, we see that $\mathfrak{D}(A)$ is dense. Moreover, if $f_n \in \mathfrak{D}(A)$ and $f_n \to f$, $Af_n \to g$ then

$$T(t)f_n - f_n = \int_0^t T(s)Af_n ds.$$

Taking $n \to \infty$ and dividing by t we obtain

$$\frac{1}{t}\big(T(t)f - f\big) = \frac{1}{t}\int_0^t T(s)g\, ds.$$

Taking $t \downarrow 0$ finally shows $f \in \mathfrak{D}(A)$ and $Af = g$. □

Note that by the closed graph theorem we have $\mathfrak{D}(A) = X$ if and only if A is bounded. Moreover, since a C_0-semigroup provides the unique solution of the abstract Cauchy problem for A we obtain

Corollary 7.10. *A C_0-semigroup is uniquely determined by its generator.*

Proof. Suppose T and S have the same generator A. Then by uniqueness for (7.17) we have $T(t)g = S(t)g$ for all $g \in \mathfrak{D}(A)$. Since $\mathfrak{D}(A)$ is dense this implies $T(t) = S(t)$ as both operators are continuous. □

Finally, as in the uniformly continuous case, the inhomogeneous problem can be solved by Duhamel's formula. However, now it is not so clear when this will actually be a solution.

Lemma 7.11. *If the inhomogeneous problem*

$$\dot{u} = Au + f, \qquad u(0) = g, \tag{7.21}$$

has a solution it is necessarily given by ***Duhamel's formula***

$$u(t) = T(t)\left(g + \int_0^t T(-s)f(s)ds\right) = T(t)g + \int_0^t T(t-s)f(s)ds. \tag{7.22}$$

Conversely, this formula gives a solution if $g \in \mathfrak{D}(A)$ and $f \in C([0,\infty), X)$ with $\mathrm{Ran}(f) \subset \mathfrak{D}(A)$ and $Af \in C([0,\infty), X)$.

Proof. Clearly $T(t)g$ is a solution of the homogenous equation if $g \in \mathfrak{D}(A)$. Hence it remains to investigate the integral, which we will denote by $F(t)$. Then

$$\frac{1}{\varepsilon}(F(t+\varepsilon) - F(t)) = T(t+\varepsilon)\frac{1}{\varepsilon}\int_t^{t+\varepsilon} T(-s)f(s)ds - \frac{1}{\varepsilon}(T(\varepsilon) - \mathbb{I})F(t),$$

where the left-hand side converges to $T(t)T(-t)f(t) = f(t)$ and the right-hand side converges to $AF(t)$ since $F(t) \in \mathfrak{D}(A)$ by Problem 7.3. □

Problem* 7.5. *Show that a uniformly continuous semigroup has a bounded generator. (Hint: Write $T(t) = V(t_0)^{-1}V(t_0)T(t) = \dots$ with $V(t) := \int_0^t T(s)ds$ and conclude that it is C^1.)*

Problem 7.6. *Let $T(t)$ be a C_0-semigroup. Show that if $T(t_0)$ has a bounded inverse for one $t_0 > 0$ then it extends to a strongly continuous group.*

Problem 7.7. *Define a semigroup on $L^1(-1,1)$ via*

$$(T(t)f)(s) = \begin{cases} 2f(s-t), & 0 < s \le t, \\ f(s-t), & \text{else}, \end{cases}$$

where we set $f(s) = 0$ for $s < 0$. Show that the estimate from Lemma 7.6 does not hold with $M < 2$.

7.4. Generator theorems

Of course in practice the abstract Cauchy problem, that is the operator A, is given and the question is if A generates a corresponding C_0-semigroup. Corollary 7.9 already gives us some necessary conditions but this alone is not enough.

It tuns out that it is crucial to understand the resolvent of A. As in the case of bounded operators (cf. Section 6.1) we define the **resolvent set** via

$$\rho(A) := \{z \in \mathbb{C} | A - z \text{ is bijective with a bounded inverse}\} \tag{7.23}$$

and call

$$R_A(z) := (A - z)^{-1}, \qquad z \in \rho(A) \tag{7.24}$$

the **resolvent** of A. The complement $\sigma(A) = \mathbb{C} \setminus \rho(A)$ is called the **spectrum** of A. As in the case of Banach algebras it follows that the resolvent is analytic and that the resolvent set is open (Problem 7.9). However, the spectrum will no longer be bounded in general and both $\sigma(A) = \emptyset$ as well as $\sigma(A) = \mathbb{C}$ are possible (cf. the example in Problem 7.11). Note that if A is closed, then bijectivity implies boundedness of the inverse (see Corollary 4.9). Moreover, by Lemma 4.8 an operator with nonempty resolvent set must be closed.

Using an operator-valued version of the elementary integral $\int_0^\infty \mathrm{e}^{t(a-z)}dt = -(a-z)^{-1}$ (for $\mathrm{Re}(a-z) < 0$) we can make the connection between the resolvent and the semigroup.

Lemma 7.12. *Let T be a C_0-semigroup with generator A satisfying* (7.15). *Then $\{z|\mathrm{Re}(z) > \omega\} \subseteq \rho(A)$ and*

$$R_A(z) = -\int_0^\infty \mathrm{e}^{-zt}T(t)dt, \quad \mathrm{Re}(z) > \omega, \tag{7.25}$$

were the right-hand side is defined as

$$\left(\int_0^\infty \mathrm{e}^{-zt}T(t)dt\right) f := \lim_{s\to\infty} \int_0^s \mathrm{e}^{-zt}T(t)f\,dt. \tag{7.26}$$

Moreover,

$$\|R_A(z)\| \le \frac{M}{\mathrm{Re}(z)-\omega}, \quad \mathrm{Re}(z) > \omega. \tag{7.27}$$

Proof. Let us abbreviate $R_s(z)f := -\int_0^s \mathrm{e}^{-zt}T(t)f\,dt$. Then, by virtue of (7.15), $\|\mathrm{e}^{-zt}T(t)f\| \le M\mathrm{e}^{\omega-\mathrm{Re}(z)t}\|f\|$ shows that $R_s(z)$ is a bounded operator satisfying $\|R_s(z)\| \le M(\mathrm{Re}(z)-\omega)^{-1}$. Moreover, this estimates also shows that the limit $R(z) := \lim_{s\to\infty} R_s(z)$ exists (and still satisfies $\|R(z)\| \le M(\mathrm{Re}(z)-\omega)^{-1}$). Next note that $S(t) = \mathrm{e}^{-zt}T(t)$ is a semigroup with generator $A - z$ (Problem 7.8) and hence for $f \in \mathfrak{D}(A)$ we have

$$R_s(z)(A-z)f = -\int_0^s S(t)(A-z)f\,dt = -\int_0^s \dot{S}(t)f\,dt = f - S(s)f.$$

In particular, taking the limit $s \to \infty$, we obtain $R(z)(A-z)f = f$ for $f \in \mathfrak{D}(A)$. Similarly, still for $f \in \mathfrak{D}(A)$, by Problem 7.3

$$(A-z)R_s(z)f = -\int_0^s (A-z)S(t)f\,dt = -\int_0^s \dot{S}(t)f\,dt = f - S(s)f$$

and taking limits, using closedness of A, implies $(A-z)R(z)f = f$ for $f \in \mathfrak{D}(A)$. Finally, if $f \in X$ choose $f_n \in \mathfrak{D}(A)$ with $f_n \to f$. Then $R(z)f_n \to R(z)f$ and $(A-z)R(z)f_n = f_n \to f$ proving $R(z)f \in \mathfrak{D}(A)$ and $(A-z)R(z)f = f$ for $f \in X$. □

Corollary 7.13. *Let T be a C_0-semigroup with generator A satisfying* (7.15). *Then*

$$R_A(z)^{n+1} = \frac{(-1)^{n+1}}{n!} \int_0^\infty t^n \mathrm{e}^{-zt} T(t) dt, \quad \mathrm{Re}(z) > \omega, \tag{7.28}$$

and

$$\|R_A(z)^n\| \le \frac{M}{(\mathrm{Re}(z) - \omega)^n}, \quad \mathrm{Re}(z) > \omega,\ n \in \mathbb{N}. \tag{7.29}$$

Proof. Abbreviate $R_n(z) := \int_0^\infty t^n \mathrm{e}^{-zt} T(t) dt$ and note that

$$\frac{R_n(z+\varepsilon) - R_n(z)}{\varepsilon} = -R_{n+1}(z) + \varepsilon \int_0^\infty t^{n+2} \phi(\varepsilon t) \mathrm{e}^{-zt} T(t) dt$$

where $|\phi(\varepsilon)| \le \frac{1}{2}\mathrm{e}^{|\varepsilon|}$ from which we see $\frac{d}{dz} R_n(z) = -R_{n+1}(z)$ and hence $\frac{d^n}{dz^n} R_A(z) = -\frac{d^n}{dz^n} R_0(z) = (-1)^{n+1} R_n(z)$. Now the first claim follows using $R_A(z)^{n+1} = \frac{1}{n!} \frac{d^n}{dz^n} R_A(z)$ (Problem 7.10). Estimating the integral using (7.15) establishes the second claim. □

Given these preparations we can now try to answer the question when A generates a semigroup. In fact, we will be constructive and obtain the corresponding semigroup by approximation. To this end we introduce the **Yosida approximation**

$$A_n := -nAR_A(\omega + n) = -n - n(\omega + n) R_A(\omega + n) \in \mathscr{L}(X). \tag{7.30}$$

Of course this is motivated by the fact that this is a valid approximation for numbers $\lim_{n\to\infty} \frac{-n}{a-\omega-n} = 1$. That we also get a valid approximation for operators is the content of the next lemma.

Lemma 7.14. *Suppose A is a densely defined closed operator with $(\omega, \infty) \subset \rho(A)$ satisfying*

$$\|R_A(\omega + n)\| \le \frac{M}{n}. \tag{7.31}$$

Then

$$\lim_{n\to\infty} -nR_A(\omega + n) f = f, \quad f \in X, \qquad \lim_{n\to\infty} A_n f = Af, \quad f \in \mathfrak{D}(A). \tag{7.32}$$

Proof. If $f \in \mathfrak{D}(A)$ we have $-nR_A(\omega + n) f = f - R_A(\omega + n)(A - \omega) f$ which shows $-nR_A(\omega + n) f \to f$ if $f \in \mathfrak{D}(A)$. Since $\mathfrak{D}(A)$ is dense and $\|nR_A(\omega + n)\| \le M$ this even holds for all $f \in X$. Moreover, for $f \in \mathfrak{D}(A)$ we have $A_n f = -nAR_A(\omega + n) f = -nR_A(\omega + n)(Af) \to Af$ by the first part. □

Moreover, A_n can also be used to approximate the corresponding semigroup under suitable assumptions.

Theorem 7.15 (Feller–Miyadera–Phillips)**.** *A linear operator A is the generator of a C_0-semigroup satisfying* (7.15) *if and only if it is densely defined, closed, $(\omega,\infty)\subseteq\rho(A)$, and*

$$\|R_A(\lambda)^n\|\le\frac{M}{(\lambda-\omega)^n},\quad \lambda>\omega,\ n\in\mathbb{N}. \tag{7.33}$$

Proof. Necessity has already been established in Corollaries 7.9 and 7.13.

For the converse we use the semigroups

$$T_n(t):=\exp(tA_n)$$

corresponding to the Yosida approximation (7.30). We note (using $\mathrm{e}^{A+B}=\mathrm{e}^A\mathrm{e}^B$ for commuting operators A, B)

$$\|T_n(t)\|\le\mathrm{e}^{-tn}\sum_{j=0}^{\infty}\frac{(tn(\omega+n))^j}{j!}\|R_A(\omega+n)^j\|\le M\mathrm{e}^{-tn}\mathrm{e}^{t(\omega+n)}=M\mathrm{e}^{\omega t}.$$

Moreover, since $R_A(\omega+m)$ and $R_A(\omega+n)$ commute by the first resolvent identity (Problem 7.10), we conclude that the same is true for A_m, A_n as well as for $T_m(t)$, $T_n(t)$ (by the very definition as a power series). Consequently

$$\begin{aligned}\|T_n(t)f-T_m(t)f\|&=\left\|\int_0^1\frac{d}{ds}T_n(st)T_m((1-s)t)f\,ds\right\|\\&\le t\int_0^1\|T_n(st)T_m((1-s)t)(A_n-A_m)f\|ds\\&\le tM^2\mathrm{e}^{\omega t}\|(A_n-A_m)f\|.\end{aligned}$$

Thus, from $f\in\mathfrak{D}(A)$ we have a Cauchy sequence and can define a linear operator by $T(t)f:=\lim_{n\to\infty}T_n(t)f$. Since $\|T(t)f\|=\lim_{n\to\infty}\|T_n(t)f\|\le M\mathrm{e}^{\omega t}\|f\|$, we see that $T(t)$ is bounded and has a unique extension to all of X. Moreover, $T(0)=\mathbb{I}$ and

$$\begin{aligned}&\|T_n(t)T_n(s)f-T(t)T(s)f\|\le\\&\quad M\mathrm{e}^{\omega t}\|T_n(s)f-T(s)f\|+\|T_n(t)T(s)f-T(t)T(s)f\|\end{aligned}$$

implies $T(t+s)f=\lim_{n\to\infty}T_n(t+s)f=\lim_{n\to\infty}T_n(t)T_n(s)f=T(t)T(s)f$, that is, the semigroup property holds. Finally, by

$$\begin{aligned}\|T(\varepsilon)f-f\|&\le\|T(\varepsilon)f-T_n(\varepsilon)f\|+\|T_n(\varepsilon)f-f\|\\&\le tM^2\mathrm{e}^{\omega t}\|(A_n-A)f\|+\|T_n(\varepsilon)f-f\|\end{aligned}$$

we see $\lim_{\varepsilon\downarrow0}T(\varepsilon)f=f$ for $f\in\mathfrak{D}(A)$ and Lemma 7.8 shows that T is a C_0-semigroup. It remains to show that A is its generator. To this end let

$f \in \mathfrak{D}(A)$, then

$$\begin{aligned} T(t)f - f &= \lim_{n\to\infty} T_n(t)f - f = \lim_{n\to\infty} \int_0^t T_n(s)A_n f\, ds \\ &= \lim_{n\to\infty} \left(\int_0^t T_n(s)Af\, ds + \int_0^t T_n(s)(A_n - A)f\, ds \right) \\ &= \int_0^t T(s)Af\, ds \end{aligned}$$

which shows $\lim_{t\downarrow 0} \frac{1}{t}(T(t)f - f) = Af$ for $f \in \mathfrak{D}(A)$. Finally, note that the domain of the generator cannot be larger, since $A - \omega - 1$ is bijective and adding a vector to its domain would destroy injectivity. But then $\omega + 1$ would not be in the resolvent set contradicting Lemma 7.12. □

Note that in combination with the following lemma this also answers the question when A generates a C_0-group.

Lemma 7.16. *An operator A generates a C_0-group if and only if both A and $-A$ generate C_0-semigroups.*

Proof. Clearly, if A generates a C_0-group $T(t)$, then $S(t) := T(-t)$ is a C_0-group with generator $-A$. Conversely, let $T(t)$, $S(t)$ be the C_0-semigroups generated by A, $-A$, respectively. Then a short calculation shows

$$\frac{d}{dt} T(t)S(t)g = -T(t)AS(t)g + T(t)AS(t)g = 0, \qquad t \geq 0.$$

Consequently, $T(t)S(t) = T(0)S(0) = \mathbb{I}$ and similarly $S(t)T(t) = \mathbb{I}$, that is, $S(t) = T(t)^{-1}$. Hence it is straightforward to check that T extends to a group via $T(-t) := S(t)$, $t \geq 0$. □

The following examples show that the spectral conditions are indeed crucial. Moreover, they also show that an operator might give rise to a Cauchy problem which is uniquely solvable for a dense set of initial conditions, without generating a strongly continuous semigroup.

Example 7.3. Let

$$A = \begin{pmatrix} 0 & A_0 \\ 0 & 0 \end{pmatrix}, \qquad \mathfrak{D}(A) = X \times \mathfrak{D}(A_0).$$

Then $u(t) = \left(\begin{smallmatrix} 1 & tA_0 \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} f_0 \\ f_1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} f_0 + tA_0 f_1 \\ f_1 \end{smallmatrix}\right)$ is the unique solution of the corresponding abstract Cauchy problem for given $f \in \mathfrak{D}(A)$. Nevertheless, if A_0 is unbounded, the corresponding semigroup is not strongly continuous.

Note that in this case we have $\sigma(A) = \{0\}$ if A_0 is bounded and $\sigma(A) = \mathbb{C}$ else. In fact, since A is not injective we must have $\{0\} \subseteq \sigma(A)$. For $z \neq 0$

the inverse of $A - z$ is given by

$$(A-z)^{-1} = -\frac{1}{z}\begin{pmatrix} 1 & \frac{1}{z}A_0 \\ 0 & 1 \end{pmatrix}, \quad \mathfrak{D}((A-z)^{-1}) = \operatorname{Ran}(A-z) = X \times \mathfrak{D}(A_0),$$

which is bounded if and only if A is bounded. ⋄

Example 7.4. Let $X_0 = C_0(\mathbb{R})$ and $m(x) = \mathrm{i}x$. Then we can regard m as a multiplication operator on X_0 when defined maximally, that is, $f \mapsto m\,f$ with $\mathfrak{D}(m) = \{f \in X_0 | m\,f \in X_0\}$. Note that since $C_c(\mathbb{R}) \subseteq \mathfrak{D}(m)$ we see that m is densely defined. Moreover, it is easy to check that m is closed.

Now consider $X = X_0 \oplus X_0$ with $\|f\| = \max(\|f_0\|, \|f_1\|)$ and note that

$$A = \begin{pmatrix} m & m \\ 0 & m \end{pmatrix}, \qquad \mathfrak{D}(A) = \mathfrak{D}(m) \oplus \mathfrak{D}(m),$$

is closed. Moreover, for $z \notin \mathrm{i}\mathbb{R}$ the resolvent is given by the multiplication operator

$$R_A(z) = -\frac{1}{m-z}\begin{pmatrix} 1 & -\frac{m}{m-z} \\ 0 & 1 \end{pmatrix}.$$

For $\lambda > 0$ we compute

$$\|R_A(\lambda)f\| \le \left(\sup_{x\in\mathbb{R}} \frac{1}{|\mathrm{i}x - \lambda|} + \sup_{x\in\mathbb{R}} \frac{|x|}{|\mathrm{i}x - \lambda|^2}\right)\|f\| = \frac{3}{2\lambda}\|f\|$$

and hence A satisfies (7.33) with $M = \frac{3}{2}$, $\omega = 0$ and $n = 1$. However, by

$$\|R_A(\lambda + \mathrm{i}n)\| \ge \|R_A(\lambda + \mathrm{i}n)(0, f_n)\| \ge \left|\frac{\mathrm{i}n f_n(n)}{(\lambda - \mathrm{i}n + \mathrm{i}n)^2}\right| = \frac{n}{\lambda^2},$$

where f_n is chosen such that $f_n(n) = 1$ and $\|f_n\|_\infty = 1$, it does not satisfy (7.29). Hence A does not generate a C_0-semigroup. Indeed, the solution of the corresponding Cauchy problem is

$$T(t) = \mathrm{e}^{t\,m}\begin{pmatrix} 1 & t\,m \\ 0 & 1 \end{pmatrix}, \qquad \mathfrak{D}(T) = X_0 \oplus \mathfrak{D}(m),$$

which is unbounded. ⋄

Finally we look at the special case of **contraction semigroups** satisfying

$$\|T(t)\| \le 1. \tag{7.34}$$

By a simple transform the case $M = 1$ in Lemma 7.6 can always be reduced to this case (Problem 7.8). Moreover, observe, that in the case $M = 1$ the estimate (7.27) implies the general estimate (7.29).

Corollary 7.17 (Hille–Yosida)**.** *A linear operator A is the generator of a contraction semigroup if and only if it is densely defined, closed, $(0, \infty) \subseteq \rho(A)$, and*

$$\|R_A(\lambda)\| \le \frac{1}{\lambda}, \quad \lambda > 0. \tag{7.35}$$

Example 7.5. If A is the generator of a contraction, then clearly all eigenvalues z must satisfy $\mathrm{Re}(z) \le 0$. Moreover, for

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

we have

$$R_A(z) = -\frac{1}{z}\begin{pmatrix} 1 & 1/z \\ 0 & 1 \end{pmatrix}, \qquad T(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix},$$

which shows that the bound on the resolvent is crucial. ⋄

However, for a given operator even the simple estimate (7.35) might be difficult to establish directly. Hence we outline another criterion.

Example 7.6. Let X be a Hilbert space and observe that for a contraction semigroup the expression $\|T(t)f\|$ must be nonincreasing. Consequently, for $f \in \mathfrak{D}(A)$ we must have

$$\frac{d}{dt}\|T(t)f\|^2\Big|_{t=0} = 2\mathrm{Re}\big(\langle f, Af\rangle\big) \le 0.$$

Operators satisfying $\mathrm{Re}(\langle f, Af\rangle) \le 0$ are called dissipative an this clearly suggests to replace the resolvent estimate by dissipativity. ⋄

To formulate this condition for Banach spaces, we first introduce the **duality set**

$$\mathcal{J}(x) := \{x' \in X^* | x'(x) = \|x\|^2 = \|x'\|^2\} \tag{7.36}$$

of a given vector $x \in X$. In other words, the elements from $\mathcal{J}(x)$ are those linear functionals which attain their norm at x and are normalized to have the same norm as x. As a consequence of the Hahn–Banach theorem (Corollary 4.15) note that $\mathcal{J}(x)$ is nonempty. Moreover, it is also easy to see that $\mathcal{J}(x)$ is convex and weak-$*$ closed.

Example 7.7. Let X be a Hilbert space and identify X with X^* via $x \mapsto \langle x, .\rangle$ as usual. Then $\mathcal{J}(x) = \{x\}$. Indeed since we have equality $\langle x', x\rangle = \|x'\|\|x\|$ in the Cauchy–Schwarz inequality, we must have $x' = \alpha x$ for some $\alpha \in \mathbb{C}$ with $|\alpha| = 1$ and $\alpha^*\|x\|^2 = \langle x', x\rangle = \|x\|^2$ shows $\alpha = 1$. ⋄

Example 7.8. If X^* is strictly convex (cf. Problem 1.12), then the duality set contains only one point. In fact, suppose $x', y' \in \mathcal{J}(x)$, then $z' = \frac{1}{2}(x' + y') \in \mathcal{J}(x)$ and $\frac{\|x\|}{2}\|x' + y'\| = z'(x) = \frac{\|x\|}{2}(\|x'\| + \|y'\|)$ implying $x' = y'$ by strict convexity.

This applies for example to $X := \ell^p(\mathbb{N})$ if $1 < p < \infty$ (cf. Problem 1.12) in which case $X^* \cong \ell^q(\mathbb{N})$ with $1 < q < \infty$. In fact, for $a \in \ell^p(\mathbb{N})$ we have $\mathcal{J}(a) = \{a'\}$ with $a'_j = \|a\|_p^{2-p}\,\mathrm{sign}(a_j^*)|a_j|^{p-1}$. ⋄

Example 7.9. Let $X = C[0,1]$ and choose $x \in X$. If t_0 is chosen such that $|x(t_0)| = \|x\|$, then the functional $y \mapsto x'(y) := x(t_0)^* y(t_0)$ satisfies $x' \in \mathcal{J}(x)$. Clearly $\mathcal{J}(x)$ will contain more than one element in general. ⋄

Now a given operator $\mathfrak{D}(A) \subseteq X \to X$ is called **dissipative** if

$$\mathrm{Re}\big(x'(Ax)\big) \le 0 \quad \text{for one } x' \in \mathcal{J}(x) \text{ and all } x \in \mathfrak{D}(A). \tag{7.37}$$

Lemma 7.18. *Let $x, y \in X$. Then $\|x\| \le \|x - \alpha y\|$ for all $\alpha > 0$ if and only if there is an $x' \in \mathcal{J}(x)$ such that $\mathrm{Re}(x'(y)) \le 0$.*

Proof. Without loss of generality we can assume $x \neq 0$. If $\mathrm{Re}(x'(y)) \le 0$ for some $x' \in \mathcal{J}(x)$, then for $\alpha > 0$ we have

$$\|x\|^2 = x'(x) \le \mathrm{Re}\big(x'(x - \alpha y)\big) \le \|x'\|\|x - \alpha y\|$$

implying $\|x\| \le \|x - \alpha y\|$.

Conversely, if $\|x\| \le \|x - \alpha y\|$ for all $\alpha > 0$, let $x'_\alpha \in \mathcal{J}(x - \alpha y)$ and set $y'_\alpha = \|x'_\alpha\|^{-1} x'_\alpha$. Then

$$\begin{aligned} \|x\| \le \|x - \alpha y\| &= y'_\alpha(x - \alpha y) = \mathrm{Re}\big(y'_\alpha(x)\big) - \alpha \mathrm{Re}\big(y'_\alpha(y)\big) \\ &\le \|x\| - \alpha \mathrm{Re}\big(y'_\alpha(y)\big). \end{aligned}$$

Now by the Banach–Alaoglu theorem we can choose a subsequence $y'_{1/n_j} \to y'_0$ in the weak-$*$ sense. (Note that the use of the Banach–Alaoglu theorem could be avoided by restricting y'_α to the two dimensional subspace spanned by x , y, passing to the limit in this subspace and then extending the limit to X^* using Hahn–Banach.) Consequently $\mathrm{Re}\big(y'_0(y)\big) \le 0$ and $y'_0(x) = \|x\|$. Whence $x'_0 = \|x\| y'_0 \in \mathcal{J}(x)$ and $\mathrm{Re}\big(x'_0(y)\big) \le 0$. □

As a straightforward consequence we obtain:

Corollary 7.19. *A linear operator is dissipative if and only if*

$$\|(A - \lambda)f\| \ge \lambda \|f\|, \qquad \lambda > 0,\ f \in \mathfrak{D}(A). \tag{7.38}$$

In particular, for a dissipative operator $A - \lambda$ is injective for $\lambda > 0$ and $(A-\lambda)^{-1}$ is bounded with $\|(A-\lambda)^{-1}\| \le \lambda^{-1}$. However, this does not imply that λ is in the resolvent set of A since $\mathfrak{D}((A - \lambda)^{-1}) = \mathrm{Ran}(A - \lambda)$ might not be all of X.

Now we are ready to show

Theorem 7.20 (Lumer–Phillips)**.** *A linear operator A is the generator of a contraction semigroup if and only if it is densely defined, dissipative, and $A - \lambda_0$ is surjective for one $\lambda_0 > 0$. Moreover, in this case* (7.37) *holds for all $x' \in \mathcal{J}(x)$.*

Proof. Let A generate a contraction semigroup $T(t)$ and let $x \in \mathfrak{D}(A)$, $x' \in \mathcal{J}(x)$. Then

$$\mathrm{Re}\big(x'(T(t)x - x)\big) \le |x'(T(t)x)| - \|x\|^2 \le \|x'\|\|x\| - \|x\|^2 = 0$$

and dividing by t and letting $t \downarrow 0$ shows $\operatorname{Re}\big(x'(Ax)\big) \leq 0$. Hence A is dissipative and by Corollary 7.17 $(0,\infty) \subseteq \rho(A)$, that is $A-\lambda$ is bijective for $\lambda > 0$.

Conversely, by Corollary 7.19 $A-\lambda$ has a bounded inverse satisfying $\|(A-\lambda)^{-1}\| \leq \lambda^{-1}$ for all $\lambda > 0$. In particular, for λ_0 the inverse is defined on all of X and hence closed. Thus A is also closed and $\lambda_0 \in \rho(A)$. Moreover, from $\|R_A(\lambda_0)\| \leq \lambda_0^{-1}$ (cf. Problem 7.9) we even get $(0, 2\lambda_0) \subseteq \rho(A)$ and iterating this argument shows $(0,\infty) \subseteq \rho(A)$ as well as $\|R_A(\lambda)\| \leq \lambda^{-1}$, $\lambda > 0$. Hence the requirements from Corollary 7.17 are satisfied. □

Note that generators of contraction semigroups are maximal dissipative in the sense that they do not have any dissipative extensions. In fact, if we extend A to a larger domain we must destroy injectivity of $A-\lambda$ and thus the extension cannot be dissipative.

Example 7.10. Let $X = C[0,1]$ and consider the one-dimensional heat equation

$$\frac{\partial}{\partial t}u(t,x) = \frac{\partial^2}{\partial x^2}u(t,x)$$

on a finite interval $x \in [0,1]$ with the boundary conditions $u(0) = u(1) = 0$ and the initial condition $u(0,x) = u_0(x)$. The corresponding operator is

$$Af = f'', \qquad \mathfrak{D}(A) = \{f \in C^2[0,1] | f(0) = f(1) = 0\} \subseteq C[0,1].$$

For $\ell \in \mathcal{J}(f)$ we can choose $\ell(g) = f(x_0)^* g(x_0)$ where x_0 is chosen such that $|f(x_0)| = \|f\|_\infty$. Then $\operatorname{Re}(f(x_0)^* f(x))$ has a global maximum at $x = x_0$ and if $f \in C^2[0,1]$ we must have $\operatorname{Re}(f(x_0)^* f''(x)) \leq 0$ provided this maximum is in the interior of $(0,1)$. Consequently $\operatorname{Re}(f(x_0)^* f''(x_0)) \leq 0$, that is, A is dissipative. That $A-\lambda$ is surjective follows using the Green's function as in Section 3.3. ⋄

Finally, we note that the condition that $A-\lambda_0$ is surjective can be weakened to the condition that $\operatorname{Ran}(A-\lambda_0)$ is dense. To this end we need:

Lemma 7.21. *Suppose A is a densely defined dissipative operator. Then A is closable and the closure $\overline{A}$ is again dissipative.*

Proof. Recall that A is closable if and only if for every $x_n \in \mathfrak{D}(A)$ with $x_n \to 0$ and $Ax_n \to y$ we have $y = 0$. So let x_n be such a sequence and chose another sequence $y_n \in \mathfrak{D}(A)$ such that $y_n \to y$ (which is possible since $\mathfrak{D}(A)$ is assumed dense). Then by dissipativity (specifically Corollary 7.19)

$$\|(A-\lambda)(\lambda x_n + y_m)\| \geq \lambda \|\lambda x_n + y_m\|, \quad \lambda > 0$$

and letting $n \to \infty$ and dividing by λ shows

$$\|y + (\lambda^{-1}A - 1)y_m\| \geq \|y_m\|.$$

Finally $\lambda \to \infty$ implies $\|y - y_m\| \geq \|y_m\|$ and $m \to \infty$ yields $0 \geq \|y\|$, that is, $y = 0$ and A is closable. To see that $\overline{A}$ is dissipative choose $x \in \mathfrak{D}(\overline{A})$ and $x_n \in \mathfrak{D}(A)$ with $x_n \to x$ and $Ax_n \to \overline{A}x$. Then (again using Corollary 7.19) taking the limit in $\|(A - \lambda)x_n\| \geq \lambda\|x_n\|$ shows $\|(\overline{A} - \lambda)x\| \geq \lambda\|x\|$ as required. □

Consequently:

Corollary 7.22. *Suppose the linear operator A is densely defined, dissipative, and $\operatorname{Ran}(A - \lambda_0)$ is dense for one $\lambda_0 > 0$. Then A is closable and $\overline{A}$ is the generator of a contraction semigroup.*

Proof. By the previous lemma A is closable with $\overline{A}$ again dissipative. In particular, $\overline{A}$ is injective and by Lemma 4.8 we have $(\overline{A}-\lambda_0)^{-1} = \overline{(A - \lambda_0)^{-1}}$. Since $(A - \lambda_0)^{-1}$ is bounded its closure is defined on the closure of its domain, that is, $\operatorname{Ran}(\overline{A} - \lambda_0) = \overline{\operatorname{Ran}(A - \lambda_0)} = X$. The rest follows from the Lumer–Phillips theorem. □

Problem* 7.8. *Let $T(t)$ be a C_0-semigroup and $\alpha > 0$, $\lambda \in \mathbb{C}$. Show that $S(t) := \mathrm{e}^{\lambda t}T(\alpha t)$ is a C_0-semigroup with generator $B = \alpha A + \lambda$, $\mathfrak{D}(B) = \mathfrak{D}(A)$.*

Problem* 7.9. *Let A be a closed operator. Show that if $z_0 \in \rho(A)$, then*

$$R_A(z) = \sum_{n=0}^{\infty} (z - z_0)^n R_A(z_0)^{n+1}, \quad |z - z_0| < \|R_A(z_0)\|^{-1}.$$

In particular, the resolvent is analytic and

$$\|(A - z)^{-1}\| \geq \frac{1}{\operatorname{dist}(z, \sigma(A))}.$$

Problem* 7.10. *Let A be a closed operator. Show the* ***first resolvent identity***

$$\begin{aligned} R_A(z_0) - R_A(z_1) &= (z_0 - z_1)R_A(z_0)R_A(z_1) \\ &= (z_0 - z_1)R_A(z_1)R_A(z_0), \end{aligned}$$

for $z_0, z_1 \in \rho(A)$. Moreover, conclude

$$\frac{d^n}{dz^n} R_A(z) = n! R_A(z)^{n+1}, \qquad \frac{d}{dz} R_A(z)^n = n R_A(z)^{n+1}.$$

Problem* 7.11. *Consider $X = C[0, 1]$ and $A = \frac{d}{dx}$ with $\mathfrak{D}(A) = C^1[0, 1]$. Compute $\sigma(A)$. Do the same for $A_0 = \frac{d}{dx}$ with $\mathfrak{D}(A) = \{x \in C^1[0, 1] | x(0) = 0\}$.*

Problem 7.12. *Suppose $z_0 \in \rho(A)$ (in particular A is closed; also note that $0 \in \sigma(R_A(z_0))$ if and only if A is unbounded). Show that for $z \neq 0$ we have $\sigma(A) = z_0 + (\sigma(R_A(z_0)) \setminus \{0\})^{-1}$ and $R_{R_A(z_0)}(z) = -\frac{1}{z} - \frac{1}{z^2} R_A(z_0 + \frac{1}{z})$ for*

$z \in \rho(R_A(z_0)) \setminus \{0\}$. *Moreover,* $\operatorname{Ker}(R_A(z_0) - z)^n = \operatorname{Ker}(A - z_0 - \frac{1}{z})^n$ *and* $\operatorname{Ran}(R_A(z_0) - z)^n = \operatorname{Ran}(A - z_0 - \frac{1}{z})^n$ *for every* $n \in \mathbb{N}_0$.

Problem 7.13. *Show that* $R_A(z) \in \mathscr{C}(X)$ *for one* $z \in \rho(A)$ *if and only this holds for all* $z \in \rho(A)$. *Moreover, in this case the spectrum of* A *consists only of discrete eigenvalues with finite (geometric and algebraic) multiplicity. (Hint: Use the previous problem to reduce it to Theorem 6.14.)*

Part 2

Real Analysis

Chapter 8

Measures

8.1. The problem of measuring sets

The Riemann integral starts straight with the definition of the integral by considering functions which can be sandwiched between step functions. This is based on the idea that for a function defined on an interval (or a rectangle in higher dimensions) the domain can be easily subdivided into smaller intervals (or rectangles). Moreover, for nice functions the variation of the values (difference between maximum and minimum) should decrease with the length of the intervals. Of course, this fails for rough functions whose variations cannot be controlled by subdividing the domain into sets of decreasing size. The Lebesgue integral remedies this by subdividing the range of the function. This shifts the problem from controlling the variations of the function to defining the content of the preimage of the subdivisions for the range. Note that this problem does not occur in the Riemann approach since only the length of an interval (or the area of an rectangle) is needed. Consequently, the outset of Lebesgue theory is the problem of defining the content for a sufficiently large class of sets.

The Riemann-style approach to this problem in $\mathbb{R}^n$ is to start with a big rectangle containing the set under consideration and then take subdivisions thereby approximating the measure of the set from the inside and outside by the measure of the rectangles which lie inside and those which cover the set, respectively. If the difference tends to zero, the set is called *measurable* and the common limit is it measure.

To this end let $\mathcal{S}^n$ be the set of all **half-closed rectangles** of the form $(a, b] := (a_1, b_1] \times \cdots \times (a_n, b_n] \subseteq \mathbb{R}^n$ with $a < b$ augmented by the empty set. Here $a < b$ should be read as $a_j < b_j$ for all $1 \leq j \leq n$ (and similarly for

$a \le b$). Moreover, we allow the intervals to be unbounded, that is $a, b \in \overline{\mathbb{R}}^n$ (with $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$).

Of course one could as well take open or closed rectangles but our half-closed rectangles have the advantage that they tile nicely. In particular, one can subdivide a half-closed rectangle into smaller ones without gaps or overlap.

The somewhat dissatisfactory situation with the Riemann integral alluded to before led Cantor, Peano, and particularly Jordan to the following attempt of measuring arbitrary sets: Define the measure of a rectangle via

$$|(a, b]| = \prod_{j=1}^{n} (b_j - a_j). \tag{8.1}$$

Note that the measure will be infinite if the rectangle is unbounded. Furthermore, define the **inner, outer Jordan content** of a set $A \subseteq \mathbb{R}^n$ as

$$J_*(A) := \sup \Big\{ \sum_{j=1}^{m} |R_j| \,\Big|\, \dot{\bigcup_{j=1}^{m}} R_j \subseteq A,\ R_j \in \mathcal{S}^n \Big\}, \tag{8.2}$$

$$J^*(A) := \inf \Big\{ \sum_{j=1}^{m} |R_j| \,\Big|\, A \subseteq \dot{\bigcup_{j=1}^{m}} R_j,\ R_j \in \mathcal{S}^n \Big\}, \tag{8.3}$$

respectively. Here the dot inside the union indicates that we only allow unions of mutually disjoint sets (for the outer measure this is not relevant since overlap between rectangles will only lead to a larger sum). If $J_*(A) = J^*(A)$ the set A is called **Jordan measurable**.

Unfortunately this approach turned out to have several shortcomings (essentially identical to those of the Riemann integral). Its limitation stems from the fact that one only allows finite covers. Switching to countable covers will produce the much more flexible Lebesgue measure.

Example 8.1. To understand this limitation let us look at the classical example of a non Riemann integrable function, the characteristic function of the rational numbers inside $[0, 1]$. If we want to cover $\mathbb{Q} \cap [0, 1]$ by a finite number of intervals, we always end up covering all of $[0, 1]$ since the rational numbers are dense. Conversely, if we want to find the inner content, no single (nontrivial) interval will fit into our set since it has empty interior. In summary, $J_*(\mathbb{Q} \cap [0, 1]) = 0 \neq 1 = J^*(\mathbb{Q} \cap [0, 1])$.

On the other hand, if we are allowed to take a countable number of intervals, we can enumerate the points in $\mathbb{Q} \cap [0, 1]$ and cover the j'th point by an interval of length $\varepsilon 2^{-j}$ such that the total length of this cover is less than ε, which can be arbitrarily small. ⋄

The previous example also hints at what is going on in general. When computing the outer content you will always end up covering the closure of A and when computing the inner content you will never get more than the interior of A. Hence a set should be Jordan measurable if the difference between the closure and the interior, which is by definition the boundary $\partial A = \overline{A} \setminus A^\circ$, is small. However, we do not want to pursue this further at this point and hence we defer it to Appendix 8.7.

Rather we will make the anticipated change and define the **Lebesgue outer measure** via

$$\lambda^{n,*}(A) := \inf \Big\{ \sum_{j=1}^{\infty} |R_j| \,\Big|\, A \subseteq \bigcup_{j=1}^{\infty} R_j,\ R_j \in \mathcal{S}^n \Big\}. \tag{8.4}$$

In particular, we will call N a Lebesgue **null set** if $\lambda^{n,*}(N) = 0$.

Example 8.2. As shown in the previous example, the set of rational numbers inside $[0,1]$ is a null set. In fact, the same argument shows that every countable set is a null set.

Consequently we expect the irrational numbers inside $[0,1]$ to be a set of measure one. But if we try to approximate this set from the inside by half-closed intervals we are bound to fail as no single (nonempty) interval will fit into this set. This explains why we did not define a corresponding inner measure. ⋄

So we have defined Lebesgue outer measure and we have seen that it has some basic properties. Moreover, there are some further properties. For example, let $f(x) = Mx + a$ be an affine transformation, then

$$\lambda^{n,*}(MA + a) = \det(M) \lambda^{n,*}(A). \tag{8.5}$$

In fact, that translations do not change the outer measure is immediate since $\mathcal{S}^n$ is invariant under translations and the same is true for $|R|$. Moreover, every matrix can be written as $M = O_1 D O_2$, where O_j are orthogonal and D is diagonal (Problem 9.22). So it reduces the problem to showing this for diagonal matrices and for orthogonal matrices. The case of diagonal matrices follows as before but the case of orthogonal matrices is more involved (it can be shown by showing that rectangles can be replaced by open balls in the definition of the outer measure). Hence we postpone this to Section 8.8.

For now we will use this fact only to explain why our outer measure is still not good enough. The reason is that it lacks one key property, namely additivity! Of course this will be crucial for the corresponding integral to be linear and hence is indispensable. Now here comes the bad news: A classical paradox by Banach and Tarski shows that one can break the unit ball in $\mathbb{R}^3$ into a finite number of (wild – choosing the pieces uses the Axiom of Choice and cannot be done with a jigsaw;-) pieces, and reassemble them using only

rotations and translations to get two copies of the unit ball. Hence our outer measure (as well as any other reasonable notion of size which is translation and rotation invariant) cannot be additive when defined for all sets! If you think that the situation in one dimension is better, I have to disappoint you as well: Problem 8.1.

So our only hope left is that additivity at least holds on a suitable class of sets. In fact, even finite additivity is not sufficient for us since limiting operations will require that we are able to handle countable operations. To this end we will introduce some abstract concepts first.

So the remaining question is if we can extend the family of sets on which σ-additivity holds. A convenient family clearly should be invariant under countable set operations and hence we will call an algebra a σ-algebra if it is closed under countable unions (and hence also under countable intersections by De Morgan's rules). But how to construct such a σ-algebra?

It was Lebesgue who eventually was successful with the following idea: As pointed out before there is no corresponding inner Lebesgue measure since approximation by intervals from the inside does not work well. However, instead you can try to approximate the complement from the outside thereby setting

$$\lambda^1_*(A) = (b-a) - \lambda^{1,*}([a,b] \setminus A) \tag{8.6}$$

for every bounded set $A \subseteq [a,b]$. Now you can call a bounded set A measurable if $\lambda^1_*(A) = \lambda^{1,*}(A)$. We will however use a somewhat different approach due to Carathéodory. In this respect note that if we set $E = [a,b]$ then $\lambda^1_*(A) = \lambda^{1,*}(A)$ can be written as

$$\lambda^{1,*}(E) = \lambda^{1,*}(A \cap E) + \lambda^{1,*}(A' \cap E) \tag{8.7}$$

which should be compared with the Carathéodory condition (8.12).

Problem* 8.1 (Vitali set)**.** *Call two numbers $x, y \in [0,1)$ equivalent if $x-y$ is rational. Construct the set V by choosing one representative from each equivalence class. Show that V cannot be measurable with respect to any nontrivial finite translation invariant measure on $[0,1)$. (Hint: How can you build up $[0,1)$ from translations of V?)*

Problem 8.2. *Show that $J_*(A) \le \lambda^{n,*}(A) \le J^*(A)$ and hence $J(A) = \lambda^{n,*}(A)$ for every Jordan measurable set.*

Problem 8.3. *Let $X := \mathbb{N}$ and define the set function*

$$\mu(A) := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \chi_A(n) \in [0,1]$$

on the collection $\mathcal{S}$ of all sets for which the above limit exists. Show that this collection is closed under disjoint unions and complements but not under

intersections. Show that there is an extension to the σ-algebra generated by $\mathcal{S}$ (i.e. to $\mathfrak{P}(\mathbb{N})$) which is additive but no extension which is σ-additive. (Hints: To show that $\mathcal{S}$ is not closed under intersections take a set of even numbers $A_1 \notin \mathcal{S}$ and let A_2 be the missing even numbers. Then let $A = A_1 \cup A_2 = 2\mathbb{N}$ and $B = A_1 \cup \tilde{A}_2$, where $\tilde{A}_2 = A_2 + 1$. To obtain an extension to $\mathfrak{P}(\mathbb{N})$ consider χ_A, $A \in \mathcal{S}$, as vectors in $\ell^\infty(\mathbb{N})$ and μ as a linear functional and see Problem 4.24.)

8.2. Sigma algebras and measures

If an algebra is closed under countable intersections, it is called a **σ-algebra**. Hence a σ-algebra is a family of subsets Σ of a given set X such that

- $\emptyset \in \Sigma$,
- Σ is closed under countable intersections, and
- Σ is closed under complements.

By De Morgan's rule Σ is also closed under countable unions.

Moreover, the intersection of any family of (σ-)algebras $\{\Sigma_\alpha\}$ is again a (σ-)algebra (check this) and for any collection S of subsets there is a unique smallest (σ-)algebra $\Sigma(S)$ containing S (namely the intersection of all (σ-)algebras containing S). It is called the (σ-)algebra generated by S.

Example 8.3. For a given set X and a subset $A \subseteq X$ we have $\Sigma(\{A\}) = \{\emptyset, A, A', X\}$. Moreover, every finite algebra is also a σ-algebra and if S is finite, so will be $\Sigma(S)$ (Problem 8.5).

The power set $\mathfrak{P}(X)$ is clearly the largest σ-algebra and $\{\emptyset, X\}$ is the smallest. ⋄

If X is a topological space, the **Borel σ-algebra** $\mathfrak{B}(X)$ of X is defined to be the σ-algebra generated by all open (equivalently, all closed) sets. In fact, if X is second countable, any countable base will suffice to generate the Borel σ-algebra (recall Lemma B.1). Sets in the Borel σ-algebra are called **Borel sets**.

Example 8.4. In the case $X = \mathbb{R}^n$ the Borel σ-algebra will be denoted by $\mathfrak{B}^n$ and we will abbreviate $\mathfrak{B} := \mathfrak{B}^1$. Note that in order to generate $\mathfrak{B}^n$, open balls with rational center and rational radius suffice. In fact, any base for the topology will suffice. Moreover, since open balls can be written as a countable union of smaller closed balls with increasing radii, we could also use compact balls instead. ⋄

Example 8.5. If X is a topological space, then any Borel set $Y \subseteq X$ is also a topological space equipped with the relative topology and its Borel σ-algebra is given by $\mathfrak{B}(Y) = \mathfrak{B}(X) \cap Y := \{A | A \in \mathfrak{B}(X), A \subseteq Y\}$ (show this). ⋄

In the sequel we will frequently meet unions of disjoint sets and hence we will introduce the following short hand notation for the union of mutually disjoint sets:

$$\bigcupdot_{j\in J} A_j := \bigcup_{j\in J} A_j \quad \text{with} \quad A_j \cap A_k = \emptyset \text{ for all } j \neq k. \tag{8.8}$$

Now let us turn to the definition of a measure: A set X together with a σ-algebra Σ is called a **measurable space**. A **measure** μ is a map $\mu : \Sigma \to [0, \infty]$ on a σ-algebra Σ such that

- $\mu(\emptyset) = 0$,
- $\mu(\bigcupdot_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$, $A_n \in \Sigma$ (σ-additivity).

Here the sum is set equal to ∞ if one of the summands is ∞ or if it diverges.

The measure μ is called **finite** if $\mu(X) < \infty$ and a **probability measure** if $\mu(X) = 1$. It is called **σ-finite** if there is a countable cover $\{X_n\}_{n=1}^{\infty}$ of X such that $X_n \in \Sigma$ and $\mu(X_n) < \infty$ for all n. (Note that it is no restriction to assume $X_n \subseteq X_{n+1}$.) The sets in Σ are called **measurable sets** and the triple (X, Σ, μ) is referred to as a **measure space**.

Example 8.6. Take a set X with $\Sigma = \mathfrak{P}(X)$ and set $\mu(A)$ to be the number of elements of A (respectively, ∞ if A is infinite). This is the so-called **counting measure**. It will be finite if and only if X is finite and σ-finite if and only if X is countable. ⋄

Example 8.7. Take a set X and $\Sigma := \mathfrak{P}(X)$. Fix a point $x \in X$ and set $\mu(A) = 1$ if $x \in A$ and $\mu(A) = 0$ else. This is the **Dirac measure** centered at x. It is also frequently written as δ_x. ⋄

Example 8.8. Let μ_1, μ_2 be two measures on (X, Σ) and $\alpha_1, \alpha_2 \geq 0$. Then $\mu = \alpha_1\mu_1 + \alpha_2\mu_2$ defined via

$$\mu(A) := \alpha_1\mu_1(A) + \alpha_2\mu_2(A)$$

is again a measure. Furthermore, given a countable number of measures μ_n and numbers $\alpha_n \geq 0$, then $\mu := \sum_n \alpha_n\mu_n$ is again a measure (show this). ⋄

Example 8.9. Let μ be a measure on (X, Σ) and $Y \subseteq X$ a measurable subset. Then

$$\nu(A) := \mu(A \cap Y)$$

is again a measure on (X, Σ) (show this). ⋄

Example 8.10. Let X be some set with a σ-algebra Σ. Then every subset $Y \subseteq X$ has a natural σ-algebra $\Sigma \cap Y := \{A \cap Y | A \in \Sigma\}$ (show that this is indeed a σ-algebra) known as the **relative σ-algebra** (also **trace σ-algebra**).

Note that if S generates Σ, then $S \cap Y$ generates $\Sigma \cap Y$: $\Sigma(S) \cap Y = \Sigma(S \cap Y)$. Indeed, since $\Sigma \cap Y$ is a σ-algebra containing $S \cap Y$, we have $\Sigma(S \cap Y) \subseteq \Sigma(S) \cap Y = \Sigma \cap Y$. Conversely, consider $\{A \in \Sigma | A \cap Y \in \Sigma(S \cap Y)\}$ which is a σ-algebra (check this). Since this last σ-algebra contains S it must be equal to $\Sigma = \Sigma(S)$ and thus $\Sigma \cap Y \subseteq \Sigma(S \cap Y)$. $\diamond$

Example 8.11. If $Y \in \Sigma$ we can restrict the σ-algebra $\Sigma|_Y = \{A \in \Sigma | A \subseteq Y\}$ such that $(Y, \Sigma|_Y, \mu|_Y)$ is again a measurable space. It will be σ-finite if (X, Σ, μ) is. $\diamond$

Finally, we will show that σ-additivity implies some crucial continuity properties for measures which eventually will lead to powerful limiting theorems for the corresponding integrals. We will write $A_n \nearrow A$ if $A_n \subseteq A_{n+1}$ with $A = \bigcup_n A_n$ and $A_n \searrow A$ if $A_{n+1} \subseteq A_n$ with $A = \bigcap_n A_n$.

Theorem 8.1. *Any measure μ satisfies the following properties:*

(i) *$A \subseteq B$ implies $\mu(A) \leq \mu(B)$ (monotonicity).*

(ii) *$\mu(A_n) \to \mu(A)$ if $A_n \nearrow A$ (continuity from below).*

(iii) *$\mu(A_n) \to \mu(A)$ if $A_n \searrow A$ and $\mu(A_1) < \infty$ (continuity from above).*

Proof. The first claim is obvious from $\mu(B) = \mu(A) + \mu(B \setminus A)$. To see the second define $\tilde{A}_1 = A_1$, $\tilde{A}_n = A_n \setminus A_{n-1}$ and note that these sets are disjoint and satisfy $A_n = \bigcup_{j=1}^n \tilde{A}_j$. Hence $\mu(A_n) = \sum_{j=1}^n \mu(\tilde{A}_j) \to \sum_{j=1}^\infty \mu(\tilde{A}_j) = \mu(\bigcup_{j=1}^\infty \tilde{A}_j) = \mu(A)$ by σ-additivity. The third follows from the second using $\tilde{A}_n = A_1 \setminus A_n \nearrow A_1 \setminus A$ implying $\mu(\tilde{A}_n) = \mu(A_1) - \mu(A_n) \to \mu(A_1 \setminus A) = \mu(A_1) - \mu(A)$. $\square$

Example 8.12. Consider the counting measure on $\mathbb{N}$ and let $A_n = \{j \in \mathbb{N} | j \geq n\}$. Then $\mu(A_n) = \infty$, but $\mu(\bigcap_n A_n) = \mu(\emptyset) = 0$ which shows that the requirement $\mu(A_1) < \infty$ in item (iii) of Theorem 8.1 is not superfluous. $\diamond$

Problem 8.4. *Find all algebras over $X := \{1, 2, 3\}$.*

Problem 8.5. *Let $\{A_j\}_{j=1}^n$ be a finite family of subsets of a given set X. Show that $\Sigma(\{A_j\}_{j=1}^n)$ has at most 4^n elements. (Hint: Let X have 2^n elements and look at the case $n = 2$ to get an idea. Consider sets of the form $B_1 \cap \cdots \cap B_n$ with $B_j \in \{A_j, A_j'\}$.)*

Problem 8.6. *Show that $\mathcal{A} := \{A \subseteq X | A$ or $X \setminus A$ is finite$\}$ is an algebra (with X some fixed set). Show that $\Sigma := \{A \subseteq X | A$ or $X \setminus A$ is countable$\}$ is a σ-algebra. (Hint: To verify closedness under unions consider the cases where all sets are finite and where one set has finite complement.)*

Problem 8.7. *Take some set X and $\Sigma := \{A \subseteq X | A$ or $X \setminus A$ is countable$\}$. Show that*

$$\nu(A) := \begin{cases} 0, & \text{if } A \text{ is countable,} \\ 1, & \text{else.} \end{cases}$$

is a measure

Problem 8.8. *Show that if X is finite, then every algebra is a σ-algebra. Show that this is not true in general if X is countable.*

8.3. Extending a premeasure to a measure

Now we are ready to show how to construct a measure starting from a small collection of sets which generate the σ-algebra and for which it is clear what the measure should be. The crucial requirement for such a collection of sets is the requirement that it is closed under intersections as the following example shows.

Example 8.13. Consider $X := \{1, 2, 3\}$ together with the measures $\mu(\{1\}) := \mu(\{2\}) := \mu(\{3\}) := \frac{1}{3}$ and $\nu(\{1\}) := \frac{1}{6}$, $\nu(\{2\}) := \frac{1}{2}$, $\nu(\{3\}) := \frac{1}{6}$. Then μ and ν agree on $S := \{\{1, 2\}, \{2, 3\}\}$ but not on $\Sigma(S) = \mathfrak{P}(X)$. Note that S is not closed under intersections. If we take $S := \{\emptyset, \{1\}, \{2, 3\}\}$, which is closed under intersections, and $\nu(\{1\}) := \frac{1}{3}$, $\nu(\{2\}) := \frac{1}{2}$, $\nu(\{3\}) := \frac{1}{6}$, then μ and ν agree on $\Sigma(S) = \{\emptyset, \{1\}, \{2, 3\}, X\}$. But they don't agree on $\mathfrak{P}(X)$. ◇

Hence we begin with the question when a family of sets determines a measure uniquely. To this end we need a better criterion to check when a given system of sets is in fact a σ-algebra. In many situations it is easy to show that a given set is closed under complements and under countable unions of disjoint sets. Hence we call a collection of sets $\mathcal{D}$ with these properties a **Dynkin system** (also **λ-system**) if it also contains X.

Note that a Dynkin system is closed under proper relative complements since $A, B \in \mathcal{D}$ implies $B \setminus A = (B' \cup A)' \in \mathcal{D}$ provided $A \subseteq B$. Moreover, if it is also closed under finite intersections (or arbitrary finite unions) then it is an algebra and hence also a σ-algebra. To see the last claim note that if $A = \bigcup_j A_j$ then also $A = \biguplus_j B_j$ where the sets $B_j = A_j \setminus \bigcup_{k<j} A_k$ are disjoint.

Example 8.14. Let $X := \{1, 2, 3, 4\}$. Then $\mathcal{D} := \{A \subseteq X | \#A$ is even$\}$ is a Dynkin system but no algebra. ◇

As with σ-algebras, the intersection of Dynkin systems is a Dynkin system and every collection of sets S generates a smallest Dynkin system $\mathcal{D}(S)$. The important observation is that if S is closed under finite intersections (in

which case it is sometimes called a π-**system**), then so is $\mathcal{D}(S)$ and hence will be a σ-algebra.

Lemma 8.2 (Dynkin's π-λ theorem)**.** *Let S be a collection of subsets of X which is closed under finite intersections (or unions). Then $\mathcal{D}(S) = \Sigma(S)$.*

Proof. It suffices to show that $\mathcal{D} := \mathcal{D}(S)$ is closed under finite intersections. To this end consider the set $D(A) := \{B \in \mathcal{D} | A \cap B \in \mathcal{D}\}$ for $A \in \mathcal{D}$. I claim that $D(A)$ is a Dynkin system.

First of all $X \in D(A)$ since $A \cap X = A \in \mathcal{D}$. Next, if $B \in D(A)$ then $A \cap B' = A \setminus (B \cap A) \in \mathcal{D}$ (since $\mathcal{D}$ is closed under proper relative complements) implying $B' \in D(A)$. Finally if $B = \bigcup\!\!\!\cdot_j B_j$ with $B_j \in D(A)$ disjoint, then $A \cap B = \bigcup\!\!\!\cdot_j (A \cap B_j) \in \mathcal{D}$ with $A \cap B_j \in \mathcal{D}$ disjoint, implying $B \in D(A)$.

Now if $A \in S$ we have $S \subseteq D(A)$ implying $D(A) = \mathcal{D}$. Consequently $A \cap B \in \mathcal{D}$ if at least one of the sets is in S. But this shows $S \subseteq D(A)$ and hence $D(A) = \mathcal{D}$ for every $A \in \mathcal{D}$. So $\mathcal{D}$ is closed under finite intersections and thus a σ-algebra. The case of unions is analogous. □

The typical use of this lemma is as follows: First verify some property for sets in a collection S which is closed under finite intersections and generates the σ-algebra. In order to show that it holds for every set in $\Sigma(S)$, it suffices to show that the collection of sets for which it holds is a Dynkin system. This is illustrated in our first result.

Theorem 8.3 (Uniqueness of measures)**.** *Let $S \subseteq \Sigma$ be a collection of sets which generates Σ and which is closed under finite intersections and contains a sequence of increasing sets $X_n \nearrow X$ of finite measure $\mu(X_n) < \infty$. Then μ is uniquely determined by the values on S.*

Proof. Let $\tilde{\mu}$ be a second measure and note $\mu(X) = \lim_{n\to\infty} \mu(X_n) = \lim_{n\to\infty} \tilde{\mu}(X_n) = \tilde{\mu}(X)$. We first suppose $\mu(X) < \infty$.

Then

$$\mathcal{D} := \{A \in \Sigma | \mu(A) = \tilde{\mu}(A)\}$$

is a Dynkin system. In fact, by $\mu(A') = \mu(X) - \mu(A) = \tilde{\mu}(X) - \tilde{\mu}(A) = \tilde{\mu}(A')$ for $A \in \mathcal{D}$ we see that $\mathcal{D}$ is closed under complements. Furthermore, by continuity of measures from below it is also closed under countable disjoint unions. Since $\mathcal{D}$ contains S by assumption, we conclude $\mathcal{D} = \Sigma(S) = \Sigma$ from Lemma 8.2. This finishes the finite case.

To extend our result to the general case observe that the finite case implies $\mu(A \cap X_j) = \tilde{\mu}(A \cap X_j)$ (just restrict μ, $\tilde{\mu}$ to X_j). Hence

$$\mu(A) = \lim_{j\to\infty} \mu(A \cap X_j) = \lim_{j\to\infty} \tilde{\mu}(A \cap X_j) = \tilde{\mu}(A)$$

and we are done. □

Example 8.15. The counting measure as well as the measure which assigns every nonempty set the value ∞ both agree on $\mathcal{S}^1$. Hence the finiteness assumption in the previous theorem/corollary is crucial. ◇

Inspired by the collection of rectangles $\mathcal{S}^n$ we call a collection of subsets $\mathcal{S}$ of a given set X is called a **semialgebra** if $\emptyset \in \mathcal{S}$, $\mathcal{S}$ is closed under finite intersections, and the complement of a set in $\mathcal{S}$ can be written as a finite union of sets from $\mathcal{S}$. A semialgebra $\mathcal{A}$ which is closed under complements is called an **algebra**.

Note that $X \in \mathcal{A}$ and that $\mathcal{A}$ is also closed under finite unions and relative complements: $X = \emptyset'$, $A \cup B = (A' \cap B')'$ (De Morgan), and $A \setminus B = A \cap B'$, where $A' = X \setminus A$ denotes the complement.

Example 8.16. Let $X := \{1, 2, 3\}$, then $\mathcal{A} := \{\emptyset, \{1\}, \{2, 3\}, X\}$ is an algebra. ◇

In fact, considering finite disjoint unions from a semialgebra we always have a corresponding algebra.

Lemma 8.4. *Let $\mathcal{S}$ be a semialgebra, then the set of all finite disjoint unions $\bar{\mathcal{S}} := \{\bigcupdot_{j=1}^n A_j | A_j \in \mathcal{S}\}$ is an algebra.*

Proof. Suppose $A = \bigcupdot_{j=1}^n A_j \in \bar{\mathcal{S}}$ and $B = \bigcupdot_{k=1}^m B_k \in \bar{\mathcal{S}}$. Then $A \cap B = \bigcupdot_{j,k} (A_j \cap B_k) \in \bar{\mathcal{S}}$. Concerning complements we have $A' = \bigcap_j A_j' \in \bar{\mathcal{S}}$ since $A_j' \in \bar{\mathcal{S}}$ by definition of a semialgebra and since $\bar{\mathcal{S}}$ is closed under finite intersections by the first part. □

Example 8.17. The collection of all intervals (augmented by the empty set) clearly is a semialgebra. Moreover, the collection $\mathcal{S}^1$ of half-open intervals of the form $(a, b] \subseteq \mathbb{R}$, $-\infty \le a < b \le \infty$ augmented by the empty set is a semialgebra. Since the product of semialgebras is again a semialgebra (Problem 8.9), the same is true for the collection of rectangles $\mathcal{S}^n$. ◇

A set function $\mu : \mathcal{A} \to [0, \infty]$ on an algebra is called a **premeasure** if it satisfies

- $\mu(\emptyset) = 0$,
- $\mu(\bigcupdot_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j)$, if $A_j \in \mathcal{A}$ and $\bigcupdot_{j=1}^{\infty} A_j \in \mathcal{A}$ (σ-additivity).

In terms of a premeasure Theorem 8.3 reads as follows:

Corollary 8.5. *Let μ be a σ-finite premeasure on an algebra $\mathcal{A}$. Then there is at most one extension to $\Sigma(\mathcal{A})$.*

The following lemma gives conditions when the natural extension of a set function on a semialgebra $\mathcal{S}$ to its associated algebra $\bar{\mathcal{S}}$ will be a premeasure.

Lemma 8.6. *Let $\mathcal{S}$ be a semialgebra and let $\mu : \mathcal{S} \to [0,\infty]$ be additive, that is, $A = \biguplus_{j=1}^n A_j$ with $A, A_j \in \mathcal{S}$ implies $\mu(A) = \sum_{j=1}^n \mu(A_j)$. Then the natural extension $\mu : \bar{\mathcal{S}} \to [0,\infty]$ given by*

$$\mu(A) := \sum_{j=1}^n \mu(A_j), \qquad A = \bigcupdot_{j=1}^n A_j, \tag{8.9}$$

is (well defined and) additive on $\bar{\mathcal{S}}$. Moreover, it will be a premeasure if

$$\mu(\bigcupdot_{j=1}^\infty A_j) \le \sum_{j=1}^\infty \mu(A_j) \tag{8.10}$$

whenever $\biguplus_j A_j \in \mathcal{S}$ and $A_j \in \mathcal{S}$.

Proof. We begin by showing that μ is well defined. To this end let $A = \biguplus_{j=1}^n A_j = \biguplus_{k=1}^m B_k$ and set $C_{jk} := A_j \cap B_k$. Then

$$\sum_j \mu(A_j) = \sum_j \mu(\bigcupdot_k C_{jk}) = \sum_{j,k} \mu(C_{jk}) = \sum_k \mu(\bigcupdot_j C_{jk}) = \sum_k \mu(B_k)$$

by additivity on $\mathcal{S}$. Moreover, if $A = \biguplus_{j=1}^n A_j$ and $B = \biguplus_{k=1}^m B_k$ are two disjoint sets from $\bar{\mathcal{S}}$, then

$$\mu(A \dot\cup B) = \mu(\left(\bigcupdot_{j=1}^n A_j\right) \dot\cup \left(\bigcupdot_{k=1}^m B_k\right)) = \sum_j \mu(A_j) + \sum_k \mu(B_k) = \mu(A) + \mu(B)$$

which establishes additivity. Finally, let $A = \biguplus_{j=1}^\infty A_j \in \mathcal{S}$ with $A_j \in \mathcal{S}$ and observe $B_n = \biguplus_{j=1}^n A_j \in \bar{\mathcal{S}}$. Hence

$$\sum_{j=1}^n \mu(A_j) = \mu(B_n) \le \mu(B_n) + \mu(A \setminus B_n) = \mu(A)$$

and combining this with our assumption (8.10) shows σ-additivity when all sets are from $\mathcal{S}$. By finite additivity this extends to the case of sets from $\bar{\mathcal{S}}$. □

Using this lemma our set function $|R|$ for rectangles is easily seen to be a premeasure.

Lemma 8.7. *The set function* (8.1) *for rectangles extends to a premeasure on $\bar{\mathcal{S}}^n$.*

Proof. Finite additivity is left at an exercise (see the proof of Lemma 8.12) and it remains to verify (8.10). We can cover each $A_j := (a_j, b_j]$ by some slightly larger rectangle $B_j := (a^j, b^j + \delta^j]$ such that $|B_j| \le |A_j| + \frac{\varepsilon}{2^j}$. Then for any $r > 0$ we can find an m such that the open intervals $\{(a^j, b^j + \delta^j)\}_{j=1}^m$ cover the compact set $\overline{A \cap Q_r}$, where Q_r is a half-open cube of side length r. Hence

$$|A \cap Q_r| \le \Big|\bigcup_{j=1}^m B_j\Big| = \sum_{j=1}^m \mu(B_j) \le \sum_{j=1}^\infty |A_j| + \varepsilon.$$

Letting $r \to \infty$ and since $\varepsilon > 0$ is arbitrary, we are done. □

The next step is to extend a premeasure to an outer measure: A function $\mu^* : \mathfrak{P}(X) \to [0, \infty]$ is an **outer measure** if it has the properties

- $\mu^*(\emptyset) = 0$,
- $A \subseteq B \;\Rightarrow\; \mu^*(A) \le \mu^*(B)$ (monotonicity), and
- $\mu^*(\bigcup_{n=1}^\infty A_n) \le \sum_{n=1}^\infty \mu^*(A_n)$ (subadditivity).

Here $\mathfrak{P}(X)$ is the power set (i.e., the collection of all subsets) of X. Note that null sets N (i.e., $\mu^*(N) = 0$) do not change the outer measure: $\mu^*(A) \le \mu^*(A \cup N) \le \mu^*(A) + \mu^*(N) = \mu^*(A)$.

It turns out that it is quite easy to get an outer measure from an arbitrary set function.

Lemma 8.8. *Let $\mathcal{E}$ be some family of subsets of X containing $\emptyset$. Suppose we have a set function $\rho : \mathcal{E} \to [0, \infty]$ such that $\rho(\emptyset) = 0$. Then*

$$\mu^*(A) := \inf\Big\{ \sum_{j=1}^\infty \rho(A_j) \Big| A \subseteq \bigcup_{j=1}^\infty A_j,\ A_j \in \mathcal{E} \Big\}$$

is an outer measure. Here the infimum extends over all countable covers from $\mathcal{E}$ with the convention that the infimum is infinite if no such cover exists.

Proof. $\mu^*(\emptyset) = 0$ is trivial since we can choose $A_j = \emptyset$ as a cover.

To see see monotonicity let $A \subseteq B$ and note that if $\{A_j\}$ is a cover for B then it is also a cover for A (if there is no cover for B, there is nothing to do). Hence

$$\mu^*(A) \le \sum_{j=1}^\infty \rho(A_j)$$

and taking the infimum over all covers for B shows $\mu^*(A) \le \mu^*(B)$.

To see subadditivity note that we can assume that all sets A_j have a cover (otherwise there is nothing to do) $\{B_{jk}\}_{k=1}^\infty$ for A_j such that $\sum_{k=1}^\infty \mu(B_{jk}) \le$

$\mu^*(A_j) + \frac{\varepsilon}{2^j}$. Since $\{B_{jk}\}_{j,k=1}^\infty$ is a cover for $\bigcup_j A_j$ we obtain

$$\mu^*(A) \le \sum_{j,k=1}^\infty \rho(B_{jk}) \le \sum_{j=1}^\infty \mu^*(A_j) + \varepsilon$$

and since $\varepsilon > 0$ is arbitrary subadditivity follows. □

Consequently, for any premeasure μ we define its corresponding outer measure $\mu^* : \mathfrak{P}(X) \to [0, \infty]$ (Lemma 8.8) as

$$\mu^*(A) := \inf \Big\{ \sum_{n=1}^\infty \mu(A_n) \Big| A \subseteq \bigcup_{n=1}^\infty A_n,\ A_n \in \mathcal{A} \Big\}, \tag{8.11}$$

where the infimum extends over all countable covers from $\mathcal{A}$. Replacing A_n by $\tilde{A}_n = A_n \setminus \bigcup_{m=1}^{n-1} A_m$ we see that we could even require the covers to be disjoint. Note that $\mu^*(A) = \mu(A)$ for $A \in \mathcal{A}$ (Problem 8.10).

Theorem 8.9 (Extensions via outer measures)**.** *Let μ^* be an outer measure. Then the set Σ of all sets A satisfying the Carathéodory condition*

$$\mu^*(E) = \mu^*(A \cap E) + \mu^*(A' \cap E), \quad \forall E \subseteq X, \tag{8.12}$$

(where $A' := X \setminus A$ is the complement of A) forms a σ-algebra and μ^* restricted to this σ-algebra is a measure.

Proof. We first show that Σ is an algebra. It clearly contains X and is closed under complements. Concerning unions let $A, B \in \Sigma$. Applying Carathéodory's condition twice shows

$$\begin{aligned}\mu^*(E) =& \mu^*(A \cap B \cap E) + \mu^*(A' \cap B \cap E) + \mu^*(A \cap B' \cap E) \\ &+ \mu^*(A' \cap B' \cap E) \\ \ge& \mu^*((A \cup B) \cap E) + \mu^*((A \cup B)' \cap E),\end{aligned}$$

where we have used De Morgan and

$$\mu^*(A \cap B \cap E) + \mu^*(A' \cap B \cap E) + \mu^*(A \cap B' \cap E) \ge \mu^*((A \cup B) \cap E)$$

which follows from subadditivity and $(A \cup B) \cap E = (A \cap B \cap E) \cup (A' \cap B \cap E) \cup (A \cap B' \cap E)$. Since the reverse inequality is just subadditivity, we conclude that Σ is an algebra.

Next, let A_n be a sequence of sets from Σ. Without restriction we can assume that they are disjoint (compare the argument for item (ii) in the proof of Theorem 8.1). Abbreviate $\tilde{A}_n = \bigcup_{k \le n} A_k$, $A = \bigcup_n A_n$. Then for

every set E we have

$$\begin{aligned}\mu^*(\tilde{A}_n \cap E) &= \mu^*(A_n \cap \tilde{A}_n \cap E) + \mu^*(A_n' \cap \tilde{A}_n \cap E)\\ &= \mu^*(A_n \cap E) + \mu^*(\tilde{A}_{n-1} \cap E)\\ &= \ldots = \sum_{k=1}^{n} \mu^*(A_k \cap E).\end{aligned}$$

Using $\tilde{A}_n \in \Sigma$ and monotonicity of μ^*, we infer

$$\begin{aligned}\mu^*(E) &= \mu^*(\tilde{A}_n \cap E) + \mu^*(\tilde{A}_n' \cap E)\\ &\geq \sum_{k=1}^{n} \mu^*(A_k \cap E) + \mu^*(A' \cap E).\end{aligned}$$

Letting $n \to \infty$ and using subadditivity finally gives

$$\begin{aligned}\mu^*(E) &\geq \sum_{k=1}^{\infty} \mu^*(A_k \cap E) + \mu^*(A' \cap E)\\ &\geq \mu^*(A \cap E) + \mu^*(A' \cap E) \geq \mu^*(E)\end{aligned} \tag{8.13}$$

and we infer that Σ is a σ-algebra.

Finally, setting $E = A$ in (8.13), we have

$$\mu^*(A) = \sum_{k=1}^{\infty} \mu^*(A_k \cap A) + \mu^*(A' \cap A) = \sum_{k=1}^{\infty} \mu^*(A_k)$$

and we are done. □

Remark: The constructed measure μ is **complete**; that is, for every measurable set A of measure zero, every subset of A is again measurable. In fact, every **null set** A, that is, every set with $\mu^*(A) = 0$, is measurable (Problem 8.11).

The only remaining question is whether there are any nontrivial sets satisfying the Carathéodory condition.

Lemma 8.10. *Let μ be a premeasure on $\mathcal{A}$ and let μ^* be the associated outer measure. Then every set in $\mathcal{A}$ satisfies the Carathéodory condition.*

Proof. Let $A_n \in \mathcal{A}$ be a countable cover for E. Then for every $A \in \mathcal{A}$ we have

$$\sum_{n=1}^{\infty} \mu(A_n) = \sum_{n=1}^{\infty} \mu(A_n \cap A) + \sum_{n=1}^{\infty} \mu(A_n \cap A') \geq \mu^*(E \cap A) + \mu^*(E \cap A')$$

since $A_n \cap A \in \mathcal{A}$ is a cover for $E \cap A$ and $A_n \cap A' \in \mathcal{A}$ is a cover for $E \cap A'$. Taking the infimum, we have $\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \cap A')$, which finishes the proof. □

Thus the Lebesgue premeasure on $\bar{\mathcal{S}}^n$ gives rise to Lebesgue measure λ^n when the outer measure $\lambda^{*,n}$ is restricted to the Borel σ-algebra $\mathcal{B}^n$. In fact, with the very same procedure we can obtain a large class of measures in $\mathbb{R}^n$ as will be demonstrated in the next section.

To end this section, let me emphasize that in our approach we started from a premeasure, which gave rise to an outer measure, which eventually lead to a measure via Carathéodory's theorem. However, while some effort was required to get the premeasure in our case, an outer measure often can be obtained much easier (recall Lemma 8.8). While this immediately leads again to a measure, one is faced with the problem if any nontrivial sets satisfy the Carathéodory condition.

To address this problem let (X, d) be a metric space and call an outer measure μ^* on X a **metric outer measure** if

$$\mu^*(A_1 \cup A_2) = \mu^*(A_1) + \mu^*(A_2)$$

whenever

$$\operatorname{dist}(A_1, A_2) := \inf_{(x_1,x_2)\in A_1\times A_2} d(x_1, x_2) > 0.$$

Lemma 8.11. *Let X be a metric space. An outer measure is metric if and only if all Borel sets satisfy the Carathéodory condition* (8.12).

Proof. To show that all Borel sets satisfy the Carathéodory condition it suffices to show this is true for all closed sets. First of all note that we have $G_n := \{x \in F' \cap E | d(x, F) \geq \frac{1}{n}\} \nearrow F' \cap E$ since F is closed. Moreover, $d(G_n, F) \geq \frac{1}{n}$ and hence $\mu^*(F \cap E) + \mu^*(G_n) = \mu^*((E \cap F) \cup G_n) \leq \mu^*(E)$ by the definition of a metric outer measure. Hence it suffices to shows $\mu^*(G_n) \to \mu^*(F' \cap E)$. Moreover, we can also assume $\mu^*(E) < \infty$ since otherwise there is noting to show. Now consider $\tilde{G}_n = G_{n+1} \backslash G_n$. Then $d(\tilde{G}_{n+2}, \tilde{G}_n) > 0$ and hence $\sum_{j=1}^m \mu^*(\tilde{G}_{2j}) = \mu^*(\cup_{j=1}^m \tilde{G}_{2j}) \leq \mu^*(E)$ as well as $\sum_{j=1}^m \mu^*(\tilde{G}_{2j-1}) = \mu^*(\cup_{j=1}^m \tilde{G}_{2j-1}) \leq \mu^*(E)$ and consequently $\sum_{j=1}^\infty \mu^*(\tilde{G}_j) \leq 2\mu^*(E) < \infty$. Now subadditivity implies $\mu^*(F' \cap E) \leq \mu(G_n) + \sum_{j\geq n} \mu^*(\tilde{G}_n)$ and thus

$$\mu^*(F' \cap E) \leq \liminf_{n\to\infty} \mu^*(G_n) \leq \limsup_{n\to\infty} \mu^*(G_n) \leq \mu^*(F' \cap E)$$

as required.

Conversely, suppose $\varepsilon := \operatorname{dist}(A_1, A_2) > 0$ and consider ε neighborhood of A_1 given by $O_\varepsilon = \bigcup_{x\in A_1} B_\varepsilon(x)$. Then $\mu^*(A_1 \cup A_2) = \mu^*(O_\varepsilon \cap (A_1 \cup A_2)) + \mu^*(O'_\varepsilon \cap (A_1 \cup A_2)) = \mu^*(A_1) + \mu^*(A_2)$ as required. □

Example 8.18. For example the Lebesgue outer measure (8.4) is a metric outer measure. In fact, given two sets A_1, A_2 as in the definition note that any cover of rectangles can be assumed to have only rectangles of side length smaller than $n^{-1/2} \operatorname{dist}(A_1, A_2)$ by taking subdivisions (here we use that $|.|$

is additive on rectangles). Hence every rectangle in the cover can intersect at most on of the two sets and we can partition the cover into two covers, on covering A_1 and the other covering A_2. From this we get $\mu^*(A_1 \cup A_2) \geq \mu^*(A_1) + \mu^*(A_2)$ and the other direction follows from subadditivity.

Note that at first sight it might look like this approach avoids the intermediate step of showing that $|.|$ is a premeasure. However, this is not quite true since you need this in order to show $\lambda^{n,*}(R) = |R|$ for rectangles $R \in \mathcal{S}^n$. ◇

Problem* 8.9. *Suppose $\mathcal{S}_1$, $\mathcal{S}_2$ are semialgebras in X_1, X_2. Then $\mathcal{S} := \mathcal{S}_1 \otimes \mathcal{S}_2 := \{A_1 \times A_2 | A_j \in \mathcal{S}_j\}$ is a semialgebra in $X := X_1 \times X_2$.*

Problem* 8.10. *Show that μ^* defined in* (8.11) *extends μ. (Hint: For the cover A_n it is no restriction to assume $A_n \cap A_m = \emptyset$ and $A_n \subseteq A$.)*

Problem* 8.11. *Show that every null set satisfies the Carathéodory condition* (8.12)*. Conclude that the measure constructed in Theorem 8.9 is complete.*

Problem* 8.12 (Completion of a measure)**.** *Show that every measure has an extension which is complete as follows:*

Denote by $\mathcal{N}$ the collection of subsets of X which are subsets of sets of measure zero. Define $\bar{\Sigma} := \{A \cup N | A \in \Sigma, N \in \mathcal{N}\}$ and $\bar{\mu}(A \cup N) := \mu(A)$ for $A \cup N \in \bar{\Sigma}$.

Show that $\bar{\Sigma}$ is a σ-algebra and that $\bar{\mu}$ is a well-defined complete measure. Moreover, show $\mathcal{N} = \{N \in \bar{\Sigma} | \bar{\mu}(N) = 0\}$ and $\bar{\Sigma} = \{B \subseteq X | \exists A_1, A_2 \in \Sigma$ with $A_1 \subseteq B \subseteq A_2$ and $\mu(A_2 \setminus A_1) = 0\}$.

Problem 8.13. *Let μ be a finite measure. Show that*

$$d(A, B) := \mu(A \Delta B), \qquad A \Delta B := (A \cup B) \setminus (A \cap B) \tag{8.14}$$

is a metric on Σ if we identify sets differing by sets of measure zero. Show that if $\mathcal{A}$ is an algebra, then it is dense in $\Sigma(\mathcal{A})$. (Hint: Show that the sets which can be approximated by sets in $\mathcal{A}$ form a Dynkin system.)

8.4. Borel measures

In this section we want to construct a large class of important measures on $\mathbb{R}^n$. We begin with a few abstract definitions.

Let X be a topological space. A measure on the Borel σ-algebra is called a **Borel measure** if $\mu(K) < \infty$ for every compact set K. Note that some authors do not require this last condition.

Example 8.19. Let $X := \mathbb{R}$ and $\Sigma := \mathfrak{B}$. The Dirac measure is a Borel measure. The counting measure is no Borel measure since $\mu([a, b]) = \infty$ for $a < b$. ◇

A measure on the Borel σ-algebra is called **outer regular** if

$$\mu(A) = \inf_{O \supseteq A, O \text{ open}} \mu(O) \tag{8.15}$$

and **inner regular** if

$$\mu(A) = \sup_{K \subseteq A, K \text{ compact}} \mu(K). \tag{8.16}$$

It is called **regular** if it is both outer and inner regular.

Example 8.20. Let $X := \mathbb{R}$ and $\Sigma := \mathfrak{B}$. The counting measure is inner regular but not outer regular (every nonempty open set has infinite measure). The Dirac measure is a regular Borel measure. ◇

Example 8.21. Let $X := \mathbb{R}^n$ and $\Sigma := \mathfrak{B}^n$. Then lebesgue measure λ^n is outer regular as can be seen from (8.4) be replacing the rectangles by slightly larger open ones. In fact we will show that it is regular below. ◇

A set $A \in \Sigma$ is called a **support** for μ if $\mu(X \setminus A) = 0$. Note that a support is not unique (see the examples below). If X is a topological space and $\Sigma = \mathfrak{B}(X)$, one defines **the support** (also **topological support**) of μ via

$$\operatorname{supp}(\mu) := \{x \in X | \mu(O) > 0 \text{ for every open neighborhood } O \text{ of } x\}. \tag{8.17}$$

Equivalently one obtains $\operatorname{supp}(\mu)$ by removing all points which have an open neighborhood of measure zero. In particular, this shows that $\operatorname{supp}(\mu)$ is closed. If X is second countable, then $\operatorname{supp}(\mu)$ is indeed a support for μ: For every point $x \notin \operatorname{supp}(\mu)$ let O_x be an open neighborhood of measure zero. These sets cover $X \setminus \operatorname{supp}(\mu)$ and by the Lindelöf theorem there is a countable subcover, which shows that $X \setminus \operatorname{supp}(\mu)$ has measure zero.

Example 8.22. Let $X := \mathbb{R}$, $\Sigma := \mathfrak{B}$. The support of the Lebesgue measure λ is all of $\mathbb{R}$. However, every single point has Lebesgue measure zero and so has every countable union of points (by σ-additivity). Hence any set whose complement is countable is a support. There are even uncountable sets of Lebesgue measure zero (see the Cantor set below) and hence a support might even lack an uncountable number of points.

The support of the Dirac measure centered at 0 is the single point 0. Any set containing 0 is a support of the Dirac measure.

In general, the support of a Borel measure on $\mathbb{R}$ is given by

$$\operatorname{supp}(d\mu) = \{x \in \mathbb{R} | \mu(x - \varepsilon) < \mu(x + \varepsilon), \forall \varepsilon > 0\}.$$

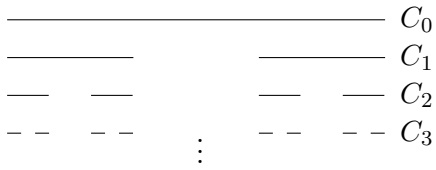
Here we have used $d\mu$ to emphasize that we are interested in the support of the measure $d\mu$ which is different from the support of its distribution function $\mu(x)$. ◇

A property is said to hold **μ-almost everywhere** (a.e.) if it holds on a support for μ or, equivalently, if the set where it does not hold is contained in a set of measure zero.

Example 8.23. The set of rational numbers is countable and hence has Lebesgue measure zero, $\lambda(\mathbb{Q}) = 0$. So, for example, the characteristic function of the rationals $\mathbb{Q}$ is zero almost everywhere with respect to Lebesgue measure.

Any function which vanishes at 0 is zero almost everywhere with respect to the Dirac measure centered at 0. ◇

Example 8.24. The **Cantor set** is an example of a closed uncountable set of Lebesgue measure zero. It is constructed as follows: Start with $C_0 := [0, 1]$ and remove the middle third to obtain $C_1 := [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Next, again remove the middle third's of the remaining sets to obtain $C_2 := [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$:



Proceeding like this, we obtain a sequence of nesting sets C_n and the limit $C := \bigcap_n C_n$ is the Cantor set. Since C_n is compact, so is C. Moreover, C_n consists of 2^n intervals of length 3^{-n}, and thus its Lebesgue measure is $\lambda(C_n) = (2/3)^n$. In particular, $\lambda(C) = \lim_{n\to\infty} \lambda(C_n) = 0$. Using the ternary expansion, it is extremely simple to describe: C is the set of all $x \in [0, 1]$ whose ternary expansion contains no one's, which shows that C is uncountable (why?). It has some further interesting properties: it is totally disconnected (i.e., it contains no subintervals) and perfect (it has no isolated points). Concerning the first note that the length of the subintervals forming C_n go to zero, and concerning the second note that a given point $x \in C$ is in a unique subinterval of C_n and hence we can find a sequence by choosing one of the boundary points of the corresponding subinterval.

Finally, note that if we change the construction by removing $\frac{1}{4^n}$ from the middle of the intervals in the nth step almost everything works as before. We get a sequence of closed sets V_n consisting of 2^n subintervals of length $2^{-2n-1}(1 + 2^n)$. The limiting set V is totally disconnected, perfect and compact but now has Lebesgue measure $\frac{1}{2}$. It is known as **Smith–Volterra–Cantor set** or **fat Cantor set**. ◇

But how can we obtain some more interesting Borel measures? We will restrict ourselves to the case of $X = \mathbb{R}^n$ and begin with the case of Borel measures on $X = \mathbb{R}$ which are also known as **Lebesgue–Stieltjes measures**. By what we have seen so far it is clear that our strategy is as

follows: Start with some simple sets and then work your way up to all Borel sets. Hence let us first show how we should define μ for intervals: To every Borel measure on $\mathfrak{B}$ we can assign its **distribution function**

$$\mu(x) := \begin{cases} -\mu((x,0]), & x < 0, \\ 0, & x = 0, \\ \mu((0,x]), & x > 0, \end{cases} \tag{8.18}$$

which is right continuous and nondecreasing as can be easily checked.

Example 8.25. The distribution function of the Dirac measure centered at 0 is

$$\mu(x) := \begin{cases} 0, & x \geq 0, \\ -1, & x < 0. \end{cases} \quad \diamond$$

For a finite measure the alternate normalization $\tilde{\mu}(x) = \mu((-\infty, x])$ can be used. The resulting distribution function differs from our above definition by a constant $\mu(x) = \tilde{\mu}(x) - \mu((-\infty, 0])$. In particular, this is the normalization used for probability measures.

Conversely, to obtain a measure from a nondecreasing function $m : \mathbb{R} \to \mathbb{R}$ we proceed as follows: Recall that an interval is a subset of the real line of the form

$$I = (a,b], \quad I = [a,b], \quad I = (a,b), \quad \text{or} \quad I = [a,b), \tag{8.19}$$

with $a \leq b$, $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$. Note that (a,a), $[a,a)$, and $(a,a]$ denote the empty set, whereas $[a,a]$ denotes the singleton $\{a\}$. For any proper interval with different endpoints (i.e. $a < b$) we can define its measure to be

$$\mu(I) := \begin{cases} m(b+) - m(a+), & I = (a,b], \\ m(b+) - m(a-), & I = [a,b], \\ m(b-) - m(a+), & I = (a,b), \\ m(b-) - m(a-), & I = [a,b), \end{cases} \tag{8.20}$$

where $m(a\pm) = \lim_{\varepsilon \downarrow 0} m(a \pm \varepsilon)$ (which exist by monotonicity). If one of the endpoints is infinite we agree to use $m(\pm\infty) = \lim_{x \to \pm\infty} m(x)$. For the empty set we of course set $\mu(\emptyset) = 0$ and for the singletons we set

$$\mu(\{a\}) := m(a+) - m(a-) \tag{8.21}$$

(which agrees with (8.20) except for the case $I = (a,a)$ which would give a negative value for the empty set if μ jumps at a). Note that $\mu(\{a\}) = 0$ if and only if $m(x)$ is continuous at a and that there can be only countably many points with $\mu(\{a\}) > 0$ since a nondecreasing function can have at most countably many jumps. Moreover, observe that the definition of μ does not involve the actual value of m at a jump. Hence any function $\tilde{m}$ with $m(x-) \leq \tilde{m}(x) \leq m(x+)$ gives rise to the same μ. We will frequently

assume that m is right continuous such that it coincides with the distribution function up to a constant, $\mu(x) = m(x+) - m(0+)$. In particular, μ determines m up to a constant and the value at the jumps.

Once we have defined μ on $\mathcal{S}^1$ we can now show that (8.20) gives a premeasure.

Lemma 8.12. *Let $m : \mathbb{R} \to \mathbb{R}$ be right continuous and nondecreasing. Then the set function defined via*

$$\mu((a, b]) := m(b) - m(a) \tag{8.22}$$

on $\mathcal{S}^1$ gives rise to a unique σ-finite premeasure on the algebra $\bar{\mathcal{S}}^1$ of finite unions of disjoint half-open intervals.

Proof. If $(a, b] = \bigcup\!\!\!\cdot_{j=1}^{n}(a_j, b_j]$ then we can assume that the a_j's are ordered. Moreover, in this case we must have $b_j = a_{j+1}$ for $1 \le j < n$ and hence our set function is additive on $\mathcal{S}^1$: $\sum_{j=1}^{n} \mu((a_j, b_j]) = \sum_{j=1}^{n}(\mu(b_j) - \mu(a_j)) = \mu(b_n) - \mu(a_1) = \mu(b) - \mu(a) = \mu((a, b])$.

So by Lemma 8.6 it remains to verify (8.10). By right continuity we can cover each $A_j := (a_j, b_j]$ by some slightly larger interval $B_j := (a_j, b_j + \delta_j]$ such that $\mu(B_j) \le \mu(A_j) + \frac{\varepsilon}{2^j}$. Then for any $x > 0$ we can find an n such that the open intervals $\{(a_j, b_j + \delta_j)\}_{j=1}^{n}$ cover the compact set $\overline{A} \cap [-x, x]$ and hence

$$\mu(A \cap (-x, x]) \le \mu(\bigcup_{j=1}^{n} B_j) = \sum_{j=1}^{n} \mu(B_j) \le \sum_{j=1}^{\infty} \mu(A_j) + \varepsilon.$$

Letting $x \to \infty$ and since $\varepsilon > 0$ is arbitrary, we are done. □

And extending this premeasure to a measure we finally obtain:

Theorem 8.13. *For every nondecreasing function $m : \mathbb{R} \to \mathbb{R}$ there exists a unique regular Borel measure μ which extends* (8.20)*. Two different functions generate the same measure if and only if the difference is a constant away from the discontinuities.*

Proof. Except for regularity existence follows from Theorem 8.9 together with Lemma 8.10 and uniqueness follows from Corollary 8.5. Regularity will be postponed until Section 8.6. □

We remark, that in the previous theorem we could as well consider $m : (a, b) \to \mathbb{R}$ to obtain regular Borel measures on (a, b).

Example 8.26. Suppose $\Theta(x) := 0$ for $x < 0$ and $\Theta(x) := 1$ for $x \ge 0$. Then we obtain the so-called **Dirac measure** at 0, which is given by $\Theta(A) = 1$ if $0 \in A$ and $\Theta(A) = 0$ if $0 \notin A$. ⋄

Example 8.27. Suppose $\lambda(x) := x$. Then the associated measure is the ordinary **Lebesgue measure** on $\mathbb{R}$. We will abbreviate the Lebesgue measure of a Borel set A by $\lambda(A) = |A|$. ◇

Finally, we show how to extend Theorem 8.13 to $\mathbb{R}^n$. We will write $x \le y$ if $x_j \le y_j$ for $1 \le j \le n$ and $(a, b) = (a_1, b_1) \times \cdots \times (a_n, b_n)$, $(a, b] = (a_1, b_1] \times \cdots \times (a_n, b_n]$, etc.

The analog of (8.18) is given by

$$\mu(x) := \operatorname{sign}(x)\mu\Big(\bigtimes_{j=1}^{n} \big(\min(0, x_j), \max(0, x_j)\big]\Big), \tag{8.23}$$

where $\operatorname{sign}(x) = \prod_{j=1}^{n} \operatorname{sign}(x_j)$.

Example 8.28. The distribution function of the Dirac measure $\mu = \delta_0$ centered at 0 is

$$\mu(x) := \begin{cases} 0, & x \ge 0, \\ -1, & \text{else.} \end{cases}$$ ◇

Again, for a finite measure the alternative normalization $\tilde{\mu}(x) = \mu((-\infty, x])$ can be used.

To recover a measure μ from its distribution function we consider the difference with respect to the j'th coordinate

$$\begin{aligned} \Delta^j_{a^1,a^2} m(x) &:= m(x_1, \dots, x_{j-1}, a^2, x_{j+1}, \dots, x_n) \\ &\quad - m(x_1, \dots, x_{j-1}, a^1, x_{j+1}, \dots, x_n) \\ &= \sum_{j \in \{1,2\}} (-1)^j m(x_1, \dots, x_{j-1}, a^j, x_{j+1}, \dots, x_n) \end{aligned} \tag{8.24}$$

and define

$$\begin{aligned} \Delta_{a^1,a^2} m &:= \Delta^1_{a^1_1,a^2_1} \cdots \Delta^n_{a^1_n,a^2_n} m(x) \\ &= \sum_{j \in \{1,2\}^n} (-1)^{j_1} \cdots (-1)^{j_n} m(a^{j_1}_1, \dots, a^{j_n}_n). \end{aligned} \tag{8.25}$$

Note that the above sum is taken over all vertices of the rectangle $(a^1, a^2]$ weighted with $+1$ if the vertex contains an even number of left endpoints and weighted with -1 if the vertex contains an odd number of left endpoints.

Then

$$\mu((a, b]) = \Delta_{a,b}\mu. \tag{8.26}$$

Of course in the case $n = 1$ this reduces to $\mu((a, b]) = \mu(b) - \mu(a)$. In the case $n = 2$ we have $\mu((a, b]) = \mu(b_1, b_2) - \mu(b_1, a_2) - \mu(a_1, b_2) + \mu(a_1, a_2)$ which (for $0 \le a \le b$) is the measure of the rectangle with corners 0, b, minus the measure of

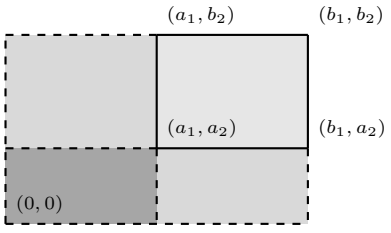


Figure 8.1. A partition and its regular refinement

the rectangle on the left with corners 0, (a_1, b_2), minus the measure of the rectangle below with corners 0, (b_1, a_2), plus the measure of the rectangle with corners 0, a which has been subtracted twice. The general case can be handled recursively (Problem 8.14).

Hence we will again assume that a nondecreasing function $m : \mathbb{R}^n \to \mathbb{R}^n$ is given (i.e. $m(a) \le m(b)$ for $a \le b$). However, this time monotonicity is not enough as the following example shows.

Example 8.29. Let $\mu := \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(1,0)} - \frac{1}{2}\delta_{(1,1)}$. Then the corresponding distribution function is increasing in each coordinate direction as the decrease due to the last term is compensated by the other two. However, (8.23) will give $-\frac{1}{2}$ for any rectangle containing $(1, 1)$ but not the other two points $(1, 0)$ and $(0, 1)$. ⋄

Now we can show how to get a premeasure on $\bar{\mathcal{S}}^n$.

Lemma 8.14. *Let $m : \mathbb{R}^n \to \mathbb{R}^n$ be right continuous such that μ defined via $\mu((a, b]) = \Delta_{a,b} m$ is a nonnegative set function. Then μ gives rise to a unique σ-finite premeasure on the algebra $\bar{\mathcal{S}}^n$ of finite unions of disjoint half-open rectangles.*

Proof. We first need to show finite additivity. We call $A = \bigcup\!\!\!\cdot_k A_k$ a regular partition of $A = (a, b]$ if there are sequences $a_j = c_{j,0} < c_{j,1} < \cdots < c_{j,m_j} = b_j$ such that each rectangle A_k is of the form

$$(c_{1,i-1}, c_{1,i}] \times \cdots \times (c_{n,i-1}, c_{n,i}].$$

That is, the sets A_k are obtained by intersecting A with the hyperplanes $x_j = c_{j,i}$, $1 \le i \le m_j - 1$, $1 \le j \le n$. Now let A be bounded. Then additivity holds when partitioning A into two sets by one hyperplane $x_j = c_{j,i}$ (note that in the sum over all vertices, the one containing $c_{j,i}$ instead of a_j cancels with the one containing $c_{j,i}$ instead of b_j as both have opposite signs by the very definition of $\Delta_{a,b} m$). Hence applying this case recursively shows that additivity holds for regular partitions. Finally, for every partition we have a corresponding regular subpartition. Moreover, the sets in this subpartions can be lumped together into regular subpartions for each of the sets in the

original partition. Hence the general case follows from the regular case. Finally, the case of unbounded sets A follows by taking limits.

The rest follows verbatim as in the previous lemma. □

Again this premeasure gives rise to a measure.

Theorem 8.15. *For every right continuous function $m : \mathbb{R}^n \to \mathbb{R}$ such that*

$$\mu((a,b]) := \Delta_{a,b} m \geq 0, \quad \forall a \leq b, \tag{8.27}$$

there exists a unique regular Borel measure μ which extends the above definition.

Proof. As in the one-dimensional case existence follows from Theorem 8.9 together with Lemma 8.10 and uniqueness follows from Corollary 8.5. Again regularity will be postponed until Section 8.6. □

Example 8.30. Choosing $m(x) := \prod_{j=1}^n x_j$ we obtain the Lebesgue measure λ^n in $\mathbb{R}^n$, which is the unique Borel measure satisfying

$$\lambda^n((a,b]) = \prod_{j=1}^n (b_j - a_j).$$

We collect some simple properties below.

(i) λ^n is regular.

(ii) λ^n is uniquely defined by its values on $\mathcal{S}^n$.

(iii) For every measurable set we have

$$\lambda^n(A) = \inf \Big\{ \sum_{m=1}^{\infty} \lambda^n(A_m) \Big| A \subseteq \bigcup_{m=1}^{\infty} A_m,\ A_m \in \mathcal{S}^n \Big\}$$

where the infimum extends over all countable disjoint covers.

(iv) λ^n is translation invariant and up to normalization the only Borel measure with this property.

(i) and (ii) are part of Theorem 8.15. (iii). This will follow from the construction of λ^n via its outer measure in the following section. (iv). The previous item implies that λ^n is translation invariant. Moreover, let μ be a second translation invariant measure. Denote by Q_r a cube with side length $r > 0$. Without loss we can assume $\mu(Q_1) = 1$. Since we can split Q_1 into m^n cubes of side length $1/m$, we see that $\mu(Q_{1/m}) = m^{-n}$ by translation invariance and additivity. Hence we obtain $\mu(Q_r) = r^n$ for every rational r and thus for every r by continuity from below. Proceeding like this we see that λ^n and μ coincide on $\mathcal{S}^n$ and equality follows from item (ii). ◇

Example 8.31. If $m_j : \mathbb{R} \to \mathbb{R}$ are nondecreasing right continuous functions, then $m(x) := \prod_{j=1}^n m_j(x_j)$ satisfies

$$\Delta_{a,b} m = \prod_{j=1}^n \big(m_j(b_j) - m_j(a_j)\big)$$

and hence the requirements of Theorem 8.15 are fulfilled. ◇

Problem* 8.14. *Let μ be a Borel measure on $\mathbb{R}^n$. For $a, b \in \mathbb{R}^n$ set*

$$m(a,b) := \operatorname{sign}(b-a)\mu\Big(\bigtimes_{j=1}^n \big(\min(a_j,b_j), \max(a_j,b_j)\big]\Big)$$

and $m(x) := m(0,x)$. In particular, for $a \le b$ we have $m(a,b) = \mu((a,b])$. Show that

$$m(a,b) = \Delta_{a,b} m(c,\cdot)$$

for arbitrary $c \in \mathbb{R}^n$. (Hint: Start with evaluating $\Delta^j_{a_j,b_j} m(c,\cdot)$.)

Problem* 8.15. *Let μ be a premeasure such that outer regularity* (8.15) *holds for every set in the algebra. Then the corresponding measure μ from Theorem 8.9 is outer regular.*

8.5. Measurable functions

The Riemann integral works by splitting the x coordinate into small intervals and approximating $f(x)$ on each interval by its minimum and maximum. The problem with this approach is that the difference between maximum and minimum will only tend to zero (as the intervals get smaller) if $f(x)$ is sufficiently nice. To avoid this problem, we can force the difference to go to zero by considering, instead of an interval, the set of x for which $f(x)$ lies between two given numbers $a < b$. Now we need the size of the set of these x, that is, the size of the preimage $f^{-1}((a,b))$. For this to work, preimages of intervals must be measurable.

Let (X, Σ_X) and (Y, Σ_Y) be measurable spaces. A function $f : X \to Y$ is called **measurable** if $f^{-1}(A) \in \Sigma_X$ for every $A \in \Sigma_Y$. When checking this condition it is useful to note that the collection of sets for which it holds, $\{A \subseteq Y | f^{-1}(A) \in \Sigma_X\}$, forms a σ-algebra on Y by $f^{-1}(Y \setminus A) = X \setminus f^{-1}(A)$ and $f^{-1}(\bigcup_j A_j) = \bigcup_j f^{-1}(A_j)$. Hence it suffices to check this condition for every set A in a collection of sets which generates Σ_Y.

We will be mainly interested in the case where $(Y, \Sigma_Y) = (\mathbb{R}^n, \mathfrak{B}^n)$.

Lemma 8.16. *A function $f : X \to \mathbb{R}^n$ is measurable if and only if*

$$f^{-1}((a,\infty)) \in \Sigma \qquad \forall\, a \in \mathbb{R}^n, \tag{8.28}$$

where $(a,\infty) := \bigtimes_{j=1}^n (a_j,\infty)$. *In particular, a function* $f : X \to \mathbb{R}^n$ *is measurable if and only if every component is measurable. Regarding* $\mathbb{C} \cong \mathbb{R}^2$ *we also get that a complex-valued function* $f : X \to \mathbb{C}$ *is measurable if and only if both its real and imaginary parts are and similarly for* $\mathbb{C}^n$*-valued functions.*

Proof. We need to show that $\mathfrak{B}^n$ is generated by rectangles of the above form. The σ-algebra generated by these rectangles also contains all open rectangles of the form $(a,b) := \bigtimes_{j=1}^n (a_j,b_j)$, which form a base for the topology. □

Clearly the intervals (a,∞) can also be replaced by $[a,\infty)$, $(-\infty,a)$, or $(-\infty,a]$.

If X is a topological space and Σ the corresponding Borel σ-algebra, we will also call a measurable function **Borel function**. Note that, in particular,

Lemma 8.17. *Let* (X,Σ_X), (Y,Σ_Y), (Z,Σ_Z) *be topological spaces with their corresponding Borel* σ*-algebras. Any continuous function* $f : X \to Y$ *is measurable. Moreover, if* $f : X \to Y$ *and* $g : Y \to Z$ *are measurable functions, then the composition* $g \circ f$ *is again measurable.*

Warning: Some authors call a function $f : \mathbb{R} \to \mathbb{R}$ Lebesgue measurable if $f^{-1}((a,\infty))$ is a Lebesgue measurable set for every $a \in \mathbb{R}$ (with respect to the complete Lebesgue measure; cf. Problem 8.11). While this class is larger than the Borel functions (consider the characteristic function of a Lebesgue measurable set which is not Borel), it has the disadvantage that the composition of Lebesgue measurable functions might not be Lebesgue measurable. By restricting ourselves to Borel functions we avoid this nuisance.

The set of all measurable functions forms an algebra.

Lemma 8.18. *Let* X *be a topological space and* Σ *its Borel* σ*-algebra. Suppose* $f, g : X \to \mathbb{R}$ *are measurable functions. Then the sum* $f+g$ *and the product* fg *are measurable. Similarly for complex-valued functions.*

Proof. Note that addition and multiplication are continuous functions from $\mathbb{R}^2 \to \mathbb{R}$ and hence the claim follows from the previous lemma. □

Sometimes it is also convenient to allow $\pm\infty$ as possible values for f, that is, functions $f : X \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty,\infty\}$. In this case $A \subseteq \overline{\mathbb{R}}$ is called Borel if $A \cap \mathbb{R}$ is. This implies that $f : X \to \overline{\mathbb{R}}$ will be Borel if and only if $f^{-1}(\pm\infty)$ are Borel and $f : X \setminus f^{-1}(\{-\infty,\infty\}) \to \mathbb{R}$ is Borel. Since

$$\{+\infty\} = \bigcap_n (n,+\infty], \qquad \{-\infty\} = \overline{\mathbb{R}} \setminus \bigcup_n (-n,+\infty], \tag{8.29}$$

we see that $f : X \to \overline{\mathbb{R}}$ is measurable if and only if

$$f^{-1}((a,\infty]) \in \Sigma \qquad \forall\, a \in \mathbb{R}. \tag{8.30}$$

Again the intervals $(a,\infty]$ can also be replaced by $[a,\infty]$, $[-\infty,a)$, or $[-\infty,a]$. Moreover, we can generate a corresponding topology on $\overline{\mathbb{R}}$ by intervals of the form $[-\infty,b)$, (a,b), and $(a,\infty]$ with $a,b \in \mathbb{R}$.

Hence it is not hard to check that the previous lemma still holds if one either avoids undefined expressions of the type $\infty-\infty$ and $\pm\infty\cdot 0$ or makes a definite choice, e.g., $\infty-\infty=0$ and $\pm\infty\cdot 0=0$.

Moreover, the set of all measurable functions is closed under all important limiting operations.

Lemma 8.19. *Suppose $f_n : X \to \overline{\mathbb{R}}$ is a sequence of measurable functions. Then*

$$\inf_{n\in\mathbb{N}} f_n, \quad \sup_{n\in\mathbb{N}} f_n, \quad \liminf_{n\to\infty} f_n, \quad \limsup_{n\to\infty} f_n \tag{8.31}$$

are measurable as well.

Proof. It suffices to prove that $\sup f_n$ is measurable since the rest follows from $\inf f_n = -\sup(-f_n)$, $\liminf f_n := \sup_n \inf_{k\geq n} f_k$, and $\limsup f_n := \inf_n \sup_{k\geq n} f_k$. But $(\sup f_n)^{-1}((a,\infty]) = \bigcup_n f_n^{-1}((a,\infty])$ are measurable and we are done. □

A few immediate consequences are worthwhile noting: It follows that if f and g are measurable functions, so are $\min(f,g)$, $\max(f,g)$, $|f| = \max(f,-f)$, and $f^{\pm} = \max(\pm f, 0)$. Furthermore, the pointwise limit of measurable functions is again measurable. Moreover, the set where the limit exists,

$$\{x \in X | \lim_{n\to\infty} f(x) \text{ exists}\} = \{x \in X | \limsup_{n\to\infty} f(x) - \liminf_{n\to\infty} f(x) = 0\}, \tag{8.32}$$

is measurable.

Sometimes the case of arbitrary suprema and infima is also of interest. In this respect the following observation is useful: Let X be a topological space. Recall that a function $f : X \to \overline{\mathbb{R}}$ is **lower semicontinuous** if the set $f^{-1}((a,\infty])$ is open for every $a \in \mathbb{R}$. Then it follows from the definition that the sup over an arbitrary collection of lower semicontinuous functions

$$\overline{f}(x) := \sup_\alpha f_\alpha(x) \tag{8.33}$$

is again lower semicontinuous (and hence measurable). Similarly, f is **upper semicontinuous** if the set $f^{-1}([-\infty,a))$ is open for every $a \in \mathbb{R}$. In this case the infimum

$$\underline{f}(x) := \inf_\alpha f_\alpha(x) \tag{8.34}$$

is again upper semicontinuous. Note that f is lower semicontinuous if and only if $-f$ is upper semicontinuous.

Problem 8.16 (preimage σ-algebra). *Let $\mathcal{S} \subseteq \mathfrak{P}(Y)$. Show that $f^{-1}(\mathcal{S}) := \{f^{-1}(A) | A \in \mathcal{S}\}$ is a σ-algebra if $\mathcal{S}$ is. Conclude that $f^{-1}(\Sigma_Y)$ is the smallest σ-algebra on X for which f is measurable.*

Problem 8.17. *Suppose $f : X \to \overline{\mathbb{R}}$ is measurable. Show that $\frac{1}{f}$ is also measurable.*

Problem 8.18. *Let $\{A_n\}_{n\in\mathbb{N}}$ be a partition for X, $X = \biguplus_{n\in\mathbb{N}} A_n$. Let $\Sigma = \Sigma(\{A_n\}_{n\in\mathbb{N}})$ be the σ-algebra generated by these sets. Show that $f : X \to \mathbb{R}$ is measurable if and only if it is constant on the sets A_n.*

Problem* 8.19. *Show that the supremum over lower semicontinuous functions is again lower semicontinuous.*

Problem* 8.20. *Let X be a topological space and $f : X \to \overline{\mathbb{R}}$. Show that f is lower semicontinuous if and only if*

$$\liminf_{x\to x_0} f(x) \geq f(x_0), \qquad x_0 \in X.$$

Similarly, f is upper semicontinuous if and only if

$$\limsup_{x\to x_0} f(x) \leq f(x_0), \qquad x_0 \in X.$$

Show that a lower semicontinuous function is also sequentially lower semicontinuous

$$\liminf_{n\to\infty} f(x_n) \geq f(x_0), \qquad x_n \to x_0,\ x_0 \in X.$$

Show that the converse is also true if X is a metric space. (Hint: Problem B.15.)

8.6. How wild are measurable objects

In this section we want to investigate how far measurable objects are away from well-understood ones. The situation is intuitively summarized in what is known as **Littlewood's three principles of real analysis**:

- Every (measurable) set is nearly a finite union of intervals.
- Every (measurable) function is nearly continuous.
- Every convergent sequence of (measurable) functions is nearly uniformly convergent.

As our first task we want to look at the first item and show that measurable sets can be well approximated by using closed sets from the inside and open sets from the outside in nice spaces like $\mathbb{R}^n$.

Lemma 8.20. *Let X be a metric space and μ a finite Borel measure. Then for every $A \in \mathfrak{B}(X)$ and any given $\varepsilon > 0$ there exists an open set O and a closed set C such that*

$$C \subseteq A \subseteq O \quad \text{and} \quad \mu(O \setminus C) \leq \varepsilon. \tag{8.35}$$

The same conclusion holds for arbitrary Borel measures if there is a sequence of open sets $U_n \nearrow X$ such that $\overline{U}_n \subseteq U_{n+1}$ and $\mu(U_n) < \infty$ (note that μ is also σ-finite in this case).

Proof. To see that (8.35) holds we begin with the case when μ is finite. Denote by $\mathcal{A}$ the set of all Borel sets satisfying (8.35). Then $\mathcal{A}$ contains every closed set C: Given C define $O_n := \{x \in X | d(x, C) < 1/n\}$ and note that O_n are open sets which satisfy $O_n \searrow C$. Thus by Theorem 8.1 (iii) $\mu(O_n \setminus C) \to 0$ and hence $C \in \mathcal{A}$.

Moreover, $\mathcal{A}$ is even a σ-algebra. That it is closed under complements is easy to see (note that $\tilde{O} := X \setminus C$ and $\tilde{C} := X \setminus O$ are the required sets for $\tilde{A} = X \setminus A$). To see that it is closed under countable unions consider $A = \bigcup_{n=1}^{\infty} A_n$ with $A_n \in \mathcal{A}$. Then there are C_n, O_n such that $\mu(O_n \setminus C_n) \leq \varepsilon 2^{-n-1}$. Now $O := \bigcup_{n=1}^{\infty} O_n$ is open and $C := \bigcup_{n=1}^{N} C_n$ is closed for any finite N. Since $\mu(A)$ is finite we can choose N sufficiently large such that $\mu(\bigcup_{N+1}^{\infty} C_n \setminus C) \leq \varepsilon/2$. Then we have found two sets of the required type: $\mu(O \setminus C) \leq \sum_{n=1}^{\infty} \mu(O_n \setminus C_n) + \mu(\bigcup_{n=N+1}^{\infty} C_n \setminus C) \leq \varepsilon$. Thus $\mathcal{A}$ is a σ-algebra containing the open sets, hence it is the entire Borel σ-algebra.

Now suppose μ is not finite. Set $X_1 := U_2$ and $X_n := U_{n+1} \setminus \overline{U_{n-1}}$, $n \geq 2$. Note that $X_{n+1} \cap X_n = U_{n+1} \setminus \overline{U_n}$ and $X_n \cap X_m = \emptyset$ for $|n - m| > 1$. Let $A_n = A \cap X_n$ and note that $A = \bigcup_{n=1}^{\infty} A_n$. By the finite case we can choose $C_n \subseteq A_n \subseteq O_n \subseteq X_n$ such that $\mu(O_n \setminus C_n) \leq \varepsilon 2^{-n}$. Now set $C := \bigcup_n C_n$ and $O := \bigcup_n O_n$ and observe that C is closed. Indeed, let $x \in \overline{C}$ and let x_j be some sequence from C converging to x. Then $x \in U_n$ for some n and hence the sequence must eventually lie in $C \cap U_n \subseteq \bigcup_{m \leq n} C_m$. Hence $x \in \overline{\bigcup_{m \leq n} C_m} = \bigcup_{m \leq n} C_m \subseteq C$. Finally, $\mu(O \setminus C) \leq \sum_{n=1}^{\infty} \mu(O_n \setminus C_n) \leq \varepsilon$ as required. □

This result immediately gives us outer regularity.

Corollary 8.21. *Under the assumptions of the previous lemma*

$$\mu(A) = \inf_{O \supseteq A, O \text{ open}} \mu(O) = \sup_{C \subseteq A, C \text{ closed}} \mu(C) \tag{8.36}$$

and μ is outer regular.

Proof. This follows from $\mu(A) = \mu(O) - \mu(O \setminus A) = \mu(C) + \mu(A \setminus C)$. □

If we strengthen our assumptions, we also get inner regularity. In fact, if we assume the sets U_n to be relatively compact, then the assumptions for the second case are equivalent to X being locally compact and separable by Lemma B.25.

Corollary 8.22. *If X is a σ-compact metric space, then every finite Borel measure is regular. If X is a locally compact separable metric space, then every Borel measure is regular.*

Proof. By assumption there is a sequence of compact sets $K_n \nearrow X$ and for every increasing sequence of closed sets C_n with $\mu(C_n) \to \mu(A)$ we also have compact sets $C_n \cap K_n$ with $\mu(C_n \cap K_n) \to \mu(A)$. In the second case we can choose relatively compact open sets U_n as in Lemma B.25 (iv) such that the assumptions of the previous theorem hold. Now argue as before using $K_n = \overline{U_n}$. □

In particular, on a locally compact and separable space every Borel measure is automatically regular and σ-finite. For example this hols for $X = \mathbb{R}^n$ (or $X = \mathbb{C}^n$).

An inner regular measure on a Hausdorff space which is locally finite (every point has a neighborhood of finite measure) is called a **Radon measure**. Accordingly every Borel measure on $\mathbb{R}^n$ (or $\mathbb{C}^n$) is automatically a Radon measure.

Example 8.32. Since Lebesgue measure on $\mathbb{R}$ is regular, we can cover the rational numbers by an open set of arbitrary small measure (it is also not hard to find such a set directly) but we cannot cover it by an open set of measure zero (since any open set contains an interval and hence has positive measure). However, if we slightly extend the family of admissible sets, this will be possible. ⋄

Looking at the Borel σ-algebra the next general sets after open sets are countable intersections of open sets, known as G_δ sets (here G and δ stand for the German words *Gebiet* and *Durchschnitt*, respectively). The next general sets after closed sets are countable unions of closed sets, known as F_σ sets (here F and σ stand for the French words *fermé* and *somme*, respectively). Of course the complement of a G_δ set is an F_σ set and vice versa.

Example 8.33. The irrational numbers are a G_δ set in $\mathbb{R}$ and the rational numbers are an F_σ set.. To see this, let x_n be an enumeration of the rational numbers and consider the intersection of the open sets $O_n := \mathbb{R} \setminus \{x_n\}$. ⋄

Corollary 8.23. *Suppose X is locally compact and separable and μ a Borel measure. A set in X is Borel if and only if it differs from a G_δ set by a*

Borel set of measure zero. Similarly, a set in X is Borel if and only if it differs from an F_σ set by a Borel set of measure zero.

Proof. Since G_δ sets are Borel, only the converse direction is nontrivial. By Lemma 8.20 we can find open sets O_n such that $\mu(O_n \setminus A) \leq 1/n$. Now let $G := \bigcap_n O_n$. Then $\mu(G \setminus A) \leq \mu(O_n \setminus A) \leq 1/n$ for any n and thus $\mu(G \setminus A) = 0$. The second claim is analogous. □

A similar result holds for convergence.

Theorem 8.24 (Egorov)**.** *Let μ be a finite measure and f_n be a sequence of complex-valued measurable functions converging pointwise to a function f for a.e. $x \in X$. Then fore every $\varepsilon > 0$ there is a set A of size $\mu(A) < \varepsilon$ such that f_n converges uniformly on $X \setminus A$.*

Proof. Let A_0 be the set where f_n fails to converge. Set

$$A_{N,k} := \bigcup_{n \geq N} \{x \in X | \, |f_n(x) - f(x)| \geq \frac{1}{k}\}, \qquad A_k := \bigcap_{N \in \mathbb{N}} A_{N,k}$$

and note that $A_{N,k} \searrow A_k \subseteq A_0$ as $N \to \infty$ (with k fixed). Hence $\mu(A_{N_k,k}) \to \mu(A_k) = 0$ by continuity from above. So for every k there is some N_k such that $\mu(A_{N_k,k}) < \frac{\varepsilon}{2^k}$. Then $A := \bigcup_{k \in \mathbb{N}} A_{N_k,k}$ satisfies $\mu(A) < \varepsilon$. Now note that $x \notin A$ implies that $x \notin A_{N_k,k}$ for every k and thus $|f_n(x) - f(x)| < \frac{1}{k}$ for $n \geq N_k$. Thus f_n converges uniformly away from A. □

Example 8.34. The example $f_n := \chi_{[n,n+1]} \to 0$ on $X = \mathbb{R}$ with Lebesgue measure shows that the finiteness assumption is important. In fact, suppose there is a set A of size less than 1 (say). Then every interval $[m, m+1]$ contains a point x_m not in A and thus $|f_m(x_m) - 0| = 1$. ⋄

To end this section let us briefly discuss the third principle, namely that bounded measurable functions can be well approximated by continuous functions (under suitable assumptions on the measure). We will discuss this in detail in Section 10.4. At this point we only mention that in such a situation Egorov's theorem implies that the convergence is uniform away from a small set and hence our original function will be continuous restricted to the complement of this small set. This is known as Luzin's theorem (cf. Theorem 10.17). Note however that this does not imply that measurable functions are continuous at every point of this complement! The characteristic function of the irrational numbers is continuous when restricted to the irrational numbers but it is not continuous at any point when regarded as a function of $\mathbb{R}$.

Problem 8.21. *Show directly (without using regularity) that for every $\varepsilon > 0$ there is an open set O of Lebesgue measure $|O| < \varepsilon$ which covers the rational numbers.*

Problem* 8.22. *A finite Borel measure is regular if and only if for every Borel set A and every $\varepsilon > 0$ there is an open set O and a compact set K such that $K \subseteq A \subseteq O$ and $\mu(O \setminus K) < \varepsilon$.*

Problem 8.23. *A sequence of measurable functions f_n* ***converges in measure*** *to a measurable function f if $\lim_{n\to\infty} \mu(\{x| \, |f_n(x) - f(x)| \geq \varepsilon\}) = 0$ for every $\varepsilon > 0$. Show that if μ is finite and $f_n \to f$ a.e. then $f_n \xrightarrow{\mu} f$ in measure. Show that the finiteness assumption is necessary. Show that the converse fails. (Hint for the last part: Every $n \in \mathbb{N}$ can be uniquely written as $n = 2^m + k$ with $0 \leq m$ and $0 \leq k < 2^m$. Now consider the characteristic functions of the intervals $I_{m,k} := [k2^{-m}, (k+1)2^{-m}]$.)*

8.7. Appendix: Jordan measurable sets

In this short appendix we want to establish the criterion for Jordan measurability alluded to in Section 8.1. We begin with a useful geometric fact.

Lemma 8.25. *Every open set $O \subseteq \mathbb{R}^n$ can be partitioned into a countable number of half-open cubes from $\mathcal{S}^n$.*

Proof. Partition $\mathbb{R}^n$ into cubes of side length one with vertices from $\mathbb{Z}^n$. Start by selecting all cubes which are fully inside O and discard all those which do not intersect O. Subdivide the remaining cubes into 2^n cubes of half the side length and repeat this procedure. This gives a countable set of cubes contained in O. To see that we have covered all of O, let $x \in O$. Since x is an inner point there it will be a δ such that every cube containing x with smaller side length will be fully covered by O. Hence x will be covered at the latest when the side length of the subdivisions drops below δ. □

Now we can establish the connection between the Jordan content and the Lebesgue measure λ^n in $\mathbb{R}^n$.

Theorem 8.26. *Let $A \subseteq \mathbb{R}^n$. We have $J_*(A) = \lambda^n(A^\circ)$ and $J^*(A) = \lambda^n(\overline{A})$. Hence A is Jordan measurable if and only if its boundary $\partial A = \overline{A} \setminus A^\circ$ has Lebesgue measure zero.*

Proof. First of all note, that for the computation of $J_*(A)$ it makes no difference when we take open rectangles instead of half-open ones. But then every R with $R \subseteq A$ will also satisfy $R \subseteq A^\circ$ implying $J_*(A) = J_*(A^\circ)$. Moreover, from Lemma 8.25 we get $J_*(A^\circ) = \lambda^n(A^\circ)$. Similarly, for the computation of $J^*(A)$ it makes no difference when we take closed rectangles and thus $J^*(A) = J^*(\overline{A})$. Next, first assume that $\overline{A}$ is compact. Then given

R, a finite number of slightly larger but open rectangles will give us the same volume up to an arbitrarily small error. Hence $J^*(\overline{A}) = \lambda^n(\overline{A})$ for bounded sets A. The general case can be reduce to this one by splitting A according to $A_m = \{x \in A | m \le |x| \le m+1\}$. □

8.8. Appendix: Equivalent definitions for the outer Lebesgue measure

In this appendix we want to show that the type of sets used in the definition of the outer Lebesgue measure $\lambda^{*,n}$ on $\mathbb{R}^n$ play no role. You can cover the set by half-closed, closed, open rectangles (which is easy to see) or even replace rectangles by balls (which follows from the following lemma). To this end observe that by (8.5)

$$\lambda^n(B_r(x)) = V_n r^n \tag{8.37}$$

where $V_n := \lambda^n(B_1(0))$ is the volume of the unit ball which is computed explicitly in Section 9.3. Will will write $|A| := \lambda^n(A)$ for the Lebesgue measure of a Borel set for brevity. We first establish a covering lemma which is of independent interest.

Lemma 8.27 (Vitali covering lemma)**.** *Let $O \subseteq \mathbb{R}^n$ be an open set and $\delta > 0$ fixed. Let $\mathcal{C}$ be a collection of balls such that every open subset of O contains at least one ball from $\mathcal{C}$. Then there exists a countable set of disjoint open balls from $\mathcal{C}$ of radius at most δ such that $O = N \cup \bigcup_j B_j$ with N a Lebesgue null set.*

Proof. Let O have finite outer measure. Start with all balls which are contained in O and have radius at most δ. Let R be the supremum of the radii of all these balls and take a ball B_1 of radius more than $\frac{R}{2}$. Now consider $O \backslash \overline{B_1}$ and proceed recursively. If this procedure terminates we are done (the missing points must be contained in the boundary of the chosen balls which has measure zero). Otherwise we obtain a sequence of balls B_j whose radii must converge to zero since $\sum_{j=1}^{\infty} |B_j| \le |O|$. Now fix m and let $x \in O \setminus \bigcup_{j=1}^{m} \bar{B}_j$. Then there must be a ball $B_0 = B_r(x) \subseteq O \setminus \bigcup_{j=1}^{m} \bar{B}_j$. Moreover, there must be a first ball B_k with $B_0 \cap B_k \neq \emptyset$ (otherwise all B_k for $k > m$ must have radius larger than $\frac{r}{2}$ violating the fact that they converge to zero). By assumption $k > m$ and hence r must be smaller than two times the radius of B_k (since both balls are available in the k'th step). So the distance of x to the center of B_k must be less then three times the radius of B_k. Now if $\tilde{B}_k$ is a ball with the same center but three times the radius of B_k, then x is contained in $\tilde{B}_k$ and hence all missing points from $\bigcup_{j=1}^{m} B_j$ are either boundary points (which are of measure zero) or contained in $\bigcup_{k>m} \tilde{B}_k$ whose measure $|\bigcup_{k>m} \tilde{B}_k| \le 3^n \sum_{k>m} |B_k| \to 0$ as $m \to \infty$.

If $|O| = \infty$ consider $O_m = O \cap (B_{m+1}(0) \setminus \bar{B}_m(0))$ and note that $O = N \cup \bigcup_m O_m$ where N is a set of measure zero. □

Note that in the one-dimensional case open balls are open intervals and we have the stronger result that every open set can be written as a countable union of disjoint intervals (Problem B.20).

Now observe that in the definition of outer Lebesgue measure we could replace half-open rectangles by open rectangles (show this). Moreover, every open rectangle can be replaced by a disjoint union of open balls up to a set of measure zero by the Vitali covering lemma. Consequently, the Lebesgue outer measure can be written as

$$\lambda^{n,*}(A) = \inf\Big\{ \sum_{k=1}^{\infty} |A_k| \Big| A \subseteq \bigcup_{k=1}^{\infty} A_k,\ A_k \in \mathcal{C} \Big\}, \tag{8.38}$$

where $\mathcal{C}$ could be the collection of all closed rectangles, half-open rectangles, open rectangles, closed balls, or open balls.

Chapter 9

Integration

Now that we know how to measure sets, we are able to introduce the Lebesgue integral. As already mentioned, in the case of the Riemann integral, the domain of the function is split into intervals leading to an approximation by step functions, that is, linear combinations of characteristic functions of intervals. In the case of the Lebesgue integral we split the range into intervals and consider their preimages. This leads to an approximation by simple functions, that is, linear combinations of characteristic functions of arbitrary (measurable) sets.

9.1. Integration — Sum me up, Henri

Throughout this section (X, Σ, μ) will be a measure space. A measurable function $s : X \to \mathbb{R}$ is called **simple** if its image is finite; that is, if

$$s = \sum_{j=1}^{p} \alpha_j \, \chi_{A_j}, \qquad \mathrm{Ran}(s) =: \{\alpha_j\}_{j=1}^{p}, \quad A_j := s^{-1}(\{\alpha_j\}) \in \Sigma. \tag{9.1}$$

Here χ_A is the **characteristic function** of A; that is, $\chi_A(x) := 1$ if $x \in A$ and $\chi_A(x) := 0$ otherwise. Note that $\dot{\bigcup}_{j=1}^{p} A_j = X$. Moreover, the set of simple functions $S(X, \mu)$ is a vector space and while there are different ways of writing a simple function as a linear combination of characteristic functions, the representation (9.1) is unique.

For a nonnegative simple function s as in (9.1) we define its **integral** as

$$\int_A s \, d\mu := \sum_{j=1}^{p} \alpha_j \, \mu(A_j \cap A). \tag{9.2}$$

Here we use the convention $0 \cdot \infty = 0$.

Lemma 9.1. *The integral has the following properties:*

(i) $\int_A s\,d\mu = \int_X \chi_A\, s\,d\mu$.
(ii) $\int_{\dot\bigcup_{n=1}^\infty A_n} s\,d\mu = \sum_{n=1}^\infty \int_{A_n} s\,d\mu$.
(iii) $\int_A \alpha\, s\,d\mu = \alpha \int_A s\,d\mu$, $\alpha \geq 0$.
(iv) $\int_A (s+t)d\mu = \int_A s\,d\mu + \int_A t\,d\mu$.
(v) $A \subseteq B \;\Rightarrow\; \int_A s\,d\mu \leq \int_B s\,d\mu$.
(vi) $s \leq t \;\Rightarrow\; \int_A s\,d\mu \leq \int_A t\,d\mu$.

Proof. (i) is clear from the definition. (ii) follows from σ-additivity of μ. (iii) is obvious. (iv) Let $s = \sum_j \alpha_j \chi_{A_j}$, $t = \sum_j \beta_k \chi_{B_k}$ as in (9.1) and abbreviate $C_{jk} = (A_j \cap B_k) \cap A$. Note $\dot\bigcup_{j,k} C_{jk} = A$. Then by (ii),

$$\begin{aligned}\int_A (s+t)d\mu &= \sum_{j,k} \int_{C_{jk}} (s+t)d\mu = \sum_{j,k} (\alpha_j + \beta_k)\mu(C_{jk}) \\ &= \sum_{j,k} \left(\int_{C_{jk}} s\,d\mu + \int_{C_{jk}} t\,d\mu \right) = \int_A s\,d\mu + \int_A t\,d\mu.\end{aligned}$$

(v) follows from monotonicity of μ. (vi) follows since by (iv) we can write $s = \sum_j \alpha_j \chi_{C_j}$, $t = \sum_j \beta_j \chi_{C_j}$ where, by assumption, $\alpha_j \leq \beta_j$. □

Our next task is to extend this definition to nonnegative measurable functions by

$$\int_A f\,d\mu := \sup_{\text{simple functions } s \leq f} \int_A s\,d\mu, \tag{9.3}$$

where the supremum is taken over all simple functions $s \leq f$. By item (vi) from our previous lemma this agrees with (9.2) if f is simple. Note that, except for possibly (ii) and (iv), Lemma 9.1 still holds for arbitrary nonnegative functions s, t.

Theorem 9.2 (Monotone convergence, Beppo Levi's theorem)**.** *Let f_n be a monotone nondecreasing sequence of nonnegative measurable functions, $f_n \nearrow f$. Then*

$$\int_A f_n\,d\mu \to \int_A f\,d\mu. \tag{9.4}$$

Proof. By property (vi), $\int_A f_n\,d\mu$ is monotone and converges to some number α. By $f_n \leq f$ and again (vi) we have

$$\alpha \leq \int_A f\,d\mu.$$

To show the converse, let s be simple such that $s \le f$ and let $\theta \in (0,1)$. Put $A_n := \{x \in A | f_n(x) \ge \theta s(x)\}$ and note $A_n \nearrow A$ (show this). Then

$$\int_A f_n \, d\mu \ge \int_{A_n} f_n \, d\mu \ge \theta \int_{A_n} s \, d\mu.$$

Letting $n \to \infty$ using (ii), we see

$$\alpha \ge \theta \int_A s \, d\mu.$$

Since this is valid for every $\theta < 1$, it still holds for $\theta = 1$. Finally, since $s \le f$ is arbitrary, the claim follows. □

In particular

$$\int_A f \, d\mu = \lim_{n\to\infty} \int_A s_n \, d\mu, \tag{9.5}$$

for every monotone sequence $s_n \nearrow f$ of simple functions. Note that there is always such a sequence, for example,

$$s_n(x) := \sum_{k=0}^{n2^n} \frac{k}{2^n} \chi_{f^{-1}(A_k)}(x), \quad A_k := [\frac{k}{2^n}, \frac{k+1}{2^n}),\ A_{n2^n} := [n, \infty). \tag{9.6}$$

By construction s_n converges uniformly if f is bounded, since $0 \le f(x) - s_n(x) < \frac{1}{2^n}$ if $f(x) \le n$.

Now what about the missing items (ii) and (iv) from Lemma 9.1? Since limits can be spread over sums, item (iv) holds, and (ii) also follows directly from the monotone convergence theorem. We even have the following result:

Lemma 9.3. *If $f \ge 0$ is measurable, then $d\nu = f \, d\mu$ defined via*

$$\nu(A) := \int_A f \, d\mu \tag{9.7}$$

is a measure such that

$$\int_A g \, d\nu = \int_A gf \, d\mu \tag{9.8}$$

for every measurable function g.

Proof. As already mentioned, additivity of ν is equivalent to linearity of the integral and σ-additivity follows from Lemma 9.1 (ii):

$$\nu(\dot{\bigcup_{n=1}^{\infty}} A_n) = \int_{\bigcup_{n=1}^{\infty} A_n} f \, d\mu = \sum_{n=1}^{\infty} \int_{A_n} f \, d\mu = \sum_{n=1}^{\infty} \nu(A_n).$$

The second claim holds for simple functions and hence for all functions by construction of the integral. □

If f_n is not necessarily monotone, we have at least

Theorem 9.4 (Fatou's lemma)**.** *If f_n is a sequence of nonnegative measurable function, then*

$$\int_A \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int_A f_n \, d\mu. \tag{9.9}$$

Proof. Set $g_n := \inf_{k\ge n} f_k$ such that $g_n \nearrow \liminf_n f_n$. Then $g_n \le f_n$ implying

$$\int_A g_n \, d\mu \le \int_A f_n \, d\mu.$$

Now take the lim inf on both sides and note that by the monotone convergence theorem

$$\liminf_{n\to\infty} \int_A g_n \, d\mu = \lim_{n\to\infty} \int_A g_n \, d\mu = \int_A \lim_{n\to\infty} g_n \, d\mu = \int_A \liminf_{n\to\infty} f_n \, d\mu,$$

proving the claim. □

Example 9.1. Consider $f_n := \chi_{[n,n+1]}$. Then $\lim_{n\to\infty} f_n(x) = 0$ for every $x \in \mathbb{R}$. However, $\int_{\mathbb{R}} f_n(x)dx = 1$. This shows that the inequality in Fatou's lemma cannot be replaced by equality in general. Another example is $f_n := \frac{1}{n}\chi_{[0,n]}$ which even converges to 0 uniformly. ⋄

If the integral is finite for both the positive and negative part $f^\pm = \max(\pm f, 0)$ of an arbitrary measurable function f, we call f **integrable** and set

$$\int_A f \, d\mu := \int_A f^+ d\mu - \int_A f^- d\mu. \tag{9.10}$$

Similarly, we handle the case where f is complex-valued by calling f integrable if both the real and imaginary part are and setting

$$\int_A f \, d\mu := \int_A \mathrm{Re}(f) d\mu + \mathrm{i} \int_A \mathrm{Im}(f) d\mu. \tag{9.11}$$

Clearly a measurable function f is integrable if and only if $|f|$ is. The set of all integrable functions is denoted by $\mathcal{L}^1(X, d\mu)$.

Lemma 9.5. *The integral is linear and Lemma 9.1 holds for integrable functions s, t.*

Furthermore, for all integrable functions f, g we have

$$\left| \int_A f \, d\mu \right| \le \int_A |f| \, d\mu \tag{9.12}$$

and (triangle inequality)

$$\int_A |f + g| \, d\mu \le \int_A |f| \, d\mu + \int_A |g| \, d\mu. \tag{9.13}$$

In the first case we have equality if and only if $f(x) = \mathrm{e}^{\mathrm{i}\theta}|f(x)|$ for a.e. x and some real number θ. In the second case we have equality if and only if

$f(x) = \mathrm{e}^{\mathrm{i}\theta(x)}|f(x)|$, $g(x) = \mathrm{e}^{\mathrm{i}\theta(x)}|g(x)|$ for a.e. x and for some real-valued function θ.

Proof. Linearity and Lemma 9.1 are straightforward to check. To see (9.12) put $\alpha := \frac{z^*}{|z|}$, where $z := \int_A f\, d\mu$ (without restriction $z \neq 0$). Then

$$|\int_A f\, d\mu| = \alpha \int_A f\, d\mu = \int_A \alpha\, f\, d\mu = \int_A \mathrm{Re}(\alpha\, f)\, d\mu \leq \int_A |f|\, d\mu,$$

proving (9.12). The second claim follows from $|f+g| \leq |f| + |g|$. The cases of equality are straightforward to check. □

Lemma 9.6. *Let f be measurable. Then*

$$\int_X |f|\, d\mu = 0 \qquad \Leftrightarrow \quad f(x) = 0 \qquad \mu - a.e. \tag{9.14}$$

Moreover, suppose f is nonnegative or integrable. Then

$$\mu(A) = 0 \qquad \Rightarrow \qquad \int_A f\, d\mu = 0. \tag{9.15}$$

Proof. Observe that we have $A := \{x | f(x) \neq 0\} = \bigcup_n A_n$, where $A_n := \{x|\, |f(x)| \geq \frac{1}{n}\}$. If $\int_X |f| d\mu = 0$ we must have $\mu(A_n) \leq n \int_{A_n} |f| d\mu = 0$ for every n and hence $\mu(A) = \lim_{n\to\infty} \mu(A_n) = 0$.

The converse will follow from (9.15) since $\mu(A) = 0$ (with A as before) implies $\int_X |f| d\mu = \int_A |f| d\mu = 0$.

Finally, to see (9.15) note that by our convention $0 \cdot \infty = 0$ it holds for any simple function and hence for any nonnegative f by definition of the integral (9.3). Since any function can be written as a linear combination of four nonnegative functions this also implies the case when f is integrable. □

Note that the proof also shows that if f is not 0 almost everywhere, there is an $\varepsilon > 0$ such that $\mu(\{x|\, |f(x)| \geq \varepsilon\}) > 0$.

In particular, the integral does not change if we restrict the domain of integration to a support of μ or if we change f on a set of measure zero. In particular, functions which are equal a.e. have the same integral.

Example 9.2. If $\mu(x) := \Theta(x)$ is the Dirac measure at 0, then

$$\int_{\mathbb{R}} f(x) d\mu(x) = f(0).$$

In fact, the integral can be restricted to any support and hence to $\{0\}$.

If $\mu(x) := \sum_n \alpha_n \Theta(x - x_n)$ is a sum of Dirac measures, $\Theta(x)$ centered at $x = 0$, then (Problem 9.2)

$$\int_{\mathbb{R}} f(x) d\mu(x) = \sum_n \alpha_n f(x_n).$$

Hence our integral contains sums as special cases. ⋄

Finally, our integral is well behaved with respect to limiting operations. We first state a simple generalization of Fatou's lemma.

Lemma 9.7 (generalized Fatou lemma)**.** *If f_n is a sequence of real-valued measurable function and g some integrable function. Then*

$$\int_A \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int_A f_n \, d\mu \tag{9.16}$$

if $g \le f_n$ and

$$\limsup_{n\to\infty} \int_A f_n \, d\mu \le \int_A \limsup_{n\to\infty} f_n \, d\mu \tag{9.17}$$

if $f_n \le g$.

Proof. To see the first apply Fatou's lemma to $f_n - g$ and subtract $\int_A g \, d\mu$ on both sides of the result. The second follows from the first using $\liminf(-f_n) = -\limsup f_n$. □

If in the last lemma we even have $|f_n| \le g$, we can combine both estimates to obtain

$$\int_A \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int_A f_n \, d\mu \le \limsup_{n\to\infty} \int_A f_n \, d\mu \le \int_A \limsup_{n\to\infty} f_n \, d\mu, \tag{9.18}$$

which is known as **Fatou–Lebesgue theorem**. In particular, in the special case where f_n converges we obtain

Theorem 9.8 (Dominated convergence)**.** *Let f_n be a convergent sequence of measurable functions and set $f := \lim_{n\to\infty} f_n$. Suppose there is an integrable function g such that $|f_n| \le g$. Then f is integrable and*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu. \tag{9.19}$$

Proof. The real and imaginary parts satisfy the same assumptions and hence it suffices to prove the case where f_n and f are real-valued. Moreover, since $\liminf f_n = \limsup f_n = f$ equation (9.18) establishes the claim. □

Remark: Since sets of measure zero do not contribute to the value of the integral, it clearly suffices if the requirements of the dominated convergence theorem are satisfied almost everywhere (with respect to μ).

Example 9.3. Note that the existence of g is crucial: The functions $f_n(x) := \frac{1}{2n}\chi_{[-n,n]}(x)$ on $\mathbb{R}$ converge uniformly to 0 but $\int_{\mathbb{R}} f_n(x)dx = 1$. ⋄

In calculus one frequently uses the notation $\int_a^b f(x)dx$. In case of general Borel measures on $\mathbb{R}$ this is ambiguous and one needs to mention to what extend the boundary points contribute to the integral. Hence we define

$$\int_a^b f\,d\mu := \begin{cases} \int_{(a,b]} f\,d\mu, & a < b, \\ 0, & a = b, \\ -\int_{(b,a]} f\,d\mu, & b < a. \end{cases} \tag{9.20}$$

such that the usual formulas

$$\int_a^b f\,d\mu = \int_a^c f\,d\mu + \int_c^b f\,d\mu \tag{9.21}$$

remain true. Note that this is also consistent with $\mu(x) = \int_0^x d\mu$.

Example 9.4. Let $f \in C[a,b]$, then the sequence of simple functions

$$s_n(x) := \sum_{j=1}^n f(x_j)\chi_{(x_{j-1},x_j]}(x), \qquad x_j = a + \frac{b-a}{n}j$$

converges to $f(x)$ and hence the integral coincides with the limit of the Riemann–Stieltjes sums:

$$\int_a^b f\,d\mu = \lim_{n\to\infty} \sum_{j=1}^n f(x_j)\big(\mu(x_j) - \mu(x_{j-1})\big).$$

Moreover, the equidistant partition could of course be replaced by an arbitrary partition $\{x_0 = a < x_1 < \cdots < x_n = b\}$ whose length $\max_{1\le j\le n}(x_j - x_{j-1})$ tends to 0. In particular, for $\mu(x) = x$ we get the usual Riemann sums and hence the Lebesgue integral coincides with the Riemann integral at least for continuous functions. Further details on the connection with the Riemann integral will be given in Section 9.6. ⋄

Even without referring to the Riemann integral, one can easily identify the Lebesgue integral as an antiderivative: Given a continuous function $f \in C(a,b)$ which is integrable over (a,b) we can introduce

$$F(x) := \int_a^x f(y)dy, \qquad x \in (a,b). \tag{9.22}$$

Then one has

$$\frac{F(x+\varepsilon) - F(x)}{\varepsilon} = f(x) + \frac{1}{\varepsilon}\int_x^{x+\varepsilon} \big(f(y) - f(x)\big)dy$$

and

$$\limsup_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_x^{x+\varepsilon} |f(y) - f(x)|dy \le \limsup_{\varepsilon\to 0} \sup_{y\in(x,x+\varepsilon]} |f(y) - f(x)| = 0$$

by the continuity of f at x. Thus $F \in C^1(a,b)$ and

$$F'(x) = f(x),$$

which is a variant of the **fundamental theorem of calculus**. This tells us that the integral of a continuous function f can be computed in terms of its antiderivative and, in particular, all tools from calculus like integration by parts or integration by substitution are readily available for the Lebesgue integral on $\mathbb{R}$. A generalization of the fundamental theorem of calculus will be given in Theorem 11.51.

Example 9.5. Another fact worthwhile mentioning is that integrals with respect to Borel measures μ on $\mathbb{R}$ can be easily computed if the distribution function is continuously differentiable. In this case $\mu([a,b)) = \mu(b) - \mu(a) = \int_a^b \mu'(x)dx$ implying that $d\mu(x) = \mu'(x)dx$ in the sense of Lemma 9.3. Moreover, it even suffices that the distribution function is piecewise continuously differentiable such that the fundamental theorem of calculus holds. ⋄

Problem 9.1. *Show the* ***inclusion exclusion principle****:*

$$\mu(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} \mu(A_{i_1} \cap \cdots \cap A_{i_k}).$$

(Hint: $\chi_{A_1 \cup \cdots \cup A_n} = 1 - \prod_{i=1}^{n} (1 - \chi_{A_i})$*.)*

Problem* 9.2. *Consider a countable set of measures μ_n and numbers $\alpha_n \ge 0$. Let $\mu := \sum_n \alpha_n \mu_n$ and show*

$$\int_A f \, d\mu = \sum_n \alpha_n \int_A f \, d\mu_n \tag{9.23}$$

for any measurable function which is either nonnegative or integrable.

Problem 9.3 (Fatou for sets)**.** *Define*

$$\liminf_{n\to\infty} A_n := \bigcup_{k\in\mathbb{N}} \bigcap_{n\ge k} A_n, \qquad \limsup_{n\to\infty} A_n := \bigcap_{k\in\mathbb{N}} \bigcup_{n\ge k} A_n.$$

That is, $x \in \liminf A_n$ if $x \in A_n$ eventually and $x \in \limsup A_n$ if $x \in A_n$ infinitely often. In particular, $\liminf A_n \subseteq \limsup A_n$*. Show*

$$\mu(\liminf_n A_n) \le \liminf_n \mu(A_n)$$

and

$$\limsup_n \mu(A_n) \le \mu(\limsup_n A_n), \qquad \text{if} \quad \mu(\bigcup_n A_n) < \infty.$$

(Hint: Show $\liminf_n \chi_{A_n} = \chi_{\liminf_n A_n}$ *and* $\limsup_n \chi_{A_n} = \chi_{\limsup_n A_n}$*.)*

Problem 9.4 (Borel–Cantelli lemma)**.** *Show that $\sum_n \mu(A_n) < \infty$ implies $\mu(\limsup_n A_n) = 0$. Give an example which shows that the converse does not hold in general.*

Problem 9.5. *Show that if $f_n \to f$ in measure (cf. Problem 8.23), then there is a subsequence which converges a.e. (Hint: Choose the subsequence such that the assumptions of the Borel–Cantelli lemma are satisfied.)*

Problem 9.6. *Consider $X = [0, 1]$ with Lebesgue measure. Show that a.e. convergence does not stem from a topology. (Hint: Start with a sequence which converges in measure to zero but not a.e. By the previous problem you can extract a subsequence which converges a.e. Now compare this with Lemma B.5.)*

Problem 9.7. *Let (X, Σ) be a measurable space. Show that the set $B(X)$ of bounded measurable functions with the sup norm is a Banach space. Show that the set $S(X)$ of simple functions is dense in $B(X)$. Show that the integral is a bounded linear functional on $B(X)$ if $\mu(X) < \infty$. (Hence Theorem 1.17 could be used to extend the integral from simple to bounded measurable functions.)*

Problem 9.8. *Show that the monotone convergence holds for nondecreasing sequences of real-valued measurable functions $f_n \nearrow f$ provided f_1 is integrable.*

Problem 9.9. *Show that the dominated convergence theorem implies (under the same assumptions)*

$$\lim_{n\to\infty} \int |f_n - f| d\mu = 0.$$

Problem 9.10 (Bounded convergence theorem)**.** *Suppose μ is a finite measure and f_n is a bounded sequence of measurable functions which converges in measure to a measurable function f (cf. Problem 8.23). Show that*

$$\lim_{n\to\infty} \int f_n d\mu = \int f \, d\mu.$$

Problem 9.11. *Consider*

$$m(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{2}, & 0 \le x < 1, \\ 1 & 1 \le x, \end{cases}$$

and let μ be the associated measure. Compute $\int_{\mathbb{R}} x \, d\mu(x)$.

Problem 9.12. *Let $\mu(A) < \infty$ and f be an integrable function satisfying $f(x) \le M$. Show that*

$$\int_A f \, d\mu \le M\mu(A)$$

with equality if and only if $f(x) = M$ for a.e. $x \in A$.

Problem 9.13. *Let μ be a nontrivial measure, $\mu(X) > 0$. Suppose f, g are integrable functions with $f < g$. Then $\int_X f \, d\mu < \int_X g \, d\mu$.*

Problem* 9.14. *Let $X \subseteq \mathbb{R}$, Y be some measure space, and $f : X \times Y \to \mathbb{C}$. Suppose $y \mapsto f(x,y)$ is measurable for every x and $x \mapsto f(x,y)$ is continuous for every y. Show that*

$$F(x) := \int_A f(x,y)\, d\mu(y) \tag{9.24}$$

is continuous if there is an integrable function $g(y)$ such that $|f(x,y)| \le g(y)$.

Problem* 9.15. *Let $X \subseteq \mathbb{R}$, Y be some measure space, and $f : X \times Y \to \mathbb{C}$. Suppose $y \mapsto f(x,y)$ is integrable for all x and $x \mapsto f(x,y)$ is differentiable for a.e. y. Show that*

$$F(x) := \int_Y f(x,y)\, d\mu(y) \tag{9.25}$$

is differentiable if there is an integrable function $g(y)$ such that $|\frac{\partial}{\partial x} f(x,y)| \le g(y)$. Moreover, $y \mapsto \frac{\partial}{\partial x} f(x,y)$ is measurable and

$$F'(x) = \int_Y \frac{\partial}{\partial x} f(x,y)\, d\mu(y) \tag{9.26}$$

in this case. (See Problem 11.49 for an extension.)

9.2. Product measures

Let μ_1 and μ_2 be two measures on Σ_1 and Σ_2, respectively. Let $\Sigma_1 \otimes \Sigma_2$ be the σ-algebra generated by **rectangles** of the form $A_1 \times A_2$.

Example 9.6. Let $\mathfrak{B}$ be the Borel sets in $\mathbb{R}$. Then $\mathfrak{B}^2 = \mathfrak{B} \otimes \mathfrak{B}$ are the Borel sets in $\mathbb{R}^2$ (since the rectangles are a basis for the product topology). $\diamond$

Any set in $\Sigma_1 \otimes \Sigma_2$ has the **section property**; that is,

Lemma 9.9. *Suppose $A \in \Sigma_1 \otimes \Sigma_2$. Then its sections*

$$A_1(x_2) := \{x_1 | (x_1, x_2) \in A\} \quad \text{and} \quad A_2(x_1) := \{x_2 | (x_1, x_2) \in A\} \tag{9.27}$$

are measurable.

Proof. Denote all sets $A \in \Sigma_1 \otimes \Sigma_2$ with the property that $A_1(x_2) \in \Sigma_1$ by S. Clearly all rectangles are in S and it suffices to show that S is a σ-algebra. Now, if $A \in S$, then $(A')_1(x_2) = (A_1(x_2))' \in \Sigma_1$ and thus S is closed under complements. Similarly, if $A_n \in S$, then $(\bigcup_n A_n)_1(x_2) = \bigcup_n (A_n)_1(x_2)$ shows that S is closed under countable unions. □

This implies that if f is a measurable function on $X_1 \times X_2$, then $f(., x_2)$ is measurable on X_1 for every x_2 and $f(x_1, .)$ is measurable on X_2 for every x_1 (observe $A_1(x_2) = \{x_1 | f(x_1, x_2) \in B\}$, where $A := \{(x_1, x_2) | f(x_1, x_2) \in B\}$).

Given two measures μ_1 on Σ_1 and μ_2 on Σ_2, we now want to construct the **product measure** $\mu_1 \otimes \mu_2$ on $\Sigma_1 \otimes \Sigma_2$ such that

$$\mu_1 \otimes \mu_2(A_1 \times A_2) := \mu_1(A_1)\mu_2(A_2), \quad A_j \in \Sigma_j,\ j = 1, 2. \tag{9.28}$$

Since the rectangles are closed under intersection, Theorem 8.3 implies that there is at most one measure on $\Sigma_1 \otimes \Sigma_2$ provided μ_1 and μ_2 are σ-finite.

Theorem 9.10. *Let μ_1 and μ_2 be two σ-finite measures on Σ_1 and Σ_2, respectively. Let $A \in \Sigma_1 \otimes \Sigma_2$. Then $\mu_2(A_2(x_1))$ and $\mu_1(A_1(x_2))$ are measurable and*

$$\int_{X_1} \mu_2(A_2(x_1))d\mu_1(x_1) = \int_{X_2} \mu_1(A_1(x_2))d\mu_2(x_2). \tag{9.29}$$

Proof. As usual, we begin with the case where μ_1 and μ_2 are finite. Let $\mathcal{D}$ be the set of all subsets for which our claim holds. Note that $\mathcal{D}$ contains at least all rectangles. Thus it suffices to show that $\mathcal{D}$ is a Dynkin system by Lemma 8.2. To see this, note that measurability and equality of both integrals follow from $A_1(x_2)' = A_1'(x_2)$ (implying $\mu_1(A_1'(x_2)) = \mu_1(X_1) - \mu_1(A_1(x_2))$) for complements and from the monotone convergence theorem for disjoint unions of sets.

If μ_1 and μ_2 are σ-finite, let $X_{i,j} \nearrow X_i$ with $\mu_i(X_{i,j}) < \infty$ for $i = 1, 2$. Now $\mu_2((A \cap X_{1,j} \times X_{2,j})_2(x_1)) = \mu_2(A_2(x_1) \cap X_{2,j})\chi_{X_{1,j}}(x_1)$ and similarly with 1 and 2 exchanged. Hence by the finite case

$$\int_{X_1} \mu_2(A_2 \cap X_{2,j})\chi_{X_{1,j}} d\mu_1 = \int_{X_2} \mu_1(A_1 \cap X_{1,j})\chi_{X_{2,j}} d\mu_2$$

and the σ-finite case follows from the monotone convergence theorem. □

Hence for given $A \in \Sigma_1 \otimes \Sigma_2$ we can define

$$\mu_1 \otimes \mu_2(A) := \int_{X_1} \mu_2(A_2(x_1))d\mu_1(x_1) = \int_{X_2} \mu_1(A_1(x_2))d\mu_2(x_2) \tag{9.30}$$

or equivalently, since $\chi_{A_1(x_2)}(x_1) = \chi_{A_2(x_1)}(x_2) = \chi_A(x_1, x_2)$,

$$\begin{aligned} \mu_1 \otimes \mu_2(A) &= \int_{X_1} \left(\int_{X_2} \chi_A(x_1, x_2)d\mu_2(x_2) \right) d\mu_1(x_1) \\ &= \int_{X_2} \left(\int_{X_1} \chi_A(x_1, x_2)d\mu_1(x_1) \right) d\mu_2(x_2). \end{aligned} \tag{9.31}$$

Then $\mu_1 \otimes \mu_2$ gives rise to a unique measure on $A \in \Sigma_1 \otimes \Sigma_2$ since σ-additivity follows from the monotone convergence theorem.

Example 9.7. Let $X_1 = X_2 = [0, 1]$ with μ_1 Lebesgue measure and μ_2 the counting measure. Let $A = \{(x, x) | x \in [0, 1]\}$ such that $\mu_2(A_2(x_1)) = 1$ and

$\mu_1(A_1(x_2)) = 0$ implying

$$1 = \int_{X_1} \mu_2(A_2(x_1))d\mu_1(x_1) \neq \int_{X_2} \mu_1(A_1(x_2))d\mu_2(x_2) = 0.$$

Hence the theorem can fail if one of the measures is not σ-finite. Note that it is still possible to define a product measure without σ-finiteness (Problem 9.17), but, as the example shows, it will lack some nice properties. ⋄

Finally we have

Theorem 9.11 (Fubini)**.** *Let f be a measurable function on $X_1 \times X_2$ and let μ_1, μ_2 be σ-finite measures on X_1, X_2, respectively.*

(i) *If $f \geq 0$, then $\int f(., x_2)d\mu_2(x_2)$ and $\int f(x_1, .)d\mu_1(x_1)$ are both measurable and*

$$\iint_{X_1\times X_2} f(x_1, x_2)d\mu_1 \otimes \mu_2(x_1, x_2) = \int_{X_2} \left(\int_{X_1} f(x_1, x_2)d\mu_1(x_1) \right) d\mu_2(x_2)$$
$$= \int_{X_1} \left(\int_{X_2} f(x_1, x_2)d\mu_2(x_2) \right) d\mu_1(x_1). \tag{9.32}$$

(ii) *If f is complex-valued, then*

$$\int_{X_1} |f(x_1, x_2)|d\mu_1(x_1) \in \mathcal{L}^1(X_2, d\mu_2), \tag{9.33}$$

respectively,

$$\int_{X_2} |f(x_1, x_2)|d\mu_2(x_2) \in \mathcal{L}^1(X_1, d\mu_1), \tag{9.34}$$

if and only if $f \in \mathcal{L}^1(X_1 \times X_2, d\mu_1 \otimes d\mu_2)$. In this case (9.32) *holds.*

Proof. By Theorem 9.10 and linearity the claim holds for simple functions. To see (i), let $s_n \nearrow f$ be a sequence of nonnegative simple functions. Then it follows by applying the monotone convergence theorem (twice for the double integrals).

For (ii) we can assume that f is real-valued by considering its real and imaginary parts separately. Moreover, splitting $f = f^+ - f^-$ into its positive and negative parts, the claim reduces to (i). □

In particular, if $f(x_1, x_2)$ is either nonnegative or integrable, then the order of integration can be interchanged. The case of nonnegative functions is also called **Tonelli's theorem**. In the general case the integrability condition is crucial, as the following example shows.

Example 9.8. Let $X := [0,1] \times [0,1]$ with Lebesgue measure and consider

$$f(x,y) = \frac{x-y}{(x+y)^3}.$$

Then

$$\int_0^1 \int_0^1 f(x,y)dx\,dy = -\int_0^1 \frac{1}{(1+y)^2}dy = -\frac{1}{2}$$

but (by symmetry)

$$\int_0^1 \int_0^1 f(x,y)dy\,dx = \int_0^1 \frac{1}{(1+x)^2}dx = \frac{1}{2}.$$

Consequently f cannot be integrable over X (verify this directly). ⋄

Lemma 9.12. *If μ_1 and μ_2 are outer regular measures, then so is $\mu_1 \otimes \mu_2$.*

Proof. Outer regularity holds for every rectangle and hence also for the algebra of finite disjoint unions of rectangles (Problem 9.16). Thus the claim follows from Problem 8.15. □

In connection with Theorem 8.3 the following observation is of interest:

Lemma 9.13. *If S_1 generates Σ_1 and S_2 generates Σ_2, then $S_1 \times S_2 := \{A_1 \times A_2 | A_j \in S_j,\ j = 1,2\}$ generates $\Sigma_1 \otimes \Sigma_2$.*

Proof. Denote the σ-algebra generated by $S_1 \times S_2$ by Σ. Consider the set $\{A_1 \in \Sigma_1 | A_1 \times X_2 \in \Sigma\}$ which is clearly a σ-algebra containing S_1 and thus equal to Σ_1. In particular, $\Sigma_1 \times X_2 \subset \Sigma$ and similarly $X_1 \times \Sigma_2 \subset \Sigma$. Hence also $(\Sigma_1 \times X_2) \cap (X_1 \times \Sigma_2) = \Sigma_1 \times \Sigma_2 \subset \Sigma$. □

Finally, note that we can iterate this procedure.

Lemma 9.14. *Suppose (X_j, Σ_j, μ_j), $j = 1,2,3$, are σ-finite measure spaces. Then $(\Sigma_1 \otimes \Sigma_2) \otimes \Sigma_3 = \Sigma_1 \otimes (\Sigma_2 \otimes \Sigma_3)$ and*

$$(\mu_1 \otimes \mu_2) \otimes \mu_3 = \mu_1 \otimes (\mu_2 \otimes \mu_3). \tag{9.35}$$

Proof. First of all note that $(\Sigma_1 \otimes \Sigma_2) \otimes \Sigma_3 = \Sigma_1 \otimes (\Sigma_2 \otimes \Sigma_3)$ is the sigma algebra generated by the rectangles $A_1 \times A_2 \times A_3$ in $X_1 \times X_2 \times X_3$. Moreover, since

$$\begin{aligned}((\mu_1 \otimes \mu_2) \otimes \mu_3)(A_1 \times A_2 \times A_3) &= \mu_1(A_1)\mu_2(A_2)\mu_3(A_3)\\ &= (\mu_1 \otimes (\mu_2 \otimes \mu_3))(A_1 \times A_2 \times A_3),\end{aligned}$$

the two measures coincide on rectangles and hence everywhere by Theorem 8.3. □

Hence we can take the product of finitely many measures. The case of infinitely many measures requires a bit more effort and will be discussed in Section 11.5.

Example 9.9. If λ is Lebesgue measure on $\mathbb{R}$, then $\lambda^n = \lambda \otimes \cdots \otimes \lambda$ is Lebesgue measure on $\mathbb{R}^n$. In fact, it satisfies $\lambda^n((a,b]) = \prod_{j=1}^n (b_j - a_j)$ and hence must be equal to Lebesgue measure which is the unique Borel measure with this property. ⋄

Example 9.10. If X_1, X_2 are topological spaces, then $\mathfrak{B}(X_1 \times X_2) = \mathfrak{B}(X_1) \otimes \mathfrak{B}(X_2)$ since open rectangles are a base for the product topology. Moreover, if μ_1, μ_2 are Borel measures and both X_1 and X_1 are locally compact, then $\mu_1 \otimes \mu_2$ is also a Borel measure. Indeed, let $K \subseteq X_1 \times X_2$ be compact. Then for every point in K there is a relatively compact open rectangle containing this point. By compactness finitely many of them suffice to cover K, that is $K \subseteq \bigcup_{j=1}^n K_{1,j} \times K_{2,j}$ implying $\mu_1 \otimes \mu_2(K) \leq \sum_{j=1}^n \mu(K_{1,j})\mu(K_{2,j}) < \infty$. ⋄

Problem* 9.16. *Show that the set of all finite union of measurable rectangles $A_1 \times A_2$ forms an algebra. Moreover, every set in this algebra can be written as a finite union of disjoint rectangles.*

Problem 9.17. *Given two measure spaces (X_1, Σ_1, μ_1) and (X_2, Σ_2, μ_2) let $\mathcal{R} = \{A_1 \times A_2 | A_j \in \Sigma_j,\ j = 1, 2\}$ be the collection of measurable rectangles. Define $\rho : \mathcal{R} \to [0, \infty]$, $\rho(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. Then*

$$(\mu_1 \otimes \mu_2)^*(A) := \inf \Big\{ \sum_{j=1}^{\infty} \rho(A_j) \Big| A \subseteq \bigcup_{j=1}^{\infty} A_j,\ A_j \in \mathcal{R} \Big\}$$

is an outer measure on $X_1 \times X_2$. Show that this constructions coincides with (9.30) for $A \in \Sigma_1 \otimes \Sigma_2$ in case μ_1 and μ_2 are σ-finite.

Problem 9.18. *Let $P : \mathbb{R}^n \to \mathbb{C}$ be a nonzero polynomial. Show that $N := \{x \in \mathbb{R}^n | P(x) = 0\}$ is a Borel set of Lebesgue zero. (Hint: Induction using Fubini.)*

Problem* 9.19. *Let $U \subseteq \mathbb{C}$ be a domain, Y be some measure space, and $f : U \times Y \to \mathbb{C}$. Suppose $y \mapsto f(z, y)$ is measurable for every z and $z \mapsto f(z, y)$ is holomorphic for every y. Show that*

$$F(z) := \int_Y f(z, y)\, d\mu(y)$$

is holomorphic if for every compact subset $V \subset U$ there is an integrable function $g(y)$ such that $|f(z, y)| \leq g(y)$, $z \in V$. (Hint: Use Fubini and Morera.)

Problem* 9.20. *Suppose $\phi : [0,\infty) \to [0,\infty)$ is integrable over every compact interval and set $\Phi(r) = \int_0^r \phi(s)ds$. Let $f : X \to \mathbb{C}$ be measurable and introduce its distribution function*

$$E_f(r) := \mu\big(\{x \in X |\, |f(x)| > r\}\big)$$

Show that

$$\int_X \Phi(|f|)d\mu = \int_0^\infty \phi(r)E_f(r)dr.$$

Moreover, show that if f is integrable, then the set of all $\alpha \in \mathbb{C}$ for which $\mu\big(\{x \in X |\, f(x) = \alpha\}\big) > 0$ is countable.

9.3. Transformation of measures and integrals

Finally we want to transform measures. Let $f : X \to Y$ be a measurable function. Given a measure μ on X we can introduce the **pushforward measure** (also image measure) $f_\star\mu$ on Y via

$$(f_\star\mu)(A) := \mu(f^{-1}(A)). \tag{9.36}$$

It is straightforward to check that $f_\star\mu$ is indeed a measure. Moreover, note that $f_\star\mu$ is supported on the range of f.

Theorem 9.15. *Let $f : X \to Y$ be measurable and let $g : Y \to \mathbb{C}$ be a Borel function. Then the Borel function $g \circ f : X \to \mathbb{C}$ is a.e. nonnegative or integrable if and only if g is and in both cases*

$$\int_Y g\, d(f_\star\mu) = \int_X g \circ f\, d\mu. \tag{9.37}$$

Proof. In fact, it suffices to check this formula for simple functions g, which follows since $\chi_A \circ f = \chi_{f^{-1}(A)}$. □

Example 9.11. Suppose $f : X \to Y$ and $g : Y \to Z$. Then

$$(g \circ f)_\star\mu = g_\star(f_\star\mu).$$

since $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$. ◇

Example 9.12. Let $f(x) = Mx + a$ be an affine transformation, where $M : \mathbb{R}^n \to \mathbb{R}^n$ is some invertible matrix. Then Lebesgue measure transforms according to

$$f_\star\lambda^n = \frac{1}{|\det(M)|}\lambda^n.$$

To see this, note that $f_\star\lambda^n$ is translation invariant and hence must be a multiple of λ^n. Moreover, for an orthogonal matrix this multiple is one (since an orthogonal matrix leaves the unit ball invariant) and for a diagonal matrix it must be the absolute value of the product of the diagonal elements (consider a rectangle). Finally, since every matrix can be written as $M =$

$O_1 D O_2$, where O_j are orthogonal and D is diagonal (Problem 9.22), the claim follows.

As a consequence we obtain

$$\int_A g(Mx+a)d^n x = \frac{1}{|\det(M)|}\int_{MA+a} g(y)d^n y,$$

which applies, for example, to shifts $f(x) = x + a$ or scaling transforms $f(x) = \alpha x$. ⋄

This result can be generalized to diffeomorphisms (one-to-one C^1 maps with inverse again C^1):

Theorem 9.16 (change of variables)**.** *Let $U, V \subseteq \mathbb{R}^n$ and suppose $f \in C^1(U,V)$ is a diffeomorphism. Then*

$$(f^{-1})_\star d^n x = |J_f(x)| d^n x, \tag{9.38}$$

where $J_f = \det(\frac{\partial f}{\partial x})$ is the Jacobi determinant of f. In particular,

$$\int_U g(f(x))|J_f(x)|d^n x = \int_V g(y)d^n y \tag{9.39}$$

whenever g is nonnegative or integrable over V.

Proof. It suffices to show

$$\int_{f(R)} d^n y = \int_R |J_f(x)| d^n x$$

for every bounded open rectangle $R \subseteq U$. By Theorem 8.3 it will then follow for characteristic functions and thus for arbitrary functions by the very definition of the integral.

To this end we consider the integral

$$I_\varepsilon := \int_{f(R)}\int_R |J_f(f^{-1}(y))|\varphi_\varepsilon(f(z)-y)d^n z\, d^n y$$

Here $\varphi := V_n^{-1}\chi_{B_1(0)}$ and $\varphi_\varepsilon(y) := \varepsilon^{-n}\varphi(\varepsilon^{-1}y)$, where V_n is the volume of the unit ball (cf. below), such that $\int \varphi_\varepsilon(x)d^n x = 1$.

We will evaluate this integral in two ways. To begin with we consider the inner integral

$$h_\varepsilon(y) := \int_R \varphi_\varepsilon(f(z)-y)d^n z.$$

For $\varepsilon < \varepsilon_0$ the integrand is nonzero only for $z \in K = f^{-1}(\overline{B_{\varepsilon_0}(y)})$, where K is some compact set containing $x = f^{-1}(y)$. Using the affine change of coordinates $z = x + \varepsilon w$ we obtain

$$h_\varepsilon(y) = \int_{W_\varepsilon(x)} \varphi\left(\frac{f(x+\varepsilon w)-f(x)}{\varepsilon}\right) d^n w, \qquad W_\varepsilon(x) = \frac{1}{\varepsilon}(K-x).$$

By

$$\left|\frac{f(x+\varepsilon w)-f(x)}{\varepsilon}\right| \geq \frac{1}{C}|w|, \qquad C := \sup_K \|df^{-1}\|$$

the integrand is nonzero only for $w \in B_C(0)$. Hence, as $\varepsilon \to 0$ the domain $W_\varepsilon(x)$ will eventually cover all of $B_C(0)$ and dominated convergence implies

$$\lim_{\varepsilon\downarrow 0} h_\varepsilon(y) = \int_{B_C(0)} \varphi(df(x)w)dw = |J_f(x)|^{-1}.$$

Consequently, $\lim_{\varepsilon\downarrow 0} I_\varepsilon = |f(R)|$ again by dominated convergence. Now we use Fubini to interchange the order of integration

$$I_\varepsilon = \int_R \int_{f(R)} |J_f(f^{-1}(y))|\varphi_\varepsilon(f(z)-y)d^n y\, d^n z.$$

Since $f(z)$ is an interior point of $f(R)$ continuity of $|J_f(f^{-1}(y))|$ implies

$$\lim_{\varepsilon\downarrow 0} \int_{f(R)} |J_f(f^{-1}(y))|\varphi_\varepsilon(f(z)-y)d^n y = |J_f(f^{-1}(f(z)))| = |J_f(z)|$$

and hence dominated convergence shows $\lim_{\varepsilon\downarrow 0} I_\varepsilon = \int_R |J_f(z)|d^n z$. □

Example 9.13. For example, we can consider **polar coordinates** $T_2 : [0,\infty)\times[0,2\pi) \to \mathbb{R}^2$ defined by

$$T_2(\rho,\varphi) := (\rho\cos(\varphi), \rho\sin(\varphi)).$$

Then

$$\det \frac{\partial T_2}{\partial(\rho,\varphi)} = \det \begin{vmatrix} \cos(\varphi) & -\rho\sin(\varphi) \\ \sin(\varphi) & \rho\cos(\varphi) \end{vmatrix} = \rho$$

and one has

$$\int_U f(\rho\cos(\varphi),\rho\sin(\varphi))\rho\, d(\rho,\varphi) = \int_{T_2(U)} f(x)d^2x.$$

Note that T_2 is only bijective when restricted to $(0,\infty)\times[0,2\pi)$. However, since the set $\{0\}\times[0,2\pi)$ is of measure zero, it does not contribute to the integral on the left. Similarly, its image $T_2(\{0\}\times[0,2\pi)) = \{0\}$ does not contribute to the integral on the right. ◇

Example 9.14. We can use the previous example to obtain the transformation formula for **spherical coordinates** in $\mathbb{R}^n$ by induction. We illustrate the process for $n = 3$. To this end let $x = (x_1, x_2, x_3)$ and start with spherical coordinates in $\mathbb{R}^2$ (which are just polar coordinates) for the first two components:

$$x = (\rho\cos(\varphi), \rho\sin(\varphi), x_3), \qquad \rho \in [0,\infty),\ \varphi \in [0,2\pi).$$

Next use polar coordinates for (ρ, x_3):

$$(\rho, x_3) = (r\sin(\theta), r\cos(\theta)), \qquad r \in [0,\infty),\ \theta \in [0,\pi].$$

Note that the range for θ follows since $\rho \geq 0$. Moreover, observe that $r^2 = \rho^2 + x_3^2 = x_1^2 + x_2^2 + x_3^2 = |x|^2$ as already anticipated by our notation. In summary,

$$x = T_3(r, \varphi, \theta) := (r\sin(\theta)\cos(\varphi), r\sin(\theta)\sin(\varphi), r\cos(\theta)).$$

Furthermore, since T_3 is the composition with T_2 acting on the first two coordinates with the last unchanged and polar coordinates P acting on the first and last coordinate, the chain rule implies

$$\det \frac{\partial T_3}{\partial(r,\varphi,\theta)} = \det \frac{\partial T_2}{\partial(\rho,\varphi,x_3)}\Big|_{\substack{\rho=r\sin(\theta)\\ x_3=r\cos(\theta)}} \det \frac{\partial P}{\partial(r,\varphi,\theta)} = r^2\sin(\theta).$$

Hence one has

$$\int_U f(T_3(r,\varphi,\theta))r^2\sin(\theta)d(r,\varphi,\theta) = \int_{T_3(U)} f(x)d^3x.$$

Again T_3 is only bijective on $(0,\infty)\times[0,2\pi)\times(0,\pi)$.

It is left as an exercise to check that the extension $T_n : [0,\infty)\times[0,2\pi)\times[0,\pi]^{n-2} \to \mathbb{R}^n$ is given by

$$x = T_n(r, \varphi, \theta_1, \ldots, \theta_{n-2})$$

with

$$\begin{aligned}
x_1 &= r\cos(\varphi)\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)\cdots\sin(\theta_{n-2}),\\
x_2 &= r\sin(\varphi)\sin(\theta_1)\sin(\theta_2)\sin(\theta_3)\cdots\sin(\theta_{n-2}),\\
x_3 &= r\cos(\theta_1)\sin(\theta_2)\sin(\theta_3)\cdots\sin(\theta_{n-2}),\\
x_4 &= r\cos(\theta_2)\sin(\theta_3)\cdots\sin(\theta_{n-2}),\\
&\vdots\\
x_{n-1} &= r\cos(\theta_{n-3})\sin(\theta_{n-2}),\\
x_n &= r\cos(\theta_{n-2}).
\end{aligned}$$

The Jacobi determinant is given by

$$\det \frac{\partial T_n}{\partial(r,\varphi,\theta_1,\ldots,\theta_{n-2})} = r^{n-1}\sin(\theta_1)\sin(\theta_2)^2\cdots\sin(\theta_{n-2})^{n-2}.$$

◇

Another useful consequence of Theorem 9.15 is the following rule for integrating radial functions.

Lemma 9.17. *There is a measure σ^{n-1} on the **unit sphere** $S^{n-1} := \partial B_1(0) = \{x \in \mathbb{R}^n \,|\, |x| = 1\}$, which is rotation invariant and satisfies*

$$\int_{\mathbb{R}^n} g(x)d^nx = \int_0^\infty \int_{S^{n-1}} g(r\omega)r^{n-1}d\sigma^{n-1}(\omega)dr, \tag{9.40}$$

for every integrable (or positive) function g.

Moreover, the surface area of S^{n-1} is given by

$$S_n := \sigma^{n-1}(S^{n-1}) = nV_n, \tag{9.41}$$

where $V_n := \lambda^n(B_1(0))$ *is the volume of the unit ball in* $\mathbb{R}^n$*, and if* $g(x) = \tilde{g}(|x|)$ *is radial we have*

$$\int_{\mathbb{R}^n} g(x)d^n x = S_n \int_0^\infty \tilde{g}(r)r^{n-1}dr. \tag{9.42}$$

Proof. Consider the measurable transformation $f : \mathbb{R}^n \to [0,\infty) \times S^{n-1}$, $x \mapsto (|x|, \frac{x}{|x|})$ (with $\frac{0}{|0|} = 1$). Let $d\mu(r) := r^{n-1}dr$ and

$$\sigma^{n-1}(A) := n\lambda^n(f^{-1}([0,1) \times A)) \tag{9.43}$$

for every $A \in \mathfrak{B}(S^{n-1}) = \mathfrak{B}^n \cap S^{n-1}$. Note that σ^{n-1} inherits the rotation invariance from λ^n. By Theorem 9.15 it suffices to show $f_\star\lambda^n = \mu \otimes \sigma^{n-1}$. This follows from

$$\begin{aligned}(f_\star\lambda^n)([0,r) \times A) &= \lambda^n(f^{-1}([0,r) \times A)) = r^n\lambda^n(f^{-1}([0,1) \times A)) \\ &= \mu([0,r))\sigma^{n-1}(A).\end{aligned}$$

since these sets determine the measure uniquely. □

Clearly in spherical coordinates the surface measure is given by

$$d\sigma^{n-1} = \sin(\theta_1)\sin(\theta_2)^2 \cdots \sin(\theta_{n-2})^{n-2}d\varphi\, d\theta_1 \cdots d\theta_{n-2}. \tag{9.44}$$

Example 9.15. Let us compute the volume of a ball in $\mathbb{R}^n$:

$$V_n(r) := \int_{\mathbb{R}^n} \chi_{B_r(0)}d^n x.$$

By the simple scaling transform $f(x) = rx$ we obtain $V_n(r) = V_n(1)r^n$ and hence it suffices to compute $V_n := V_n(1)$.

To this end we use (Problem 9.23)

$$\begin{aligned}\pi^n &= \int_{\mathbb{R}^n} \mathrm{e}^{-|x|^2}d^n x = nV_n \int_0^\infty \mathrm{e}^{-r^2}r^{n-1}dr = \frac{nV_n}{2}\int_0^\infty \mathrm{e}^{-s}s^{n/2-1}ds \\ &= \frac{nV_n}{2}\Gamma(\frac{n}{2}) = \frac{V_n}{2}\Gamma(\frac{n}{2}+1)\end{aligned}$$

where Γ is the gamma function (Problem 9.24). Hence

$$V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}. \tag{9.45}$$

By $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (see Problem 9.25) this coincides with the well-known values $V_1 = 2$, $V_2 = \pi$, $V_3 = \frac{4\pi}{3}$. ⋄

Example 9.16. The above lemma can be used to determine when a radial function is integrable. For example, we obtain

$$|x|^\alpha \in L^1(B_1(0)) \Leftrightarrow \alpha > -n, \qquad |x|^\alpha \in L^1(\mathbb{R}^n \setminus B_1(0)) \Leftrightarrow \alpha < -n. \quad \diamond$$

Problem 9.21. *Let λ be Lebesgue measure on $\mathbb{R}$, and let f be a strictly increasing function with $\lim_{x\to\pm\infty} f(x) = \pm\infty$. Show that*

$$d(f_\star\lambda) = d(f^{-1}),$$

where f^{-1} is the inverse of f extended to all of $\mathbb{R}$ by setting $f^{-1}(y) = x$ for $y \in [f(x-), f(x+)]$ (note that f^{-1} is continuous).

Moreover, if $f \in C^1(\mathbb{R})$ with $f' > 0$, then

$$d(f_\star\lambda) = \frac{1}{f'(f^{-1})} d\lambda.$$

Problem* 9.22. *Show that every invertible matrix M can be written as $M = O_1 D O_2$, where D is diagonal and O_j are orthogonal. (Hint: The matrix M^*M is nonnegative and hence there is an orthogonal matrix U which diagonalizes $M^*M = UD^2U^*$. Then one can choose $O_1 = MUD^{-1}$ and $O_2 = U^*$.)*

Problem* 9.23. *Show*

$$I_n := \int_{\mathbb{R}^n} \mathrm{e}^{-|x|^2} d^n x = \pi^{n/2}.$$

(Hint: Use Fubini to show $I_n = I_1^n$ and compute I_2 using polar coordinates.)

Problem* 9.24. *The **gamma function** is defined via*

$$\Gamma(z) := \int_0^\infty x^{z-1}\mathrm{e}^{-x} dx, \qquad \mathrm{Re}(z) > 0. \tag{9.46}$$

Verify that the integral converges and defines an analytic function in the indicated half-plane (cf. Problem 9.19). Use integration by parts to show

$$\Gamma(z+1) = z\Gamma(z), \qquad \Gamma(1) = 1. \tag{9.47}$$

Conclude $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. Show that the relation $\Gamma(z) = \Gamma(z+1)/z$ can be used to define $\Gamma(z)$ for all $z \in \mathbb{C}\backslash\{0, -1, -2, \dots\}$. Show that near $z = -n$, $n \in \mathbb{N}_0$, the Gamma functions behaves like $\Gamma(z) = \frac{(-1)^n}{n!(z+n)} + O(1)$.

Problem* 9.25. *Show that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Moreover, show*

$$\Gamma(n + \frac{1}{2}) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$$

(Hint: Use the change of coordinates $x = t^2$ and then use Problem 9.23.)

Problem 9.26. *Establish*

$$\left(\frac{d}{dz}\right)^j \Gamma(z) = \int_0^\infty \log(x)^j x^{z-1}\mathrm{e}^{-x} dx, \qquad \mathrm{Re}(z) > 0,$$

and show that Γ is log-convex, that is, $\log(\Gamma(x))$ is convex. (Hint: Regard the first derivative as a scalar product and apply Cauchy–Schwarz.)

Problem 9.27. *Show that the* ***Beta function*** *satisfies*

$$B(u,v) := \int_0^1 t^{u-1}(1-t)^{v-1}dt = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}, \quad \mathrm{Re}(u) > 0,\ \mathrm{Re}(v) > 0.$$

Use this to establish ***Euler's reflection formula***

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin(\pi z)}.$$

Conclude that the Gamma function has no zeros on $\mathbb{C}$.

(Hint: Start with $\Gamma(u)\Gamma(v)$ *and make a change of variables* $x = ts$, $y = t(1-s)$. *For the reflection formula evaluate* $B(z, 1-z)$ *using Problem 9.28.)*

Problem 9.28. *Show*

$$\int_{-\infty}^{\infty} \frac{\mathrm{e}^{zx}}{1+\mathrm{e}^x}dx = \frac{\pi}{\sin(z\pi)}, \qquad 0 < \mathrm{Re}(z) < 1.$$

(Hint: To compute the integral, use a contour consisting of the straight lines connecting the points $-R$, R, $R+2\pi\mathrm{i}$, $-R+2\pi\mathrm{i}$. *Evaluate the contour integral using the residue theorem and let* $R \to \infty$. *Show that the contributions from the vertical lines vanish in the limit and relate the integrals along the horizontal lines.)*

Problem 9.29. *Let* $U \subseteq \mathbb{R}^m$ *be open and let* $f : U \to \mathbb{R}^n$ *be locally Lipschitz (i.e., for every compact set* $K \subset U$ *there is some constant* L *such that* $|f(x) - f(y)| \le L|x-y|$ *for all* $x, y \in K$*). Show that if* $A \subset U$ *has Lebesgue measure zero, then* $f(A)$ *is contained in a set of Lebesgue measure zero. (Hint: By Lindelöf it is no restriction to assume that* A *is contained in a compact ball contained in* U. *Now approximate* A *by a union of rectangles.)*

9.4. Surface measure and the Gauss–Green theorem

We begin by recalling the definition of an m-dimensional submanifold: We will call a subset $\Sigma \subseteq \mathbb{R}^n$ an m-dimensional **submanifold** if there is a parametrization $\varphi \in C^1(U, \mathbb{R}^n)$, where $U \subseteq \mathbb{R}^n$ is open, $\Sigma = \varphi(U)$, and φ is an immersion (i.e., the Jacobian is injective at every point). Somewhat more general one extends this definition to the case where a parametrization only exists locally in the sense that every point $z \in \Sigma$ has a neighborhood W such that there is a parametrization for $W \cap \Sigma$.

Moreover, given a parametrization near a point $z^0 \in \Sigma$, the assumption that the Jacobian $\frac{\partial \varphi}{\partial x}$ is injective implies that, after a permutation of the coordinates, the first m vectors of the Jacobian are linearly independent. Hence, after restricting U, we can assume that $(\varphi_1, \dots, \varphi_m)$ is invertible and hence there is a parametrization of the form

$$\phi(x) = (x_1, \dots, x_m, \phi_{m+1}(x), \dots, \phi_n(x)) \tag{9.48}$$

(up to a permutation of the coordinates in $\mathbb{R}^n$). We will require for all parameterizations U to be so small that this is possible.

Given a submanifold Σ and a parametrization $\varphi : U \subseteq \mathbb{R}^m \to \Sigma \subseteq \mathbb{R}^n$ let

$$\Gamma(\partial\varphi) := \Gamma(\partial_1\varphi, \dots, \partial_m\varphi) = \det\left(\partial_j\varphi \cdot \partial_k\varphi\right)_{1\le j,k\le m} \tag{9.49}$$

be the **Gram determinant** of the tangent vectors (here the dot indicates a scalar product in $\mathbb{R}^n$) and define the **submanifold measure** dS via

$$\int_\Sigma g\, dS := \int_U g(\varphi(x))\sqrt{\Gamma(\partial\varphi(x))}d^m x. \tag{9.50}$$

Note that this can be geometrically motivated since the Gram determinant can be interpreted as the square of the volume of the parallelotope $\{\sum_{j=1}^m \alpha_j\partial_j\varphi | 0 \le \alpha_j \le m\}$ formed by the tangent vectors, which is just the linear approximation to the surface at the given point. Indeed, defining the volume recursively by the volume of the base $\{\sum_{j=1}^{m-1} \alpha_j\partial_j\varphi | 0 \le \alpha_j \le m-1\}$ times the height (i.e., the distance of $\partial_j\varphi_m$ from $\operatorname{span}\{\partial_1\varphi, \dots, \partial_{m-1}\varphi\}$), this follows from Problem 2.1.

If $\phi : V \subseteq \mathbb{R}^m \to \Sigma \subseteq \mathbb{R}^n$ is another parametrization, and hence $f = \phi^{-1} \circ \varphi \in C^1(U, V)$ is a diffeomorphism, the change of variables formula gives

$$\begin{aligned}\int_V g(\phi(x))\sqrt{\Gamma(\partial\phi(y))}d^m y &= \int_U g(\phi(f(x)))\sqrt{\Gamma(\partial\phi(f(x)))}|J_f(x)|d^n x\\ &= \int_U g(\varphi(x))\sqrt{\Gamma(\partial\varphi(x))}d^m x,\end{aligned}$$

where we have used the chain rule $\frac{\partial\varphi}{\partial x}(x) = \frac{\partial(\phi\circ f)}{\partial x}(f(x)) = \frac{\partial\phi}{\partial y}(f(x))\frac{\partial f}{\partial x}(x)$ in the last step. Hence our definition is independent of the parametrization chosen. If our submanifold cannot be covered by a single parametrization, we choose a partition into countably many measurable subsets A_j such that for each A_j there is a parametrization (U_j, φ_j) such that $A_j \subseteq \varphi_j(U_j)$. Then we set

$$\int_\Sigma g\, dS := \sum_j \int_{\varphi_j^{-1}(A_j)} g(\varphi_j(x))\sqrt{\Gamma(\partial\varphi_j(x))}d^m x. \tag{9.51}$$

Note that given a different splitting B_k with parameterizations (V_k, ϕ_k) we can first change to a common refinement $A_j \cap B_k$ and then conclude that the individual integrals are equal by our above calculation. Hence again our definition is independent of the splitting and the parametrization chosen.

Example 9.17. Let T_n be spherical coordinates from Example 9.14 and let $S_n(\varphi, \theta_1, \dots, \theta_{n-2}) = T_n(1, \varphi, \theta_1, \dots, \theta_{n-2})$ be the corresponding parametrization of the unit sphere. Then one computes

$$\det(\partial T_n)^2 = \Gamma(\partial T_n) = r^{2n-2}\,\Gamma(\partial S_n)$$

since $\partial_r T^n \cdot \partial_r T^n = 1$, $\partial_r T^n \cdot \partial_\varphi T^n = \partial_r T^n \cdot \partial_{\theta_j} T^n = 0$. Hence dS_n coincides with $d\sigma^{n-1}$ from (9.44). ⋄

In the case $m = n-1$ a submanifold is also known as a **(hyper-)surface**. Given a surface and a parametrization, a **normal vector** is given by

$$\tilde{\nu} = \big(\det(\partial_1\varphi, \dots, \partial_{n-1}\varphi, \delta_1), \dots, \det(\partial_1\varphi, \dots, \partial_{n-1}\varphi, \delta_n)\big), \tag{9.52}$$

where δ_j are the canonical basis vectors in $\mathbb{R}^n$ (it is straightforward to check that $\tilde{\nu} \cdot \partial_j\varphi = 0$ for $1 \le j \le n-1$). Its length is given by (Problem 9.30)

$$|\tilde{\nu}|^2 = \Gamma(\partial\varphi) \tag{9.53}$$

and the unit normal is given by

$$\nu = \frac{1}{\sqrt{\Gamma(\partial\varphi)}}\tilde{\nu}. \tag{9.54}$$

It is uniquely defined up to orientation. Moreover, given a vector field $u : \Sigma \to \mathbb{R}^n$ (or $\mathbb{C}^n$) we have

$$\int_\Sigma u \cdot \nu \, dS = \int_U \det(\partial_1\varphi, \dots, \partial_{n-1}\varphi, u \circ \varphi) d^{n-1}x. \tag{9.55}$$

Here we will mainly be interested in the case of a surface arising as the boundary of some open domain $\Omega \subset \mathbb{R}^n$. To this end we recall that $\Omega \subseteq \mathbb{R}^n$ is said to have a C^1 boundary if around any point $x^0 \in \partial\Omega$ we can find a small neighborhood $O(x^0)$ so that, after a possible permutation of the coordinates, we can write

$$\Omega \cap O(x^0) = \{x \in O(x^0) | x_n > \gamma(x_1, \dots, x_{n-1})\} \tag{9.56}$$

with $\gamma \in C^1$. Similarly we could define C^k or $C^{k,\theta}$ domains. According to our definition above $\partial\Omega$ is then a surface in $\mathbb{R}^n$ and we have

$$\partial\Omega \cap O(x^0) = \{x \in O(x^0) | x_n = \gamma(x_1, \dots, x_{n-1})\}. \tag{9.57}$$

Note that in this case we have a change of coordinates $y = \psi(x)$ such that in these coordinates the boundary is given by (part of) the hyperplane $y_n = 0$. Explicitly we have $\psi \in C^1_b(\Omega \cap O(x^0), V_+(y^0))$ given by

$$\psi(x) = (x_1, \dots, x_{n-1}, x_n - \gamma(x_1, \dots, x_{n-1})) \tag{9.58}$$

with inverse $\psi^{-1} \in C^1_b(V_+(y^0), \Omega \cap O(x^0))$ given by

$$\psi^{-1}(y) = (y_1, \dots, y_{n-1}, y_n + \gamma(y_1, \dots, y_{n-1})). \tag{9.59}$$

This is known as straightening out the boundary (see Figure 9.1). Moreover, at every point of the boundary we have the **outward pointing unit normal vector** $\nu(x^0)$ which, in the above setting, is given as

$$\nu(x^0) := \frac{1}{\sqrt{1 + (\partial_1\gamma)^2 + \cdots + (\partial_{n-1}\gamma)^2}}(\partial_1\gamma, \dots, \partial_{n-1}\gamma, -1). \tag{9.60}$$

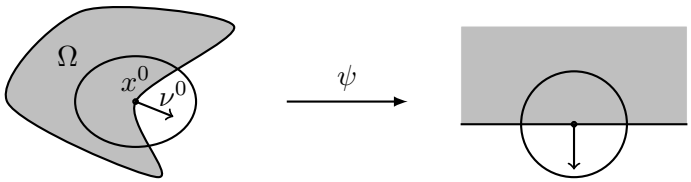


Figure 9.1. Straightening out the boundary

If we straighten out the boundary, then clearly, $\nu(y^0) = (0, \dots, 0, -1)$.

Theorem 9.18 (Gauss–Green)**.** *If Ω is a bounded C^1 domain in $\mathbb{R}^n$ and $u \in C^1(\overline{\Omega}, \mathbb{R}^n)$ is a vector field, then*

$$\int_\Omega (\operatorname{div} u) d^n x = \int_{\partial\Omega} u \cdot \nu \, dS. \tag{9.61}$$

Here $\operatorname{div} u = \sum_{j=1}^n \partial_j u_j$ *is the* ***divergence*** *of a vector field.*

Proof. By linearity it suffices to prove

$$\int_\Omega (\partial_j f) d^n x = \int_{\partial\Omega} f \nu_j dS, \qquad 1 \le j \le n, \tag{9.62}$$

for $f \in C^1(\overline{\Omega})$. We first suppose that u is supported in a neighborhood $O(x^0)$ as in (9.56). We also assume that $O(x^0)$ is a rectangle. Let $O = O(X^0) \cap \partial\Omega$. Then for $j = n$ we have

$$\begin{aligned}\int_\Omega (\partial_n f) d^n x &= \int_O \left(\int_{x_n \ge \gamma(x')} \partial_n f(x', x_n) dx_n \right) d^{n-1} x' \\ &= -\int_O f(x', \gamma(x')) d^{n-1} x' = \int_{\partial\Omega} f \nu_n dS,\end{aligned}$$

where we have used Fubini and integration by parts. For $j < n$ let us assume that we have just $n = 2$ to simplify notation (as the other coordinates will not affect the calculation). Then $O(x^0) = (a_1, b_1) \times (a_2, b_2)$ and we have (by the fundamental theorem of calculus and the Leibniz integral rule — Problem 9.31)

$$\begin{aligned}0 &= \int_{a_1}^{b_1} \partial_1 \int_{\gamma(x_1)}^{b_2} f(x_1, x_2) dx_2 dx_1 \\ &= \int_{a_1}^{b_1} \int_{\gamma(x_1)}^{b_2} \big(\partial_1 f(x_1, x_2)\big) dx_2 dx_1 - \int_{a_1}^{b_1} f(x_1, \gamma(x_1)) \partial_1 \gamma(x_1) dx_1\end{aligned}$$

from which the claim follows.

For the general case cover $\overline{\Omega}$ by rectangles which either contain no boundary points or otherwise are as in (9.56). By compactness there is a finite subcover. Choose a smooth partition of unity ζ_j subordinate to this cover

(Lemma B.31) and consider $f = \sum_j \zeta_j f$. Then for each summand having support in a rectangle intersecting the boundary, the claim holds by the above computation. Similarly, for each summand having support in an interior rectangle, Fubini and the fundamental theorem of calculus show $\int_\Omega (\partial_n \zeta_j f) d^n x = 0$. □

Applying the Gauss–Green theorem to a product fg we obtain

Corollary 9.19 (Integration by parts)**.** *We have*

$$\int_\Omega (\partial_j f) g \, d^n x = \int_{\partial\Omega} f g \nu_j dS - \int_\Omega f(\partial_j g) d^n x, \qquad 1 \le j \le n, \tag{9.63}$$

for $f, g \in C^1(\overline{\Omega})$.

Problem* 9.30. *Show* (9.53)*. (Hint: Problem 2.1)*

Problem* 9.31 (Leibniz integral rule)**.** *Suppose* $f \in C^1(R)$*, where* $R = [a_1, b_1] \times [a_2, b_2]$ *is some rectangle, and* $g \in C^1([a_1, b_1], [a_2, b_2])$*. Show*

$$\frac{d}{dx} \int_{a_2}^{g(x)} f(x, y) dy = f(x, g(x)) g'(x) + \int_{a_2}^{g(x)} \frac{\partial}{\partial x} f(x, y) dy.$$

9.5. Appendix: Transformation of Lebesgue–Stieltjes integrals

In this section we will look at Borel measures on $\mathbb{R}$. In particular, we want to derive a generalized substitution rule.

As a preparation we will need a generalization of the usual inverse which works for arbitrary nondecreasing functions. Such a generalized inverse arises, for example, as quantile functions in probability theory.

So we look at nondecreasing functions $f : \mathbb{R} \to \mathbb{R}$. By monotonicity the limits from left and right exist at every point and we will denote them by

$$f(x\pm) := \lim_{\varepsilon \downarrow 0} f(x \pm \varepsilon). \tag{9.64}$$

Clearly we have $f(x-) \le f(x+)$ and a strict inequality can occur only at a countable number of points. By monotonicity the value of f has to lie between these two values $f(x-) \le f(x) \le f(x+)$. It will also be convenient to extend f to a function on the extended reals $\mathbb{R} \cup \{-\infty, +\infty\}$. Again by monotonicity the limits $f(\pm\infty\mp) = \lim_{x \to \pm\infty} f(x)$ exist and we will set $f(\pm\infty\pm) = f(\pm\infty)$.

If we want to define an inverse, problems will occur at points where f jumps and on intervals where f is constant. Informally speaking, if f jumps, then the corresponding jump will be missing in the domain of the inverse and if f is constant, the inverse will be multivalued. For the first case there is a natural fix by choosing the inverse to be constant along the missing interval.

In particular, observe that this natural choice is independent of the actual value of f at the jump and hence the inverse *loses* this information. The second case will result in a jump for the inverse function and here there is no natural choice for the value at the jump (except that it must be between the left and right limits such that the inverse is again a nondecreasing function).

To give a precise definition it will be convenient to look at relations instead of functions. Recall that a (binary) relation R on $\mathbb{R}$ is a subset of $\mathbb{R}^2$.

To every nondecreasing function f associate the relation

$$\Gamma(f) := \{(x,y)|y \in [f(x-), f(x+)]\}. \tag{9.65}$$

Note that $\Gamma(f)$ does not depend on the values of f at a discontinuity and f can be partially recovered from $\Gamma(f)$ using $f(x-) = \inf \Gamma(f)(x)$ and $f(x+) = \sup \Gamma(f)(x)$, where $\Gamma(f)(x) := \{y|(x,y) \in \Gamma(f)\} = [f(x-), f(x+)]$. Moreover, the relation $\Gamma(f)$ is nondecreasing in the sense that $x_1 < x_2$ implies $y_1 \le y_2$ for $(x_1,y_1),(x_2,y_2) \in \Gamma(f)$ (just note $y_1 \le f(x_1+) \le f(x_2-) \le y_2$). It is uniquely defined as the largest relation containing the graph of f with this property.

The graph of any reasonable inverse should be a subset of the inverse relation

$$\Gamma(f)^{-1} := \{(y,x)|(x,y) \in \Gamma(f)\} \tag{9.66}$$

and we will call any function f^{-1} whose graph is a subset of $\Gamma(f)^{-1}$ a **generalized inverse** of f. Note that any generalized inverse is again nondecreasing since a pair of points $(y_1,x_1),(y_2,x_2) \in \Gamma(f)^{-1}$ with $y_1 < y_2$ and $x_1 > x_2$ would contradict the fact that $\Gamma(f)$ is nondecreasing. Moreover, since $\Gamma(f)^{-1}$ and $\Gamma(f^{-1})$ are two nondecreasing relations containing the graph of f^{-1}, we conclude

$$\Gamma(f^{-1}) = \Gamma(f)^{-1} \tag{9.67}$$

since both are maximal. In particular, it follows that if f^{-1} is a generalized inverse of f then f is a generalized inverse of f^{-1}.

There are two particular choices, namely the left continuous version $f_-^{-1}(y) := \inf \Gamma(f)^{-1}(y)$ and the right continuous version $f_+^{-1}(y) := \sup \Gamma(f)^{-1}(y)$. It is straightforward to verify that they can be equivalently defined via

$$\begin{aligned} f_-^{-1}(y) &:= \inf f^{-1}([y,\infty)) = \sup f^{-1}((-\infty,y)), \\ f_+^{-1}(y) &:= \inf f^{-1}((y,\infty)) = \sup f^{-1}((-\infty,y]). \end{aligned} \tag{9.68}$$

For example, $\inf f^{-1}([y,\infty)) = \inf\{x|(x,\tilde{y}) \in \Gamma(f),\ \tilde{y} \ge y\} = \inf \Gamma(f)^{-1}(y)$. The first one is typically used in probability theory, where it corresponds to the quantile function of a distribution.

If f is strictly increasing the generalized inverse f^{-1} extends the usual inverse by setting it constant on the gaps missing in the range of f. In particular we have $f^{-1}(f(x)) = x$ and $f(f^{-1}(y)) = y$ for y in the range of f. The purpose of the next lemma is to investigate to what extend this remains valid for a generalized inverse.

Note that for every y there is some x with $y \in [f(x-), f(x+)]$. Moreover, if we can find two values, say x_1 and x_2, with this property, then $f(x) = y$ is constant for $x \in (x_1, x_2)$. Hence, the set of all such x is an interval which is closed since at the left, right boundary point the left, right limit equals y, respectively.

We collect a few simple facts for later use.

Lemma 9.20. *Let f be nondecreasing.*

(i) *$f_-^{-1}(y) \le x$ if and only if $y \le f(x+)$.*

(i') *$f_+^{-1}(y) \ge x$ if and only if $y \ge f(x-)$.*

(ii) *$f_-^{-1}(f(x)) \le x \le f_+^{-1}(f(x))$ with equality on the left, right iff f is not constant to the right, left of x, respectively.*

(iii) *$f(f^{-1}(y)-) \le y \le f(f^{-1}(y)+)$ with equality on the left, right iff f^{-1} is not constant to right, left of y, respectively.*

Proof. Item (i) follows since both claims are equivalent to $y \le f(\tilde{x})$ for all $\tilde{x} > x$. Similarly for (i'). Item (ii) follows from $f_-^{-1}(f(x)) = \inf f^{-1}([f(x), \infty)) = \inf\{\tilde{x} | f(\tilde{x}) \ge f(x)\} \le x$ with equality iff $f(\tilde{x}) < f(x)$ for $\tilde{x} < x$. Similarly for the other inequality. Item (iii) follows by reversing the roles of f and f^{-1} in (ii). □

In particular, $f(f^{-1}(y)) = y$ if f is continuous. We will also need the set

$$L(f) := \{y | f^{-1}((y, \infty)) = (f_+^{-1}(y), \infty)\}. \tag{9.69}$$

Note that $y \notin L(f)$ if and only if there is some x such that $y \in [f(x-), f(x))$.

Lemma 9.21. *Let $m : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function on $\mathbb{R}$ and μ its associated measure via (8.20). Let $f(x)$ be a nondecreasing function on $\mathbb{R}$ such that $\mu((0, \infty)) < \infty$ if f is bounded above and $\mu((-\infty, 0)) < \infty$ if f is bounded below.*

Then $f_\star\mu$ is a Borel measure whose distribution function coincides up to a constant with $m_+ \circ f_+^{-1}$ at every point y which is in $L(f)$ or satisfies $\mu(\{f_+^{-1}(y)\}) = 0$. If $y \in [f(x-), f(x))$ and $\mu(\{f_+^{-1}(y)\}) > 0$, then $m_+ \circ f_+^{-1}$ jumps at $f(x-)$ and $(f_\star\mu)(y)$ jumps at $f(x)$.

Proof. First of all note that the assumptions in case f is bounded from above or below ensure that $(f_\star\mu)(K) < \infty$ for any compact interval. Moreover, we can assume $m = m_+$ without loss of generality. Now note that we have $f^{-1}((y,\infty)) = (f^{-1}(y),\infty)$ for $y \in L(f)$ and $f^{-1}((y,\infty)) = [f^{-1}(y),\infty)$ else. Hence

$$\begin{aligned}(f_\star\mu)((y_0,y_1]) &= \mu(f^{-1}((y_0,y_1])) = \mu((f^{-1}(y_0),f^{-1}(y_1)])\\ &= m(f_+^{-1}(y_1)) - m(f_+^{-1}(y_0)) = (m\circ f_+^{-1})(y_1) - (m\circ f_+^{-1})(y_0)\end{aligned}$$

if y_j is either in $L(f)$ or satisfies $\mu(\{f_+^{-1}(y_j)\}) = 0$. For the last claim observe that $f^{-1}((y,\infty))$ will jump from $(f_+^{-1}(y),\infty)$ to $[f_+^{-1}(y),\infty)$ at $y = f(x)$. □

Example 9.18. For example, consider $f(x) = \chi_{[0,\infty)}(x)$ and $\mu = \Theta$, the Dirac measure centered at 0 (note that $\Theta(x) = f(x)$). Then

$$f_+^{-1}(y) = \begin{cases} +\infty, & 1 \le y,\\ 0, & 0 \le y < 1,\\ -\infty, & y < 0,\end{cases}$$

and $L(f) = (-\infty,0)\cup[1,\infty)$. Moreover, $\mu(f_+^{-1}(y)) = \chi_{[0,\infty)}(y)$ and $(f_\star\mu)(y) = \chi_{[1,\infty)}(y)$. If we choose $g(x) = \chi_{(0,\infty)}(x)$, then $g_+^{-1}(y) = f_+^{-1}(y)$ and $L(g) = \mathbb{R}$. Hence $\mu(g_+^{-1}(y)) = \chi_{[0,\infty)}(y) = (g_\star\mu)(y)$. ⋄

For later use it is worth while to single out the following consequence:

Corollary 9.22. *Let m, f be as in the previous lemma and denote by μ, $\nu_\pm$ the measures associated with m, $m_\pm \circ f^{-1}$, respectively. Then, $(f_\mp)_\star\mu = \nu_\pm$ and hence*

$$\int g\, d(m_\pm \circ f^{-1}) = \int (g\circ f_\mp)\, dm. \tag{9.70}$$

In the special case where μ is Lebesgue measure this reduces to a way of expressing the Lebesgue–Stieltjes integral as a Lebesgue integral via

$$\int g\, dh = \int g(h^{-1}(y))dy. \tag{9.71}$$

If we choose f to be the distribution function of μ we get the following generalization of the **integration by substitution** rule. To formulate it we introduce

$$i_m(y) := m(m_-^{-1}(y)). \tag{9.72}$$

Note that $i_m(y) = y$ if m is continuous. By $\mathrm{conv}(\mathrm{Ran}(m))$ we denote the convex hull of the range of m.

Corollary 9.23. *Suppose m, m are two nondecreasing functions on $\mathbb{R}$ with n right continuous. Then we have*

$$\int_{\mathbb{R}} (g\circ m)\, d(n\circ m) = \int_{\mathrm{conv}(\mathrm{Ran}(m))} (g\circ i_m)dn \tag{9.73}$$

for any Borel function g which is either nonnegative or for which one of the two integrals is finite. Similarly, if n is left continuous and i_m is replaced by $m(m_+^{-1}(y))$.

Hence the usual $\int_{\mathbb{R}}(g\circ m)\,d(n\circ m) = \int_{\mathrm{Ran}(m)} g\,dn$ only holds if m is continuous. In fact, the right-hand side looses all point masses of μ. The above formula fixes this problem by rendering g constant along a gap in the range of m and includes the gap in the range of integration such that it makes up for the lost point mass. It should be compared with the previous example!

If one does not want to bother with i_m one can at least get inequalities for monotone g.

Corollary 9.24. *Suppose m, n are nondecreasing functions on $\mathbb{R}$ and g is monotone. Then we have*

$$\int_{\mathbb{R}}(g\circ m)\,d(n\circ m) \le \int_{\mathrm{conv}(\mathrm{Ran}(m))} g\,dn \tag{9.74}$$

if m, n are right continuous and g nonincreasing or m, n left continuous and g nondecreasing. If m, n are right continuous and g nondecreasing or m, n left continuous and g nonincreasing the inequality has to be reversed.

Proof. Immediate from the previous corollary together with $i_m(y) \le y$ if $y = f(x) = f(x+)$ and $i_m(y) \ge y$ if $y = f(x) = f(x-)$ according to Lemma 9.20. □

Problem* 9.32. *Show* (9.68).

Problem 9.33. *Show that* $\Gamma(f)\circ\Gamma(f^{-1}) = \{(y,z)|y,z\in[f(x-),f(x+)]$ *for some* $x\}$ *and* $\Gamma(f^{-1})\circ\Gamma(f) = \{(y,z)|f(y+) > f(z-)$ *or* $f(y-) < f(z_+)\}$.

Problem 9.34. *Let* $d\mu(\lambda) := \chi_{[0,1]}(\lambda)d\lambda$ *and* $f(\lambda) := \chi_{(-\infty,t]}(\lambda)$, $t\in\mathbb{R}$. *Compute* $f_\star\mu$.

9.6. Appendix: The connection with the Riemann integral

In this section we want to investigate the connection with the Riemann integral. We restrict our attention to compact intervals $[a,b]$ and bounded real-valued functions f. A **partition** of $[a,b]$ is a finite set $P = \{x_0,\dots,x_n\}$ with

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b. \tag{9.75}$$

The number

$$\|P\| := \max_{1\le j\le n} x_j - x_{j-1} \tag{9.76}$$

is called the norm of P. Given a partition P and a bounded real-valued function f we can define

$$s_{P,f,-}(x) := \sum_{j=1}^{n} m_j \chi_{[x_{j-1},x_j)}(x), \qquad m_j := \inf_{x\in[x_{j-1},x_j]} f(x), \tag{9.77}$$

$$s_{P,f,+}(x) := \sum_{j=1}^{n} M_j \chi_{[x_{j-1},x_j)}(x), \qquad M_j := \sup_{x\in[x_{j-1},x_j]} f(x), \tag{9.78}$$

Hence $s_{P,f,-}(x)$ is a step function approximating f from below and $s_{P,f,+}(x)$ is a step function approximating f from above. In particular,

$$m \le s_{P,f,-}(x) \le f(x) \le s_{P,f,+}(x) \le M, \quad m := \inf_{x\in[a,b]} f(x), M := \sup_{x\in[a,b]} f(x). \tag{9.79}$$

Moreover, we can define the upper and lower Riemann sum associated with P as

$$L(P,f) := \sum_{j=1}^{n} m_j(x_j - x_{j-1}), \qquad U(P,f) := \sum_{j=1}^{n} M_j(x_j - x_{j-1}). \tag{9.80}$$

Of course, $L(f,P)$ is just the Lebesgue integral of $s_{P,f,-}$ and $U(f,P)$ is the Lebesgue integral of $s_{P,f,+}$. In particular, $L(P,f)$ approximates the area under the graph of f from below and $U(P,f)$ approximates this area from above.

By the above inequality

$$m\,(b-a) \le L(P,f) \le U(P,f) \le M\,(b-a). \tag{9.81}$$

We say that the partition P_2 is a refinement of P_1 if $P_1 \subseteq P_2$ and it is not hard to check, that in this case

$$s_{P_1,f,-}(x) \le s_{P_2,f,-}(x) \le f(x) \le s_{P_2,f,+}(x) \le s_{P_1,f,+}(x) \tag{9.82}$$

as well as

$$L(P_1,f) \le L(P_2,f) \le U(P_2,f) \le U(P_1,f). \tag{9.83}$$

Hence we define the **lower, upper Riemann integral** of f as

$$\underline{\int} f(x)dx := \sup_P L(P,f), \qquad \overline{\int} f(x)dx := \inf_P U(P,f), \tag{9.84}$$

respectively. Since for arbitrary partitions P and Q we have

$$L(P,f) \le L(P\cup Q,f) \le U(P\cup Q,f) \le U(Q,f). \tag{9.85}$$

we obtain

$$m\,(b-a) \le \underline{\int} f(x)dx \le \overline{\int} f(x)dx \le M\,(b-a). \tag{9.86}$$

We will call f **Riemann integrable** if both values coincide and the common value will be called the **Riemann integral** of f.

Example 9.19. Let $[a,b] := [0,1]$ and $f(x) := \chi_{\mathbb{Q}}(x)$. Then $s_{P,f,-}(x) = 0$ and $s_{P,f,+}(x) = 1$. Hence $\underline{\int} f(x)dx = 0$ and $\overline{\int} f(x)dx = 1$ and f is not Riemann integrable.

On the other hand, every continuous function $f \in C[a,b]$ is Riemann integrable (Problem 9.35). ⋄

Example 9.20. Let f nondecreasing, then f is integrable. In fact, since $m_j = f(x_{j-1})$ and $M_j = f(x_j)$ we obtain

$$U(f,P) - L(f,P) \le \|P\| \sum_{j=1}^{n} (f(x_j) - f(x_{j-1})) = \|P\|(f(b) - f(a))$$

and the claim follows (cf. also the next lemma). Similarly nonincreasing functions are integrable. ⋄

Lemma 9.25. *A function f is Riemann integrable if and only if there exists a sequence of partitions P_n such that*

$$\lim_{n\to\infty} L(P_n, f) = \lim_{n\to\infty} U(P_n, f). \tag{9.87}$$

In this case the above limits equal the Riemann integral of f and P_n can be chosen such that $P_n \subseteq P_{n+1}$ and $\|P_n\| \to 0$.

Proof. If there is such a sequence of partitions then f is integrable by $\lim_n L(P_n, f) \le \sup_P L(P, f) \le \inf_P U(P, f) \le \lim_n U(P_n, f)$.

Conversely, given an integrable f, there is a sequence of partitions $P_{L,n}$ such that $\underline{\int} f(x)dx = \lim_n L(P_{L,n}, f)$ and a sequence $P_{U,n}$ such that $\overline{\int} f(x)dx = \lim_n U(P_{U,n}, f)$. By (9.83) the common refinement $P_n = P_{L,n} \cup P_{U,n}$ is the partition we are looking for. Since, again by (9.83), any refinement will also work, the last claim follows. □

Note that when computing the Riemann integral as in the previous lemma one could choose instead of m_j or M_j any value in $[m_j, M_j]$ (e.g. $f(x_{j-1})$ or $f(x_j)$).

With the help of this lemma we can give a characterization of Riemann integrable functions and show that the Riemann integral coincides with the Lebesgue integral.

Theorem 9.26 (Lebesgue)**.** *A bounded measurable function $f : [a,b] \to \mathbb{R}$ is Riemann integrable if and only if the set of its discontinuities is of Lebesgue measure zero. Moreover, in this case the Riemann and the Lebesgue integral of f coincide.*

Proof. Suppose f is Riemann integrable and let P_j be a sequence of partitions as in Lemma 9.25. Then $s_{f,P_j,-}(x)$ will be monotone and hence converge to some function $s_{f,-}(x) \le f(x)$. Similarly, $s_{f,P_j,+}(x)$ will converge to some function $s_{f,+}(x) \ge f(x)$. Moreover, by dominated convergence

$$0 = \lim_j \int \big(s_{f,P_j,+}(x) - s_{f,P_j,-}(x)\big)dx = \int \big(s_{f,+}(x) - s_{f,-}(x)\big)dx$$

and thus by Lemma 9.6 $s_{f,+}(x) = s_{f,-}(x)$ almost everywhere. Moreover, f is continuous at every x at which equality holds and which is not in any of the partitions. Since the first as well as the second set have Lebesgue measure zero, f is continuous almost everywhere and

$$\lim_j L(P_j, f) = \lim_j U(P_j, f) = \int s_{f,\pm}(x)dx = \int f(x)dx.$$

Conversely, let f be continuous almost everywhere and choose some sequence of partitions P_j with $\|P_j\| \to 0$. Then at every x where f is continuous we have $\lim_j s_{f,P_j,\pm}(x) = f(x)$ implying

$$\lim_j L(P_j, f) = \int s_{f,-}(x)dx = \int f(x)dx = \int s_{f,+}(x)dx = \lim_j U(P_j, f)$$

by the dominated convergence theorem. □

Note that if f is not assumed to be measurable, the above proof still shows that f satisfies $s_{f,-} \le f \le s_{f,+}$ for two measurable functions $s_{f,\pm}$ which are equal almost everywhere. Hence if we replace the Lebesgue measure by its completion, we can drop this assumption.

Finally, recall that if one endpoint is unbounded or f is unbounded near one endpoint, one defines the **improper Riemann integral** by taking limits towards this endpoint. More specifically, if f is Riemann integrable for every $(a, c) \subset (a, b)$ one defines

$$\int_a^b f(x)dx := \lim_{c\uparrow b} \int_a^c f(x)dx \tag{9.88}$$

with an analogous definition if f is Riemann integrable for every $(c, b) \subset (a, b)$. Note that in this case improper integrability no longer implies Lebesgue integrability unless $|f(x)|$ has a finite improper integral. The prototypical example being the Dirichlet integral

$$\int_0^\infty \frac{\sin(x)}{x}dx = \lim_{c\to\infty} \int_0^c \frac{\sin(x)}{x}dx = \frac{\pi}{2} \tag{9.89}$$

(cf. Problem 14.24) which does not exist as a Lebesgue integral since

$$\int_0^\infty \frac{|\sin(x)|}{x}dx \ge \sum_{k=0}^\infty \int_{\pi/4}^{3\pi/4} \frac{|\sin(k\pi + x)|}{k\pi + 3/4}dx \ge \frac{1}{2\sqrt{2}} \sum_{k=1}^\infty \frac{1}{k} = \infty. \tag{9.90}$$

Problem* 9.35. *Show that for any function $f \in C[a,b]$ we have*

$$\lim_{\|P\|\to 0} L(P,f) = \lim_{\|P\|\to 0} U(P,f).$$

In particular, f is Riemann integrable.

Problem 9.36. *Prove that the Riemann integral is linear: If f, g are Riemann integrable and $\alpha \in \mathbb{R}$, then αf and $f+g$ are Riemann integrable with $\int (f+g)dx = \int f\,dx + \int g\,dx$ and $\int \alpha f\,dx = \alpha \int f\,dx$.*

Problem 9.37. *Suppose f is Riemann integrable and ϕ is Lipschitz continuous on the range of f, then $\phi \circ f$ is Riemann integrable. Moreover, show that if f, g are Riemann integrable, so is fg. (Hint: The second claim can be reduced to the first using $\phi(x) = x^2$.)*

Problem 9.38. *Show that the uniform limit of Riemann integrable functions is again Riemann integrable. Conclude that in the previous problem it suffices to assume that ψ is continuous.*

Problem 9.39. *Let $\{q_n\}_{n\in\mathbb{N}}$ be an enumeration of the rational numbers in $[0,1)$. Show that*

$$f(x) := \sum_{n\in\mathbb{N}: q_n < x} \frac{1}{2^n}$$

is discontinuous at every q_n but still Riemann integrable.

Chapter 10

The Lebesgue spaces L^p

10.1. Functions almost everywhere

We fix some measure space (X, Σ, μ) and define the L^p norm by

$$\|f\|_p := \left(\int_X |f|^p \, d\mu\right)^{1/p}, \qquad 1 \le p, \tag{10.1}$$

and denote by $\mathcal{L}^p(X, d\mu)$ the set of all complex-valued measurable functions for which $\|f\|_p$ is finite. First of all note that $\mathcal{L}^p(X, d\mu)$ is a vector space, since $|f+g|^p \le 2^p \max(|f|,|g|)^p = 2^p \max(|f|^p, |g|^p) \le 2^p(|f|^p + |g|^p)$. Of course our hope is that $\mathcal{L}^p(X, d\mu)$ is a Banach space. However, Lemma 9.6 implies that there is a small technical problem (recall that a property is said to hold almost everywhere if the set where it fails to hold is contained in a set of measure zero):

Lemma 10.1. *Let f be measurable. Then*

$$\int_X |f|^p \, d\mu = 0 \tag{10.2}$$

if and only if $f(x) = 0$ almost everywhere with respect to μ.

Thus $\|f\|_p = 0$ only implies $f(x) = 0$ for almost every x, but not for all! Hence $\|.\|_p$ is not a norm on $\mathcal{L}^p(X, d\mu)$. The way out of this misery is to identify functions which are equal almost everywhere: Let

$$\mathcal{N}(X, d\mu) := \{f | f(x) = 0 \ \mu\text{-almost everywhere}\}. \tag{10.3}$$

Then $\mathcal{N}(X, d\mu)$ is a linear subspace of $\mathcal{L}^p(X, d\mu)$ and we can consider the quotient space

$$L^p(X, d\mu) := \mathcal{L}^p(X, d\mu)/\mathcal{N}(X, d\mu). \tag{10.4}$$

If $d\mu$ is the Lebesgue measure on $X \subseteq \mathbb{R}^n$, we simply write $L^p(X)$. Observe that $\|f\|_p$ is well defined on $L^p(X, d\mu)$.

Even though the elements of $L^p(X, d\mu)$ are, strictly speaking, equivalence classes of functions, we will still treat them as functions for notational convenience. However, if we do so it is important to ensure that every statement made does not depend on the representative in the equivalence classes. In particular, note that for $f \in L^p(X, d\mu)$ the value $f(x)$ is not well defined (unless there is a continuous representative and continuous functions with different values are in different equivalence classes, e.g., in the case of Lebesgue measure).

With this modification we are back in business since $L^p(X, d\mu)$ turns out to be a Banach space. We will show this in the following sections. Moreover, note that $L^2(X, d\mu)$ is a Hilbert space with scalar product given by

$$\langle f, g\rangle := \int_X f(x)^* g(x) d\mu(x). \tag{10.5}$$

But before that let us also define $L^\infty(X, d\mu)$. It should be the set of bounded measurable functions $B(X)$ together with the sup norm. The only problem is that if we want to identify functions equal almost everywhere, the supremum is no longer independent of the representative in the equivalence class. The solution is the **essential supremum**

$$\|f\|_\infty := \inf\{C \,|\, \mu(\{x| \,|f(x)| > C\}) = 0\}. \tag{10.6}$$

That is, C is an essential bound if $|f(x)| \le C$ almost everywhere and the essential supremum is the infimum over all essential bounds.

Example 10.1. If λ is the Lebesgue measure, then the essential sup of $\chi_{\mathbb{Q}}$ with respect to λ is 0. If Θ is the Dirac measure centered at 0, then the essential sup of $\chi_{\mathbb{Q}}$ with respect to Θ is 1 (since $\chi_{\mathbb{Q}}(0) = 1$, and $x = 0$ is the only point which counts for Θ). ◇

As before we set

$$L^\infty(X, d\mu) := B(X)/\mathcal{N}(X, d\mu) \tag{10.7}$$

and observe that $\|f\|_\infty$ is independent of the representative from the equivalence class.

If you wonder where the ∞ comes from, have a look at Problem 10.2.

Since the support of a *function* in L^p is also not well defined one uses the **essential support** in this case:

$$\operatorname{supp}(f) = X \setminus \bigcup\{O | f = 0 \ \mu\text{-almost everywhere on } O \subseteq X \text{ open}\}. \tag{10.8}$$

In other words, x is in the essential support if for every neighborhood the set of points where f does not vanish has positive measure. Here we use the same notation as for functions and it should be understood from the context which one is meant. Note that the essential support is always smaller than the support (since we get the latter if we require f to vanish everywhere on O in the above definition).

Example 10.2. The support of $\chi_{\mathbb{Q}}$ is $\overline{\mathbb{Q}} = \mathbb{R}$ but the essential support with respect to Lebesgue measure is $\emptyset$ since the function is 0 a.e. ⋄

If X is a locally compact Hausdorff space (together with the Borel sigma algebra), a function is called **locally integrable** if it is integrable when restricted to any compact subset $K \subseteq X$. The set of all (equivalence classes of) locally integrable functions will be denoted by $L^1_{loc}(X, d\mu)$. We will say that $f_n \to f$ in $L^1_{loc}(X, d\mu)$ if this holds on $L^1(K, d\mu)$ for all compact subsets $K \subseteq X$. Of course this definition extends to L^p for any $1 \leq p \leq \infty$.

Problem* 10.1. *Let $\|.\|$ be a seminorm on a vector space X. Show that $N := \{x \in X | \, \|x\| = 0\}$ is a vector space. Show that the quotient space X/N is a normed space with norm $\|x + N\| := \|x\|$.*

Problem* 10.2. *Suppose $\mu(X) < \infty$. Show that $L^\infty(X, d\mu) \subseteq L^p(X, d\mu)$ and*

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty, \qquad f \in L^\infty(X, d\mu).$$

Problem 10.3. *Construct a function $f \in L^p(0, 1)$ which has a singularity at every rational number in $[0, 1]$ (such that the essential supremum is infinite on every open subinterval). (Hint: Start with the function $f_0(x) = |x|^{-\alpha}$ which has a single singularity at 0, then $f_j(x) = f_0(x - x_j)$ has a singularity at x_j.)*

Problem 10.4. *Show that for a continuous function on $\mathbb{R}^n$ the support and the essential support with respect to Lebesgue measure coincide.*

10.2. Jensen ≤ Hölder ≤ Minkowski

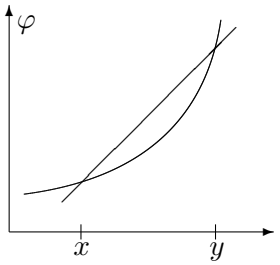
As a preparation for proving that L^p is a Banach space, we will need Hölder's inequality, which plays a central role in the theory of L^p spaces. In particular, it will imply Minkowski's inequality, which is just the triangle inequality for L^p. Our proof is based on Jensen's inequality and emphasizes the connection with convexity. In fact, the triangle inequality just states that a norm is convex:

$$\|(1 - \lambda)f + \lambda\, g\| \leq \lambda\, \|f\| + \lambda\|g\|, \quad \lambda \in (0, 1). \tag{10.9}$$

Recall that a real function φ defined on an open interval (a,b) is called **convex** if

$$\varphi((1-\lambda)x+\lambda y) \le (1-\lambda)\varphi(x)+\lambda\varphi(y), \quad \lambda\in(0,1), \tag{10.10}$$

that is, on (x,y) the graph of $\varphi(x)$ lies below or on the line connecting $(x,\varphi(x))$ and $(y,\varphi(y))$:



If the inequality is strict, then φ is called **strictly convex**. A function φ is **concave** if $-\varphi$ is convex.

Lemma 10.2. *Let $\varphi:(a,b)\to\mathbb{R}$ be convex. Then*

(i) *φ is locally Lipschitz continuous.*

(ii) *The left/right derivatives $\varphi'_\pm(x)=\lim_{\varepsilon\downarrow 0}\frac{\varphi(x\pm\varepsilon)-\varphi(x)}{\pm\varepsilon}$ exist and are monotone nondecreasing. Moreover, φ' exists except at a countable number of points.*

(iii) *For fixed x we have $\varphi(y)\ge\varphi(x)+\alpha(y-x)$ for every α with $\varphi'_-(x)\le\alpha\le\varphi'_+(x)$. The inequality is strict for $y\ne x$ if φ is strictly convex.*

Proof. Abbreviate $D(x,y)=D(y,x):=\frac{\varphi(y)-\varphi(x)}{y-x}$ and observe (use $z=(1-\lambda)x+\lambda y$) that the definition implies

$$D(x,z)\le D(x,y)\le D(y,z), \quad x<z<y,$$

where the inequalities are strict if φ is strictly convex. Hence $\varphi'_\pm(x)$ exist and we have $\varphi'_-(x)\le\varphi'_+(x)\le\varphi'_-(y)\le\varphi'_+(y)$ for $x<y$. So (ii) follows after observing that a monotone function can have at most a countable number of jumps. Next

$$\varphi'_+(x)\le D(y,x)\le\varphi'_-(y)$$

shows $\varphi(y)\ge\varphi(x)+\varphi'_\pm(x)(y-x)$ if $\pm(y-x)>0$ and proves (iii). Moreover, $\varphi'_+(z)\le D(y,x)\le\varphi'_-(\tilde z)$ for $z<x, y<\tilde z$ proves (i). □

Remark: It is not hard to see that $\varphi\in C^1$ is convex if and only if $\varphi'(x)$ is monotone nondecreasing (e.g., $\varphi''\ge 0$ if $\varphi\in C^2$) — Problem 10.5.

With these preparations out of the way we can show

Theorem 10.3 (Jensen's inequality)**.** *Let $\varphi : (a, b) \to \mathbb{R}$ be convex ($a = -\infty$ or $b = \infty$ being allowed). Suppose μ is a finite measure satisfying $\mu(X) = 1$ and $f \in \mathcal{L}^1(X, d\mu)$ with $a < f(x) < b$. Then the negative part of $\varphi \circ f$ is integrable and*

$$\varphi\Big(\int_X f\, d\mu\Big) \le \int_X (\varphi \circ f)\, d\mu. \tag{10.11}$$

For $f \ge 0$ the requirement that f is integrable can be dropped if $\varphi(b)$ is understood as $\lim_{x\to b} \varphi(x)$.

Proof. By (iii) of the previous lemma we have

$$\varphi(f(x)) \ge \varphi(I) + \alpha(f(x) - I), \qquad I = \int_X f\, d\mu \in (a, b).$$

This shows that the negative part of $\varphi \circ f$ is integrable and integrating over X finishes the proof in the case $f \in \mathcal{L}^1$. If $f \ge 0$ we note that for $X_n = \{x \in X | f(x) \le n\}$ the first part implies

$$\varphi\Big(\frac{1}{\mu(X_n)} \int_{X_n} f\, d\mu\Big) \le \frac{1}{\mu(X_n)} \int_{X_n} \varphi(f)\, d\mu.$$

Taking $n \to \infty$ the claim follows from $X_n \nearrow X$ and the monotone convergence theorem. □

Observe that if φ is strictly convex, then equality can only occur if f is constant.

Now we are ready to prove

Theorem 10.4 (Hölder's inequality)**.** *Let p and q be **dual** indices; that is,*

$$\frac{1}{p} + \frac{1}{q} = 1 \tag{10.12}$$

with $1 \le p \le \infty$. If $f \in L^p(X, d\mu)$ and $g \in L^q(X, d\mu)$, then $fg \in L^1(X, d\mu)$ and

$$\|f\, g\|_1 \le \|f\|_p \|g\|_q. \tag{10.13}$$

Proof. The case $p = 1$, $q = \infty$ (respectively $p = \infty$, $q = 1$) follows directly from the properties of the integral and hence it remains to consider $1 < p, q < \infty$.

First of all it is no restriction to assume $\|g\|_q = 1$. Let $A = \{x | \, |g(x)| > 0\}$, then (note $(1 - q)p = -q$)

$$\|f\, g\|_1^p = \Big| \int_A |f|\, |g|^{1-q} |g|^q d\mu \Big|^p \le \int_A (|f|\, |g|^{1-q})^p |g|^q d\mu = \int_A |f|^p d\mu \le \|f\|_p^p,$$

where we have used Jensen's inequality with $\varphi(x) = |x|^p$ applied to the function $h = |f|\, |g|^{1-q}$ and measure $d\nu = |g|^q d\mu$ (note $\nu(X) = \int |g|^q d\mu = \|g\|_q^q = 1$). □

Note that in the special case $p = 2$ we have $q = 2$ and Hölder's inequality reduces to the Cauchy–Schwarz inequality. For a generalization see Problem 10.8. Moreover, in the case $1 < p < \infty$ the function x^p is strictly convex and equality will occur precisely if $|f|$ is a multiple of $|g|^{q-1}$ or g is trivial. This gives us a

Corollary 10.5. *Consider $f \in L^p(X, d\mu)$ with $1 \leq p < \infty$ and let q be the corresponding dual index, $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\|f\|_p = \sup_{\|g\|_q=1} \left| \int_X f g \, d\mu \right|. \tag{10.14}$$

If every set of infinite measure has a subset of finite positive measure (e.g. if μ is σ-finite), then the claim also holds for $p = \infty$.

Proof. In the case $1 < p < \infty$ equality is attained for $g = c^{-1} \operatorname{sign}(f^*)|f|^{p-1}$, where $c = \||f|^{p-1}\|_q$ (assuming $c > 0$ w.l.o.g.). In the case $p = 1$ equality is attained for $g = \operatorname{sign}(f^*)$. Now let us turn to the case $p = \infty$. For every $\varepsilon > 0$ the set $A_\varepsilon = \{x| \, |f(x)| \geq \|f\|_\infty - \varepsilon\}$ has positive measure. Moreover, by assumption on μ we can choose a subset $B_\varepsilon \subseteq A_\varepsilon$ with finite positive measure. Then $g_\varepsilon = \operatorname{sign}(f^*)\chi_{B_\varepsilon}/\mu(B_\varepsilon)$ satisfies $\int_X f g_\varepsilon \, d\mu \geq \|f\|_\infty - \varepsilon$. □

Of course it suffices to take the sup in (10.14) over a dense set of L^q (e.g. integrable simple functions — see Problem 10.18). Moreover, note that the extra assumption for $p = \infty$ is crucial since if there is a set of infinite measure which has no subset with finite positive measure, then every integrable function must vanish on this subset.

If it is not a priori known that $f \in L^p$, the following generalization will be useful.

Lemma 10.6. *Suppose μ is σ-finite. Consider $L^p(X, d\mu)$ with if $1 \leq p \leq \infty$ and let q be the corresponding dual index, $\frac{1}{p} + \frac{1}{q} = 1$. If $fs \in L^1$ for every nonnegative simple function $s \in L^q$, then*

$$\|f\|_p = \sup_{s \text{ simple, } \|s\|_q=1} \left| \int_X f s \, d\mu \right|.$$

If f is nonnegative it suffices to take the sup over nonnegative simple functions.

Proof. If $\|f\|_p < \infty$ the claim follows from the previous corollary since simple functions are dense (Problem 10.18). Conversely, if $\|f\|_p = \infty$ we can split f into nonnegative functions $f = f_1 - f_2 + \mathrm{i}(f_3 - f_4)$ as usual and assume $\|f_1\|_p = \infty$ without loss of generality. Now choose $\hat{s}_n \nearrow f_1$ as in (9.6). Moreover, since μ is σ-finite we can find $X_n \nearrow X$ with $\mu(X_n) < \infty$.

Then $\tilde{s}_n = \chi_{X_n}\hat{s}_n$ will be in L^p and will still satisfy $\tilde{s}_n \nearrow f_1$. Now if $1 < p < \infty$ choose $s_n = (\int \tilde{s}_n^p d\mu)^{-1/q}\tilde{s}_n^{p-1}$ and observe

$$\left|\int_X f s_n d\mu\right| \geq \int_X f_1 s_n d\mu = \left(\frac{\int_X \tilde{s}_n^{p-1} f_1 d\mu}{\int_X \tilde{s}_n^p d\mu}\right)^{1/q} \left(\int_X \tilde{s}_n^{p-1} f_1 d\mu\right)^{1/p}$$
$$\geq \left(\int_X \tilde{s}_n^{p-1} f_1 d\mu\right)^{1/p} \to \|f_1\|_p = \infty$$

by monotone convergence. If $p = 1$ simply choose $s_n = \chi_{X_n \cap \operatorname{supp}(f_1)}$. If $p = \infty$ then for every n there is some m such that $\mu(A_n) > 0$, where $A_n = \{x \in X_m | f(x) \geq n\}$. Now use $s_n = \mu(A_n)^{-1}\chi_{A_n}$ to conclude that the sup is infinite. □

Again note that if there is a set of infinite measure which has no subset with finite positive measure, then every integrable function must vanish on this subset and hence the above lemma cannot work in such a situation.

As another consequence we get

Theorem 10.7 (Minkowski's integral inequality)**.** *Suppose, μ and ν are two σ-finite measures and f is $\mu \otimes \nu$ measurable. Let $1 \leq p \leq \infty$. Then*

$$\left\|\int_Y f(.,y)d\nu(y)\right\|_p \leq \int_Y \|f(.,y)\|_p d\nu(y), \tag{10.15}$$

where the p-norm is computed with respect to μ.

Proof. Let $g \in L^q(X, d\mu)$ with $g \geq 0$ and $\|g\|_q = 1$. Then using Fubini

$$\int_X g(x) \int_Y |f(x,y)| d\nu(y) d\mu(x) = \int_Y \int_X |f(x,y)| g(x) d\mu(x) d\nu(y)$$
$$\leq \int_Y \|f(.,y)\|_p d\nu(y)$$

and the claim follows from Lemma 10.6. □

In the special case where ν is supported on two points this reduces to the triangle inequality (our proof inherits the assumption that μ is σ-finite, but this can be avoided – Problem 10.7).

Corollary 10.8 (Minkowski's inequality)**.** *Let $f, g \in L^p(X, d\mu)$, $1 \leq p \leq \infty$. Then*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{10.16}$$

This shows that $L^p(X, d\mu)$ is a normed vector space.

Note that Fatou's lemma implies that the norm is lower semi continuous $\|f\|_p \leq \liminf_{n\to\infty} \|f_n\|_p$ with respect to pointwise convergence (a.e.). The next lemma sheds some light on the missing part.

Lemma 10.9 (Brezis–Lieb)**.** *Let $1 \le p < \infty$ and let $f_n \in L^p(X, d\mu)$ be a sequence which converges pointwise a.e. to f such that $\|f_n\|_p \le C$. Then $f \in L^p(X, d\mu)$ and*

$$\lim_{n\to\infty} \big(\|f_n\|_p^p - \|f_n - f\|_p^p\big) = \|f\|_p^p. \tag{10.17}$$

In the case $p = 1$ we can replace $\|f_n\|_1 \le C$ by $f \in L^1(X, d\mu)$.

Proof. As pointed out before $\|f\|_p \le \liminf_{n\to\infty} \|f_n\|_p \le C$ which shows $f \in L^p(X, d\mu)$. Moreover, one easyly checks the elementary inequality

$$\big||s+t|^p - |t|^p - |s|^p\big| \le \big||\,|s| + |t|\,|^p - |t|^p - |s|^p\big| \le \varepsilon|t|^p + C_\varepsilon|s|^p$$

(note that by scaling it suffices to consider the case $s = 1$). Setting $t = f_n - f$ and $s = f$, bringing everything to the right-hand-side and applying Fatou gives

$$\begin{aligned} C_\varepsilon\|f\|_p^p &\le \liminf_{n\to\infty} \int_X \big(\varepsilon|f_n - f|^p + C_\varepsilon|f|^p - \big||f_n|^p - |f - f_n|^p - |f|^p\big|\big)\, d\mu \\ &\le \varepsilon(2C)^p + C_\varepsilon\|f\|_p^p - \limsup_{n\to\infty} \int_X \big||f_n|^p - |f - f_n|^p - |f|^p\big| d\mu. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary the claim follows. Finally, note that in the case $p = 1$ we can choose $\varepsilon = 0$ and $C_\varepsilon = 2$. □

It might be more descriptive to write the conclusion of the lemma as

$$\|f_n\|_p^p = \|f\|_p^p + \|f_n - f\|_p^p + o(1) \tag{10.18}$$

which shows an important consequence:

Corollary 10.10. *Let $1 \le p < \infty$ and let $f_n \in L^p(X, d\mu)$ be a sequence which converges pointwise a.e. to f such that $\|f_n\|_p \le C$ if $p > 1$ or $f \in L^1(X, d\mu)$ if $p = 1$. Then $\|f_n - f\|_p \to 0$ if and only if $\|f_n\|_p \to \|f\|_p$.*

Note that it even suffices to show $\limsup \|f_n\|_p \le \|f\|_p$ since $\|f\|_p \le \liminf \|f_n\|_p$ comes for free from Fatou as pointed out before.

Note that a similar conclusion holds for weakly convergent sequences by the Radon–Riesz theorem (Theorem 5.19) since L^p is uniformly convex for $1 < p < \infty$.

Theorem 10.11 (Clarkson)**.** *Suppose $1 < p < \infty$, then $L^p(X, d\mu)$ is uniformly convex.*

Proof. As a preparation we note that strict convexity of $|.|^p$ implies that $|\frac{t+s}{2}|^p \le |\frac{|t|+|s|}{2}|^p < \frac{|t|^p+|s|^p}{2}$ and hence

$$\rho(\varepsilon) := \min\left\{ \frac{|t|^p + |s|^p}{2} - \Big|\frac{t+s}{2}\Big|^p \Big| |t|^p + |s|^p = 2,\ \Big|\frac{t-s}{2}\Big|^p \ge \varepsilon \right\} > 0.$$

Hence, by scaling,

$$\left|\frac{t-s}{2}\right|^p \geq \varepsilon\frac{|t|^p+|s|^p}{2} \quad \Rightarrow \quad \frac{|t|^p+|s|^p}{2}\rho(\varepsilon) \leq \frac{|t|^p+|s|^p}{2} - \left|\frac{t+s}{2}\right|^p.$$

Now given f, g with $\|f\|_p = \|g\|_p = 1$ and $\varepsilon > 0$ we need to find a $\delta > 0$ such that $\|\frac{f+g}{2}\|_p > 1-\delta$ implies $\|f-g\|_p < 2\varepsilon$. Introduce

$$M := \left\{ x \in X \middle| \left|\frac{f(x)-g(x)}{2}\right|^p \geq \varepsilon\frac{|f(x)|^p+|g(x)|^p}{2} \right\}.$$

Then

$$\begin{aligned}
\int_X \left|\frac{f-g}{2}\right|^p d\mu &= \int_{X\setminus M} \left|\frac{f-g}{2}\right|^p d\mu + \int_M \left|\frac{f-g}{2}\right|^p d\mu \\
&\leq \varepsilon \int_{X\setminus M} \frac{|f|^p+|g|^p}{2} d\mu + \int_M \frac{|f|^p+|g|^p}{2} d\mu \\
&\leq \varepsilon \int_{X\setminus M} \frac{|f|^p+|g|^p}{2} d\mu + \frac{1}{\rho}\int_M \left(\frac{|f|^p+|g|^p}{2} - \left|\frac{f+g}{2}\right|^p \right) d\mu \\
&\leq \varepsilon + \frac{1-(1-\delta)^p}{\rho} < 2\varepsilon
\end{aligned}$$

provided $\delta < (1+\varepsilon\rho)^{1/p} - 1$. □

In particular, by the Milman–Pettis theorem (Theorem 5.21), $L^p(X, d\mu)$ is reflexive for $1 < p < \infty$. We will give a direct proof for this fact in Corollary 12.2.

Problem* 10.5. *Show that $\varphi \in C^1(a,b)$ is (strictly) convex if and only if φ' is (strictly) increasing. Moreover, $\varphi \in C^2(a,b)$ is (strictly) convex if and only if $\varphi'' \geq 0$ ($\varphi'' > 0$ a.e.).*

Problem 10.6. *Prove*

$$\prod_{k=1}^n x_k^{\alpha_k} \leq \sum_{k=1}^n \alpha_k x_k, \qquad \text{if} \quad \sum_{k=1}^n \alpha_k = 1,$$

for $\alpha_k > 0$, $x_k > 0$. (Hint: Take a sum of Dirac-measures and use that the exponential function is convex.)

Problem 10.7. *Show Minkowski's inequality directly from Hölder's inequality. Show that $L^p(X, d\mu)$ is strictly convex for $1 < p < \infty$ but not for $p = 1, \infty$ if X contains two disjoint subsets of positive finite measure. (Hint: Start from $|f+g|^p \leq |f|\,|f+g|^{p-1} + |g|\,|f+g|^{p-1}$.)*

Problem* 10.8. *Show the* ***generalized Hölder's inequality****:*

$$\|f\,g\|_r \leq \|f\|_p \|g\|_q, \qquad \frac{1}{p}+\frac{1}{q} = \frac{1}{r}. \tag{10.19}$$

Problem* 10.9. *Show the iterated Hölder's inequality:*

$$\|f_1 \cdots f_m\|_r \le \prod_{j=1}^{m} \|f_j\|_{p_j}, \qquad \frac{1}{p_1} + \cdots + \frac{1}{p_m} = \frac{1}{r}. \tag{10.20}$$

Problem* 10.10. *Suppose μ is finite. Show that $L^p \subseteq L^{p_0}$ and*

$$\|f\|_{p_0} \le \mu(X)^{\frac{1}{p_0}-\frac{1}{p}} \|f\|_p, \qquad 1 \le p_0 \le p.$$

(Hint: Hölder's inequality.)

Problem* 10.11. *Show that if $f \in L^{p_0} \cap L^{p_1}$ for some $p_0 < p_1$ then $f \in L^p$ for every $p \in [p_0, p_1]$ and we have the* ***Lyapunov inequality***

$$\|f\|_p \le \|f\|_{p_0}^{1-\theta} \|f\|_{p_1}^{\theta},$$

where $\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$, $\theta \in (0,1)$. (Hint: Generalized Hölder inequality from Problem 10.8.)

Problem 10.12. *Let $1 < p < \infty$ and μ σ-finite. Let $f_n \in L^p(X, d\mu)$ be a sequence which converges pointwise a.e. to f such that $\|f_n\|_p \le C$. Then*

$$\int_X f_n g \, d\mu \to \int_X f g \, d\mu$$

for every $g \in L^q(X, d\mu)$. By Theorem 12.1 this implies that f_n converges weakly to f. (Hint: Theorem 8.24 and Problem 4.39.)

Problem 10.13. *Show that the Radon–Riesz theorem fails for $p = 1$: Find a sequence $f_n \in L^1(0,1)$ such that $f_n \ge 0$, $\int_0^1 f_n g \, dx \to \int_0^1 f g \, dx$ for every $g \in L^\infty(0,1)$ (i.e. $f_n \rightharpoonup f$ by Theorem 12.1), $\|f_n\|_1 \to \|f\|_1$ but $\|f_n - f\|_1 \not\to 0$. (Hint: Riemann–Lebesgue lemma.)*

10.3. Nothing missing in L^p

Finally it remains to show that $L^p(X, d\mu)$ is complete.

Theorem 10.12 (Riesz–Fischer)**.** *The space $L^p(X, d\mu)$, $1 \le p \le \infty$, is a Banach space.*

Proof. We begin with the case $1 \le p < \infty$. Suppose f_n is a Cauchy sequence. It suffices to show that some subsequence converges (show this). Hence we can drop some terms such that

$$\|f_{n+1} - f_n\|_p \le \frac{1}{2^n}.$$

Now consider $g_n = f_n - f_{n-1}$ (set $f_0 = 0$). Then

$$G(x) = \sum_{k=1}^{\infty} |g_k(x)|$$

is in L^p. This follows from

$$\Big\| \sum_{k=1}^{n} |g_k| \Big\|_p \le \sum_{k=1}^{n} \|g_k\|_p \le \|f_1\|_p + 1$$

using the monotone convergence theorem. In particular, $G(x) < \infty$ almost everywhere and the sum

$$\sum_{n=1}^{\infty} g_n(x) = \lim_{n\to\infty} f_n(x)$$

is absolutely convergent for those x. Now let $f(x)$ be this limit. Since $|f(x) - f_n(x)|^p$ converges to zero almost everywhere and $|f(x) - f_n(x)|^p \le (2G(x))^p \in L^1$, dominated convergence shows $\|f - f_n\|_p \to 0$.

In the case $p = \infty$ note that the Cauchy sequence property $|f_n(x) - f_m(x)| < \varepsilon$ for $n, m > N$ holds except for sets $A_{m,n}$ of measure zero. Since $A = \bigcup_{n,m} A_{n,m}$ is again of measure zero, we see that $f_n(x)$ is a Cauchy sequence for $x \in X\backslash A$. The pointwise limit $f(x) = \lim_{n\to\infty} f_n(x)$, $x \in X\backslash A$, is the required limit in $L^\infty(X, d\mu)$ (show this). □

In particular, in the proof of the last theorem we have seen:

Corollary 10.13. *If $\|f_n - f\|_p \to 0$, $1 \le p \le \infty$, then there is a subsequence f_{n_j} (of representatives) which converges pointwise almost everywhere and a function $g \in L^p(X, d\mu)$ such that $f_{n_j} \le g$ almost everywhere.*

Consequently, if $f_n \in L^{p_0} \cap L^{p_1}$ converges in both L^{p_0} and L^{p_1}, then the limits will be equal a.e. Be warned that the statement is not true in general without passing to a subsequence (Problem 10.14).

It even turns out that L^p is separable.

Lemma 10.14. *Suppose X is a second countable topological space (i.e., it has a countable basis) and μ is an outer regular Borel measure. Then $L^p(X, d\mu)$, $1 \le p < \infty$, is separable. In particular, for every countable base the set of characteristic functions $\chi_O(x)$ with O in this base is total.*

Proof. The set of all characteristic functions $\chi_A(x)$ with $A \in \Sigma$ and $\mu(A) < \infty$ is total by construction of the integral (Problem 10.18). Now our strategy is as follows: Using outer regularity, we can restrict A to open sets and using the existence of a countable base, we can restrict A to open sets from this base.

Fix A. By outer regularity, there is a decreasing sequence of open sets $O_n \supseteq A$ such that $\mu(O_n) \to \mu(A)$. Since $\mu(A) < \infty$, it is no restriction to assume $\mu(O_n) < \infty$, and thus $\|\chi_A - \chi_{O_n}\|_p = \mu(O_n\backslash A) = \mu(O_n) - \mu(A) \to 0$. Thus the set of all characteristic functions $\chi_O(x)$ with O open and $\mu(O) < \infty$ is total. Finally let $\mathcal{B}$ be a countable base for the topology. Then, every

open set O can be written as $O = \bigcup_{j=1}^{\infty} \tilde{O}_j$ with $\tilde{O}_j \in \mathcal{B}$. Moreover, by considering the set of all finite unions of elements from $\mathcal{B}$, it is no restriction to assume $\bigcup_{j=1}^{n} \tilde{O}_j \in \mathcal{B}$. Hence there is an increasing sequence $\tilde{O}_n \nearrow O$ with $\tilde{O}_n \in \mathcal{B}$. By monotone convergence, $\|\chi_O - \chi_{\tilde{O}_n}\|_p \to 0$ and hence the set of all characteristic functions $\chi_{\tilde{O}}$ with $\tilde{O} \in \mathcal{B}$ is total. □

Finally, we can generalize Theorem 1.13 and give a characterization of relatively compact sets. To this end let $X \subseteq \mathbb{R}^n$, $f \in L^p(X)$ and consider the **translation operator**

$$T_a(f)(x) = \begin{cases} f(x-a), & x-a \in X, \\ 0, & \text{else}, \end{cases} \tag{10.21}$$

for fixed $a \in \mathbb{R}^n$. Then one checks $\|T_a\| = 1$ (unless $|(X-a) \cap X| = 0$ in which case $T_a \equiv 0$) and $T_a f \to f$ as $a \to 0$ for $1 \le p < \infty$ (Problem 10.15).

Theorem 10.15 (Kolmogorov–Riesz–Sudakov)**.** *Let $X \subseteq \mathbb{R}^n$ be open. A subset F of $L^p(X)$, $1 \le p < \infty$, is relatively compact if and only if*

(i) *for every $\varepsilon > 0$ there is some $\delta > 0$ such that $\|T_a f - f\|_p \le \varepsilon$ for all $|a| \le \delta$ and $f \in F$.*

(ii) *for every $\varepsilon > 0$ there is some $r > 0$ such that $\|(1 - \chi_{B_r(0)})f\|_p \le \varepsilon$ for all $f \in F$.*

Of course the last condition is void if X is bounded.

Proof. We first show that F is bounded. For this fix $\varepsilon = 1$ and choose δ, r according to (i), (ii), respectively. Then

$$\|f\chi_{B_r(x)}\|_p \le \|(T_y f - f)\chi_{B_r(x)}\|_p + \|f\chi_{B_r(x+y)}\|_p \le 1 + \|f\chi_{B_r(x+y)}\|_p$$

for $f \in F$ and $|y| \le \delta$. Hence by induction $\|f\chi_{B_r(0)}\|_p \le m + \|f\chi_{B_r(my)}\|_p$ and choosing $|y| = \delta$ and $m \ge \frac{2r}{\delta}$ such that $B_r(my) \cap B_r(0) = \emptyset$ we obtain

$$\|f\|_p = \|f\chi_{B_r(0)}\|_p + \|f\chi_{\mathbb{R}^n \setminus B_r(0)}\|_p \le 2 + m.$$

Now our strategy is to use Lemma 1.12. To this end we fix ε and choose a cube Q centered at 0 with side length δ according to (i) and finitely many disjoint cubes $\{Q_j\}_{j=1}^m$ of side length $\delta/2$ such that they cover $B_r(0)$ with r as in (ii). Now let Y be the finite dimensional subspace spanned by the characteristic functions of the cubes Q_j and let

$$P_m f := \sum_{j=1}^{m} \left(\frac{1}{|Q_j|} \int_{Q_j} f(y) d^n y \right) \chi_{Q_j}$$

be the projection from $L^p(X)$ onto Y. Note that using the triangle inequality and then Hölders inequality

$$\|P_m f\|_p \le \sum_{j=1}^m \left(\frac{1}{|Q_j|} \int_{Q_j} |f(y)| d^n y \right) \|\chi_{Q_j}\|_p \le \sum_{j=1}^m \|f\|_{L^p(Q_j)} \le \|f\|_p$$

and since F is bounded so is $P_m F$. Moreover, for $f \in F$ we have

$$\|(1-P_m)f\|_p^p \le \varepsilon^p + \sum_{j=1}^m \int_{Q_j} |f(x) - P_m f(x)|^p d^n x$$

and using Jensen's inequality we further get

$$\begin{aligned}
\|(1-P_m)f\|_p^p &\le \varepsilon^p + \sum_{j=1}^m \int_{Q_j} \frac{1}{|Q_j|} \int_{Q_j} |f(x) - f(y)|^p d^n y \, d^n x \\
&\le \varepsilon^p + \sum_{j=1}^m \int_{Q_j} \frac{2^n}{|Q|} \int_{Q} |f(x) - f(x-y)|^p d^n y \, d^n x \\
&\le \varepsilon^p + \frac{2^n}{|Q|} \int_Q \|f - T_y f\|_p^p d^n y \le (1+2^n)\varepsilon^p,
\end{aligned}$$

since $x, y \in Q_j$ implies $x - y \in Q$.

Conversely, suppose F is relatively compact. To see (i) and (ii) pick an ε-cover $\{B_\varepsilon(f_j)\}_{j=1}^m$ and choose δ such that $\|f_j - T_a f_j\|_p \le \varepsilon$ for all $|a| \le \delta$ and $1 \le j \le m$. Then for every f there is some j such that $f \in B_\varepsilon(f_j)$ and hence $\|f - T_a f\|_p \le \|f - f_j\|_p + \|f_j - T_a f_j\|_p + \|T_a(f_j - f)\|_p \le 3\varepsilon$ implying (i). For (ii) choose r such that $\|(1-\chi_{B_r(0)})f_j\|_p \le \varepsilon$ for $1 \le j \le m$ implying $\|(1-\chi_{B_r(0)})f\|_p \le 3\varepsilon$ as before. □

Example 10.3. Choosing a fixed $f_0 \in L^p(X)$ condition (ii) is for example satisfied if $|f(x)| \le |f_0(x)|$ for all $f \in F$. Similarly, if $f_0(x) \ge 1$ with $\lim_{|x|\to\infty} f_0(x) = \infty$ such that $\|f_0 f\|_p \le C$ for all $f \in F$, then (ii) holds (to see this let ε be given and choose r such that $f_0(x) \ge \frac{C}{\varepsilon}$ for $|x| \ge r$). Condition (i) is for example satisfied if F is equicontinuous. ⋄

Problem* 10.14. *Find a sequence f_n which converges to 0 in $L^p([0,1], dx)$, $1 \le p < \infty$, but for which $f_n(x) \to 0$ for a.e. $x \in [0,1]$ does not hold. (Hint: Every $n \in \mathbb{N}$ can be uniquely written as $n = 2^m + k$ with $0 \le m$ and $0 \le k < 2^m$. Now consider the characteristic functions of the intervals $I_{m,k} = [k2^{-m}, (k+1)2^{-m}]$.)*

Problem* 10.15. *Let $f \in L^p(X)$, $X \subseteq \mathbb{R}^n$ open, $1 \le p < \infty$ and show that $T_a f \to f$ in L^p as $a \to 0$. (Hint: Start with $f \in C_c(\mathbb{R}^n)$ and use Theorem 10.16 below.)*

Problem 10.16. *Show that L^p convergence implies convergence in measure (cf. Problem 8.23). Show that the converse fails.*

Problem 10.17. *Let X_1, X_2 be second countable topological spaces and let μ_1, μ_2 be outer regular σ-finite Borel measures. Let $\mathcal{B}_1$, $\mathcal{B}_2$ be bases for X_1, X_2, respectively. Show that the set of all functions $\chi_{O_1\times O_2}(x_1,x_2) = \chi_{O_1}(x_1)\chi_{O_2}(x_2)$ for $O_1 \in \mathcal{B}_1$, $O_2 \in \mathcal{B}_2$ is total in $L^2(X_1 \times X_2, \mu_1 \otimes \mu_2)$. (Hint: Lemma 9.12 and 10.14.)*

Problem* 10.18. *Show that for any $f \in L^p(X, d\mu)$, $1 \le p \le \infty$ there exists a sequence of simple functions s_n such that $|s_n| \le |f|$ and $s_n \to f$ in $L^p(X, d\mu)$. If $p < \infty$ then s_n will be integrable. (Hint: Split f into the sum of four nonnegative functions and use* (9.6).*)*

10.4. Approximation by nicer functions

Since measurable functions can be quite wild, they are sometimes hard to work with. In fact, in many situations some properties are much easier to prove for a dense set of *nice* functions and the general case can then be reduced to the nice case by an *approximation argument*. But for such a strategy to work one needs to identify suitable sets of nice functions which are dense in L^p.

Theorem 10.16. *Let X be a locally compact metric space and let μ be a regular Borel measure. Then the set $C_c(X)$ of continuous functions with compact support is dense in $L^p(X, d\mu)$, $1 \le p < \infty$.*

Proof. As in the proof of Lemma 10.14 the set of all characteristic functions $\chi_K(x)$ with K compact is total (using inner regularity). Hence it suffices to show that $\chi_K(x)$ can be approximated by continuous functions. By outer regularity there is an open set $O \supset K$ such that $\mu(O \setminus K) \le \varepsilon$. By Urysohn's lemma (Lemma B.28) there is a continuous function $f_\varepsilon : X \to [0,1]$ with compact support which is 1 on K and 0 outside O. Since

$$\int_X |\chi_K - f_\varepsilon|^p d\mu = \int_{O\setminus K} |f_\varepsilon|^p d\mu \le \mu(O \setminus K) \le \varepsilon,$$

we have $\|f_\varepsilon - \chi_K\|_p \to 0$ and we are done. □

Clearly this result has to fail in the case $p = \infty$ (in general) since the uniform limit of continuous functions is again continuous. In fact, the closure of $C_c(\mathbb{R}^n)$ in the infinity norm is the space $C_0(\mathbb{R}^n)$ of continuous functions vanishing at ∞ (Problem 1.46). Another variant of this result is

Theorem 10.17 (Luzin)**.** *Let X be a locally compact metric space and let μ be a finite regular Borel measure. Let f be integrable. Then for every $\varepsilon > 0$ there is an open set O_ε with $\mu(O_\varepsilon) < \varepsilon$ and $X \setminus O_\varepsilon$ compact and a continuous function g which coincides with f on $X \setminus O_\varepsilon$.*

Proof. From the proof of the previous theorem we know that the set of all characteristic functions $\chi_K(x)$ with K compact is total. Hence we can restrict our attention to a sufficiently large compact set K. By Theorem 10.16 we can find a sequence of continuous functions f_n which converges to f in L^1. After passing to a subsequence we can assume that f_n converges a.e. and by Egorov's theorem there is a subset A_ε on which the convergence is uniform. By outer regularity we can replace A_ε by a slightly larger open set O_ε such that $C = K \setminus O_\varepsilon$ is compact. Now Tietze's extension theorem implies that $f|_C$ can be extended to a continuous function g on X. □

If X is some subset of $\mathbb{R}^n$, we can do even better and approximate integrable functions by smooth functions. The idea is to replace the value $f(x)$ by a suitable average computed from the values in a neighborhood. This is done by choosing a nonnegative bump function ϕ, whose area is normalized to 1, and considering the **convolution**

$$(\phi * f)(x) := \int_{\mathbb{R}^n} \phi(x-y) f(y) d^n y = \int_{\mathbb{R}^n} \phi(y) f(x-y) d^n y. \tag{10.22}$$

For example, if we choose $\phi_r = |B_r(0)|^{-1} \chi_{B_r(0)}$ to be the characteristic function of a ball centered at 0, then $(\phi_r * f)(x)$ will be precisely the average of the values of f in the ball $B_r(x)$. In the general case we can think of $(\phi * f)(x)$ as an weighted average. Moreover, if we choose ϕ differentiable, we can interchange differentiation and integration to conclude that $\phi * f$ will also be differentiable. Iterating this argument shows that $\phi * f$ will have as many derivatives as ϕ. Finally, if the set over which the average is computed (i.e., the support of ϕ) shrinks, we expect $(\phi * f)(x)$ to get closer and closer to $f(x)$.

To make these ideas precise we begin with a few properties of the convolution.

Lemma 10.18. *The convolution has the following properties:*

(i) $f(x-.)g(.)$ *is integrable if and only if* $f(.)g(x-.)$ *is and*

$$(f * g)(x) = (g * f)(x) \tag{10.23}$$

in this case.

(ii) *Suppose* $\phi \in C_c^k(\mathbb{R}^n)$ *and* $f \in L^1_{loc}(\mathbb{R}^n)$*, then* $\phi * f \in C^k(\mathbb{R}^n)$ *and*

$$\partial_\alpha(\phi * f) = (\partial_\alpha \phi) * f \tag{10.24}$$

for any partial derivative of order at most k*.*

(iii) *We have* $\operatorname{supp}(f * g) \subseteq \overline{\operatorname{supp}(f) + \operatorname{supp}(g)}$*. In particular, if* $\phi \in C_c^k(\mathbb{R}^n)$ *and* $f \in L^1_c(\mathbb{R}^n)$*, then* $\phi * f \in C_c^k(\mathbb{R}^n)$*.*

(iv) *Suppose* $\phi \in L^1(\mathbb{R}^n)$ *and* $f \in L^p(\mathbb{R}^n)$, $1 \le p \le \infty$, *then their convolution is in* $L^p(\mathbb{R}^n)$ *and satisfies* ***Young's inequality***

$$\|\phi * f\|_p \le \|\phi\|_1 \|f\|_p. \tag{10.25}$$

(v) *Suppose* $\phi \ge 0$ *with* $\|\phi\|_1 = 1$ *and* $f \in L^\infty(\mathbb{R}^n)$ *real-valued, then*

$$\inf_{x\in\mathbb{R}^n} f(x) \le (\phi * f)(x) \le \sup_{x\in\mathbb{R}^n} f(x). \tag{10.26}$$

Proof. (i) is a simple affine change of coordinates. (ii) follows by interchanging differentiation with the integral using Problems 9.14 and 9.15. (iii) If $x \notin \operatorname{supp}(f) + \operatorname{supp}(g)$, then $x - y \notin \operatorname{supp}(f)$ for $y \in \operatorname{supp}(g)$ and hence $f(x-y)g(y)$ vanishes on $\operatorname{supp}(g)$. This establishes the claim about the support and the rest follows from the previous item. (iv) The case $p = \infty$ follows from Hölder's inequality and we can assume $1 \le p < \infty$. Without loss of generality let $\|\phi\|_1 = 1$. Then

$$\begin{aligned}\|\phi * f\|_p^p &\le \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} |f(y-x)||\phi(y)| d^n y \right|^p d^n x \\ &\le \int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |f(y-x)|^p |\phi(y)| d^n y\, d^n x = \|f\|_p^p,\end{aligned}$$

where we have use Jensen's inequality with $\varphi(x) = |x|^p$, $d\mu = |\phi| d^n y$ in the first and Fubini in the second step. (v) Immediate from integrating $\phi(y) \inf_{x\in\mathbb{R}^n} f(x) \le \phi(y) f(x-y) \le \phi(y) \sup_{x\in\mathbb{R}^n} f(x)$. □

Note that (iv) also extends to Hölder continuous functions since it is straightforward to show

$$[\phi * f]_\gamma \le \|\phi\|_1 [f]_\gamma. \tag{10.27}$$

Example 10.4. For $f := \chi_{[0,1]} - \chi_{[-1,0]}$ and $g := \chi_{[-2,2]}$ we have that $f * g$ is the difference of two triangles supported on $[1,3]$ and $[-3,-1]$ whereas $\operatorname{supp}(f) + \operatorname{supp}(g) = [-3,3]$, which shows that the inclusion in (iii) is strict in general. ⋄

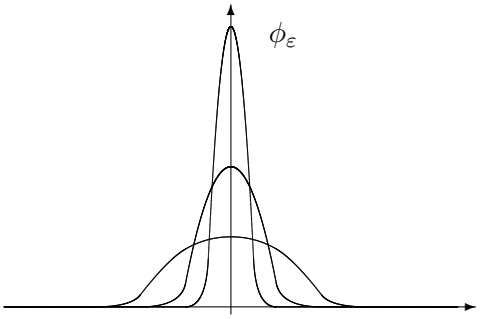
Next we turn to approximation of f. To this end we call a family of integrable functions ϕ_ε, $\varepsilon \in (0,1]$, an **approximate identity** if it satisfies the following three requirements:

(i) $\|\phi_\varepsilon\|_1 \le C$ for all $\varepsilon > 0$.
(ii) $\int_{\mathbb{R}^n} \phi_\varepsilon(x) d^n x = 1$ for all $\varepsilon > 0$.
(iii) For every $r > 0$ we have $\lim_{\varepsilon\downarrow 0} \int_{|x|\ge r} \phi_\varepsilon(x) d^n x = 0$.

Moreover, a nonnegative function $\phi \in C_c^\infty(\mathbb{R}^n)$ satisfying $\|\phi\|_1 = 1$ is called a **mollifier**. Note that if the support of ϕ is within a ball of radius s, then the support of $\phi * f$ will be within $\{x \in \mathbb{R}^n | \operatorname{dist}(x, \operatorname{supp}(f)) \le s\}$.

Example 10.5. The standard (also Friedrichs) mollifier is $\phi(x) := c^{-1}\exp(\frac{1}{|x|^2-1})$ for $|x| < 1$ and $\phi(x) = 0$ otherwise. Here the normalization constant $c := \int_{B_1(0)} \exp(\frac{1}{|x|^2-1})dx^n = S_n \int_0^1 \exp(\frac{1}{r^2-1})r^{n-1}dr$ is chosen such that $\|\phi\|_1 = 1$. To show that this function is indeed smooth it suffices to show that all left derivatives of $f(r) = \exp(\frac{1}{r-1})$ at $r = 1$ vanish, which can be done using l'Hôpital's rule. ◇

Example 10.6. Scaling a mollifier according to $\phi_\varepsilon(x) = \varepsilon^{-n}\phi(\frac{x}{\varepsilon})$ such that its mass is preserved ($\|\phi_\varepsilon\|_1 = 1$) and it concentrates more and more around the origin as $\varepsilon \downarrow 0$ we obtain an approximate identity:



In fact, (i), (ii) are obvious from $\|\phi_\varepsilon\|_1 = 1$ and the integral in (iii) will be identically zero for $\varepsilon \geq \frac{r}{s}$, where s is chosen such that $\operatorname{supp}\phi \subseteq \overline{B_s(0)}$. ◇

Now we are ready to show that an approximate identity deserves its name.

Lemma 10.19. *Let ϕ_ε be an approximate identity. If $f \in L^p(\mathbb{R}^n)$ with $1 \leq p < \infty$. Then*

$$\lim_{\varepsilon\downarrow 0} \phi_\varepsilon * f = f \tag{10.28}$$

with the limit taken in L^p. In the case $p = \infty$ the claim holds for $f \in C_0(\mathbb{R}^n)$.

Proof. We begin with the case where $f \in C_c(\mathbb{R}^n)$. Fix some small $\delta > 0$. Since f is uniformly continuous we know $|f(x-y) - f(x)| \to 0$ as $y \to x$ uniformly in x. Since the support of f is compact, this remains true when taking the L^p norm and thus we can find some r such that

$$\|f(.-y) - f(.)\|_p \leq \frac{\delta}{2C}, \qquad |y| \leq r.$$

(Here the C is the one for which $\|\phi_\varepsilon\|_1 \leq C$ holds.) Now we use

$$(\phi_\varepsilon * f)(x) - f(x) = \int_{\mathbb{R}^n} \phi_\varepsilon(y)(f(x-y) - f(x))d^n y$$

Splitting the domain of integration according to $\mathbb{R}^n = \{y||y| \leq r\} \cup \{y||y| > r\}$, we can estimate the L^p norms of the individual integrals using the

Minkowski inequality as follows:

$$\left\| \int_{|y|\le r} \phi_\varepsilon(y)(f(x-y)-f(x))d^n y \right\|_p \le \int_{|y|\le r} |\phi_\varepsilon(y)| \|f(.-y)-f(.)\|_p d^n y \le \frac{\delta}{2}$$

and

$$\left\| \int_{|y|>r} \phi_\varepsilon(y)(f(x-y)-f(x))d^n y \right\|_p \le 2\|f\|_p \int_{|y|>r} |\phi_\varepsilon(y)| d^n y \le \frac{\delta}{2}$$

provided ε is sufficiently small such that the integral in (iii) is less than $\delta/2$.

This establishes the claim for $f \in C_c(\mathbb{R}^n)$. Since these functions are dense in L^p for $1 \le p < \infty$ and in $C_0(\mathbb{R}^n)$ for $p = \infty$ the claim follows from Lemma 4.32 and Young's inequality. □

Note that in case of a mollifier with support in $B_r(0)$ this result implies a corresponding local version since the value of $(\phi_\varepsilon * f)(x)$ is only affected by the values of f on $B_{\varepsilon r}(x)$. The question when the pointwise limit exists will be addressed in Problem 10.25.

Example 10.7. The Fejér kernel introduced in (2.52) is an approximate identity if we set it equal to 0 outside $[-\pi, \pi]$ (see the proof of Theorem 2.20). Then taking a 2π periodic function and setting it equal to 0 outside $[-2\pi, 2\pi]$ Lemma 10.19 shows that for $f \in L^p(-\pi, \pi)$ the mean values of the partial sums of a Fourier series $\bar{S}_n(f)$ converge to f in $L^p(-\pi, \pi)$ for every $1 \le p < \infty$. For $p = \infty$ we recover Theorem 2.20. Note also that this shows that the map $f \mapsto \hat{f}$ is injective on L^p.

Another classical example it the Poisson kernel, see Problem 10.24. Note that the Dirichlet kernel D_n from (2.46) is no approximate identity since $\|D_n\|_1 \to \infty$ as was shown in Example 4.3. ⋄

Now we are ready to prove

Theorem 10.20. *If $X \subseteq \mathbb{R}^n$ is open and μ is a regular Borel measure, then the set $C_c^\infty(X)$ of all smooth functions with compact support is dense in $L^p(X, d\mu)$, $1 \le p < \infty$.*

Proof. By Theorem 10.16 it suffices to show that every continuous function $f(x)$ with compact support can be approximated by smooth ones. By setting $f(x) = 0$ for $x \notin X$, it is no restriction to assume $X = \mathbb{R}^n$. Now choose a mollifier ϕ and observe that $\phi_\varepsilon * f$ has compact support inside X for ε

sufficiently small (since the distance from $\operatorname{supp}(f)$ to the boundary ∂X is positive by compactness). Moreover, $\phi_\varepsilon * f \to f$ uniformly by the previous lemma and hence also in $L^p(X, d\mu)$. □

Our final result is known as the **fundamental lemma of the calculus of variations**.

Lemma 10.21. *Suppose $X \subseteq \mathbb{R}^n$ is open and $f \in L^1_{loc}(X)$. (i) If f is real-valued then*

$$\int_X \varphi(x) f(x) d^n x \geq 0, \qquad \forall \varphi \in C_c^\infty(X),\ \varphi \geq 0, \tag{10.29}$$

if and only if $f(x) \geq 0$ (a.e.). (ii) Moreover,

$$\int_X \varphi(x) f(x) d^n x = 0, \qquad \forall \varphi \in C_c^\infty(X),\ \varphi \geq 0, \tag{10.30}$$

if and only if $f(x) = 0$ (a.e.).

Proof. (i) Choose a compact set $K \subset X$ and some $\varepsilon_0 > 0$ such that $K_{\varepsilon_0} := K + B_{\varepsilon_0}(0) \subseteq X$. Set $\tilde{f} := f\chi_{K_{\varepsilon_0}}$ and let ϕ be the standard mollifier. Then $(\phi_\varepsilon * \tilde{f})(x) = (\phi_\varepsilon * f)(x) \geq 0$ for $x \in K$, $\varepsilon < \varepsilon_0$ and since $\phi_\varepsilon * \tilde{f} \to \tilde{f}$ in $L^1(X)$ we have $(\phi_\varepsilon * \tilde{f})(x) \to f(x) \geq 0$ for a.e. $x \in K$ for an appropriate subsequence. Since $K \subset X$ is arbitrary the first claim follows. (ii) The first part shows that $\operatorname{Re}(f) \geq 0$ as well as $-\operatorname{Re}(f) \geq 0$ and hence $\operatorname{Re}(f) = 0$. Applying the same argument to $\operatorname{Im}(f)$ establishes the claim. □

The following variant is also often useful

Lemma 10.22 (du Bois-Reymond)**.** *Suppose $U \subseteq \mathbb{R}^n$ is open and connected. If $f \in L^1_{loc}(U)$ with*

$$\int_U f(x) \partial_j \varphi(x) d^n x = 0, \qquad \forall \varphi \in C_c^\infty(U),\ 1 \leq j \leq n, \tag{10.31}$$

then f is constant a.e. on U.

Proof. Choose a compact set $K \subset U$ and $\tilde{f}$, ϕ as in the proof of the previous lemma but additionally assume that K is connected. Then by Lemma 10.18 (ii)

$$\partial_j(\phi_\varepsilon * \tilde{f})(x) = ((\partial_j \phi_\varepsilon) * \tilde{f})(x) = ((\partial_j \phi_\varepsilon) * f)(x) = 0, \qquad x \in K,\ \varepsilon \leq \varepsilon_0.$$

Hence $(\phi_\varepsilon * \tilde{f})(x) = c_\varepsilon$ for $x \in K$ and as $\varepsilon \to 0$ there is a subsequence which converges a.e. on K. Clearly this limit function must also be constant: $(\phi_\varepsilon * \tilde{f})(x) = c_\varepsilon \to f(x) = c$ for a.e. $x \in K$. Now write U as a countable union of open balls whose closure is contained in U. If the corresponding constants for these balls were not all the same, we could find a partition

into two union of open balls which were disjoint. This contradicts that U is connected. □

Of course the last result can be extended to higher derivatives. For the one-dimensional case this is outlined in Problem 10.28.

Problem* 10.19 (Smooth Urysohn lemma)**.** *Suppose K and C are disjoint closed subsets of $\mathbb{R}^n$ with K compact. Then there is a smooth function $f \in C_c^\infty(\mathbb{R}^n, [0,1])$ such that f is zero on C and one on K.*

Problem 10.20. *Let $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Show that $f * g \in C_0(\mathbb{R}^n)$ with*

$$\|f * g\|_\infty \leq \|f\|_p \|g\|_q.$$

Problem 10.21. *Show that the convolution on $L^1(\mathbb{R}^n)$ is associative. Conclude that $L^1(\mathbb{R}^n)$ together with convolution as a product is a commutative Banach algebra (without identity). (Hint: It suffices to verify associativity for nice functions.)*

Problem 10.22. *Let μ be a finite measure on $\mathbb{R}$. Then the set of all exponentials $\{\mathrm{e}^{\mathrm{i}tx}\}_{t\in\mathbb{R}}$ is total in $L^p(\mathbb{R}, d\mu)$ for $1 \leq p < \infty$.*

Problem 10.23. *Let ϕ be integrable and normalized such that $\int_{\mathbb{R}^n} \phi(x) d^n x = 1$. Show that $\phi_\varepsilon(x) = \varepsilon^{-n}\phi(\frac{x}{\varepsilon})$ is an approximate identity.*

Problem 10.24. *Show that the* ***Poisson kernel***

$$P_\varepsilon(x) := \frac{1}{\pi} \frac{\varepsilon}{x^2 + \varepsilon^2}$$

is an approximate identity on $\mathbb{R}$.

Show that the ***Cauchy transform*** *(also* ***Borel transform****)*

$$F(z) := \frac{1}{\pi} \int_{\mathbb{R}} \frac{f(\lambda)}{\lambda - z} d\lambda$$

of a real-valued function $f \in L^p(\mathbb{R})$ is analytic in the upper half-plane with imaginary part given by

$$\mathrm{Im}(F(x + \mathrm{i}y)) = (P_y * f)(x).$$

In particular, by Young's inequality $\|\mathrm{Im}(F(. + \mathrm{i}y))\|_p \leq \|f\|_p$ and thus $\sup_{y>0} \|\mathrm{Im}(F(. + \mathrm{i}y))\|_p = \|f\|_p$. Such analytic functions are said to be in the ***Hardy space*** *$H^p(\mathbb{C}_+)$.*

(Hint: To see analyticity of F use Problem 9.19 plus the estimate

$$\left|\frac{1}{\lambda - z}\right| \leq \frac{1}{1 + |\lambda|} \frac{1 + |z|}{|\mathrm{Im}(z)|}.)$$

Problem* 10.25. *Let ϕ be bounded with support in $\overline{B_1(0)}$ and normalized such that $\int_{\mathbb{R}^n} \phi(x)d^n x = 1$. Set $\phi_\varepsilon(x) = \varepsilon^{-n}\phi(\frac{x}{\varepsilon})$.*

For f locally integrable show

$$|(\phi_\varepsilon * f)(x) - f(x)| \leq \frac{V_n\|\phi\|_\infty}{|B_\varepsilon(x)|}\int_{B_\varepsilon(x)} |f(y) - f(x)|d^n y.$$

Hence at every Lebesgue point (cf. Theorem 11.6) x we have

$$\lim_{\varepsilon\downarrow 0}(\phi_\varepsilon * f)(x) = f(x).$$

If f is uniformly continuous then the above limit will be uniform. See Problem 15.8 for the case when ϕ is not compactly supported.

Problem 10.26. *Let f, g be integrable (or nonnegative). Show that*

$$\int_{\mathbb{R}^n}(f * g)(x)d^n x = \int_{\mathbb{R}^n} f(x)d^n x \int_{\mathbb{R}^n} g(x)d^n x.$$

Problem 10.27. *Let μ, ν be two complex measures on $\mathbb{R}^n$ and set $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, $(x, y) \mapsto x + y$. Define the convolution of μ and ν by*

$$\mu * \nu := S_\star(\mu \otimes \nu).$$

Show

- *$\mu * \nu$ is a complex measure given by*

$$(\mu * \nu)(A) = \int_{\mathbb{R}^n\times\mathbb{R}^n} \chi_A(x + y)d\mu(x)d\nu(y) = \int_{\mathbb{R}^n} \mu(A - y)d\nu(y)$$

  and satisfying $|\mu * \nu|(\mathbb{R}^n) \leq |\mu|(\mathbb{R}^n)|\nu|(\mathbb{R}^n)$ with equality for positive measures.

- *$\mu * \nu = \nu * \mu$.*
- *If $d\nu(x) = g(x)d^n x$ then $d(\mu * \nu)(x) = h(x)d^n x$ with $h(x) = \int_{\mathbb{R}^n} g(x - y)d\mu(y)$.*

In particular the last item shows that this definition agrees with our definition for functions if both measures have a density.

Problem 10.28 (du Bois-Reymond lemma). *Let f be a locally integrable function on the interval (a, b) and let $n \in \mathbb{N}_0$. If*

$$\int_a^b f(x)\varphi^{(n)}(x)dx = 0, \qquad \forall\varphi \in C_c^\infty(a, b),$$

then f is a polynomial of degree at most $n-1$ a.e. (Hint: Begin by showing that there exists $\{\phi_{n,j}\}_{0\leq j\leq n} \subset C_c^\infty(a, b)$ such that

$$\int_a^b x^k\phi_{n,j}(x)dx = \delta_{j,k}, \qquad 0 \leq j, k \leq n.$$

The case $n = 0$ is easy and for the general case note that one can choose $\phi_{n+1,n+1} = (n+1)^{-1}\phi'_{n,n}$. Then, for given $\phi \in C_c^\infty(a,b)$, look at $\varphi(x) = \frac{1}{n!}\int_a^x (\phi(y) - \tilde{\phi}(y))(x-y)^n dy$ where $\tilde{\phi}$ is chosen such that this function is in $C_c^\infty(a,b)$.)

10.5. Integral operators

Using Hölder's inequality, we can also identify a class of bounded operators from $L^p(Y, d\nu)$ to $L^p(X, d\mu)$. We will assume all measures to be σ-finite throughout this section.

Lemma 10.23 (Schur criterion)**.** *Let μ, ν be measures on X, Y, respectively, and let $\frac{1}{p} + \frac{1}{q} = 1$. Suppose that $K(x,y)$ is measurable and there are nonnegative measurable functions $K_1(x,y)$, $K_2(x,y)$ such that $|K(x,y)| \leq K_1(x,y)K_2(x,y)$ and*

$$\|K_1(x,.)\|_{L^q(Y,d\nu)} \leq C_1, \qquad \|K_2(.,y)\|_{L^p(X,d\mu)} \leq C_2 \tag{10.32}$$

for μ-almost every x, respectively, for ν-almost every y. Then the operator $K : L^p(Y, d\nu) \to L^p(X, d\mu)$, defined by

$$(Kf)(x) := \int_Y K(x,y)f(y)d\nu(y), \tag{10.33}$$

for μ-almost every x is bounded with $\|K\| \leq C_1C_2$.

Proof. Choose $f \in L^p(Y, d\nu)$. By Fubini's theorem $\int_Y |K(x,y)f(y)|d\nu(y)$ is measurable and by Hölder's inequality we have

$$\begin{aligned}\int_Y |K(x,y)f(y)|d\nu(y) &\leq \int_Y K_1(x,y)K_2(x,y)|f(y)|d\nu(y)\\ &\leq C_1\|K_2(x,.)f(.)\|_{L^p(X,d\mu)}\end{aligned}$$

for μ a.e. x (if $K_2(x,.)f(.) \notin L^p(Y, d\nu)$, the inequality is trivially true). If $p < \infty$ take this inequality to the p'th power and integrate with respect to x using Fubini

$$\begin{aligned}\int_X \left(\int_Y |K(x,y)f(y)|d\nu(y)\right)^p d\mu(x) &\leq C_1^p \int_X \int_Y |K_2(x,y)f(y)|^p d\nu(y)d\mu(x)\\ = C_1^p \int_Y \int_X |K_2(x,y)f(y)|^p d\mu(x)d\nu(y) &\leq C_1^p C_2^p \|f\|_p^p.\end{aligned}$$

Hence $\int_Y |K(x,y)f(y)|d\nu(y) \in L^p(X, d\mu)$ and, in particular, it is finite for μ-almost every x. Thus $K(x,.)f(.)$ is ν integrable for μ-almost every x and $\int_Y K(x,y)f(y)d\nu(y)$ is measurable. If $p = \infty$ just take the supremum and note that integrability again follows from Fubini by restricting to subsets of finite μ-measure. □

Note that the assumptions are, for example, satisfied if $\|K(x,.)\|_{L^1(Y,d\nu)} \le C$ and $\|K(.,y)\|_{L^1(X,d\mu)} \le C$ which follows by choosing $K_1(x,y) = |K(x,y)|^{1/q}$ and $K_2(x,y) = |K(x,y)|^{1/p}$. For related results see also Problems 10.31 and 15.2.

Another case of special importance is the case of integral operators

$$(Kf)(x) := \int_X K(x,y) f(y) d\mu(y), \qquad f \in L^2(X, d\mu), \tag{10.34}$$

where $K(x,y) \in L^2(X \times X, d\mu \otimes d\mu)$. Such an operator is called a **Hilbert–Schmidt operator**.

Lemma 10.24. *Let K be a Hilbert–Schmidt operator in $L^2(X, d\mu)$. Then*

$$\int_X \int_X |K(x,y)|^2 d\mu(x) d\mu(y) = \sum_{j \in J} \|K u_j\|^2 \tag{10.35}$$

for every orthonormal basis $\{u_j\}_{j\in J}$ in $L^2(X, d\mu)$.

Proof. Since $K(x,.) \in L^2(X, d\mu)$ for μ-almost every x we infer from Parseval's relation

$$\sum_j \left| \int_X K(x,y) u_j(y) d\mu(y) \right|^2 = \int_X |K(x,y)|^2 d\mu(y)$$

for μ-almost every x and thus

$$\begin{aligned} \sum_j \|K u_j\|^2 &= \sum_j \int_X \left| \int_X K(x,y) u_j(y) d\mu(y) \right|^2 d\mu(x) \\ &= \int_X \sum_j \left| \int_X K(x,y) u_j(y) d\mu(y) \right|^2 d\mu(x) \\ &= \int_X \int_X |K(x,y)|^2 d\mu(x) d\mu(y) \end{aligned}$$

as claimed. □

Hence in combination with Lemma 3.23 this shows that our definition for integral operators agrees with our previous definition from Section 3.6. In particular, this gives us an easy to check criterion for compactness of an integral operator.

Example 10.8. Let $[a,b]$ be some compact interval and suppose $K(x,y)$ is bounded. Then the corresponding integral operator in $L^2(a,b)$ is Hilbert–Schmidt and thus compact. This generalizes Lemma 3.4. ⋄

In combination with the spectral theorem for compact operators (compare in particular Corollary 3.8) we obtain the classical Hilbert–Schmidt theorem:

Theorem 10.25 (Hilbert–Schmidt)**.** *Let K be a self-adjoint Hilbert–Schmidt operator in $L^2(X, d\mu)$. Let $\{u_j\}$ be an orthonormal set of eigenfunctions with corresponding nonzero eigenvalues $\{\kappa_j\}$ from the spectral theorem for compact operators (Theorem 3.7). Then*

$$K(x, y) = \sum_j \kappa_j u_j(x) u_j(y)^*, \tag{10.36}$$

where the sum converges in $L^2(X \times X, d\mu \otimes d\mu)$.

Proof. First of all we can extend $\{u_j\}$ to an orthonormal basis (setting the corresponding κ_j equal to 0). Now by Problem 10.32 $\{u_j(x)u_j(y)^*\}$ is an orthogonal basis for $L^2(X \times X, d\mu \otimes d\mu)$ and the expansion coefficients of $K(x, y)$ are given by $\int_X \int_X u_j(x)^* u_k(y) K(x, y) d\mu(y) d\mu(x) = \langle u_j, K u_k \rangle = \kappa_k \delta_{j,k}$. □

In this context the above theorem is known as second Hilbert–Schmidt theorem and the spectral theorem for compact operators is known as first Hilbert–Schmidt theorem.

If an integral operator is positive we can say more. But first we will discuss two equivalent definitions of positivity in this context. First of all recall that an operator $K \in \mathscr{L}(L^2(X, d\mu))$ is positive if $\langle f, Kf \rangle \geq 0$ for all $f \in L^2(X, d\mu)$. Recall that positive operators are in particular self-adjoint. Secondly we call a kernel **symmetric** if it satisfies $K(x, y)^* = K(y, x)$ (cf. Problem 10.29) and a symmetric continuous kernel **positive semidefinite** on $U \subseteq X$ if

$$\sum_{j,k=1}^{n} \alpha_j^* \alpha_k K(x_j, x_k) \geq 0 \tag{10.37}$$

for all $(\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ and $\{x_j\}_{j=1}^n \subseteq U$.

Both conditions have their advantages. For example, note that for a positive semidefinite kernel the case $n = 1$ shows that $K(x, x) \geq 0$ for $x \in U$ and the case $n = 2$ shows (look at the determinant) $|K(x, y)|^2 \leq K(x, x)K(y, y)$ for $x, y \in U$. On the other hand, note that for a positive operator all eigenvalues are nonnegative.

Lemma 10.26. *Suppose X is a locally compact metric space and μ a Borel measure. Let $K \in \mathscr{L}(L^2(X, d\mu))$ be an integral operator with a continuous kernel. Then, if K is positive, its kernel is positive semidefinite on* $\operatorname{supp}(\mu)$*. If μ is regular, the converse is also true.*

Proof. Let $x_0 \in \operatorname{supp}(\mu)$ and consider $\delta_{x_0,\varepsilon} = \mu(B_\varepsilon(x_0))^{-1}\chi_{B_\varepsilon(x_0)}$. Then for $f_\varepsilon = \sum_{j=1}^n \alpha_j \delta_{x_0,\varepsilon}$

$$0 \le \lim_{\varepsilon\downarrow 0}\langle f_\varepsilon, Kf_\varepsilon\rangle = \lim_{\varepsilon\downarrow 0}\sum_{j,k=1}^n \alpha_j^*\alpha_k \int_{B_\varepsilon(x_j)}\int_{B_\varepsilon(x_k)} K(x,y)\frac{d\mu(x)}{\mu(B_\varepsilon(x_j))}\frac{d\mu(y)}{\mu(B_\varepsilon(x_k))}$$

$$= \sum_{j,k=1}^n \alpha_j^*\alpha_k K(x_j,x_k).$$

Conversely, let $f \in C_c(X)$ and let $S := \operatorname{supp}(f)\cap\operatorname{supp}(\mu)$. Then the function $f(x)^*K(x,y)f(y) \in C_c(X\times X)$ is uniformly continuous and for every $\varepsilon > 0$ we can partition the compact set S into a finite number of sets U_j which are contained in a ball $B_\delta(x_j)$ such that

$$|f(x)^*K(x,y)f(y) - \sum_{j,k}\chi_{U_j}(x)f(x_j)^*K(x_j,x_k)f(x_k)\chi_{U_k}(y)| \le \varepsilon, \quad x,y\in S.$$

Hence

$$\Big|\langle f, Kf\rangle - \sum_{j,k}\mu(U_j)f(x_j)^*K(x_j,x_k)f(x_k)\mu(U_k)\Big| \le \varepsilon\mu(S)^2$$

and since $\varepsilon > 0$ is arbitrary we have $\langle f, Kf\rangle \ge 0$. Since $C_c(X)$ is dense in $L^2(X,d\mu)$ if μ is regular by Theorem 10.16, we get this for all $f \in L^2(X,d\mu)$ by taking limits. □

Now we are ready for the following classical result:

Theorem 10.27 (Mercer)**.** *Let K be a positive integral operator with a continuous kernel on $L^2(X,d\mu)$ with X a locally compact metric space. Let μ a Borel measure such that the diagonal $K(x,x)$ is integrable. Then K is trace class, all eigenfunctions u_j corresponding to positive eigenvalues κ_j are continuous, and* (10.36) *converges uniformly on compact subsets of the support of μ. Moreover,*

$$\operatorname{tr}(K) = \sum_j \kappa_j = \int_X K(x,x)d\mu(x). \tag{10.38}$$

Proof. Define $k(x)^2 := K(x,x)$ such that $|K(x,y)| \le k(x)k(y)$ for $x,y \in \operatorname{supp}(\mu)$. Since by assumption $k \in L^2(X,d\mu)$ we see that K is Hilbert–Schmidt and we have the representation (10.36) with $\kappa_j > 0$. Moreover, dominated convergence shows that $u_j(x) = \kappa_j^{-1}\int_X K(x,y)u_j(y)d\mu(y)$ is continuous.

Now note that the operator corresponding to the kernel

$$K_n(x,y) = K(x,y) - \sum_{j\le n}\kappa_j u_j(x)^*u_j(y) = \sum_{j>n}\kappa_j u_j(x)^*u_j(y)$$

is also positive (its nonzero eigenvalues are κ_j, $j > n$). Hence $K_n(x,x) \geq 0$ implying

$$\sum_{j\leq n} \kappa_j |u_j(x)|^2 \leq k(x)^2$$

for $x \in \operatorname{supp}(\mu)$ and hence (10.36) converges absolutely for $x, y \in \operatorname{supp}(\mu)$. Denote the limit by $\tilde{K}(x,y)$ and observe

$$\begin{aligned}\left|\sum_{m<j<n} \kappa_j u_j(x)^* u_j(y)\right| &\leq \left(\sum_{m<j<n} \kappa_j |u_j(x)|^2\right)^{1/2} \left(\sum_{m<j<n} \kappa_j |u_j(y)|^2\right)^{1/2} \\ &\leq \left(\sum_{m<j<n} \kappa_j |u_j(x)|^2\right)^{1/2} k(y)\end{aligned}$$

shows that the series converges uniformly for fixed x with respect to y in compact sets. Hence $\tilde{K}(x,.)$ is continuous for fixed x and satisfies $|\tilde{K}(x,y)| \leq k(x)k(y)$. Moreover, for fixed x we have $K(x,.), \tilde{K}(x,.) \in L^2(X, d\mu)$ and

$$\int_X K(x,y)u(y)d\mu(y) = \kappa u(x) = \int_X \tilde{K}(x,y)u(y)d\mu(y)$$

for either $u \in \operatorname{Ker}(K)$ with $\kappa = 0$ or $u = u_j$ with $\kappa = \kappa_j$ (use dominated convergence to interchange the sum and the integral). Since these functions are total we get $K(x,y) = \tilde{K}(x,y)$ for fixed x and a.e. y. But since both functions are continuous for fixed x, we get equality for all y. In particular, we have $\sum_{j\leq n} \kappa_j |u_j(x)|^2 \nearrow k(x)^2$ and the convergence is uniform on compact subsets of $\operatorname{supp}(\mu)$ by Dini's theorem (Problem B.29). By our estimate this also shows uniform convergence of (10.36) on compact sets.

Finally, integrating (10.36) for $x = y$ using $\|u_j\| = 1$ shows the last claim. □

Example 10.9. Let k be a periodic function which is square integrable over $[-\pi, \pi]$. Then the integral operator

$$(Kf)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} k(y-x) f(y) dy$$

has the eigenfunctions $u_j(x) = (2\pi)^{-1/2} \mathrm{e}^{-\mathrm{i}jx}$ with corresponding eigenvalues $\hat{k}_j$, $j \in \mathbb{Z}$, where $\hat{k}_j$ are the Fourier coefficients of k. Since $\{u_j\}_{j\in\mathbb{Z}}$ is an ONB we have found all eigenvalues. Moreover, in this case (10.36) is just the Fourier series

$$k(y-x) = \sum_{j\in\mathbb{Z}} \hat{k}_j \mathrm{e}^{\mathrm{i}j(y-x)}.$$

Choosing a continuous function for which the Fourier series does not converge absolutely shows that the positivity assumption in Mercer's theorem is crucial. ◇

Of course given a kernel this raises the question if it is positive semi-definite. This can often be done by reverse engineering basend on constructions which produce new positive semi-definite kernels out of old ones.

Lemma 10.28. *Let $K_j(x,y)$ be positive semi-definite kernels on X. Then the following kernels are also positive semi-definite.*

(i) $\alpha_1 K_1 + \alpha_2 K_2$ *provided* $\alpha_j \geq 0$.
(ii) $\lim_j K_j(x,y)$ *provided the limit exists pointwise.*
(iii) $K_1(\phi(x),\phi(y))$ *for* $\phi : X \to X$.
(iv) $f(x)^* K_1(x,y) f(y)$ *for* $f : X \to \mathbb{C}$.
(v) $K_1(x,y)K_2(x,y)$.

Proof. Only the last item is not straightforward. However, it boils down to the Schur product theorem from linear algebra given below. □

Theorem 10.29 (Schur)**.** *Let A and B be two $n \times n$ matrices and let $A \circ B$ be their* ***Hadamard product*** *defined by multiplying the entries pointwise: $(A \circ B)_{jk} := A_{jk} B_{jk}$. If A and B are positive (semi-) definite, then so is $A \circ B$.*

Proof. By the spectral theorem we can write $A = \sum_{j=1}^n \alpha_j \langle u_j, . \rangle u_j$, where $\alpha_j > 0$ ($\alpha_j \geq 0$) are the eigenvalues and u_j are a corresponding ONB of eigenvectors. Similarly $B = \sum_{j=1}^n \beta_j \langle v_j, . \rangle v_j$. Then $A \circ B = \sum_{j,k=1}^n \alpha_j \beta_k (\langle u_j, . \rangle u_j) \circ (\langle v_k, . \rangle v_k) = \sum_{j,k=1}^n \alpha_j \beta_k \langle u_j \circ v_k, . \rangle u_j \circ v_k$, which proves the claim. □

Example 10.10. Let $X = \mathbb{R}^n$. The most basic kernel is $K(x,y) = x^* \cdot y$ which is positive semi-definite since $\sum_{j,k} \alpha_j^* \alpha_k K(x_j, x_k) = \left| \sum_j \alpha_j x_j \right|^2$. Slightly more general is $K(x,y) = \langle x, Ay \rangle$, where A is a positive semi-definite matrix. Indeed writing $A = B^2$ this follows from the previous observation using item (iii) with $\phi = B$. Another famous kernel is the Gaussian kernel

$$K(x,y) = \mathrm{e}^{-|x-y|^2/\sigma}, \qquad \sigma > 0.$$

To see that it is positive semi-definite start by observing that $K_1(x,y) = \exp(2x^* \cdot y/\sigma)$ is by items (i) and (ii) since the Taylor series of the exponential function has positive coefficients. Finally observe that our kernel is of the form (vi) with $f(x) = \exp(-|x|^2/\sigma)$. ◇

Example 10.11. A **reproducing kernel Hilbert space** is a Hilbert space $\mathfrak{H}$ of complex-valued functions $X \to \mathbb{C}$ such that point evaluations are continuous linear functionals. In this case the Riesz lemma implies that for every $x \in X$ there is a corresponding function $K_x \in \mathfrak{H}$ such that $f(x) = \langle K_x, f \rangle$. Applying this to $f = K_y$ we get $K_y(x) = \langle K_x, K_y \rangle$ which suggests the more symmetric notation

$$K(x,y) := \langle K_x, K_y \rangle.$$

The kernel K is called the **reproducing kernel** for $\mathfrak{H}$. A short calculation

$$\sum_{j,k} \alpha_j^* \alpha_k K(x_j, x_k) = \sum_{j,k} \alpha_j^* \alpha_k \langle K_{x_j}, K_{x_k} \rangle = \Big\| \sum_j \alpha_j K_{x_j} \Big\|^2$$

verifies that it is a positive semi-definite kernel. An explicit example is given by the quadratic form domains of positive Sturm–Liouville operators; Problem 3.9. ⋄

Problem* 10.29. *Suppose K is a Hilbert–Schmidt operator in $L^2(X, d\mu)$ with kernel $K(x, y)$. Show that the adjoint operator is given by*

$$(K^* f)(x) = \int_X K(y, x)^* f(y) d\mu(y), \qquad f \in L^2(X, d\mu).$$

In particular, K is self-adjoint if and only if its kernel is symmetric.

Problem 10.30. *Obtain Young's inequality* (10.25) *from Schur's criterion.*

Problem 10.31 (Schur test)**.** *Let $K(x, y)$ be given and suppose there are positive measurable functions $a(x)$ and $b(y)$ such that*

$$\|K(x, .)b(.)\|_{L^1(Y,d\nu)} \le C_1 a(x), \qquad \|a(.)K(., y)\|_{L^1(X,d\mu)} \le C_2 b(y).$$

Then the operator $K : L^2(Y, d\nu) \to L^2(X, d\mu)$, defined by

$$(Kf)(x) := \int_Y K(x, y) f(y) d\nu(y),$$

for μ-almost every x is bounded with $\|K\| \le \sqrt{C_1 C_2}$. Show that the adjoint operator is given by

$$(K^* f)(x) = \int_X K(y, x)^* f(y) d\mu(y), \qquad f \in L^2(X, d\mu).$$

(Hint: Estimate $|(Kf)(x)|^2 \le |\int_Y |K(x, y)|^{1/2} b(y) |K(x, y)|^{1/2} \frac{|f(y)|}{b(y)} d\nu(y)|^2$ using Cauchy–Schwarz and integrate the result with respect to x.)

Problem* 10.32. *Let $(X, d\mu)$ and $(Y, d\nu)$ be two measure spaces and $\{u_j(x)\}_{j\in J}$, $\{v_k(y)\}_{k\in K}$ be orthonormal bases for $L^2(X, d\mu)$, $L^2(Y, d\nu)$, respectively. Then $\{u_j(x) v_k(y)\}_{(j,k)\in J\times K}$ is an orthonormal base for $L^2(X \times Y, d\mu \otimes d\nu)$.*

Problem 10.33. *Let K be an self-adjoint integral operator with continuous kernel satisfying the estimate $|K(x, y)| \le k(x)k(y)$ with $k \in L^2(X, d\mu) \cap L^\infty(X, d\mu)$ (e.g. X is compact). Show that the conclusion from Mercer's theorem still hold if K has only a finite number of negative eigenvalues.*

Problem 10.34. *Show that the Fourier transform of a finite (positive) measure*

$$\hat{\mu}(p) := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}px} d\mu(x)$$

gives rise to a positive semi-definite kernel $K(x, y) = \hat{\mu}(x - y)$.

Chapter 11

More measure theory

11.1. Decomposition of measures

Let μ, ν be two measures on a measurable space (X, Σ). They are called **mutually singular** (in symbols $\mu \perp \nu$) if they are supported on disjoint sets. That is, there is a measurable set N such that $\mu(N) = 0$ and $\nu(X \setminus N) = 0$.

Example 11.1. Let λ be the Lebesgue measure and Θ the Dirac measure (centered at 0). Then $\lambda \perp \Theta$: Just take $N = \{0\}$; then $\lambda(\{0\}) = 0$ and $\Theta(\mathbb{R} \setminus \{0\}) = 0$. ⋄

On the other hand, ν is called **absolutely continuous** with respect to μ (in symbols $\nu \ll \mu$) if $\mu(A) = 0$ implies $\nu(A) = 0$.

Example 11.2. The prototypical example is the measure $d\nu := f\, d\mu$ (compare Lemma 9.3). Indeed by Lemma 9.6 $\mu(A) = 0$ implies

$$\nu(A) = \int_A f\, d\mu = 0 \tag{11.1}$$

and shows that ν is absolutely continuous with respect to μ. In fact, we will show below that every absolutely continuous measure is of this form. ⋄

The two main results will follow as simple consequence of the following result:

Theorem 11.1. *Let μ, ν be σ-finite measures. Then there exists a nonnegative function f and a set N of μ measure zero, such that*

$$\nu(A) = \nu(A \cap N) + \int_A f\, d\mu. \tag{11.2}$$

Proof. We first assume μ, ν to be finite measures. Let $\alpha := \mu + \nu$ and consider the Hilbert space $L^2(X, d\alpha)$. Then

$$\ell(h) := \int_X h \, d\nu$$

is a bounded linear functional on $L^2(X, d\alpha)$ by Cauchy–Schwarz:

$$\begin{aligned} |\ell(h)|^2 = \left| \int 1 \cdot h \, d\nu \right|^2 &\leq \left(\int |1|^2 \, d\nu \right) \left(\int |h|^2 d\nu \right) \\ &\leq \nu(X) \left(\int |h|^2 d\alpha \right) = \nu(X) \|h\|^2. \end{aligned}$$

Hence by the Riesz lemma (Theorem 2.10) there exists a $g \in L^2(X, d\alpha)$ such that

$$\ell(h) = \int_X hg \, d\alpha.$$

By construction

$$\nu(A) = \int \chi_A \, d\nu = \int \chi_A g \, d\alpha = \int_A g \, d\alpha. \tag{11.3}$$

In particular, g must be positive a.e. (take A the set where g is negative). Moreover,

$$\mu(A) = \alpha(A) - \nu(A) = \int_A (1 - g) d\alpha$$

which shows that $g \leq 1$ a.e. Now chose $N := \{x | g(x) = 1\}$ such that $\mu(N) = 0$ and set

$$f := \frac{g}{1-g} \chi_{N'}, \qquad N' = X \setminus N.$$

Then, since (11.3) implies $d\nu = g \, d\alpha$, respectively, $d\mu = (1 - g) d\alpha$, we have

$$\int_A f d\mu = \int \chi_A \frac{g}{1-g} \chi_{N'} \, d\mu = \int \chi_{A \cap N'} g \, d\alpha = \nu(A \cap N')$$

as desired.

To see the σ-finite case, observe that $Y_n \nearrow X$, $\mu(Y_n) < \infty$ and $Z_n \nearrow X$, $\nu(Z_n) < \infty$ implies $X_n := Y_n \cap Z_n \nearrow X$ and $\alpha(X_n) < \infty$. Now we set $\tilde{X}_n := X_n \setminus X_{n-1}$ (where $X_0 = \emptyset$) and consider $\mu_n(A) := \mu(A \cap \tilde{X}_n)$ and $\nu_n(A) := \nu(A \cap \tilde{X}_n)$. Then there exist corresponding sets N_n and functions f_n such that

$$\nu_n(A) = \nu_n(A \cap N_n) + \int_A f_n d\mu_n = \nu(A \cap N_n) + \int_A f_n d\mu,$$

where for the last equality we have assumed $N_n \subseteq \tilde{X}_n$ and $f_n(x) = 0$ for $x \in \tilde{X}_n'$ without loss of generality. Now set $N := \bigcup_n N_n$ as well as

$f := \sum_n f_n$, then $\mu(N) = 0$ and

$$\nu(A) = \sum_n \nu_n(A) = \sum_n \nu(A \cap N_n) + \sum_n \int_A f_n d\mu = \nu(A \cap N) + \int_A f d\mu,$$

which finishes the proof. □

Note that another set $\tilde{N}$ will give the same decomposition as long as $\mu(\tilde{N}) = 0$ and $\nu(\tilde{N}' \cap N) = 0$ since in this case $\nu(A) = \nu(A \cap \tilde{N}) + \nu(A \cap \tilde{N}') = \nu(A \cap \tilde{N}) + \nu(A \cap \tilde{N}' \cap N) + \int_{A \cap \tilde{N}'} f d\mu = \nu(A \cap \tilde{N}) + \int_A f d\mu$. Hence we can increase N by sets of μ measure zero and decrease N by sets of ν measure zero.

Now the anticipated results follow with no effort:

Theorem 11.2 (Radon–Nikodym)**.** *Let μ, ν be two σ-finite measures on a measurable space (X, Σ). Then ν is absolutely continuous with respect to μ if and only if there is a nonnegative measurable function f such that*

$$\nu(A) = \int_A f \, d\mu \tag{11.4}$$

for every $A \in \Sigma$. The function f is determined uniquely a.e. with respect to μ and is called the ***Radon–Nikodym derivative*** $\frac{d\nu}{d\mu}$ *of ν with respect to μ.*

Proof. Just observe that in this case $\nu(A \cap N) = 0$ for every A. Uniqueness will be shown in the next theorem. □

Example 11.3. Take $X := \mathbb{R}$. Let μ be the counting measure and ν Lebesgue measure. Then $\nu \ll \mu$ but there is no f with $d\nu = f \, d\mu$. If there were such an f, there must be a point $x_0 \in \mathbb{R}$ with $f(x_0) > 0$ and we have $0 = \nu(\{x_0\}) = \int_{\{x_0\}} f \, d\mu = f(x_0) > 0$, a contradiction. Hence the Radon–Nikodym theorem can fail if μ is not σ-finite. ◇

Theorem 11.3 (Lebesgue decomposition)**.** *Let μ, ν be two σ-finite measures on a measurable space (X, Σ). Then ν can be uniquely decomposed as $\nu = \nu_{sing} + \nu_{ac}$, where μ and ν_{sing} are mutually singular and ν_{ac} is absolutely continuous with respect to μ.*

Proof. Taking $\nu_{sing}(A) := \nu(A \cap N)$ and $d\nu_{ac} := f \, d\mu$ from the previous theorem, there is at least one such decomposition. To show uniqueness assume there is another one, $\nu = \tilde{\nu}_{ac} + \tilde{\nu}_{sing}$, and let $\tilde{N}$ be such that $\mu(\tilde{N}) = 0$ and $\tilde{\nu}_{sing}(\tilde{N}') = 0$. Then $\nu_{sing}(A) - \tilde{\nu}_{sing}(A) = \int_A (\tilde{f} - f) d\mu$. In particular, $\int_{A \cap N' \cap \tilde{N}'} (\tilde{f} - f) d\mu = 0$ and hence, since A is arbitrary, $\tilde{f} = f$ a.e. away from $N \cup \tilde{N}$. Since $\mu(N \cup \tilde{N}) = 0$, we have $\tilde{f} = f$ a.e. and hence $\tilde{\nu}_{ac} = \nu_{ac}$ as well as $\tilde{\nu}_{sing} = \nu - \tilde{\nu}_{ac} = \nu - \nu_{ac} = \nu_{sing}$. □

Problem* 11.1. *Let μ be a Borel measure on $\mathfrak{B}$ and suppose its distribution function $\mu(x)$ is continuously differentiable. Show that the Radon–Nikodym derivative equals the ordinary derivative $\mu'(x)$.*

Problem 11.2. *Suppose ν is an inner regular measures. Show that $\nu \ll \mu$ if and only if $\mu(K) = 0$ implies $\nu(K) = 0$ for every compact set.*

Problem 11.3. *Suppose $\nu(A) \leq C\mu(A)$ for all $A \in \Sigma$. Then $d\nu = f\, d\mu$ with $0 \leq f \leq C$ a.e.*

Problem 11.4. *Let $d\nu = f\, d\mu$. Suppose $f > 0$ a.e. with respect to μ. Then $\mu \ll \nu$ and $d\mu = f^{-1} d\nu$.*

Problem 11.5 (Chain rule)**.** *Show that $\nu \ll \mu$ is a transitive relation. In particular, if $\omega \ll \nu \ll \mu$, show that*

$$\frac{d\omega}{d\mu} = \frac{d\omega}{d\nu}\frac{d\nu}{d\mu}.$$

Problem 11.6. *Suppose $\nu \ll \mu$. Show that for every measure ω we have*

$$\frac{d\omega}{d\mu} d\mu = \frac{d\omega}{d\nu} d\nu + d\zeta,$$

where ζ is a positive measure (depending on ω) which is singular with respect to ν. Show that $\zeta = 0$ if and only if $\mu \ll \nu$.

11.2. Derivatives of measures

If μ is a Borel measure on $\mathfrak{B}$ and its distribution function $\mu(x)$ is continuously differentiable, then the Radon–Nikodym derivative is just the ordinary derivative $\mu'(x)$ (Problem 11.1). Our aim in this section is to generalize this result to arbitrary Borel measures on $\mathfrak{B}^n$.

Let μ be a Borel measure on $\mathbb{R}^n$. We call

$$(D\mu)(x) := \lim_{\varepsilon \downarrow 0} \frac{\mu(B_\varepsilon(x))}{|B_\varepsilon(x)|} \tag{11.5}$$

the derivative of μ at $x \in \mathbb{R}^n$ provided the above limit exists. (Here $B_r(x) \subset \mathbb{R}^n$ is a ball of radius r centered at $x \in \mathbb{R}^n$ and $|A|$ denotes the Lebesgue measure of $A \in \mathfrak{B}^n$.)

Example 11.4. Consider a Borel measure on $\mathfrak{B}$ and suppose its distribution $\mu(x)$ (as defined in (8.18)) is differentiable at x. Then

$$(D\mu)(x) = \lim_{\varepsilon \downarrow 0} \frac{\mu((x+\varepsilon, x-\varepsilon))}{2\varepsilon} = \lim_{\varepsilon \downarrow 0} \frac{\mu(x+\varepsilon) - \mu(x-\varepsilon)}{2\varepsilon} = \mu'(x).$$

$\diamond$

To compute the derivative of μ, we introduce the **upper** and **lower derivative**,

$$(\overline{D}\mu)(x) := \limsup_{\varepsilon\downarrow 0} \frac{\mu(B_\varepsilon(x))}{|B_\varepsilon(x)|} \quad \text{and} \quad (\underline{D}\mu)(x) := \liminf_{\varepsilon\downarrow 0} \frac{\mu(B_\varepsilon(x))}{|B_\varepsilon(x)|}. \tag{11.6}$$

Clearly μ is differentiable at x if $(\underline{D}\mu)(x) = (\overline{D}\mu)(x) < \infty$. Next note that they are measurable: In fact, this follows from

$$(\overline{D}\mu)(x) = \lim_{n\to\infty} \sup_{0<\varepsilon<1/n} \frac{\mu(B_\varepsilon(x))}{|B_\varepsilon(x)|} \tag{11.7}$$

since the supremum on the right-hand side is lower semicontinuous with respect to x (cf. Problem 8.19) as $x \mapsto \mu(B_\varepsilon(x))$ is lower semicontinuous (Problem 11.7). Similarly for $(\underline{D}\mu)(x)$.

Next, the following geometric fact of $\mathbb{R}^n$ will be needed.

Lemma 11.4 (Wiener covering lemma)**.** *Given open balls $B_1 := B_{r_1}(x_1)$, ..., $B_m := B_{r_m}(x_m)$ in $\mathbb{R}^n$, there is a subset of disjoint balls B_{j_1}, ..., B_{j_k} such that*

$$\bigcup_{j=1}^{m} B_j \subseteq \bigcup_{\ell=1}^{k} B_{3r_{j_\ell}}(x_{j_\ell}) \tag{11.8}$$

Proof. Assume that the balls B_j are ordered by decreasing radius. Start with $B_{j_1} = B_1$ and remove all balls from our list which intersect B_{j_1}. Observe that the removed balls are all contained in $B_{3r_1}(x_1)$. Proceeding like this, we obtain the required subset. □

The upshot of this lemma is that we can select a disjoint subset of balls which still controls the Lebesgue volume of the original set up to a universal constant 3^n (recall $|B_{3r}(x)| = 3^n|B_r(x)|$).

Now we can show

Lemma 11.5. *Let $\alpha > 0$. For every Borel set A we have*

$$|\{x \in A \,|\, (\overline{D}\mu)(x) > \alpha\}| \le 3^n \frac{\mu(A)}{\alpha} \tag{11.9}$$

and

$$|\{x \in A \,|\, (\overline{D}\mu)(x) > 0\}| = 0, \text{ whenever } \mu(A) = 0. \tag{11.10}$$

Proof. Let $A_\alpha := \{x \in A | (\overline{D}\mu)(x) > \alpha\}$. We will show

$$|K| \le 3^n \frac{\mu(O)}{\alpha}$$

for every open set O with $A \subseteq O$ and every compact set $K \subseteq A_\alpha$. The first claim then follows from outer regularity of μ and inner regularity of the Lebesgue measure.

Given fixed K, O, for every $x \in K$ there is some r_x such that $B_{r_x}(x) \subseteq O$ and $|B_{r_x}(x)| < \alpha^{-1}\mu(B_{r_x}(x))$. Since K is compact, we can choose a finite subcover of K from these balls. Moreover, by Lemma 11.4 we can refine our set of balls such that

$$|K| \le 3^n \sum_{i=1}^{k} |B_{r_i}(x_i)| < \frac{3^n}{\alpha} \sum_{i=1}^{k} \mu(B_{r_i}(x_i)) \le 3^n \frac{\mu(O)}{\alpha}.$$

To see the second claim, observe that $A_0 = \cup_{j=1}^{\infty} A_{1/j}$ and by the first part $|A_{1/j}| = 0$ for every j if $\mu(A) = 0$. □

Theorem 11.6 (Lebesgue)**.** *Let f be (locally) integrable, then for a.e. $x \in \mathbb{R}^n$ we have*

$$\lim_{r \downarrow 0} \frac{1}{|B_r(x)|} \int_{B_r(x)} |f(y) - f(x)| d^n y = 0. \tag{11.11}$$

The points where (11.11) *holds are called* ***Lebesgue points*** *of f.*

Proof. Decompose f as $f = g + h$, where g is continuous and $\|h\|_1 < \varepsilon$ (Theorem 10.16) and abbreviate

$$D_r(f)(x) := \frac{1}{|B_r(x)|} \int_{B_r(x)} |f(y) - f(x)| d^n y.$$

Then, since $\lim D_r(g)(x) = 0$ (for every x) and $D_r(f) \le D_r(g) + D_r(h)$, we have

$$\limsup_{r \downarrow 0} D_r(f)(x) \le \limsup_{r \downarrow 0} D_r(h)(x) \le (\overline{D}\mu)(x) + |h(x)|,$$

where $d\mu = |h| d^n x$. This implies

$$\{x \,|\, \limsup_{r \downarrow 0} D_r(f)(x) \ge 2\alpha\} \subseteq \{x | (\overline{D}\mu)(x) \ge \alpha\} \cup \{x \,|\, |h(x)| \ge \alpha\}$$

and using the first part of Lemma 11.5 plus $|\{x \,|\, |h(x)| \ge \alpha\}| \le \alpha^{-1}\|h\|_1$ (Problem 11.10), we see

$$|\{x \,|\, \limsup_{r \downarrow 0} D_r(f)(x) \ge 2\alpha\}| \le (3^n + 1)\frac{\varepsilon}{\alpha}.$$

Since ε is arbitrary, the Lebesgue measure of this set must be zero for every α. That is, the set where the lim sup is positive has Lebesgue measure zero. □

Example 11.5. It is easy to see that every point of continuity of f is a Lebesgue point. However, the converse is not true. To see this consider a function $f : \mathbb{R} \to [0, 1]$ which is given by a sum of non-overlapping spikes centered at $x_j = 2^{-j}$ with base length $2b_j$ and height 1. Explicitly

$$f(x) = \sum_{j=1}^{\infty} \max(0, 1 - b_j^{-1}|x - x_j|)$$

with $b_{j+1}+b_j \le 2^{-j-1}$ (such that the spikes don't overlap). By construction $f(x)$ will be continuous except at $x = 0$. However, if we let the b_j's decay sufficiently fast such that the area of the spikes inside $(-r, r)$ is $o(r)$, the point $x = 0$ will nevertheless be a Lebesgue point. For example, $b_j = 2^{-2j-1}$ will do. ⋄

Note that the balls can be replaced by more general sets: A sequence of sets $A_j(x)$ is said to shrink to x nicely if there are balls $B_{r_j}(x)$ with $r_j \to 0$ and a constant $\varepsilon > 0$ such that $A_j(x) \subseteq B_{r_j}(x)$ and $|A_j| \ge \varepsilon |B_{r_j}(x)|$. For example, $A_j(x)$ could be some balls or cubes (not necessarily containing x). However, the portion of $B_{r_j}(x)$ which they occupy must not go to zero! For example, the rectangles $(0, \frac{1}{j}) \times (0, \frac{2}{j}) \subset \mathbb{R}^2$ do shrink nicely to 0, but the rectangles $(0, \frac{1}{j}) \times (0, \frac{2}{j^2})$ do not.

Lemma 11.7. *Let f be (locally) integrable. Then at every Lebesgue point we have*

$$f(x) = \lim_{j\to\infty} \frac{1}{|A_j(x)|} \int_{A_j(x)} f(y) d^n y \tag{11.12}$$

whenever $A_j(x)$ shrinks to x nicely.

Proof. Let x be a Lebesgue point and choose some nicely shrinking sets $A_j(x)$ with corresponding $B_{r_j}(x)$ and ε. Then

$$\frac{1}{|A_j(x)|} \int_{A_j(x)} |f(y) - f(x)| d^n y \le \frac{1}{\varepsilon |B_{r_j}(x)|} \int_{B_{r_j}(x)} |f(y) - f(x)| d^n y$$

and the claim follows. □

If we take f to be the Radon–Nykodym derivative of an absolutely continuous measure, then these results can also be understood as differentiation results for measures. To complement these results we provide the corresponding statement for singular measures.

Lemma 11.8. *Let μ be singular with respect to Lebesgue measure. Then at almost every point x (with respect to Lebesgue measure) we have*

$$\lim_{j\to\infty} \frac{\mu(A_j(x))}{|A_j(x)|} = 0, \tag{11.13}$$

whenever $A_j(x)$ shrinks to x nicely. In particular, we have $(\overline{D}\mu)(x) = 0$ for almost every x.

Proof. By the Lebesgue decomposition theorem μ is supported on a set N of Lebesgue measure zero. Hence choosing $A = \mathbb{R}^n \setminus N$ in the second part of Lemma 11.5 shows $(\overline{D}\mu)(x) = 0$ for $x \in A$. The remaining claim now follows from $\limsup_{j\to\infty} \frac{\mu(A_j(x))}{|A_j(x)|} \le \varepsilon \limsup_{j\to\infty} \frac{\mu(B_{r_j}(x))}{|B_{r_j}(x)|} \le \varepsilon (\overline{D}\mu)(x)$. □

Combining these two lemmas with Theorem 11.6 shows

Theorem 11.9. *Let μ be a Borel measure on $\mathbb{R}^n$. The derivative $D\mu$ exists a.e. with respect to Lebesgue measure and equals the Radon–Nikodym derivative of the absolutely continuous part of μ with respect to Lebesgue measure; that is,*

$$\mu_{ac}(A) = \int_A (D\mu)(x) d^n x. \tag{11.14}$$

In particular, μ is singular with respect to Lebesgue measure if and only if $D\mu = 0$ a.e. with respect to Lebesgue measure.

In the special case of Borel measures on $\mathbb{R}$ the last result reads

Corollary 11.10. *Let μ be a Borel measure on $\mathbb{R}$. Then its distribution function is differentiable a.e. with respect to Lebesgue measure and the derivative equals the Radon–Nykodym derivative.*

Proof. Let f be the Radon–Nykodym derivative. Since the sets $(x, x+r)$ shrink nicely to x as $r \to 0$, Lemma 11.7 implies

$$\lim_{r\to 0} \frac{\mu((x, x+r))}{r} = \lim_{r\to 0} \frac{\mu(x+r) - \mu(x)}{r} = f(x)$$

a.e. Since the same is true for the sets $(x-r, x)$, $\mu(x)$ is differentiable a.e. and $\mu'(x) = f(x)$. □

Using the upper and lower derivatives, we can also give supports for the absolutely and singularly continuous parts.

Theorem 11.11. *The set $\{x|0 < (D\mu)(x) < \infty\}$ is a support for the absolutely continuous and $\{x|(\underline{D}\mu)(x) = \infty\}$ is a support for the singular part.*

Proof. The first part is immediate from the previous theorem. For the second part first note that by $(\underline{D}\mu)(x) \geq (\underline{D}\mu_{sing})(x)$ we can assume that μ is purely singular. It suffices to show that the set $A_k := \{x \,|\, (\underline{D}\mu)(x) < k\}$ satisfies $\mu(A_k) = 0$ for every $k \in \mathbb{N}$.

Let $K \subset A_k$ be compact, and let $V_j \supset K$ be some open set such that $|V_j \setminus K| \leq \frac{1}{j}$. For every $x \in K$ there is some $\varepsilon = \varepsilon(x)$ such that $B_\varepsilon(x) \subseteq V_j$ and $\mu(B_{3\varepsilon}(x)) \leq k|B_{3\varepsilon}(x)|$. By compactness, finitely many of these balls cover K and hence

$$\mu(K) \leq \sum_i \mu(B_{\varepsilon_i}(x_i)).$$

Selecting disjoint balls as in Lemma 11.4 further shows

$$\mu(K) \leq \sum_\ell \mu(B_{3\varepsilon_{i_\ell}}(x_{i_\ell})) \leq k3^n \sum_\ell |B_{\varepsilon_{i_\ell}}(x_{i_\ell})| \leq k3^n |V_j|.$$

Letting $j \to \infty$, we see $\mu(K) \le k3^n|K|$ and by regularity we even have $\mu(A) \le k3^n|A|$ for every $A \subseteq A_k$. Hence μ is absolutely continuous on A_k and since we assumed μ to be singular, we must have $\mu(A_k) = 0$. □

Finally, we note that these supports are minimal. Here a support M of some measure μ is called a **minimal support** (it is sometimes also called an **essential support**) if every subset $M_0 \subseteq M$ which does not support μ (i.e., $\mu(M_0) = 0$) has Lebesgue measure zero.

Example 11.6. Let $X := \mathbb{R}$, $\Sigma := \mathfrak{B}$. If $d\mu(x) := \sum_n \alpha_n d\theta(x - x_n)$ is a sum of Dirac measures, then the set $\{x_n\}$ is clearly a minimal support for μ. Moreover, it is clearly the smallest support as none of the x_n can be removed. If we choose $\{x_n\}$ to be the rational numbers, then $\operatorname{supp}(\mu) = \mathbb{R}$, but $\mathbb{R}$ is not a minimal support, as we can remove the irrational numbers.

On the other hand, if we consider the Lebesgue measure λ, then $\mathbb{R}$ is a minimal support. However, the same is true if we remove any set of measure zero, for example, the Cantor set. In particular, since we can remove any single point, we see that, just like supports, minimal supports are not unique. ⋄

Lemma 11.12. *The set $M_{ac} := \{x|0 < (D\mu)(x) < \infty\}$ is a minimal support for μ_{ac}.*

Proof. Suppose $M_0 \subseteq M_{ac}$ and $\mu_{ac}(M_0) = 0$. Set $M_\varepsilon = \{x \in M_0|\varepsilon < (D\mu)(x)\}$ for $\varepsilon > 0$. Then $M_\varepsilon \nearrow M_0$ and

$$|M_\varepsilon| = \int_{M_\varepsilon} d^n x \le \frac{1}{\varepsilon}\int_{M_\varepsilon} (D\mu)(x)dx = \frac{1}{\varepsilon}\mu_{ac}(M_\varepsilon) \le \frac{1}{\varepsilon}\mu_{ac}(M_0) = 0$$

shows $|M_0| = \lim_{\varepsilon\downarrow 0} |M_\varepsilon| = 0$. □

Note that the set $M = \{x|0 < (D\mu)(x)\}$ is a minimal support of μ (Problem 11.8).

Example 11.7. The **Cantor function** is constructed as follows: Take the sets C_n used in the construction of the Cantor set C: C_n is the union of 2^n closed intervals with $2^n - 1$ open gaps in between. Set f_n equal to $j/2^n$ on the j'th gap of C_n and extend it to $[0, 1]$ by linear interpolation. Note that, since we are creating precisely one new gap between every old gap when going from C_n to C_{n+1}, the value of f_{n+1} is the same as the value of f_n on the gaps of C_n. Explicitly, we have $f_0(x) = x$ and $f_{n+1} = K(f_n)$, where

$$K(f)(x) := \begin{cases} \frac{1}{2}f(3x), & 0 \le x \le \frac{1}{3}, \\ \frac{1}{2}, & \frac{1}{3} \le x \le \frac{2}{3}, \\ \frac{1}{2}(1 + f(3x - 2)), & \frac{2}{3} \le x \le 1. \end{cases}$$

Since $\|f_{n+1} - f_n\|_\infty \le \frac{1}{2}\|f_{n+1} - f_n\|_\infty$ we can define the Cantor function as $f = \lim_{n\to\infty} f_n$. By construction f is a continuous function which is constant on every subinterval of $[0,1] \setminus C$. Since C is of Lebesgue measure zero, this set is of full Lebesgue measure and hence $f' = 0$ a.e. in $[0,1]$. In particular, the corresponding measure, the **Cantor measure**, is supported on C and is purely singular with respect to Lebesgue measure. $\diamond$

Problem* 11.7. *Show that*

$$\mu(B_\varepsilon(x)) \le \liminf_{y\to x} \mu(B_\varepsilon(y)) \le \limsup_{y\to x} \mu(B_\varepsilon(y)) \le \mu(\overline{B_\varepsilon(x)}).$$

In particular, conclude that $x \mapsto \mu(B_\varepsilon(x))$ is lower semicontinuous for $\varepsilon > 0$.

Problem* 11.8. *Show that $M := \{x | 0 < (D\mu)(x)\}$ is a minimal support of μ.*

Problem 11.9. *Suppose $\overline{D}\mu \le \alpha$. Show that $d\mu = f\, d^n x$ with $0 \le f \le \alpha$.*

Problem* 11.10 (Markov (also Chebyshev) inequality)**.** *For $f \in L^1(\mathbb{R}^n)$ and $\alpha > 0$ show*

$$|\{x \in A |\, |f(x)| \ge \alpha\}| \le \frac{1}{\alpha} \int_A |f(x)| d^n x.$$

Somewhat more general, assume $g(x) \ge 0$ is nondecreasing and $g(\alpha) > 0$. Then

$$\mu(\{x \in A |\, f(x) \ge \alpha\}) \le \frac{1}{g(\alpha)} \int_A g \circ f\, d\mu.$$

Problem* 11.11. *Let $f \in L^p_{loc}(\mathbb{R}^n)$, $1 \le p < \infty$, then for a.e. $x \in \mathbb{R}^n$ we have*

$$\lim_{r\downarrow 0} \frac{1}{|B_r(x)|} \int_{B_r(x)} |f(y) - f(x)|^p d^n y = 0.$$

The same conclusion hols if the balls are replaced by sets $A_j(x)$ which shrink nicely to x.

Problem 11.12. *Show that the Cantor function is Hölder continuous $|f(x) - f(y)| \le |x-y|^\alpha$ with exponent $\alpha = \log_3(2)$. (Hint: Show that if a bijection $g : [0,1] \to [0,1]$ satisfies a Hölder estimate $|g(x) - g(y)| \le M|x-y|^\alpha$, then so does $K(g)$: $|K(g)(x) - K(g)(y)| \le \frac{3^\alpha}{2} M |x-y|^\alpha$.)*

11.3. Complex measures

Let (X, Σ) be some measurable space. A map $\nu : \Sigma \to \mathbb{C}$ is called a **complex measure** if it is σ-additive:

$$\nu(\bigcupdot_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \nu(A_n), \quad A_n \in \Sigma. \tag{11.15}$$

Choosing $A_n = \emptyset$ for all n in (11.15) shows $\nu(\emptyset) = 0$.

Note that a positive measure is a complex measure only if it is finite (the value ∞ is not allowed for complex measures). Moreover, the definition implies that the sum is independent of the order of the sets A_n, that is, it converges unconditionally and thus absolutely by the Riemann series theorem.

Example 11.8. Let μ be a positive measure. For every $f \in L^1(X, d\mu)$ we have that $f\,d\mu$ is a complex measure (compare the proof of Lemma 9.3 and use dominated in place of monotone convergence). In fact, we will show that every complex measure is of this form. $\diamond$

Example 11.9. Let ν_1, ν_2 be two complex measures and α_1, α_2 two complex numbers. Then $\alpha_1\nu_1 + \alpha_2\nu_2$ is again a complex measure. Clearly we can extend this to any finite linear combination of complex measures. $\diamond$

When dealing with complex functions f an important object is the positive function $|f|$. Given a complex measure ν it seems natural to consider the set function $A \mapsto |\nu(A)|$. However, considering the simple example $d\nu(x) := \operatorname{sign}(x)dx$ on $X := [-1, 1]$ one sees that this set function is not additive and this simple approach does not provide a positive measure associated with ν. However, using $|\nu(A \cap [-1, 0))| + |\nu(A \cap [0, 1])|$ we do get a positive measure. Motivated by this we introduce the **total variation** of a measure defined as

$$|\nu|(A) := \sup\Big\{ \sum_{k=1}^{n} |\nu(A_k)| \Big| A_k \in \Sigma \text{ disjoint}, \; A = \dot{\bigcup_{k=1}^{n}} A_k \Big\}. \tag{11.16}$$

Note that by construction we have

$$|\nu(A)| \leq |\nu|(A). \tag{11.17}$$

Moreover, the total variation is monotone $|\nu|(A) \leq |\nu|(B)$ if $A \subseteq B$ and for a positive measure μ we have of course $|\mu|(A) = \mu(A)$.

Theorem 11.13. *The total variation $|\nu|$ of a complex measure ν is a finite positive measure.*

Proof. We begin by showing that $|\nu|$ is a positive measure. We need to show $|\nu|(A) = \sum_{n=1}^{\infty} |\nu|(A_n)$ for any partition of A into disjoint sets A_n. If $|\nu|(A_n) = \infty$ for some n it is not hard to see that $|\nu|(A) = \infty$ and hence we can assume $|\nu|(A_n) < \infty$ for all n.

Let $\varepsilon > 0$ be fixed and for each A_n choose a disjoint partition $B_{n,k}$ of A_n such that

$$|\nu|(A_n) \leq \sum_{k=1}^{m_n} |\nu(B_{n,k})| + \frac{\varepsilon}{2^n}.$$

Then

$$\sum_{n=1}^{N} |\nu|(A_n) \le \sum_{n=1}^{N} \sum_{k=1}^{m_n} |\nu(B_{n,k})| + \varepsilon \le |\nu|(\bigcup_{n=1}^{N} A_n) + \varepsilon \le |\nu|(A) + \varepsilon$$

since $\bigcup_{n=1}^{N} \bigcup_{k=1}^{m_n} B_{n,k} = \bigcup_{n=1}^{N} A_n$. As ε was arbitrary this shows $|\nu|(A) \ge \sum_{n=1}^{\infty} |\nu|(A_n)$.

Conversely, given a finite partition B_k of A, then

$$\begin{aligned}\sum_{k=1}^{m} |\nu(B_k)| &= \sum_{k=1}^{m} \Big| \sum_{n=1}^{\infty} \nu(B_k \cap A_n) \Big| \le \sum_{k=1}^{m} \sum_{n=1}^{\infty} |\nu(B_k \cap A_n)| \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^{m} |\nu(B_k \cap A_n)| \le \sum_{n=1}^{\infty} |\nu|(A_n).\end{aligned}$$

Taking the supremum over all partitions B_k shows $|\nu|(A) \le \sum_{n=1}^{\infty} |\nu|(A_n)$.

Hence $|\nu|$ is a positive measure and it remains to show that it is finite. Splitting ν into its real and imaginary part, it is no restriction to assume that ν is real-valued since $|\nu|(A) \le |\mathrm{Re}(\nu)|(A) + |\mathrm{Im}(\nu)|(A)$.

The idea is as follows: Suppose we can split any given set A with $|\nu|(A) = \infty$ into two subsets B and $A \setminus B$ such that $|\nu(B)| \ge 1$ and $|\nu|(A \setminus B) = \infty$. Then we can construct a sequence B_n of disjoint sets with $|\nu(B_n)| \ge 1$ for which $\sum_{n=1}^{\infty} \nu(B_n)$ diverges (the terms of a convergent series must converge to zero). But σ-additivity requires that the sum converges to $\nu(\bigcup_n B_n)$, a contradiction.

It remains to show existence of this splitting. Let A with $|\nu|(A) = \infty$ be given. Then there is a partition of disjoint sets A_j such that

$$\sum_{j=1}^{n} |\nu(A_j)| \ge 2 + |\nu(A)|.$$

Now let $A_+ := \bigcup\{A_j | \nu(A_j) \ge 0\}$ and $A_- := A \setminus A_+ = \bigcup\{A_j | \nu(A_j) < 0\}$. Then the above inequality reads $\nu(A_+) + |\nu(A_-)| \ge 2 + |\nu(A_+) - |\nu(A_-)||$ implying (show this) that for both of them we have $|\nu(A_\pm)| \ge 1$ and by $|\nu|(A) = |\nu|(A_+) + |\nu|(A_-)$ either A_+ or A_- must have infinite $|\nu|$ measure. □

Note that this implies that every complex measure ν can be written as a linear combination of four positive measures. In fact, first we can split ν into its real and imaginary part

$$\nu = \nu_r + \mathrm{i}\nu_i, \qquad \nu_r(A) := \mathrm{Re}(\nu(A)), \quad \nu_i(A) := \mathrm{Im}(\nu(A)). \tag{11.18}$$

Second we can split every real (also called **signed**) measure according to

$$\nu = \nu_+ - \nu_-, \qquad \nu_\pm(A) := \frac{|\nu|(A) \pm \nu(A)}{2}. \tag{11.19}$$

By (11.17) both ν_- and ν_+ are positive measures. This splitting is also known as **Jordan decomposition** of a signed measure. In summary, we can split every complex measure ν into four positive measures

$$\nu = \nu_{r,+} - \nu_{r,-} + \mathrm{i}(\nu_{i,+} - \nu_{i,-}) \tag{11.20}$$

which is also known as Jordan decomposition.

Of course such a decomposition of a signed measure is not unique (we can always add a positive measure to both parts), however, the Jordan decomposition is unique in the sense that it is the smallest possible decomposition.

Lemma 11.14. *Let ν be a complex measure and μ a positive measure satisfying $|\nu(A)| \le \mu(A)$ for all measurable sets A. Then $|\nu| \le \mu$. (Here $|\nu| \le \mu$ has to be understood as $|\nu|(A) \le \mu(A)$ for every measurable set A.)*

Furthermore, let ν be a signed measure and $\nu = \tilde\nu_+ - \tilde\nu_-$ a decomposition into positive measures. Then $\tilde\nu_\pm \ge \nu_\pm$, where $\nu_\pm$ is the Jordan decomposition.

Proof. It suffices to prove the first part since the second is a special case. But for every measurable set A and a corresponding finite partition A_k we have $\sum_k |\nu(A_k)| \le \sum \mu(A_k) = \mu(A)$ implying $|\nu|(A) \le \mu(A)$. □

Moreover, we also have:

Theorem 11.15. *The set of all complex measures $\mathcal{M}(X)$ together with the norm $\|\nu\| := |\nu|(X)$ is a Banach space.*

Proof. Clearly $\mathcal{M}(X)$ is a vector space and it is straightforward to check that $|\nu|(X)$ is a norm. Hence it remains to show that every Cauchy sequence ν_k has a limit.

First of all, by $|\nu_k(A)-\nu_j(A)| = |(\nu_k-\nu_j)(A)| \le |\nu_k-\nu_j|(A) \le \|\nu_k-\nu_j\|$, we see that $\nu_k(A)$ is a Cauchy sequence in $\mathbb{C}$ for every $A \in \Sigma$ and we can define

$$\nu(A) := \lim_{k\to\infty} \nu_k(A).$$

Moreover, $C_j := \sup_{k\ge j} \|\nu_k - \nu_j\| \to 0$ as $j \to \infty$ and we have

$$|\nu_j(A) - \nu(A)| \le C_j.$$

Next we show that ν satisfies (11.15). Let A_m be given disjoint sets and set $\tilde A_n := \bigcup_{m=1}^n A_m$, $A := \bigcup_{m=1}^\infty A_m$. Since we can interchange limits with

finite sums, (11.15) holds for finitely many sets. Hence it remains to show $\nu(\tilde{A}_n) \to \nu(A)$. This follows from

$$\begin{aligned}|\nu(\tilde{A}_n) - \nu(A)| &\leq |\nu(\tilde{A}_n) - \nu_k(\tilde{A}_n)| + |\nu_k(\tilde{A}_n) - \nu_k(A)| + |\nu_k(A) - \nu(A)| \\ &\leq 2C_k + |\nu_k(\tilde{A}_n) - \nu_k(A)|.\end{aligned}$$

Finally, $\nu_k \to \nu$ since $|\nu_k(A) - \nu(A)| \leq C_k$ implies $\|\nu_k - \nu\| \leq 4C_k$ (Problem 11.17). □

If μ is a positive and ν a complex measure we say that ν is absolutely continuous with respect to μ if $\mu(A) = 0$ implies $\nu(A) = 0$. We say that ν is singular with respect to μ if $|\nu|$ is, that is, there is a measurable set N such that $\mu(N) = 0$ and $|\nu|(X \setminus N) = 0$.

Lemma 11.16. *If μ is a positive and ν a complex measure then $\nu \ll \mu$ if and only if $|\nu| \ll \mu$. Similarly, $\nu \ll \mu$ if and only if this holds for all four measures in the Jordan decomposition.*

Proof. If $\nu \ll \mu$, then $\mu(A) = 0$ implies $\mu(B) = 0$ for every $B \subseteq A$ and hence $|\nu|(A) = 0$. Conversely, if $|\nu| \ll \mu$, then $\mu(A) = 0$ implies $|\nu(A)| \leq |\nu|(A) = 0$. The second claim now is immediate from Problem 11.17. □

Now we can prove the complex version of the Radon–Nikodym theorem:

Theorem 11.17 (Complex Radon–Nikodym)**.** *Let (X, Σ) be a measurable space, μ a positive σ-finite measure and ν a complex measure. Then there exists a function $f \in L^1(X, d\mu)$ and a set N of μ measure zero, such that*

$$\nu(A) = \nu(A \cap N) + \int_A f \, d\mu. \tag{11.21}$$

*The function f is determined uniquely a.e. with respect to μ and is called the **Radon–Nikodym derivative** $\frac{d\nu}{d\mu}$ of ν with respect to μ.*

In particular, ν can be uniquely decomposed as $\nu = \nu_{sing} + \nu_{ac}$, where $\nu_{sing}(A) := \nu(A \cap N)$ is singular and $d\nu_{ac} := f \, d\mu$ is absolutely continuous with respect to μ.

Proof. We start with the case where ν is a signed measure. Let $\nu = \nu_+ - \nu_-$ be its Jordan decomposition. Then by Theorem 11.1 there are sets $N_\pm$ and functions $f_\pm$ such that $\nu_\pm(A) = \nu_\pm(A \cap N_\pm) + \int_A f_\pm d\mu$. Since $\nu_\pm$ are finite measures we must have $\int_X f_\pm d\mu \leq \nu_\pm(X)$ and hence $f_\pm \in L^1(X, d\mu)$. Moreover, since $N := N_- \cup N_+$ has μ measure zero the remark after Theorem 11.1 implies $\nu_\pm(A) = \nu_\pm(A \cap N) + \int_A f_\pm d\mu$ and hence $\nu(A) = \nu(A \cap N) + \int_A f \, d\mu$ where $f := f_+ - f_- \in L^1(X, d\mu)$.

If ν is complex we can split it into real and imaginary part and use the same reasoning to reduce it to the singed case. If ν is absolutely continuous

we have $|\nu|(N) = 0$ and hence $\nu(A \cap N) = 0$. Uniqueness of the decomposition (and hence of f) follows literally as in the proof of Theorem 11.3. □

If ν is absolutely continuous with respect to μ the total variation of $d\nu = f\, d\mu$ is just $d|\nu| = |f| d\mu$:

Lemma 11.18. *Let $d\nu = d\nu_{sing} + f\, d\mu$ be the Lebesgue decomposition of a complex measure ν with respect to a positive σ-finite measure μ. Then*

$$|\nu|(A) = |\nu_{sing}|(A) + \int_A |f| d\mu. \tag{11.22}$$

Proof. We first show $|\nu| = |\nu_{sing}| + |\nu_{ac}|$. Let A be given and let A_k be a partition of A as in (11.16). By the definition of the total variation we can find a partition $A_{sing,k}$ of $A \cap N$ such that $\sum_k |\nu(A_{sing,k})| \geq |\nu_{sing}|(A) - \frac{\varepsilon}{2}$ for arbitrary $\varepsilon > 0$ (note that $\nu_{sing}(A_{sing,k}) = \nu(A_{sing,k})$ as well as $\nu_{sing}(A \cap N') = 0$). Similarly, there is such a partition $A_{ac,k}$ of $A \cap N'$ such that $\sum_k |\nu(A_{ac,k})| \geq |\nu_{ac}|(A) - \frac{\varepsilon}{2}$. Then combing both partitions into a partition A_k for A we obtain $|\nu|(A) \geq \sum_k |\nu(A_k)| \geq |\nu_{sing}|(A) + |\nu_{ac}|(A) - \varepsilon$. Since $\varepsilon > 0$ is arbitrary we conclude $|\nu|(A) \geq |\nu_{sing}|(A) + |\nu_{ac}|(A)$ and as the converse inequality is trivial the first claim follows.

It remains to show $d|\nu_{ac}| = |f|\, d\mu$. Given a partition $A = \bigcup\!\!\!\cdot_n A_n$ we have

$$\sum_n |\nu(A_n)| = \sum_n \Big| \int_{A_n} f\, d\mu \Big| \leq \sum_n \int_{A_n} |f| d\mu = \int_A |f| d\mu.$$

Hence $|\nu|(A) \leq \int_A |f| d\mu$. To show the converse define

$$A_k^n := \{x \in A | \frac{k-1}{n} < \frac{\arg(f(x)) + \pi}{2\pi} \leq \frac{k}{n}\}, \quad 1 \leq k \leq n.$$

Then the simple functions

$$s_n(x) := \sum_{k=1}^n \mathrm{e}^{-2\pi \mathrm{i} \frac{k-1}{n} + \mathrm{i}\pi} \chi_{A_k^n}(x)$$

converge to $\operatorname{sign}(f(x)^*)$ for every $x \in A$ and hence

$$\lim_{n\to\infty} \int_A s_n f\, d\mu = \int_A |f| d\mu$$

by dominated convergence. Moreover,

$$\Big| \int_A s_n f\, d\mu \Big| \leq \sum_{k=1}^n \Big| \int_{A_k^n} s_n f\, d\mu \Big| = \sum_{k=1}^n |\nu(A_k^n)| \leq |\nu|(A)$$

shows $\int_A |f| d\mu \leq |\nu|(A)$. □

As a consequence we obtain (Problem 11.13):

Corollary 11.19 (polar decomposition)**.** *If ν is a complex measure, then $d\nu = h\, d|\nu|$, where $|h| = 1$.*

If ν is a signed measure, then h is real-valued and we obtain:

Corollary 11.20. *If ν is a signed measure, then $d\nu = h\, d|\nu|$, where $h^2 = 1$. In particular, $d\nu_\pm = \chi_{X_\pm} d|\nu|$, where $X_\pm := h^{-1}(\pm 1)$.*

The decomposition $X = X_+ \dot\cup X_-$ from the previous corollary is known as **Hahn decomposition** and it is characterized by the property that $\pm\nu(A) \geq 0$ if $A \subseteq X_\pm$. This decomposition is not unique since we can shift sets of $|\nu|$ measure zero from one to the other.

We also briefly mention that the concept of regularity generalizes in a straightforward manner to complex Borel measures. If X is a topological space with its Borel σ-algebra we call ν **(outer/inner) regular** if $|\nu|$ is. It is not hard to see (Problem 11.18):

Lemma 11.21. *A complex measure is regular if and only if all measures in its Jordan decomposition are.*

The subspace of regular Borel measures will be denoted by $\mathcal{M}_{reg}(X)$. Note that it is closed and hence again a Banach space (Problem 11.19).

Clearly we can use Corollary 11.19 to define the integral of a bounded function f with respect to a complex measure $d\nu = h\, d|\nu|$ as

$$\int f\, d\nu := \int f h\, d|\nu|. \tag{11.23}$$

In fact, it suffices to assume that f is integrable with respect to $d|\nu|$ and we obtain

$$\left|\int f\, d\nu\right| \leq \int |f| d|\nu|. \tag{11.24}$$

For bounded functions this implies

$$\Big|\int_A f\, d\nu\Big| \leq \|f\|_\infty |\nu|(A). \tag{11.25}$$

Finally, there is an interesting equivalent definition of absolute continuity:

Lemma 11.22. *If μ is a positive and ν a complex measure then $\nu \ll \mu$ if and only if for every $\varepsilon > 0$ there is a corresponding $\delta > 0$ such that*

$$\mu(A) < \delta \quad \Rightarrow \quad |\nu(A)| < \varepsilon, \qquad \forall A \in \Sigma. \tag{11.26}$$

Proof. Suppose $\nu \ll \mu$ implying that it is of the from $d\nu = f\, d\mu$. Let $X_n = \{x \in X \,|\, |f(x)| \leq n\}$ and note that $|\nu|(X \setminus X_n) \to 0$ since $X_n \nearrow X$

and $|\nu|(X) < \infty$. Given $\varepsilon > 0$ we can choose n such that $|\nu|(X \setminus X_n) \leq \frac{\varepsilon}{2}$ and $\delta = \frac{\varepsilon}{2n}$. Then, if $\mu(A) < \delta$ we have

$$|\nu(A)| \leq |\nu|(A \cap X_n) + |\nu|(X \setminus X_n) \leq n\,\mu(A) + \frac{\varepsilon}{2} < \varepsilon.$$

The converse direction is obvious. □

It is important to emphasize that the fact that $|\nu|(X) < \infty$ is crucial for the above lemma to hold. In fact, it can fail for positive measures as the simple counterexample $d\nu(\lambda) = \lambda^2 d\lambda$ on $\mathbb{R}$ shows.

Problem* 11.13. *Prove Corollary 11.19. (Hint: Use the complex Radon–Nikodym theorem to get existence of h. Then show that $1 - |h|$ vanishes a.e.)*

Problem 11.14 (Markov inequality)**.** *Let ν be a complex and μ a positive measure. If f denotes the Radon–Nikodym derivative of ν with respect to μ, then show that*

$$\mu(\{x \in A \,|\, |f(x)| \geq \alpha\}) \leq \frac{|\nu|(A)}{\alpha}.$$

Problem 11.15. *Let ν be a complex and μ a positive measure and suppose $|\nu(A)| \leq C\mu(A)$ for all $A \in \Sigma$. Then $d\nu = f\,d\mu$ with $\|f\|_\infty \leq C$. (Hint: First show $|\nu|(A) \leq C\mu(A)$ and then use Problem 11.3.)*

Problem 11.16. *Let ν be a signed measure and $\nu_\pm$ its Jordan decomposition. Show*

$$\nu_+(A) = \max_{B \in \Sigma, B \subseteq A} \nu(B), \qquad \nu_-(A) = -\min_{B \in \Sigma, B \subseteq A} \nu(B).$$

Problem* 11.17. *Let ν be a complex measure with Jordan decomposition* (11.20)*. Show the estimate*

$$\frac{1}{\sqrt{2}}\nu_s(A) \leq |\nu|(A) \leq \nu_s(A), \quad \nu_s = \nu_{r,+} + \nu_{r,-} + \nu_{i,+} + \nu_{i,-}.$$

Show that $|\nu(A)| \leq C$ for all measurable sets A implies $\|\nu\| \leq 4C$.

Problem* 11.18. *Show Lemma 11.21. (Hint: Problems 8.22 and 11.17.)*

Problem* 11.19. *Let X be a topological space. Show that $\mathcal{M}_{reg}(X) \subseteq \mathcal{M}(X)$ is a closed subspace.*

Problem 11.20. *Define the convolution of two complex Borel measures μ and ν on $\mathbb{R}^n$ via*

$$(\mu * \nu)(A) := \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \chi_A(x+y) d\mu(x) d\nu(y).$$

Note $|\mu * \nu|(\mathbb{R}^n) \leq |\mu|(\mathbb{R}^n)|\nu|(\mathbb{R}^n)$. Show that this implies

$$\int_{\mathbb{R}^n} h(x) d(\nu * \nu)(x) = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} h(x+y) d\mu(x) d\nu(y)$$

for any bounded measurable function h. Conclude that it coincides with our previous definition in case μ and ν are absolutely continuous with respect to Lebesgue measure.

11.4. Hausdorff measure

Throughout this section we will assume that (X, d) is a metric space. Recall that the **diameter** of a subset $A \subseteq X$ is defined by $\operatorname{diam}(A) := \sup_{x,y\in A} d(x, y)$ with the convention that $\operatorname{diam}(\emptyset) = 0$. A cover $\{A_j\}$ of A is called a δ**-cover** if it is countable and if $\operatorname{diam}(A_j) \le \delta$ for all j.

For $A \subseteq X$ and $\alpha \ge 0$, $\delta > 0$ we define

$$h^{\alpha,*}_\delta(A) := \inf\Big\{ \sum\nolimits_j \operatorname{diam}(A_j)^\alpha \Big| \{A_j\} \text{ is a } \delta\text{-cover of } A\Big\} \in [0, \infty]. \quad (11.27)$$

which is an outer measure by Lemma 8.8. In the case $\alpha = 0$ and $A = \emptyset$ we also regard the empty cover as a valid cover such that $h^{0,*}_\delta(\emptyset) = 0$. As δ decreases the number of admissible covers decreases and hence $h^\alpha_\delta(A)$ increases as a function of δ. Thus the limit

$$h^{\alpha,*}(A) := \lim_{\delta\downarrow 0} h^{\alpha,*}_\delta(A) = \sup_{\delta>0} h^{\alpha,*}_\delta(A) \quad (11.28)$$

exists. Moreover, it is not hard to see that it is again an outer measure (Problem 11.21) and by Theorem 8.9 we get a measure. To show that the σ-algebra from Theorem 8.9 contains all Borel sets it suffices to show that μ^* is a metric outer measure (cf. Lemma 8.11).

Now if A_1, A_2 with $\operatorname{dist}(A_1, A_2) > 0$ are given and $\delta < \operatorname{dist}(A_1, A_2)$ then every set from a cover for $A_1 \cup A_1$ can have nonempty intersection with at most one of both sets. Consequently $h^{*,\alpha}_\delta(A_1 \cup A_2) = h^{*,\alpha}_\delta(A_1) + h^{*,\alpha}_\delta(A_2)$ for $\delta < \operatorname{dist}(A_1, A_2)$ implying $h^{*,\alpha}(A_1 \cup A_2) = h^{*,\alpha}(A_1) + h^{*,\alpha}(A_2)$. Hence $h^{\alpha,*}$ is a metric outer measure and the resulting measure h^α on the Borel σ-algebra is called the α-dimensional **Hausdorff measure**. Note that if X is a vector space with a translation invariant metric, then the diameter of a set is translation invariant and so will be h^α.

Example 11.10. For example, consider the case $\alpha = 0$. Suppose $A = \{x, y\}$ consists of two points. Then $h^0_\delta(A) = 1$ for $\delta \ge d(x, y)$ and $h^0_\delta(A) = 2$ for $\delta < |x - y|$. In particular, $h^0(A) = 2$. Similarly, it is not hard to see (show this) that $h^0(A)$ is just the number of points in A, that is, h^0 is the counting measure on X. ⋄

Example 11.11. At the other extreme, if $X := \mathbb{R}^n$, we have $h^n(A) = c_n|A|$, where $|A|$ denotes the Lebesgue measure of A. In fact, since the square $(0, 1]$ has $\operatorname{diam}((0, 1]) = \sqrt{n}$ and we can cover it by k^n squares of side length $\frac{1}{k}$, we see $h^n((0, 1]) \le n^{n/2}$. Thus h^n is a translation invariant Borel measure

which hence must be a multiple of Lebesgue measure. To identify c_n we will need the isodiametric inequality. ⋄

For the rest of this section we restrict ourselves to the case $X = \mathbb{R}^n$. Note that while in dimension more than one it is not true that a set of diameter d is contained in a ball of diameter d (a counter example is an equilateral triangle), we at least have the following:

Lemma 11.23 (Isodiametric inequality)**.** *For every Borel set $A \in \mathfrak{B}^n$ we have*

$$|A| \leq \frac{V_n}{2^n} \operatorname{diam}(A)^n.$$

In other words, a ball is the set with the largest volume when the diameter is kept fixed.

Proof. The trick is to transform A to a set with smaller diameter but same volume via Steiner symmetrization. To this end we build up A from slices obtained by keeping the first coordinate fixed: $A(y) = \{x \in \mathbb{R} | (x, y) \in A\}$. Now we build a new set $\tilde{A}$ by replacing $A(y)$ with a symmetric interval with the same measure, that is, $\tilde{A} = \{(x, y) | \, |x| \leq |A(y)|/2\}$. Note that by Theorem 9.10 $\tilde{A}$ is measurable with $|A| = |\tilde{A}|$. Hence the same is true for $\tilde{A} = \{(x, y) | \, |x| \leq |A(y)|/2\} \setminus \{(0, y) | \, A(y) = \emptyset\}$ if we look at the complete Lebesgue measure, since the set we subtract is contained in a set of measure zero.

Moreover, if $\bar{I}$, $\bar{J}$ are closed intervals, then $\sup_{x_1 \in \bar{I}, x_2 \in \bar{J}} |x_2 - x_1| \geq \frac{|\bar{I}|}{2} + \frac{|\bar{J}|}{2}$ (without loss both intervals are compact and $i \leq j$, where i, j are the midpoints of $\bar{I}$, $\bar{J}$; then the sup is at least $(j + \frac{|\bar{J}|}{2}) - (i - \frac{|\bar{I}|}{2})$). If I, J are Borel sets in $\mathfrak{B}^1$ and $\bar{I}$, $\bar{J}$ are their respective closed convex hulls, then $\sup_{x_1 \in I, x_2 \in J} |x_2 - x_1| = \sup_{x_1 \in \bar{I}, x_2 \in \bar{J}} |x_2 - x_1| \geq \frac{|\bar{I}|}{2} + \frac{|\bar{J}|}{2} \geq \frac{|I|}{2} + \frac{|J|}{2}$. Thus for $(\tilde{x}_1, y_1), (\tilde{x}_2, y_2) \in \tilde{A}$ we can find $(x_1, y_1), (x_2, y_2) \in A$ with $|x_1 - x_2| \geq |\tilde{x}_1 - \tilde{x}_2|$ implying $\operatorname{diam}(\tilde{A}) \leq \operatorname{diam}(A)$.

In addition, if A is symmetric with respect to $x_j \mapsto -x_j$ for some $2 \leq j \leq n$, then so is $\tilde{A}$.

Now repeat this procedure with the remaining coordinate directions, to obtain a set $\tilde{A}$ which is symmetric with respect to reflection $x \mapsto -x$ and satisfies $|A| = |\tilde{A}|$, $\operatorname{diam}(\tilde{A}) \leq \operatorname{diam}(A)$. By symmetry $\tilde{A}$ is contained in a ball of diameter $\operatorname{diam}(\tilde{A})$ and the claim follows. □

Lemma 11.24. *For every Borel set $A \in \mathfrak{B}^n$ we have*

$$h^n(A) = \frac{2^n}{V_n} |A|.$$

Proof. Using (8.38) (for $\mathcal{C}$ choose the collection of all open balls of radius at most δ) one infers $h^n_\delta(A) \le \frac{2^n}{V_n}|A|$ implying $c_n \le 2^n/V_n$. The converse inequality $h^n_\delta(A) \ge \frac{2^n}{V_n}|A|$ follows from the isodiametric inequality. □

We have already noted that the Hausdorff measure is translation invariant. Similarly, it is also invariant under orthogonal transformations (since the diameter is). Moreover, using the fact that for $\lambda > 0$ the map $\lambda : x \mapsto \lambda x$ gives rise to a bijection between δ-covers and (δ/λ)-covers, we easily obtain the following scaling property of Hausdorff measures.

Lemma 11.25. *Let $\lambda > 0$, $d \in \mathbb{R}^n$, $O \in O(n)$ an orthogonal matrix, and A be a Borel set of $\mathbb{R}^n$. Then*

$$h^\alpha(\lambda OA + d) = \lambda^\alpha h^\alpha(A). \tag{11.29}$$

Moreover, Hausdorff measures also behave nicely under uniformly Hölder continuous maps.

Lemma 11.26. *Suppose $f : A \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \le c|x - y|^\gamma \quad \text{for all } x, y \in A. \tag{11.30}$$

Then

$$h^\alpha(f(A)) \le c^\alpha h^{\alpha\gamma}(A). \tag{11.31}$$

Proof. A simple consequence of the fact that for every δ-cover $\{A_j\}$ of a Borel set A, the set $\{f(A \cap A_j)\}$ is a $(c\delta^\gamma)$-cover for the Borel set $f(A)$. □

Now we are ready to define the Hausdorff dimension. First note that h^α_δ is non increasing with respect to α for $\delta < 1$ and hence the same is true for h^α. Moreover, for $\alpha \le \beta$ we have $\sum_j \operatorname{diam}(A_j)^\beta \le \delta^{\beta-\alpha} \sum_j \operatorname{diam}(A_j)^\alpha$ and hence

$$h^\beta_\delta(A) \le \delta^{\beta-\alpha}\, h^\alpha_\delta(A) \le \delta^{\beta-\alpha}\, h^\alpha(A). \tag{11.32}$$

Thus if $h^\alpha(A)$ is finite, then $h^\beta(A) = 0$ for every $\beta > \alpha$. Hence there must be one value of α where the Hausdorff measure of a set jumps from ∞ to 0. This value is called the **Hausdorff dimension**

$$\dim_H(A) = \inf\{\alpha | h^\alpha(A) = 0\} = \sup\{\alpha | h^\alpha(A) = \infty\}. \tag{11.33}$$

It is also not hard to see that we have $\dim_H(A) \le n$ (Problem 11.23).

The following observations are useful when computing Hausdorff dimensions. First the Hausdorff dimension is monotone, that is, for $A \subseteq B$ we have $\dim_H(A) \le \dim_H(B)$. Furthermore, if A_j is a (countable) sequence of Borel sets we have $\dim_H(\bigcup_j A_j) = \sup_j \dim_H(A_j)$ (show this).

Using Lemma 11.26 it is also straightforward to show

Lemma 11.27. *Suppose $f : A \to \mathbb{R}^n$ is uniformly Hölder continuous with exponent $\gamma > 0$, that is,*

$$|f(x) - f(y)| \leq c|x - y|^\gamma \quad \text{for all } x, y \in A, \tag{11.34}$$

then

$$\dim_H(f(A)) \leq \frac{1}{\gamma} \dim_H(A). \tag{11.35}$$

Similarly, if f is bi-Lipschitz, that is,

$$a|x - y| \leq |f(x) - f(y)| \leq b|x - y| \quad \text{for all } x, y \in A, \tag{11.36}$$

then

$$\dim_H(f(A)) = \dim_H(A). \tag{11.37}$$

Example 11.12. The Hausdorff dimension of the Cantor set (see Example 8.24) is

$$\dim_H(C) = \frac{\log(2)}{\log(3)}. \tag{11.38}$$

To see this let $\delta = 3^{-n}$. Using the δ-cover given by the intervals forming C_n used in the construction of C we see $h^\alpha_\delta(C) \leq (\frac{2}{3^\alpha})^n$. Hence for $\alpha = d = \log(2)/\log(3)$ we have $h^d_\delta(C) \leq 1$ implying $\dim_H(C) \leq d$.

The reverse inequality is a little harder. Let $\{A_j\}$ be a cover and $\delta < \frac{1}{3}$. It is clearly no restriction to assume that all V_j are open intervals. Moreover, finitely many of these sets cover C by compactness. Drop all others and fix j. Furthermore, increase each interval A_j by at most ε.

For V_j there is a k such that

$$\frac{1}{3^{k+1}} \leq |A_j| < \frac{1}{3^k}.$$

Since the distance of two intervals in C_k is at least 3^{-k} we can intersect at most one such interval. For $n \geq k$ we see that V_j intersects at most $2^{n-k} = 2^n(3^{-k})^d \leq 2^n 3^d |A_j|^d$ intervals of C_n.

Now choose n larger than all k (for all A_j). Since $\{A_j\}$ covers C, we must intersect all 2^n intervals in C_n. So we end up with

$$2^n \leq \sum_j 2^n 3^d |A_j|^d,$$

which together with our first estimate yields

$$\frac{1}{2} \leq h^d(C) \leq 1.$$

Observe that this result can also formally be derived from the scaling property of the Hausdorff measure by solving the identity

$$h^\alpha(C) = h^\alpha(C \cap [0, \tfrac{1}{3}]) + h^\alpha(C \cap [\tfrac{2}{3}, 1]) = 2\, h^\alpha(C \cap [0, \tfrac{1}{3}]))$$
$$= \frac{2}{3^\alpha} h^\alpha(3(C \cap [0, \tfrac{1}{3}])) = \frac{2}{3^\alpha} h^\alpha(C) \tag{11.39}$$

for α. However, this is possible only if we already know that $0 < h^\alpha(C) < \infty$ for some α. ◇

Problem* 11.21. *Suppose $\{\mu^*_\alpha\}_\alpha$ is a family of outer measures on X. Then $\mu^* = \sup_\alpha \mu^*_\alpha$ is again an outer measure.*

Problem 11.22. *Let $L = [0,1] \times \{0\} \subseteq \mathbb{R}^2$. Show that $h^1(L) = 1$.*

Problem* 11.23. *Show that $\dim_H(U) \leq n$ for every $U \subseteq \mathbb{R}^n$.*

11.5. Infinite product measures

In Section 9.2 we have dealt with finite products of measures. However, in some situations even infnite products are of interest. For example, in probability theory one describes a single random experiment is described by a probability measure and performing n independent trials is modeled by taking the n-fold product. If one is interested in the behavior of certain quantities in the limit as $n \to \infty$ one is naturally lead to an infinite product.

Hence our goal is to to define a probability measure on the product space $\mathbb{R}^\mathbb{N} = \bigtimes_\mathbb{N} \mathbb{R}$. We can regard $\mathbb{R}^\mathbb{N}$ as the set of all sequences $x = (x_j)_{j\in\mathbb{N}}$. A **cylinder set** in $\mathbb{R}^\mathbb{N}$ is a set of the form $A \times \mathbb{R}^\mathbb{N} \subseteq \mathbb{R}^N$ with $A \subseteq \mathbb{R}^n$ for some n. We equip $\mathbb{R}^\mathbb{N}$ with the product topology, that is, the topology generated by open cylinder sets with A open (which are a base for the product topology since they are closed under intersections — note that the cylinder sets are precisely the finite intersections of preimages of projections). Then the Borel σ-algebra $\mathfrak{B}^\mathbb{N}$ on $\mathbb{R}^\mathbb{N}$ is the σ-algebra generated by cylinder sets with $A \in \mathfrak{B}^n$.

Now suppose we have probability measures μ_n on $(\mathbb{R}^n, \mathfrak{B}^n)$ which are **consistent** in the sense that

$$\mu_{n+1}(A \times \mathbb{R}) = \mu_n(A), \qquad A \in \mathfrak{B}^n. \tag{11.40}$$

Example 11.13. The prototypical example would be the case where μ is a probability measure on $(\mathbb{R}, \mathfrak{B})$ and $\mu_n = \mu \otimes \cdots \otimes \mu$ is the n-fold product. Slightly more general, one could even take probability measures ν_j on $(\mathbb{R}, \mathfrak{B})$ and consider $\mu_n = \nu_1 \otimes \cdots \otimes \nu_n$. ◇

Theorem 11.28 (Kolmogorov extension theorem)**.** *Suppose that we have a consistent family of probability measures $(\mathbb{R}^n, \mathfrak{B}^n, \mu_n)$, $n \in \mathbb{N}$. Then there exists a unique probability measure μ on $(\mathbb{R}^\mathbb{N}, \mathfrak{B}^\mathbb{N})$ such that $\mu(A \times \mathbb{R}^\mathbb{N}) = \mu_n(A)$ for all $A \in \mathfrak{B}^n$.*

Proof. Consider the algebra $\mathcal{A}$ of all Borel cylinder sets which generates $\mathfrak{B}^{\mathbb{N}}$ as noted above. Then $\mu(A \times \mathbb{R}^{\mathbb{N}}) = \mu_n(A)$ for $A \times \mathbb{R}^{\mathbb{N}} \in \mathcal{A}$ defines an additive set function on $\mathcal{A}$. Indeed, by our consistency assumption different representations of a cylinder set will give the same value and (finite) additivity follows from additivity of μ_n. Hence it remains to verify that μ is a premeasure such that we can apply the extension results from Section 8.3.

Now in order to show σ-additivity it suffices to show continuity from above, that is, for given sets $A_n \in \mathcal{A}$ with $A_n \searrow \emptyset$ we have $\mu(A_n) \searrow 0$. Suppose to the contrary that $\mu(A_n) \searrow \varepsilon > 0$. Moreover, by repeating sets in the sequence A_n if necessary, we can assume without loss of generality that $A_n = \tilde{A}_n \times \mathbb{R}^{\mathbb{N}}$ with $\tilde{A}_n \subseteq \mathbb{R}^n$. Next, since μ_n is inner regular, we can find a compact set $\tilde{K}_n \subseteq \tilde{A}_n$ such that $\mu_n(\tilde{K}_n) \geq \frac{\varepsilon}{2}$. Furthermore, since A_n is decreasing we can arrange $K_n = \tilde{K}_n \times \mathbb{R}^{\mathbb{N}}$ to be decreasing as well: $K_n \searrow \emptyset$. However, by compactness of $\tilde{K}_n$ we can find a sequence with $x \in K_n$ for all n (Problem 11.24), a contradiction. □

Example 11.14. The simplest example for the use of this theorem is a discrete **random walk** in one dimension. So we suppose we have a fictitious particle confined to the lattice $\mathbb{Z}$ which starts at 0 and moves one step to the left or right depending on whether a fictitious coin gives head or tail (the imaginative reader might also see the price of a share here). Somewhat more formal, we take $\mu_1 = (1-p)\delta_{-1} + p\delta_1$ with $p \in [0,1]$ being the probability of moving right and $1-p$ the probability of moving left. Then the infinite product will give a probability measure for the sets of all paths $x \in \{-1,1\}^{\mathbb{N}}$ and one might try to answer questions like if the location of the particle at step n, $s_n = \sum_{j=1}^{n} x_j$ remains bounded for all $n \in \mathbb{N}$, etc. ⋄

Example 11.15. Another classical example is the **Anderson model**. The discrete one-dimensional Schrödinger equation for a single electron in an external potential is given by the difference operator

$$(Hu)_n = u_{n+1} + u_{n-1} + q_n u_n, \qquad u \in \ell^2(\mathbb{Z}),$$

where the *potential* q_n is a bounded real-valued sequence. A simple model for an electron in a crystal (where the atoms are arranged in a periodic structure) is hence the case when q_n is periodic. But what happens if you introduce some random impurities (known as doping in the context of semiconductors)? This can be modeled by $q_n = q_n^0 + x_n(q_n^1 - q_n^0)$ where $x \in \{0,1\}^{\mathbb{N}}$ and we can take $\mu_1 = (1-p)\delta_0 + p\delta_1$ with $p \in [0,1]$ the probability of an impurity being present. ⋄

Problem* 11.24. *Suppose $K_n \subseteq \mathbb{R}^n$ is a sequence of nonempty compact sets which are nesting in the sense that $K_{n+1} \subseteq K_n \times \mathbb{R}$. Show that there is a sequence $x = (x_j)_{j\in\mathbb{N}}$ with $(x_1, \dots, x_n) \in K_n$ for all n. (Hint: Choose x_m*

by considering the projection of K_n onto the m'th coordinate and using the finite intersection property of compact sets.)

11.6. The Bochner integral

In this section we want to extend the Lebesgue integral to the case of functions with values in a normed space. This extension is known as **Bochner integral**. Since a normed space has no order we cannot use monotonicity and hence are restricted to finite values for the integral. Other than that, we only need some small adaptions.

Let (X, Σ, μ) be a measure space and Y a Banach space equipped with the Borel σ-algebra $\mathfrak{B}(Y)$. As in (9.1), a measurable function $s : X \to Y$ is called **simple** if its image is finite; that is, if

$$s = \sum_{j=1}^{p} \alpha_j \, \chi_{A_j}, \qquad \operatorname{Ran}(s) =: \{\alpha_j\}_{j=1}^{p}, \quad A_j := s^{-1}(\alpha_j) \in \Sigma. \tag{11.41}$$

Also the integral of a simple function can be defined as in (9.2) provided we ensure that it is finite. To this end we call s **integrable** if $\mu(A_j) < \infty$ for all j with $\alpha_j \neq 0$. Now, for an integrable simple function s as in (11.41) we define its **Bochner integral** as

$$\int_A s \, d\mu := \sum_{j=1}^{p} \alpha_j \, \mu(A_j \cap A). \tag{11.42}$$

As before we use the convention $0 \cdot \infty = 0$ (where the 0 on the left is the zero vector from Y).

Lemma 11.29. *The integral has the following properties:*

(i) $\int_A s \, d\mu = \int_X \chi_A \, s \, d\mu$.
(ii) $\int_{\bigcup_{n=1}^{\infty} A_n} s \, d\mu = \sum_{n=1}^{\infty} \int_{A_n} s \, d\mu$.
(iii) $\int_A \alpha \, s \, d\mu = \alpha \int_A s \, d\mu$, $\alpha \in \mathbb{C}$.
(iv) $\int_A (s + t) d\mu = \int_A s \, d\mu + \int_A t \, d\mu$.
(v) $\| \int_A s \, d\mu \| \leq \int_A \|s\| d\mu$.

Proof. The first four items follow literally as in Lemma 9.1. (v) follows from the triangle inequality. □

Now we extend the integral via approximation by simple functions. However, while a complex-valued measurable function can always be approximated by simple functions, this might no longer be true in our present setting. In fact, note that a sequence of simple functions involves only a

countable number of values from Y and since the limit must be in the closure of the span of these values, the range of f must be separable. Moreover, we also need to ensure finiteness of the integral.

If μ is finite, the latter requirement is easily satisfied by considering only bounded functions. Consequently we could equip the space of simple functions $S(X, \mu, Y)$ with the supremum norm $\|s\|_\infty = \sup_{x\in X} \|s(x)\|$ and use the fact that the integral is a bounded linear functional,

$$\| \int_A s \, d\mu\| \leq \mu(A)\|s\|_\infty, \tag{11.43}$$

to extend it to the completion of the simple functions, known as the regulated functions $R(X, \mu, Y)$. Hence the integrable functions will be the bounded functions which are uniform limits of simple functions. While this gives a theory suitable for many cases we want to do better and look at functions which are pointwise limits of simple functions.

Consequently, we call a function f **integrable** if there is a sequence of integrable simple functions s_n which converges pointwise to f such that

$$\int_X \|f - s_n\| d\mu \to 0. \tag{11.44}$$

That $f - s_n$ (and hence also $\|f - s_n\|$) is measurable will be shown in Lemma 11.31 below.

In this case item (v) from Lemma 11.29 ensures that $\int_X s_n d\mu$ is a Cauchy sequence and we can define the **Bochner integral** of f to be

$$\int_X f \, d\mu := \lim_{n\to\infty} \int_X s_n d\mu. \tag{11.45}$$

If there are two sequences of integrable simple functions as in the definition, we could combine them into one sequence (taking one as even and the other one as odd elements) to conclude that the limit of the first two sequences equals the limit of the third sequence. In other words, the definition of the integral is independent of the sequence chosen.

Lemma 11.30. *The integral is linear and Lemma 11.29 holds for integrable functions s, t.*

Proof. All items except for (ii) are immediate. (ii) is also immediate for finite unions. The general case will follow from the dominated convergence theorem to be shown below. □

Before we proceed, we try to shed some light on when a function is integrable.

Lemma 11.31. *A function $f : X \to Y$ is the pointwise limit of simple functions if and only if it is measurable and its range is separable. If its*

range is compact it is the uniform limit of simple functions. Moreover, the sequence s_n can be chosen such that $\|s_n(x)\| \le 2\|f(x)\|$ for every $x \in X$ and $\operatorname{Ran}(s_n) \subseteq \operatorname{Ran}(f) \cup \{0\}$.

Proof. Let $\{y_j\}_{j\in\mathbb{N}}$ be a dense set for the range. Note that the balls $B_{1/m}(y_j)$ will cover the range for every fixed $m \in \mathbb{N}$. Furthermore, we will augment this set by $y_0 = 0$ to make sure that any value less than $1/m$ is replaced by 0 (since otherwise one might destroy properties of f). By iteratively removing what has already been covered we get a disjoint cover $A_j^m \subseteq B_{1/m}(y_j)$ (some sets might be empty) such that $A_{n,m} := \bigcup_{j\le n} A_j^m = \bigcup_{j\le n} B_{1/m}(y_j)$. Now set $A_n := A_{n,n}$ and consider the simple function s_n defined as follows

$$s_n = \begin{cases} y_j, & \text{if } f(x) \in A_j^m \setminus \bigcup_{m<k\le n} A_k, \\ 0, & \text{else.} \end{cases}$$

That is, we first search for the largest $m \le n$ with $f(x) \in A_m$ and look for the unique j with $f(x) \in A_j^m$. If such an m exists we have $s_n(x) = y_j$ and otherwise $s_n(x) = 0$. By construction s_n is measurable and converges pointwise to f. Moreover, to see $\|s_n(x)\| \le 2\|f(x)\|$ we consider two cases. If $s_n(x) = 0$ then the claim is trivial. If $s_n(x) \ne 0$ then $f(x) \in A_j^m$ with $j > 0$ and hence $\|f(x)\| \ge 1/m$ implying

$$\|s_n(x)\| = \|y_j\| \le \|y_j - f(x)\| + \|f(x)\| \le \frac{1}{m} + \|f(x)\| \le 2\|f(x)\|.$$

If the range of f is compact, we can find a finite index N_m such that $A_m := A_{N_m,m}$ covers the whole range. Now our definition simplifies since the largest $m \le n$ with $f(x) \in A_m$ is always n. In particular, $\|s_n(x) - f(x)\| \le \frac{1}{n}$ for all x. □

Functions $f : X \to Y$ which are the pointwise limit of simple functions are also called **strongly measurable**. Since the limit of measurable functions is again measurable (Problem 11.25) and still has separable range, the limit of strongly measurable functions is strongly measurable. Moreover, our lemma also shows that continuous functions on separable spaces are strongly measurable (recall Lemma 8.17 and Problem B.14).

Now we are in the position to give a useful characterization of integrability.

Lemma 11.32 (Bochner)**.** *A function $f : X \to Y$ is integrable if and only if it is strongly measurable and $\|f(x)\|$ is integrable.*

Proof. We have already seen that an integrable function has the stated properties. Conversely, the sequence s_n from the previous lemma satisfies (11.44) by the dominated convergence theorem. □

Another useful observation is the fact that the integral behaves nicely under bounded linear transforms.

Theorem 11.33 (Hille)**.** *Let $A : \mathfrak{D}(A) \subseteq Y \to Z$ be closed. Suppose $f : X \to Y$ is integrable with $\operatorname{Ran}(f) \subseteq \mathfrak{D}(A)$ and Af is also integrable. Then*

$$A \int_X f \, d\mu = \int_X (Af) d\mu, \qquad f \in \mathfrak{D}(A). \tag{11.46}$$

If $A \in \mathscr{L}(Y, Z)$, then f integrable implies Af is integrable.

Proof. By assumption $(f, Af) : X \to Y \times Z$ is integrable and by the proof of Lemma 11.32 the sequence of simple functions can be chosen to have its range in the graph of A. In other words, there exists a sequence of simple functions (s_n, As_n) such that

$$\int_X s_n d\mu \to \int_X f \, d\mu, \qquad A \int_X s_n d\mu = \int_X As_n d\mu \to \int_X Af \, d\mu.$$

Hence the claim follows from closedness of A.

If A is bounded and s_n is a sequence of simple functions satisfying (11.44), then $t_n = As_n$ is a corresponding sequence of simple functions for Af since $\|t_n - Af\| \le \|A\| \|s_n - f\|$. This shows the last claim. □

Clearly the assumptions of this theorem are satisfied if $A \in \mathscr{L}(Y, Z)$, for example if we choose a bounded linear functional from Y^*.

Next, we note that the dominated convergence theorem holds for the Bochner integral.

Theorem 11.34. *Suppose f_n are integrable and $f_n \to f$ pointwise. Moreover suppose there is an integrable function g with $\|f_n(x)\| \le \|g(x)\|$. Then f is integrable and*

$$\lim_{n \to \infty} \int_X f_n d\mu = \int_X f \, d\mu \tag{11.47}$$

Proof. If $s_{n,m}(x) \to f_n(x)$ as $m \to \infty$ then $s_{n,n}(x) \to f(x)$ as $n \to \infty$ which together with $\|f(x)\| \le \|g(x)\|$ shows that f is integrable. Moreover, the usual dominated convergence theorem shows $\int_X \|f_n - f\| d\mu \to 0$ from which the claim follows. □

There is also yet another useful characterization of strong measurability. To this end we call a function $f : X \to Y$ **weakly measurable** if $\ell \circ f$ is measurable for every $\ell \in Y^*$.

Theorem 11.35 (Pettis)**.** *A function $f : X \to Y$ is strongly measurable if and only if it is weakly measurable and its range is separable.*

Proof. Since every measurable function is weakly measurable, Lemma 11.31 shows one direction. To see the converse direction let $\{y_k\}_{k\in\mathbb{N}} \subseteq f(X)$ be dense. Define $s_n : \overline{f(X)} \to \{y_1, \dots, y_n\}$ via $s_n(y) = y_k$, where $k = k_{y,n}$ is the smallest k such that $\|y - y_k\| = \min_{1\le j\le n} \|y - y_j\|$. By density we have $\lim_{n\to\infty} s_n(y) = y$ for every $y \in \overline{f(X)}$. Now set $f_n = s_n \circ f$ and note that $f_n \to f$ pointwise. Moreover, for $1 \le k \le n$ we have

$$\begin{aligned} f_n^{-1}(y_k) =& \{x \in X | \|f(x) - y_k\| = \min_{1\le j\le n} \|f(x) - y_j\|\} \cap \\ & \{x \in X | \|f(x) - y_l\| < \min_{1\le j\le n} \|f(x) - y_j\|,\ 1 \le l < k\} \end{aligned}$$

and f_n will be measurable once we show that $\|f - y\|$ is measurable for every $y \in \overline{f(Y)}$. To show this last claim choose (Problem 11.26) a countable set $\{y_k'\} \in Y^*$ of unit vectors such that $\|y\| = \sup_k y_k'(y)$. Then $\|f - y\| = \sup_k y'(f - y)$ from which the claim follows. □

We close with the remark that since the integral does not see null sets, one could also work with functions which satisfy the above requirements only away from null sets.

Finally, we briefly discuss the associated Lebesgue spaces. As in the complex-valued case we define the L^p norm by

$$\|f\|_p := \left(\int_X \|f\|^p \, d\mu\right)^{1/p}, \qquad 1 \le p, \tag{11.48}$$

and denote by $\mathcal{L}^p(X, d\mu, Y)$ the set of all strongly measurable functions for which $\|f\|_p$ is finite. Note that $\mathcal{L}^p(X, d\mu, Y)$ is a vector space, since $\|f + g\|^p \le 2^p \max(\|f\|, \|g\|)^p = 2^p \max(\|f\|^p, \|g\|^p) \le 2^p(\|f\|^p + \|g\|^p)$. Again Lemma 9.6 (which still holds in this situation) implies that we need to identify functions which are equal almost everywhere: Let

$$\mathcal{N}(X, d\mu, Y) := \{f \text{ strongly measurable} | f(x) = 0 \ \mu\text{-almost everywhere}\} \tag{11.49}$$

and consider the quotient space

$$L^p(X, d\mu, Y) := \mathcal{L}^p(X, d\mu, Y)/\mathcal{N}(X, d\mu, Y). \tag{11.50}$$

If $d\mu$ is the Lebesgue measure on $X \subseteq \mathbb{R}^n$, we simply write $L^p(X, Y)$. Observe that $\|f\|_p$ is well defined on $L^p(X, d\mu, Y)$ and hence we have a normed space (the triangle will be established below).

Lemma 11.36. *The integrable simple functions are dense in $L^p(X, d\mu, Y)$.*

Proof. Let $f \in L^p(X, d\mu, Y)$. By Lemma 11.31 there is a sequence of simple functions such that $s_n(x) \to f(x)$ pointwise and $\|s_n(x)\| \le 2\|f(x)\|$. In particular, $s_n \in L^p(X, d\mu, Y)$ and thus integrable since $\|s_n\|_p^p = \|s_n\|_1$.

Moreover, $\|f(x)-s_n(x)\|^p \le 3^p\|f(x)\|^p$ and thus $s_n \to f$ in $L^p(X,d\mu,Y)$ by dominated convergence. □

Similarly we define $L^\infty(X,d\mu,Y)$ together with the **essential supremum**

$$\|f\|_\infty := \inf\{C \,|\, \mu(\{x|\,\|f(x)\| > C\}) = 0\}. \tag{11.51}$$

From Hölder's inequality for complex-valued functions we immediately get the following generalized version:

Theorem 11.37 (Hölder's inequality)**.** *Let p and q be dual indices, $\frac{1}{p}+\frac{1}{q} = 1$, with $1 \le p \le \infty$. If*

(i) *$f \in L^p(X,d\mu,Y^*)$ and $g \in L^q(X,d\mu,Y)$, or*

(ii) *$f \in L^p(X,d\mu)$ and $g \in L^q(X,d\mu,Y)$, or*

(iii) *$f \in L^p(X,d\mu,Y)$ and $g \in L^q(X,d\mu,Y)$ and Y is a Banach algebra,*

then fg is integrable and

$$\|f\,g\|_1 \le \|f\|_p\|g\|_q. \tag{11.52}$$

As a consequence we get

Corollary 11.38 (Minkowski's inequality)**.** *Let $f,g \in L^p(X,d\mu,Y)$, $1 \le p \le \infty$. Then*

$$\|f+g\|_p \le \|f\|_p + \|g\|_p. \tag{11.53}$$

Proof. Since the cases $p = 1,\infty$ are straightforward, we only consider $1 < p < \infty$. Using $\|f(x)+g(x)\|^p \le \|f(x)\|\,\|f(x)+g(x)\|^{p-1} + \|g(x)\|\,\|f(x)+g(x)\|^{p-1}$ we obtain from Hölder's inequality (note $(p-1)q = p$)

$$\begin{aligned}\|f+g\|_p^p &\le \|f\|_p\||f+g|^{p-1}\|_q + \|g\|_p\||f+g|^{p-1}\|_q\\ &= (\|f\|_p + \|g\|_p)\|(f+g)\|_p^{p-1}.\end{aligned}$$

□

Moreover, literally the same proof as for the complex-valued case shows:

Theorem 11.39 (Riesz–Fischer)**.** *The space $L^p(X,d\mu,Y)$, $1 \le p \le \infty$, is a Banach space.*

Of course, if Y is a Hilbert space, then $L^2(X,d\mu,Y)$ is a Hilbert space with scalar product

$$\langle f,g\rangle = \int_X \langle f(x),g(x)\rangle_Y d\mu(x). \tag{11.54}$$

Problem* 11.25. *Let Y be a metric space equipped with the Borel σ-algebra. Show that the pointwise limit $f : X \to Y$ of measurable functions $f_n : X \to Y$ is measurable. (Hint: Show that for $U \subseteq Y$ open we have that $f^{-1}(U) = \bigcup_{m=1}^\infty \bigcup_{n=1}^\infty \bigcap_{k=n}^\infty f_k^{-1}(U_m)$, where $U_m := \{y \in U | d(y, Y\setminus U) > \frac{1}{m}\}$.)*

Problem* 11.26. *Let X be a separable Banach space. Show that there is a countable set $\ell_k \in X^*$ such that $\|x\| = \sup_k |\ell_k(x)|$ for all x.*

Problem 11.27. *Let $Y = C[a,b]$ and $f : [0,1] \to Y$ integrable. Compute the Bochner Integral $\int_0^1 f(x)dx$.*

11.7. Weak and vague convergence of measures

In this section X will be a metric space equipped with the Borel σ-algebra. We say that a sequence of finite Borel measures μ_n **converges weakly** to a finite Borel measure μ if

$$\int_X f d\mu_n \to \int_X f d\mu \tag{11.55}$$

for every $f \in C_b(X)$. Since by the Riesz representation theorem the set of (complex) measures is the dual of $C(X)$ (under appropriate assumptions on X), this is what would be denoted by weak-$*$ convergence in functional analysis. However, we will not need this connection here. Nevertheless we remark that the weak limit is unique. To see this let C be a nonempty closed set and consider

$$f_n(x) := \max(0, 1 - n \operatorname{dist}(x, C)). \tag{11.56}$$

Then $f \in C_b(X)$, $f_n \downarrow \chi_C$, and (dominated convergence)

$$\mu(C) = \lim_{n\to\infty} \int_X g_n d\mu \tag{11.57}$$

shows that μ is uniquely determined by the integral for continuous functions (recall Lemma 8.20 and its corollary). For later use observe that f_n is even Lipschitz continuous, $|f_n(x) - f_n(y)| \le n\, d(x,y)$ (cf. Lemma B.27).

Moreover, choosing $f \equiv 1$ shows

$$\mu(X) = \lim_{n\to\infty} \mu_n(X). \tag{11.58}$$

However, this might not be true for arbitrary sets in general. To see this look at the case $X = \mathbb{R}$ with $\mu_n = \delta_{1/n}$. Then μ_n converges weakly to $\mu = \delta_0$. But $\mu_n((0,1)) = 1 \not\to 0 = \mu((0,1))$ as well as $\mu_n([-1,0]) = 0 \not\to 1 = \mu([-1,0])$. So mass can appear/disappear at boundary points but this is the worst that can happen:

Theorem 11.40 (Portmanteau)**.** *Let X be a metric space and μ_n, μ finite Borel measures. The following are equivalent:*

(i) *μ_n converges weakly to μ.*
(ii) *$\int_X f\, d\mu_n \to \int_X f\, d\mu$ for every bounded Lipschitz continuous f.*
(iii) *$\limsup_n \mu_n(C) \le \mu(C)$ for all closed sets C and $\mu_n(X) \to \mu(X)$.*
(iv) *$\liminf_n \mu_n(O) \ge \mu(O)$ for all open sets O and $\mu_n(X) \to \mu(X)$.*

(v) $\mu_n(A) \to \mu(A)$ *for all Borel sets* A *with* $\mu(\partial A) = 0$.

(vi) $\int_X f\, d\mu_n \to \int_X f\, d\mu$ *for all bounded functions* f *which are continuous at* μ*-a.e.* $x \in X$.

Proof. (i) $\Rightarrow$ (ii) is trivial. (ii) $\Rightarrow$ (iii): Define f_n as in (11.56) and start by observing

$$\limsup_n \mu_n(C) = \limsup_n \int \chi_F \mu_n \le \limsup_n \int f_m \mu_n = \int f_m \mu.$$

Now taking $m \to \infty$ establishes the claim. Moreover, $\mu_n(X) \to \mu(X)$ follows choosing $f \equiv 1$. (iii) $\Leftrightarrow$ (iv): Just use $O = X \setminus C$ and $\mu_n(O) = \mu_n(X) - \mu_n(C)$. (iii) and (iv) $\Rightarrow$ (v): By $A^\circ \subseteq A \subseteq \overline{A}$ we have

$$\begin{aligned}\limsup_n \mu_n(A) &\le \limsup_n \mu_n(\overline{A}) \le \mu(\overline{A}) \\ &= \mu(A^\circ) \le \liminf_n \mu_n(A^\circ) \le \liminf_n \mu_n(A)\end{aligned}$$

provided $\mu(\partial A) = 0$. (v) $\Rightarrow$ (vi): By considering real and imaginary parts separately we can assume f to be real-valued. moreover, adding an appropriate constant we can even assume $0 \le f_n \le M$. Set $A_r = \{x \in X | f(x) > r\}$ and denote by D_f the set of discontinuities of f. Then $\partial A_r \subseteq D_f \cup \{x \in X | f(x) = r\}$. Now the first set has measure zero by assumption and the second set is countable by Problem 9.20. Thus the set of all r with $\mu(\partial A_r) > 0$ is countable and thus of Lebesgue measure zero. Then by Problem 9.20 (with $\phi(r) = 1$) and dominated convergence

$$\int_X f\, d\mu_n = \int_0^M \mu_n(A_r) dr \to \int_0^M \mu(A_r) dr = \int_X f\, d\mu$$

Finally, (vi) $\Rightarrow$ (i) is trivial. □

Next we want to extend our considerations to unbounded measures. In this case boundedness of f will not be sufficient to guarantee integrability and hence we will require f to have compact support. If $x \in X$ is such that $f(x) \neq 0$, then $f^{-1}(B_r(f(x))$ will be a relatively compact neighborhood of x whenever $0 < r < |f(x)|$. Hence $C_b(X)$ will not have sufficiently many functions with compact support unless we assume X to be locally compact, which we will do for the remainder of this section.

Let μ_n be a sequence of Borel measures on a locally compact metric space X. We will say that μ_n **converges vaguely** to a Borel measure μ if

$$\int_X f\, d\mu_n \to \int_X f\, d\mu \tag{11.59}$$

for every $f \in C_c(X)$. As with weak convergence (cf. Problem 12.4) we can conclude that the integral over functions with compact supports determines $\mu(K)$ for every compact set. Hence the vague limit will be unique if μ is

inner regular (which we already know to always hold if X is locally compact and separable by Corollary 8.22).

We first investigate the connection with weak convergence.

Lemma 11.41. *Let X be a locally compact separable metric space and suppose $\mu_m \to \mu$ vaguely. Then $\mu(X) \leq \liminf_n \mu_n(X)$ and (11.59) holds for every $f \in C_0(X)$ in case $\mu_n(X) \leq M$. If in addition $\mu_n(X) \to \mu(X)$, then (11.59) holds for every $f \in C_b(X)$, that is, $\mu_n \to \mu$ weakly.*

Proof. For every compact set K_m we can find a nonnegative function $g_m \in C_c(X)$ which is one on K_m by Urysohn's lemma. Hence $\mu(K_m) \leq \int g_m d\mu = \lim_n \int g_m d\mu_n \leq \liminf_n \mu_n(X)$. Letting $K_m \nearrow X$ shows $\mu(X) \leq \liminf_n \mu_n(X)$. Next, let $f \in C_0(X)$ and fix $\varepsilon > 0$. Then there is a compact set K such that $|f(x)| \leq \varepsilon$ for $x \in X \setminus K$. Choose g for K as before and set $f = f_1 + f_2$ with $f_1 = gf$. Then $|\int f d\mu - \int f d\mu_n| \leq |\int f_1 d\mu - \int f_1 d\mu_n| + 2\varepsilon M$ and the first claim follows.

Similarly, for the second claim, let $|f| \leq C$ and choose a compact set K such that $\mu(X \backslash K) < \frac{\varepsilon}{2}$. Then we have $\mu_n(X \backslash K) < \varepsilon$ for $n \geq N$.Choose g for K as before and set $f = f_1 + f_2$ with $f_1 = gf$. Then $|\int f d\mu - \int f d\mu_n| \leq |\int f_1 d\mu - \int f_1 d\mu_n| + 2\varepsilon C$ and the second claim follows. □

Example 11.16. The example $X = \mathbb{R}$ with $\mu_n = \delta_n$ shows that in the first claim f cannot be replaced by a bounded continuous function. Moreover, the example $\mu_n = n\,\delta_n$ also shows that the uniform bound cannot be dropped. ◇

The analog of Theorem 11.40 reads as follows.

Theorem 11.42. *Let X be a locally compact metric space and μ_n, μ Borel measures. The following are equivalent:*

(i) *μ_n converges vagly to μ.*

(ii) *$\int_X f\, d\mu_n \to \int_X f\, d\mu$ for every Lipschitz continuous f with compact support.*

(iii) *$\limsup_n \mu_n(C) \leq \mu(C)$ for all compact sets K and $\liminf_n \mu_n(O) \geq \mu(O)$ for all relatively compact open sets O.*

(iv) *$\mu_n(A) \to \mu(A)$ for all relative compact Borel sets A with $\mu(\partial A) = 0$.*

(v) *$\int_X f\, d\mu_n \to \int_X f\, d\mu$ for all bounded functions f with compact support which are continuous at μ-a.e. $x \in X$.*

Proof. (i) ⇒ (ii) is trivial. (ii) ⇒ (iii): The claim for compact sets follows as in the proof of Theorem 11.40. To see the case of open sets let $K_n = \{x \in$

$X|\operatorname{dist}(x, X \setminus O) \geq n^{-1}\}$. Then $K_n \subseteq O$ is compact and we can look at

$$g_n(x) := \frac{\operatorname{dist}(x, X \setminus O)}{\operatorname{dist}(x, X \setminus O) + \operatorname{dist}(x, K_n)}.$$

Then g_n is supported on O and $g_n \nearrow \chi_O$. Then

$$\liminf_n \mu_n(O) = \liminf_n \int \chi_O \mu_n \geq \liminf_n \int g_m \mu_n = \int g_m \mu.$$

Now taking $m \to \infty$ establishes the claim.

The remaining directions follow literally as in the proof of Theorem 11.40 (concerning (iv) $\Rightarrow$ (v) note that A_r is relatively compact). □

Finally we look at Borel measures on $\mathbb{R}$. In this case, we have the following equivalent characterization of vague convergence.

Lemma 11.43. *Let μ_n be a sequence of Borel measures on $\mathbb{R}$. Then $\mu_n \to \mu$ vaguely if and only if the distribution functions (normalized at a point of continuity of μ) converge at every point of continuity of μ.*

Proof. Suppose $\mu_n \to \mu$ vaguely. Then the distribution functions converge at every point of continuity of μ by item (iv) of Theorem 11.42.

Conversely, suppose that the distribution functions converge at every point of continuity of μ. To see that in fact $\mu_n \to \mu$ vaguely, let $f \in C_c(\mathbb{R})$. Fix some $\varepsilon > 0$ and note that, since f is uniformly continuous, there is a $\delta > 0$ such that $|f(x) - f(y)| \leq \varepsilon$ whenever $|x - y| \leq \delta$. Next, choose some points $x_0 < x_1 < \cdots < x_k$ such that $\operatorname{supp}(f) \subset (x_0, x_k)$, μ is continuous at x_j, and $x_j - x_{j-1} \leq \delta$ (recall that a monotone function has at most countable discontinuities). Furthermore, there is some N such that $|\mu_n(x_j) - \mu(x_j)| \leq \frac{\varepsilon}{2k}$ for all j and $n \geq N$. Then

$$\begin{aligned}\left|\int f d\mu_n - \int f d\mu\right| \leq & \sum_{j=1}^{k} \int_{(x_{j-1}, x_j]} |f(x) - f(x_j)| d\mu_n(x) \\ & + \sum_{j=1}^{k} |f(x_j)||\mu((x_{j-1}, x_j]) - \mu_n((x_{j-1}, x_j])| \\ & + \sum_{j=1}^{k} \int_{(x_{j-1}, x_j]} |f(x) - f(x_j)| d\mu(x).\end{aligned}$$

Now, for $n \geq N$, the first and the last terms on the right-hand side are both bounded by $(\mu((x_0, x_k]) + \frac{\varepsilon}{k})\varepsilon$ and the middle term is bounded by $\max|f|\varepsilon$. Thus the claim follows. □

Moreover, every bounded sequence of measures has a vaguely convergent subsequence (this is a special case of **Helly's selection theorem** — a generalization will be provided in Theorem 12.11).

Lemma 11.44. *Suppose μ_n is a sequence of finite Borel measures on $\mathbb{R}$ such that $\mu_n(\mathbb{R}) \leq M$. Then there exists a subsequence n_j which converges vaguely to some measure μ with $\mu(\mathbb{R}) \leq \liminf_j \mu_{n_j}(\mathbb{R})$.*

Proof. Let $\mu_n(x) = \mu_n((-\infty, x])$ be the corresponding distribution functions. By $0 \leq \mu_n(x) \leq M$ there is a convergent subsequence for fixed x. Moreover, by the standard diagonal series trick (enumerate the rationals, extract a subsequence which converges for the first rational, form this subsequence extract another one which converges also on the second rational, etc.; finally choose the diagonal sequence, i.e., the fisrt element form the first subsequence, the second element from the second subsequence, etc.), we can assume that $\mu_n(y)$ converges to some number $\mu(y)$ for each rational y. For irrational x we set $\mu(x) = \inf\{\mu(y) | x < y \in \mathbb{Q}\}$. Then $\mu(x)$ is monotone, $0 \leq \mu(x_1) \leq \mu(x_2) \leq M$ for $x_1 < x_2$. Indeed for $x_1 \leq y_1 < x_2 \leq y_2$ with $y_j \in \mathbb{Q}$ we have $\mu(x_1) \leq \mu(y_1) = \lim_n \mu_n(y_1) \leq \lim_n \mu_n(y_2) = \mu(y_2)$. Taking the infimum over y_2 gives the result.

Furthermore, for every $\varepsilon > 0$ and $x - \varepsilon < y_1 \leq x \leq y_2 < x + \varepsilon$ with $y_j \in \mathbb{Q}$ we have

$$\mu(x-\varepsilon) \leq \lim_n \mu_n(y_1) \leq \liminf \mu_n(x) \leq \limsup \mu_n(x) \leq \lim_n \mu_n(y_2) \leq \mu(x+\varepsilon)$$

and thus

$$\mu(x-) \leq \liminf \mu_n(x) \leq \limsup \mu_n(x) \leq \mu(x+)$$

which shows that $\mu_n(x) \to \mu(x)$ at every point of continuity of μ. So we can redefine μ to be right continuous without changing this last fact. The bound for $\mu(\mathbb{R})$ follows since for every point of continuity x of μ we have $\mu(x) = \lim_n \mu_n(x) \leq \liminf_n \mu_n(\mathbb{R})$. □

Example 11.17. The example $d\mu_n(x) = d\Theta(x - n)$ for which $\mu_n(x) = \Theta(x - n) \to 0$ shows that we can have $\mu(\mathbb{R}) = 0 < 1 = \mu_n(\mathbb{R})$. ⋄

Problem 11.28. *Suppose $\mu_n \to \mu$ vaguely and let I be a bounded interval with boundary points x_0 and x_1. Then*

$$\limsup_n \left| \int_I f d\mu_n - \int_I f d\mu \right| \leq |f(x_1)| \mu(\{x_1\}) + |f(x_0)| \mu(\{x_0\})$$

for any $f \in C([x_0, x_1])$.

Problem 11.29. *Let $\mu_n(X) \leq M$ and suppose* (11.59) *holds for all $f \in U \subseteq C(X)$. Then* (11.59) *holds for all f in the closed linear span of U.*

Problem 11.30. *Let $\mu_n(\mathbb{R}), \mu(\mathbb{R}) \leq M$ and suppose the Cauchy transforms converge*

$$\int_{\mathbb{R}} \frac{1}{x-z} d\mu_n(x) \to \int_{\mathbb{R}} \frac{1}{x-z} d\mu(x)$$

for $z \in U$, where $U \subseteq \mathbb{C}\backslash\mathbb{R}$ is a set which has a limit point. Then $\mu_n \to \mu$ vaguely. (Hint: Problem B.31.)

Problem 11.31. *Let X be a proper metric space. A sequence of finite measures is called **tight** if for every $\varepsilon > 0$ there is a compact set $K \subseteq X$ such that $\sup_n \mu_n(X \setminus K) \leq \varepsilon$. Show that a vaguely convergent sequence is tight if and only if $\mu_n(X) \to \mu(X)$.*

Problem 11.32. *Let ϕ be a mollifier and μ a finite measure. Show that $\phi_\varepsilon * \mu$ converges weakly to μ. Here $\phi * \mu$ is the absolutely continuous measure with density $\int_{\mathbb{R}} \phi(x-y) d\mu(y)$ (cf. Problem 10.27).*

11.8. Appendix: Functions of bounded variation and absolutely continuous functions

Let $[a,b] \subseteq \mathbb{R}$ be some compact interval and $f : [a,b] \to \mathbb{C}$. Given a partition $P = \{a = x_0, \ldots, x_n = b\}$ of $[a,b]$ we define the **variation** of f with respect to the partition P by

$$V(P,f) := \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|. \tag{11.60}$$

Note that the triangle inequality implies that adding points to a partition increases the variation: if $P_1 \subseteq P_2$ then $V(P_1,f) \leq V(P_2,f)$. The supremum over all partitions

$$V_a^b(f) := \sup_{\text{partitions } P \text{ of } [a,b]} V(P,f) \tag{11.61}$$

is called the **total variation** of f over $[a,b]$. If the total variation is finite, f is called of **bounded variation**. Since we clearly have

$$V_a^b(\alpha f) = |\alpha| V_a^b(f), \qquad V_a^b(f+g) \leq V_a^b(f) + V_a^b(g) \tag{11.62}$$

the space $BV[a,b]$ of all functions of finite total variation is a vector space. However, the total variation is not a norm since (consider the partition $P = \{a, x, b\}$)

$$V_a^b(f) = 0 \quad \Leftrightarrow \quad f(x) \equiv c. \tag{11.63}$$

Moreover, any function of bounded variation is in particular bounded (consider again the partition $P = \{a, x, b\}$)

$$\sup_{x \in [a,b]} |f(x)| \leq |f(a)| + V_a^b(f). \tag{11.64}$$

Theorem 11.45. *The functions of bounded variationt $BV[a,b]$ together with the norm*

$$\|f\|_{BV} := |f(a)| + V_a^b(f) \tag{11.65}$$

are a Banach space. Moreover, by (11.64) *we have* $\|f\|_\infty \le \|f\|_{BV}$.

Proof. By (11.63) we have $\|f\|_{BV} = 0$ if and only if f is constant and $|f(a)| = 0$, that is $f = 0$. Moreover, by (11.62) the norm is homogenous and satisfies the triangle inequality. So let f_n be a Cauchy sequence. Then f_n converges uniformly and pointwise to some bounded function f. Moreover, choose N such that $\|f_n - f_m\|_{BV} < \varepsilon$ whenever $m, n \ge N$. Then for $n \ge N$ and for any fixed partition

$$\begin{aligned} |f(a) - f_n(a)| + V(P, f - f_n) &= \lim_{m\to\infty} \big(|f_m(a) - f_n(a)| + V(P, f_m - f_n)\big) \\ &\le \sup_{m\ge N} \|f_n - f_m\|_{BV} < \varepsilon. \end{aligned}$$

Consequently $\|f - f_n\|_{BV} < \varepsilon$ which shows $f \in BV[a,b]$ as well as $f_n \to f$ in $BV[a,b]$. □

Observe $V_a^a(f) = 0$ as well as (Problem 11.34)

$$V_a^b(f) = V_a^c(f) + V_c^b(f), \qquad c \in [a,b], \tag{11.66}$$

and it will be convenient to set

$$V_b^a(f) = -V_a^b(f). \tag{11.67}$$

Example 11.18. Every Lipschitz continuous function is of bounded variation. In fact, if $|f(x) - f(y)| \le L|x - y|$ for $x, y \in [a,b]$, then $V_a^b(f) \le L(b-a)$. However, (Hölder) continuity is not sufficient (cf. Problems 11.35 and 11.36). ◇

Example 11.19. By the inverse triangle inequality we have

$$V_a^b(|f|) \le V_a^b(f)$$

whereas the converse is not true: The function $f : [0,1] \to \{-1, 1\}$ which is 1 on the rational and -1 on the irrational numbers satisfies $V_0^1(f) = \infty$ (show this) and $V_0^1(|f|) = V_0^1(1) = 0$.

From $2^{-1/2}(|\mathrm{Re}(z)| + |\mathrm{Im}(z)|) \le |z| \le |\mathrm{Re}(z)| + |\mathrm{Im}(z)|$ we infer

$$2^{-1/2}\big(V_a^b(\mathrm{Re}(f)) + V_a^b(\mathrm{Im}(f))\big) \le V_a^b(f) \le V_a^b(\mathrm{Re}(f)) + V_a^b(\mathrm{Im}(f))$$

which shows that f is of bounded variation if and only if $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ are. ◇

Example 11.20. Any real-valued nondecreasing function f is of bounded variation with variation given by $V_a^b(f) = f(b) - f(a)$. Similarly, every real-valued nonincreasing function g is of bounded variation with variation given by $V_a^b(g) = g(a) - g(b)$. Moreover, the sum $f + g$ is of bounded variation

with variation given by $V_a^b(f+g) \le V_a^b(f) + V_a^b(g)$. The following theorem shows that the converse is also true. $\diamond$

Theorem 11.46 (Jordan)**.** *Let $f : [a,b] \to \mathbb{R}$ be of bounded variation, then f can be decomposed as*

$$f(x) = f_+(x) - f_-(x), \qquad f_\pm(x) := \frac{1}{2}\left(V_a^x(f) \pm f(x)\right), \tag{11.68}$$

where $f_\pm$ are nondecreasing functions. Moreover, $V_a^b(f_\pm) \le V_a^b(f)$.

Proof. From

$$f(y) - f(x) \le |f(y) - f(x)| \le V_x^y(f) = V_a^y(f) - V_a^x(f)$$

for $x \le y$ we infer $V_a^x(f) - f(x) \le V_a^y(f) - f(y)$, that is, f_+ is nondecreasing. Moreover, replacing f by $-f$ shows that f_- is nondecreasing and the claim follows. $\square$

In particular, we see that functions of bounded variation have at most countably many discontinuities and at every discontinuity the limits from the left and right exist.

For functions $f : (a,b) \to \mathbb{C}$ (including the case where (a,b) is unbounded) we will set

$$V_a^b(f) := \lim_{c \downarrow a, d \uparrow b} V_c^d(f). \tag{11.69}$$

In this respect the following lemma is of interest (whose proof is left as an exercise):

Lemma 11.47. *Suppose $f \in BV[a,b]$. We have $\lim_{c\uparrow b} V_a^c(f) = V_a^b(f)$ if and only if $f(b) = f(b-)$ and $\lim_{c\downarrow a} V_c^b(f) = V_a^b(f)$ if and only if $f(a) = f(a+)$. In particular, $V_a^x(f)$ is left, right continuous if and only f is.*

If $f : \mathbb{R} \to \mathbb{C}$ is of bounded variation, then we can write it as a linear combination of four nondecreasing functions and hence associate a complex measure df with f via Theorem 8.13 (since all four functions are bounded, so are the associated measures).

Theorem 11.48. *There is a one-to-one correspondence between functions in $f \in BV(\mathbb{R})$ which are right continuous and normalized by $f(0) = 0$ and complex Borel measures ν on $\mathbb{R}$ such that f is the distribution function of ν as defined in* (8.18)*. Moreover, in this case the distribution function of the total variation of ν is $|\nu|(x) = V_0^x(f)$.*

Proof. We have already seen how to associate a complex measure df with a function of bounded variation. If f is right continuous and normalized, it will be equal to the distribution function of df by construction. Conversely,

let $d\nu$ be a complex measure with distribution function ν. Then for every $a < b$ we have

$$\begin{aligned} V_a^b(\nu) &= \sup_{P=\{a=x_0,\ldots,x_n=b\}} V(P,\nu) \\ &= \sup_{P=\{a=x_0,\ldots,x_n=b\}} \sum_{k=1}^{n} |\nu((x_{k-1},x_k])| \le |\nu|((a,b]) \end{aligned}$$

and thus the distribution function is of bounded variation. Furthermore, consider the measure μ whose distribution function is $\mu(x) = V_0^x(\nu)$. Then we see $|\nu((a,b])| = |\nu(b) - \nu(a)| \le V_a^b(\nu) = \mu((a,b]) \le |\nu|((a,b])$. Hence we obtain $|\nu(A)| \le \mu(A) \le |\nu|(A)$ for all intervals A, thus for all open sets (by Problem B.20), and thus for all Borel sets by outer regularity. Hence Lemma 11.14 implies $\mu = |\nu|$ and hence $|\nu|(x) = V_0^x(f)$. □

As a consequence of Corollary 11.10 we note:

Corollary 11.49. *Functions of bounded variation are differentiable a.e. with respect to Lebesgue measure.*

We will call a function $f : [a,b] \to \mathbb{C}$ **absolutely continuous** if for every $\varepsilon > 0$ there is a corresponding $\delta > 0$ such that

$$\sum_k |y_k - x_k| < \delta \quad \Rightarrow \quad \sum_k |f(y_k) - f(x_k)| < \varepsilon \tag{11.70}$$

for every countable collection of pairwise disjoint intervals $(x_k, y_k) \subset [a,b]$. The set of all absolutely continuous functions on $[a,b]$ will be denoted by $AC[a,b]$. The special choice of just one interval shows that every absolutely continuous function is (uniformly) continuous, $AC[a,b] \subset C[a,b]$.

Example 11.21. Every Lipschitz continuous function is absolutely continuous. In fact, if $|f(x) - f(y)| \le L|x-y|$ for $x, y \in [a,b]$, then we can choose $\delta = \frac{\varepsilon}{L}$. In particular, $C^1[a,b] \subset AC[a,b]$. Note that Hölder continuity is neither sufficient (cf. Problem 11.36 and Theorem 11.51 below) nor necessary (cf. Problem 11.50). ⋄

Theorem 11.50. *A complex Borel measure ν on $\mathbb{R}$ is absolutely continuous with respect to Lebesgue measure if and only if its distribution function is locally absolutely continuous (i.e., absolutely continuous on every compact subinterval). Moreover, in this case the distribution function $\nu(x)$ is differentiable almost everywhere and*

$$\nu(x) = \nu(0) + \int_0^x \nu'(y)dy \tag{11.71}$$

with ν' integrable, $\int_{\mathbb{R}} |\nu'(y)|dy = |\nu|(\mathbb{R})$.

Proof. Suppose the measure ν is absolutely continuous. Now (11.70) follows from (11.26) in the special case where A is a union of pairwise disjoint intervals.

Conversely, suppose $\nu(x)$ is absolutely continuous on $[a, b]$. We will verify (11.26). To this end fix ε and choose δ such that $\nu(x)$ satisfies (11.70). By outer regularity it suffices to consider the case where A is open. Moreover, by Problem B.20, every open set $O \subset (a, b)$ can be written as a countable union of disjoint intervals $I_k = (x_k, y_k)$ and thus $|O| = \sum_k |y_k - x_k| \leq \delta$ implies

$$|\nu(O)| = \Big| \sum_k \big(\nu(y_k) - \nu(x_k)\big) \Big| \leq \sum_k |\nu(y_k) - \nu(x_k)| \leq \varepsilon$$

as required.

The rest follows from Corollary 11.10. □

As a simple consequence of this result we can give an equivalent definition of absolutely continuous functions as precisely the functions for which the **fundamental theorem of calculus** holds.

Theorem 11.51. *A function $f : [a, b] \to \mathbb{C}$ is absolutely continuous if and only if it is of the form*

$$f(x) = f(a) + \int_a^x g(y)dy \tag{11.72}$$

for some integrable function g. Moreover, in this case f is differentiable a.e with respect to Lebesgue measure and $f'(x) = g(x)$. In addition, f is of bounded variation and

$$V_a^x(f) = \int_a^x |g(y)|dy. \tag{11.73}$$

Proof. This is just a reformulation of the previous result. To see the last claim combine the last part of Theorem 11.48 with Lemma 11.18. □

In particular, since the fundamental theorem of calculus fails for the Cantor function, this function is an example of a continuous function which is not absolutely continuous. Note that even if f is differentiable everywhere the fundamental theorem of calculus might fail (Problem 11.51).

Finally, we note that in this case the **integration by parts formula** continues to hold.

Lemma 11.52. *Let $f, g \in BV[a, b]$, then*

$$\int_{[a,b)} f(x-)dg(x) = f(b-)g(b-) - f(a-)g(a-) - \int_{[a,b)} g(x+)df(x) \tag{11.74}$$

as well as

$$\int_{(a,b]} f(x+)dg(x) = f(b+)g(b+) - f(a+)g(a+) - \int_{[a,b)} g(x-)df(x). \quad (11.75)$$

Proof. Since the formula is linear in f and holds if f is constant, we can assume $f(a-) = 0$ without loss of generality. Similarly, we can assume $g(b-) = 0$. Plugging $f(x-) = \int_{[a,x)} df(y)$ into the left-hand side of the first formula we obtain from Fubini

$$\begin{aligned}\int_{[a,b)} f(x-)dg(x) &= \int_{[a,b)}\int_{[a,x)} df(y)dg(x)\\ &= \int_{[a,b)}\int_{[a,b)} \chi_{\{(x,y)|y<x\}}(x,y)df(y)dg(x)\\ &= \int_{[a,b)}\int_{[a,b)} \chi_{\{(x,y)|y<x\}}(x,y)dg(x)df(y)\\ &= \int_{[a,b)}\int_{(y,b)} dg(x)df(y) = -\int_{[a,b)} g(y+)df(y).\end{aligned}$$

The second formula is shown analogously. □

If both $f, g \in AC[a,b]$ this takes the usual form

$$\int_a^b f(x)g'(x)dx = f(b)g(b) - f(a)g(a) - \int_a^b g(x)f'(x)dx. \quad (11.76)$$

For further generalizations of classical calculus rules to absolutely continuous functions see the problems.

Problem 11.33. *Compute $V_a^b(f)$ for $f(x) = \operatorname{sign}(x)$ on $[a,b] = [-1,1]$.*

Problem* 11.34. *Show* (11.66).

Problem* 11.35. *Consider $f_j(x) := x^j \cos(\pi/x)$ for $j \in \mathbb{N}$. Show that $f_j \in C[0,1]$ if we set $f_j(0) = 0$. Show that f_j is of bounded variation for $j \geq 2$ but not for $j = 1$.*

Problem* 11.36. *Let $\alpha \in (0,1)$ and $\beta > 1$ with $\alpha\beta \leq 1$. Set $M := \sum_{k=1}^{\infty} k^{-\beta}$, $x_0 := 0$, and $x_n := M^{-1}\sum_{k=1}^{n} k^{-\beta}$. Then we can define a function on $[0,1]$ as follows: Set $g(0) := 0$, $g(x_n) := n^{-\beta}$, and*

$$g(x) := c_n \left| x - t_n x_n - (1-t_n)x_{n+1} \right|, \qquad x \in [x_n, x_{n+1}],$$

where c_n and $t_n \in [0,1]$ are chosen such that g is (Lipschitz) continuous. Show that $f = g^\alpha$ is Hölder continuous of exponent α but not of bounded variation. (Hint: What is the variation on each subinterval?)

Problem* 11.37 (Takagi function)**.** *The Takagi function (also blancmange function) is the fixed point of the contraction $K(f)(x) := s(x) + \frac{1}{2}f(2x)$ for one-periodic functions, where $s(x) := \operatorname{dist}(x, \mathbb{Z}) = \min_{n\in\mathbb{Z}} |x - n|$:*

$$b(x) := \lim_{n\to\infty} b_n(x), \qquad b_n(x) = \sum_{k=0}^{n} \frac{s(2^k x)}{2^k}. \tag{11.77}$$

(i) Show that the Takagi function is Hölder continuous $|b(x)-b(y)| \le M_\alpha |x-y|^\alpha$ with $M_\alpha = (2^{1-\alpha} - 1)^{-1}$ for any $\alpha \in (0, 1)$. (Hint: Show that the contraction leaves the space of periodic functions satisfying such an estimate invariant. Note $|s(x) - s(y)| \le 2^{\alpha-1}|x - y|^\alpha$.)

(ii) Show that the Takagi function is not of bounded variation. (Hint: Note that b_n is piecewise linear on the intervals of the partition $P_n = \{k2^{-n-1} | 0 \le k \le 2^{n+1}\}$. Now investigate the derivative on this intervals and in particular how they change in each step. You should get $V_0^1(b_n) = V(P_n, b_n) = \sum_{k=0}^{\lfloor (n+1)/2 \rfloor} \binom{2k}{k} 2^{-2k}$.)

(iii) Show that the Takagi function is nowhere differentiable. By Corollary 11.49 this also shows that b is not ob bounded variation. (Hint: Consider a difference quotient $\frac{b(z)-b(y)}{z-y}$, where y, z are dyadic rationals.)

Problem 11.38. *Show that if $f \in BV[a, b]$ then so is f^*, $|f|$ and*

$$V_a^b(f^*) = V_a^b(f), \qquad V_a^b(|f|) \le V_a^b(f).$$

Moreover, show

$$V_a^b(\operatorname{Re}(f)) \le V_a^b(f), \qquad V_a^b(\operatorname{Im}(f)) \le V_a^b(f).$$

Problem 11.39. *Show that if $f, g \in BV[a, b]$ then so is $f\,g$ and*

$$V_a^b(f\,g) \le V_a^b(f) \sup |g| + V_a^b(g) \sup |f|.$$

Hence, together with the norm $\|f\|_{BV} := \|f\|_\infty + V_a^b(f)$ the space $BV[a, b]$ is a Banach algebra.

Problem 11.40. *Show that for $f \in BV(\mathbb{R})$ we have*

$$\int_{\mathbb{R}} |f(x + y) - f(x)| dx \le V_{-\infty}^{\infty}(f)|y|.$$

Problem 11.41. *Suppose $f \in AC[a, b]$. Show that f' vanishes a.e. on $f^{-1}(c)$ for every c. (Hint: Split the set $f^{-1}(c)$ into isolated points and the rest.)*

Problem 11.42. *A function $f : [a, b] \to \mathbb{R}$ is said to have the **Luzin N property** if it maps Lebesgue null sets to Lebesgue null sets. Show that absolutely continuous functions have the Luzin N property. Show that the Cantor function does not have the Luzin N property. (Hint: Use* (11.70)

and recall that: A set $A \subseteq \mathbb{R}$ is a null set if and only if for every ε there exists a countable set of intervals I_j which cover A and satisfy $\sum_j |I_j| < \varepsilon$.)

Problem* 11.43. *The variation of a function $f : [a,b] \to X$ with X a Banach space is defined as*

$$V_a^b(f) := \sup_{\text{partitions } P \text{ of } [a,b]} V(P,f), \quad V(P,f) := \sum_{k=1}^{m} \|f(x_k) - f(x_{k-1})\|.$$

Show that $f : [a,b] \to \mathbb{R}^n$ (with Euclidean norm) is of bounded variation if and only if every component is of bounded variation.

*Recall that a curve $\gamma : [a,b] \to \mathbb{R}^n$ is **rectifiable** if $V_a^b(\gamma) < \infty$. In this case $V_a^b(\gamma)$ is called the **arc length** of γ. Conclude that γ is rectifiable if and only if each of its coordinate functions is of bounded variation. Show that if each coordinate function is absolutely continuous, then*

$$V_a^b(\gamma) = \int_a^b |\gamma'(t)| dt.$$

(Hint: For the last part note that one inequality is easy. Then reduce it to the case when γ' is a step function.)

Problem 11.44. *Show that if $f, g \in AC[a,b]$ then so is $f\,g$ and the **product rule** $(f\,g)' = f'g + f\,g'$ holds. Conclude that $AC[a,b]$ is a closed subalgebra of the Banach algebra $BV[a,b]$. (Hint: Integration by parts. For the second part use Problem 11.39 and* (11.73)*.)*

Problem 11.45. *Show that $f \in AC[a,b]$ is nondecreasing iff $f' \geq 0$ a.e. and prove the **substitution rule***

$$\int_a^b g(f(x)) f'(x) dx = \int_{f(a)}^{f(b)} g(y) dy$$

*in this case. Conclude that if h is absolutely continuous, then so is $h \circ f$ and the **chain rule** $(h \circ f)' = (h' \circ f) f'$ holds.*

*Moreover, show that $f \in AC[a,b]$ is strictly increasing iff $f' > 0$ a.e. In this case f^{-1} is also absolutely continuous and the **inverse function rule***

$$(f^{-1})' = \frac{1}{f'(f^{-1})}$$

holds. (Hint: (9.71)*.)*

Problem 11.46. *Consider $f(x) := x^2 \sin(\frac{\pi}{x})$ on $[0,1]$ (here $f(0) = 0$) and $g(x) = \sqrt{|x|}$. Show that both functions are absolutely continuous, but $g \circ f$ is not. Hence the monotonicity assumption in Problem 11.45 is important. Show that if f is absolutely continuous and g Lipschitz, then $g \circ f$ is absolutely continuous.*

Problem 11.47. *Consider $f(x) := \frac{1}{2}(c(x) + x)$ on $[0,1]$, where c is the Cantor function. Show that the inverse f^{-1} is Lipschitz continuous with Lipschitz constant $\frac{1}{2}$. In particular f^{-1} is absolutely continuous while f is not. Hence the $f' > 0$ assumption in Problem 11.45 is important.*

Problem 11.48 (Characterization of the exponential function)**.** *Show that every nontrivial locally integrable function $f : \mathbb{R} \to \mathbb{C}$ satisfying*

$$f(x+y) = f(x)f(y), \qquad x, y \in \mathbb{R},$$

is of the from $f(x) = \mathrm{e}^{\alpha x}$ for some $\alpha \in \mathbb{C}$. (Hint: Start with $F(x) = \int_0^x f(t)dt$ and show $F(x+y) - F(x) = F(y)f(x)$. Conclude that f is absolutely continuous.)

Problem 11.49. *Let $X \subseteq \mathbb{R}$ be an interval, Y some measure space, and $f : X \times Y \to \mathbb{C}$ some measurable function. Suppose $x \mapsto f(x,y)$ is absolutely continuous for a.e. y such that*

$$\int_a^b \int_Y \left| \frac{\partial}{\partial x} f(x,y) \right| d\mu(y) dx < \infty \tag{11.78}$$

for every compact interval $[a,b] \subseteq X$ and $\int_Y |f(c,y)| d\mu(y) < \infty$ for one $c \in X$.

Show that

$$F(x) := \int_Y f(x,y)\, d\mu(y) \tag{11.79}$$

is absolutely continuous and

$$F'(x) = \int_Y \frac{\partial}{\partial x} f(x,y)\, d\mu(y) \tag{11.80}$$

in this case. (Hint: Fubini.)

Problem* 11.50. *Show that if $f \in AC[a,b]$ and $f' \in L^p(a,b)$, then f is Hölder continuous:*

$$|f(x) - f(y)| \le \|f'\|_p |x-y|^{1-\frac{1}{p}}.$$

Show that the function $f(x) = -\log(x)^{-1}$ is absolutely continuous but not Hölder continuous on $[0, \frac{1}{2}]$.

Problem* 11.51. *Consider $f(x) := x^2 \sin(\frac{\pi}{x^2})$ on $[0,1]$ (here $f(0) = 0$). Show that f is differentiable everywhere and compute its derivative. Show that its derivative is* not *integrable. In particular, this function is not absolutely continuous and the fundamental theorem of calculus does not hold for this function.*

Problem 11.52. *Show that the function from the previous problem is Hölder continuous of exponent $\frac{1}{2}$. (Hint: Consider $0 < x < y$. There is an $x' < y$ with $f(x') = f(x)$ and $(x')^{-2} - y^{-2} \le 2\pi$. Hence $(x')^{-1} - y^{-1} \le \sqrt{2\pi}$).*

Now use the Cauchy–Schwarz inequality to estimate $|f(y) - f(x)| = |f(y) - f(x')| = |\int_{x'}^{y} 1 \cdot f'(t)dt|$.*)*

Chapter 12

The dual of L^p

12.1. The dual of L^p, $p < \infty$

By the Hölder inequality every $g \in L^q(X, d\mu)$ gives rise to a linear functional on $L^p(X, d\mu)$ and this clearly raises the question if every linear functional is of this form. For $1 \leq p < \infty$ this is indeed the case:

Theorem 12.1. *Consider $L^p(X, d\mu)$ with some σ-finite measure μ and let q be the corresponding dual index, $\frac{1}{p} + \frac{1}{q} = 1$. Then the map $g \in L^q \mapsto \ell_g \in (L^p)^*$ given by*

$$\ell_g(f) := \int_X gf \, d\mu \tag{12.1}$$

is an isometric isomorphism for $1 \leq p < \infty$. If $p = \infty$ it is at least isometric.

Proof. Given $g \in L^q$ it follows from Hölder's inequality that ℓ_g is a bounded linear functional with $\|\ell_g\| \leq \|g\|_q$. Moreover, $\|\ell_g\| = \|g\|_q$ follows from Corollary 10.5.

To show that this map is surjective if $1 \leq p < \infty$, first suppose $\mu(X) < \infty$ and choose some $\ell \in (L^p)^*$. Since $\|\chi_A\|_p = \mu(A)^{1/p}$, we have $\chi_A \in L^p$ for every $A \in \Sigma$ and we can define

$$\nu(A) := \ell(\chi_A).$$

Suppose $A := \bigcup_{j=1}^{\infty} A_j$, where the A_j's are disjoint. Then, by dominated convergence, $\| \sum_{j=1}^{n} \chi_{A_j} - \chi_A \|_p \to 0$ (this is false for $p = \infty$!) and hence

$$\nu(A) = \ell(\sum_{j=1}^{\infty} \chi_{A_j}) = \sum_{j=1}^{\infty} \ell(\chi_{A_j}) = \sum_{j=1}^{\infty} \nu(A_j).$$

Thus ν is a complex measure. Moreover, $\mu(A) = 0$ implies $\chi_A = 0$ in L^p and hence $\nu(A) = \ell(\chi_A) = 0$. Thus ν is absolutely continuous with respect to μ and by the complex Radon–Nikodym theorem $d\nu = g\,d\mu$ for some $g \in L^1(X, d\mu)$. In particular, we have

$$\ell(f) = \int_X f g\, d\mu$$

for every simple function f. Next let $A_n = \{x||g(x)| < n\}$, then $g_n = g\chi_{A_n} \in L^q$ and by Corollary 10.5 we conclude $\|g_n\|_q \le \|\ell\|$. Letting $n \to \infty$ shows $g \in L^q$ and finishes the proof for finite μ.

If μ is σ-finite, let $X_n \nearrow X$ with $\mu(X_n) < \infty$. Then for every n there is some g_n on X_n and by uniqueness of g_n we must have $g_n = g_m$ on $X_n \cap X_m$. Hence there is some g and by $\|g_n\|_q \le \|\ell\|$ independent of n, we have $g \in L^q$. By construction $\ell(f\chi_{X_n}) = \ell_g(f\chi_{X_n})$ for every $f \in L^p$ and letting $n \to \infty$ shows $\ell(f) = \ell_g(f)$. □

Corollary 12.2. *Let μ be some σ-finite measure. Then $L^p(X, d\mu)$ is reflexive for $1 < p < \infty$.*

Proof. Identify $L^p(X, d\mu)^*$ with $L^q(X, d\mu)$ and choose $h \in L^p(X, d\mu)^{**}$. Then there is some $f \in L^p(X, d\mu)$ such that

$$h(g) = \int g(x)f(x)d\mu(x), \qquad g \in L^q(X, d\mu) \cong L^p(X, d\mu)^*.$$

But this implies $h(g) = g(f)$, that is, $h = J(f)$, and thus J is surjective. □

Note that in the case $0 < p < 1$, where L^p fails to be a Banach space, the dual might even be empty (see Problem 12.1)!

Problem* 12.1. *Show that $L^p(0, 1)$ is a quasinormed space if $0 < p < 1$ (cf. Problem 1.14). Moreover, show that $L^p(0, 1)^* = \{0\}$ in this case. (Hint: Suppose there were a nontrivial $\ell \in L^p(0, 1)^*$. Start with $f_0 \in L^p$ such that $|\ell(f_0)| \ge 1$. Set $g_0 = \chi_{(0,s]}f$ and $h_0 = \chi_{(s,1]}f$, where $s \in (0, 1)$ is chosen such that $\|g_0\|_p = \|h_0\|_p = 2^{-1/p}\|f_0\|_p$. Then $|\ell(g_0)| \ge \frac{1}{2}$ or $|\ell(h_0)| \ge \frac{1}{2}$ and we set $f_1 = 2g_0$ in the first case and $f_1 = 2h_0$ else. Iterating this procedure gives a sequence f_n with $|\ell(f_n)| \ge 1$ and $\|f_n\|_p = 2^{-n(1/p-1)}\|f_0\|_p$.)*

Problem 12.2. *Suppose $K : L^p(Y, d\nu) \to L^p(X, d\mu)$ is an integral operator with kernel $K(x, y)$ satisfying the Schur criterion (Lemma 10.23). Show that for $1 \le p < \infty$ the adjoint operator $K' : L^q(X, d\mu) \to L^q(Y, d\nu)$ is given by*

$$(K'f)(x) = \int_X K(y, x)f(y)d\mu(y), \qquad f \in L^q(X, d\mu).$$

12.2. The dual of L^∞ and the Riesz representation theorem

In the last section we have computed the dual space of L^p for $p < \infty$. Now we want to investigate the case $p = \infty$. Recall that we already know that the dual of L^∞ is much larger than L^1 since it cannot be separable in general.

Example 12.1. Let ν be a complex measure. Then

$$\ell_\nu(f) := \int_X f d\nu \tag{12.2}$$

is a bounded linear functional on $B(X)$ (the Banach space of bounded measurable functions) with norm

$$\|\ell_\nu\| = |\nu|(X) \tag{12.3}$$

by (11.25) and Corollary 11.19. If ν is absolutely continuous with respect to μ, then it will even be a bounded linear functional on $L^\infty(X, d\mu)$ since the integral will be independent of the representative in this case. ⋄

So the dual of $B(X)$ contains all complex measures. However, this is still not all of $B(X)^*$. In fact, it turns out that it suffices to require only finite additivity for ν.

Let (X, Σ) be a measurable space. A **complex content** ν is a map $\nu : \Sigma \to \mathbb{C}$ such that (finite additivity)

$$\nu(\bigcupdot_{k=1}^{n} A_k) = \sum_{k=1}^{n} \nu(A_k), \qquad A_j \cap A_k = \emptyset,\ j \neq k. \tag{12.4}$$

A content is called positive if $\nu(A) \geq 0$ for all $A \in \Sigma$ and given ν we can define its **total variation** $|\nu|(A)$ as in (11.16). The same proof as in Theorem 11.13 shows that $|\nu|$ is a positive content. However, since we do not require σ-additivity, it is not clear that $|\nu|(X)$ is finite. Hence we will call ν **finite** if $|\nu|(X) < \infty$. As in (11.18), (11.19) we can split every content into a complex linear combination of four positive contents.

Given a content ν we can define the corresponding integral for simple functions $s(x) = \sum_{k=1}^n \alpha_k \chi_{A_k}$ as usual

$$\int_A s\, d\nu := \sum_{k=1}^{n} \alpha_k \nu(A_k \cap A). \tag{12.5}$$

As in the proof of Lemma 9.1 one shows that the integral is linear. Moreover,

$$|\int_A s\, d\nu| \leq |\nu|(A)\, \|s\|_\infty \tag{12.6}$$

and for a finite content this integral can be extended to all of $B(X)$ such that

$$|\int_X f\, d\nu| \leq |\nu|(X)\, \|f\|_\infty \tag{12.7}$$

by Theorem 1.17 (compare Problem 9.7). However, note that our convergence theorems (monotone convergence, dominated convergence) will no longer hold in this case (unless ν happens to be a measure).

In particular, every complex content gives rise to a bounded linear functional on $B(X)$ and the converse also holds:

Theorem 12.3. *Every bounded linear functional $\ell \in B(X)^*$ is of the form*

$$\ell(f) = \int_X f \, d\nu \tag{12.8}$$

for some unique finite complex content ν and $\|\ell\| = |\nu|(X)$.

Proof. Let $\ell \in B(X)^*$ be given. If there is a content ν at all, it is uniquely determined by $\nu(A) := \ell(\chi_A)$. Using this as definition for ν, we see that finite additivity follows from linearity of ℓ. Moreover, (12.8) holds for characteristic functions and by

$$\ell(\sum_{k=1}^{n} \alpha_k \chi_{A_k}) = \sum_{k=1}^{n} \alpha_k \nu(A_k) = \sum_{k=1}^{n} |\nu(A_k)|, \qquad \alpha_k = \operatorname{sign}(\nu(A_k)),$$

we see $|\nu|(X) \le \|\ell\|$.

Since the characteristic functions are total, (12.8) holds everywhere by continuity and (12.7) shows $\|\ell\| \le |\nu|(X)$. □

It is also easy to tell when ν is positive. To this end call ℓ a **positive functional** if $\ell(f) \ge 0$ whenever $f \ge 0$.

Corollary 12.4. *Let $\ell \in B^*(X)$ be associated with the finite content ν. Then ν will be a positive content if and only if ℓ is a positive functional. Moreover, every $\ell \in B^*(X)$ can be written as a complex linear combination of four positive functionals.*

Proof. Clearly $\ell \ge 0$ implies $\nu(A) = \ell(\chi_A) \ge 0$. Conversely $\nu(A) \ge 0$ implies $\ell(s) \ge 0$ for every simple $s \ge 0$. Now for $f \ge 0$ we can find a sequence of simple functions s_n such that $\|s_n - f\|_\infty \to 0$. Moreover, by $||s_n| - f| \le |s_n - f|$ we can assume s_n to be nonnegative. But then $\ell(f) = \lim_{n\to\infty} \ell(s_n) \ge 0$ as required.

The last part follows by splitting the content ν into a linear combination of positive contents. □

Remark: To obtain the dual of $L^\infty(X, d\mu)$ from this you just need to restrict to those linear functionals which vanish on $\mathcal{N}(X, d\mu)$ (cf. Problem 12.3), that is, those whose content is *absolutely continuous* with respect to μ (note that the Radon–Nikodym theorem does not hold unless the content is a measure).

Example 12.2. Consider $B(\mathbb{R})$ and define

$$\ell(f) = \lim_{\varepsilon\downarrow 0}(\lambda f(-\varepsilon) + (1-\lambda)f(\varepsilon)), \qquad \lambda \in [0,1], \tag{12.9}$$

for f in the subspace of bounded measurable functions which have left and right limits at 0. Since $\|\ell\| = 1$ we can extend it to all of $B(\mathbb{R})$ using the Hahn–Banach theorem. Then the corresponding content ν is no measure:

$$\lambda = \nu([-1,0)) = \nu(\bigcup_{n=1}^{\infty}[-\frac{1}{n}, -\frac{1}{n+1})) \neq \sum_{n=1}^{\infty}\nu([-\frac{1}{n}, -\frac{1}{n+1})) = 0. \tag{12.10}$$

Observe that the corresponding distribution function (defined as in (8.18)) is nondecreasing but not right continuous! If we render ν right continuous, we get the the distribution function of the Dirac measure (centered at 0). In addition, the Dirac measure has the same integral at least for continuous functions! ⋄

Based on this observation we can give a simple proof of the Riesz representation for compact intervals. The general version will be shown in the next section.

Theorem 12.5 (Riesz representation)**.** *Let $I = [a,b] \subseteq \mathbb{R}$ be a compact interval. Every bounded linear functional $\ell \in C(I)^*$ is of the form*

$$\ell(f) = \int_I f\, d\nu \tag{12.11}$$

for some unique complex Borel measure ν and $\|\ell\| = |\nu|(I)$.

Proof. By the Hahn–Banach theorem we can extend ℓ to a bounded linear functional $\bar{\ell} \in B(I)^*$ and we have a corresponding content $\tilde{\nu}$. Splitting this content into positive parts it is no restriction to assume $\tilde{\nu}$ is positive.

Now the idea is as follows: Define a distribution function for $\tilde{\nu}$ as in (8.18). By finite additivity of $\tilde{\nu}$ it will be nondecreasing and we can use Theorem 8.13 to obtain an associated measure ν whose distribution function coincides with $\tilde{\nu}$ except possibly at points where ν is discontinuous. It remains to show that the corresponding integral coincides with ℓ for continuous functions.

Let $f \in C(I)$ be given. Fix points $a < x_0^n < x_1^n < \dots x_n^n < b$ such that $x_0^n \to a$, $x_n^n \to b$, and $\sup_k |x_{k-1}^n - x_k^n| \to 0$ as $n \to \infty$. Then the sequence of simple functions

$$f_n(x) = f(x_0^n)\chi_{[x_0^n, x_1^n)} + f(x_1^n)\chi_{[x_1^n, x_2^n)} + \cdots + f(x_{n-1}^n)\chi_{[x_{n-1}^n, x_n^n]}.$$

converges uniformly to f by continuity of f. Moreover,

$$\begin{aligned}\int_I f\,d\nu &= \lim_{n\to\infty}\int_I f_n\,d\nu = \lim_{n\to\infty}\sum_{k=1}^{n} f(x_{k-1}^n)(\nu(x_k^n)-\nu(x_{k-1}^n))\\ &= \lim_{n\to\infty}\sum_{k=1}^{n} f(x_{k-1}^n)(\tilde\nu(x_k^n)-\tilde\nu(x_{k-1}^n)) = \lim_{n\to\infty}\int_I f_n\,d\tilde\nu\\ &= \int_I f\,d\tilde\nu = \ell(f)\end{aligned}$$

provided the points x_k^n are chosen to stay away from all discontinuities of $\nu(x)$ (recall that there are at most countably many).

To see $\|\ell\| = |\nu|(I)$ recall $d\nu = hd|\nu|$ where $|h| = 1$ (Corollary 11.19). Now choose continuous functions $h_n(x) \to h(x)$ pointwise a.e. (Theorem 10.16). Using $\tilde h_n = \frac{h_n}{\max(1,|h_n|)}$ we even get such a sequence with $|\tilde h_n| \le 1$. Hence $\ell(\tilde h_n) = \int \tilde h_n^* h\,d|\nu| \to \int |h|^2 d|\nu| = |\nu|(I)$ implying $\|\ell\| \ge |\nu|(I)$. The converse follows from (12.7). □

Problem* 12.3. *Let M be a closed subspace of a Banach space X. Show that $(X/M)^* \cong \{\ell \in X^* | M \subseteq \operatorname{Ker}(\ell)\}$ (cf. Theorem 4.27).*

12.3. The Riesz–Markov representation theorem

In this section section we want to generalize Theorem 12.5. To this end X will be a metric space with the Borel σ-algebra. Given a Borel measure μ the integral

$$\ell(f) := \int_X f\,d\mu \tag{12.12}$$

will define a linear functional ℓ on the set of continuous functions with compact support $C_c(X)$. If μ were bounded we could drop the requirement for f to have compact support, but we do not want to impose this restriction here. However, in an arbitrary metric space there might not be many continuous functions with compact support. In fact, if $f \in C_c(X)$ and $x \in X$ is such that $f(x) \neq 0$, then $f^{-1}(B_r(f(x)))$ will be a relatively compact neighborhood of x whenever $0 < r < |f(x)|$. So in order to be able to *see* all of X, we will assume that every point has a relatively compact neighborhood, that is, X is locally compact.

Moreover, note that positivity of μ implies that ℓ is **positive** in the sense that $\ell(f) \ge 0$ if $f \ge 0$. This raises the question if there are any other requirements for a linear functional to be of the form (12.12). The purpose of this section is to prove that there are none, that is, there is a one-to-one connection between positive linear functionals on $C_c(X)$ and positive Borel measures on X.

As a preparation let us reflect how μ could be recovered from ℓ as in (12.12). Given a Borel set A it seems natural to try to approximate the characteristic function χ_A by continuous functions form the inside or the outside. However, if you try this for the rational numbers in the case $X = \mathbb{R}$, then this works neither from the inside nor the outside. So we have to be more modest. If K is a compact set, we can choose a sequence $f_n \in C_c(X)$ with $f_n \downarrow \chi_K$ (Problem 12.4). In particular,

$$\mu(K) = \lim_{n\to\infty} \int_X f_n d\mu \tag{12.13}$$

by dominated convergence. So we can recover the measure of compact sets from ℓ and hence μ if it is inner regular. In particular, for every positive linear functional there can be at most one inner regular measure. This shows how μ should be defined given a linear functional ℓ. Nevertheless it will be more convenient for us to approximate characteristic functions of open sets from the inside since we want to define an outer measure and use the Carathéodory construction. Hence, given a positive linear functional ℓ we define

$$\rho(O) := \sup\{\ell(f) | f \in C_c(X),\ f \prec O\} \tag{12.14}$$

for any open set O. Here $f \prec O$ is short hand for $f \le \chi_O$ and $\operatorname{supp}(f) \subseteq O$. Since ℓ is positive, so is ρ. Note that it is not clear that this definition will indeed coincide with $\mu(O)$ if ℓ is given by (12.12) unless O has a compact exhaustion. However, this is of no concern for us at this point.

Lemma 12.6. *Given a positive linear functional ℓ on $C_c(X)$ the set function ρ defined in* (12.14) *has the following properties:*

(i) $\rho(\emptyset) = 0$,

(ii) *monotonicity* $\rho(O_1) \le \rho(O_2)$ *if* $O_1 \subseteq O_2$,

(iii) ρ *is finite for relatively compact sets,*

(iv) $\rho(O) \le \sum_n \rho(O_n)$ *for every countable open cover* $\{O_n\}$ *of* O, *and*

(v) *additivity* $\rho(O_1 \uplus O_2) = \rho(O_1) + \rho(O_2)$ *if* $O_1 \cap O_2 = \emptyset$.

Proof. (i) and (ii) are clear. To see (iii) note that if $\overline{O}$ is compact, then by Urysohn's lemma there is a function $f \in C_c(X)$ which is one on $\overline{O}$ implying $\rho(O) \le \ell(f)$. To see (iv) let $f \in C_c(X)$ with $f \prec O$. Then finitely many of the sets $O_1, \dots, O_N$ will cover $K := \operatorname{supp}(f)$. Hence we can choose a partition of unity $h_1, \dots, h_{N+1}$ (Lemma B.30) subordinate to the cover $O_1, \dots, O_N, X \setminus K$. Then $\chi_K \le h_1 + \cdots + h_N$ and hence

$$\ell(f) = \sum_{j=1}^{N} \ell(h_j f) \le \sum_{j=1}^{N} \rho(O_j) \le \sum_n \rho(O_n).$$

To see (v) note that $f_1 \prec O_1$ and $f_2 \prec O_2$ implies $f_1 + f_2 \prec O_1 \uplus O_2$ and hence $\ell(f_1) + \ell(f_2) = \ell(f_1 + f_2) \leq \rho(O_1 \uplus O_2)$. Taking the supremum over f_1 and f_2 shows $\rho(O_1) + \rho(O_2) \leq \rho(O_1 \uplus O_2)$. The reverse inequality follows from the previous item. □

Lemma 12.7. *Let ℓ be a positive linear functional on $C_c(X)$ and let ρ be defined as in* (12.14)*. Then*

$$\mu^*(A) := \inf\Big\{\rho(O)\Big|A \subseteq O,\ O \text{ open}\Big\}. \tag{12.15}$$

defines a metric outer measure on X.

Proof. Consider the outer measure (Lemma 8.8)

$$\nu^*(A) := \inf\Big\{\sum_{n=1}^{\infty}\rho(O_n)\Big|A \subseteq \bigcup_{n=1}^{\infty} O_n,\ O_n \text{ open}\Big\}.$$

Then we clearly have $\nu^*(A) \leq \mu^*(A)$. Moreover, if $\nu^*(A) < \mu^*(A)$ we can find an open cover $\{O_n\}$ such that $\sum_n \rho(O_n) < \mu^*(A)$. But for $O = \bigcup_n O_n$ we have $\mu^*(A) \leq \rho(O) \leq \sum_n \rho(O_n)$, a contradiction. Hence $\mu^* = \nu^*$ and we have an outer measure.

To see that μ^* is a metric outer measure let A_1, A_2 with $\operatorname{dist}(A_1, A_2) > 0$ be given. Then, there are disjoint open sets $O_1 \supseteq A_1$ and $O_2 \supseteq A_2$. Hence for $A_1 \uplus A_2 \subseteq O$ we have $\mu^*(A_1) + \mu^*(A_2) \leq \rho(O_1 \cap O) + \rho(O_2 \cap O) \leq \rho(O)$ and taking the infimum over all O we have $\mu^*(A_1) + \mu^*(A_2) \leq \mu^*(A_1 \uplus A_2)$. The converse follows from subadditivity and hence μ^* is a metric outer measure. □

So Theorem 8.9 gives us a corresponding measure μ defined on the Borel σ-algebra by Lemma 8.11. By construction this Borel measure will be outer regular and it will also be inner regular as the next lemma shows. Note that if one is willing to make the extra assumption of separability for X, this will come for free from Corollary 8.22.

Lemma 12.8. *The Borel measure μ associated with μ^* from* (12.15) *is regular.*

Proof. Since μ is outer regular by construction it suffices to show

$$\mu(O) = \sup_{K \subseteq O,\, K \text{ compact}} \mu(K)$$

for every open set $O \subseteq X$. Now denote the last supremum by α and observe $\alpha \leq \mu(O)$ by monotonicity. For the converse we can assume $\alpha < \infty$ without loss of generality. Then, by the definition of $\mu(O) = \rho(O)$ we can find some $f \in C_c(X)$ with $f \prec O$ such that $\mu(O) \leq \ell(f) + \varepsilon$. Since $K := \operatorname{supp}(f) \subseteq O$ is compact we have $\mu(O) \leq \ell(f) + \varepsilon \leq \mu(K) + \varepsilon \leq \alpha + \varepsilon$ and as $\varepsilon > 0$ is arbitrary this establishes the claim. □

Now we are ready to show

Theorem 12.9 (Riesz–Markov representation)**.** *Let X be a locally compact metric space. Then every positive linear functional $\ell : C_c(X) \to \mathbb{C}$ gives rise to a unique regular Borel measure μ such that* (12.12) *holds.*

Proof. We have already constructed a corresponding Borel measure μ and it remains to show that ℓ is given by (12.12). To this end observe that if $f \in C_c(X)$ satisfies $\chi_O \le f \le \chi_C$, where O is open and C closed, then $\mu(O) \le \ell(f) \le \mu(C)$. In fact, every $g \prec O$ satisfies $\ell(g) \le \ell(f)$ and hence $\mu(O) = \rho(O) \le \ell(f)$. Similarly, for every $\tilde{O} \supseteq C$ we have $f \prec \tilde{O}$ and hence $\ell(f) \le \rho(\tilde{O})$ implying $\ell(f) \le \mu(C)$.

Now the next step is to split f into smaller pieces for which this estimate can be applied. To this end suppose $0 \le f \le 1$ and define $g_k^n := \min(f, \frac{k}{n})$ for $0 \le k \le n$. Clearly $g_0^n = 0$ and $g_n^n = f$. Setting $f_k^n := g_k^n - g_{k-1}^n$ for $1 \le k \le n$ we have $f = \sum_{k=1}^n f_k^n$ and $\frac{1}{n}\chi_{C_k^n} \le f_k^n \le \frac{1}{n}\chi_{O_{k-1}^n}$ where $O_k^n = \{x \in X | f(x) > \frac{k}{n}\}$ and $C_k^n = \overline{O_k^n} = \{x \in X | f(x) \ge \frac{k}{n}\}$. Summing over k we have $\frac{1}{n}\sum_{k=1}^n \chi_{C_k^n} \le f \le \frac{1}{n}\sum_{k=0}^{n-1} \chi_{O_k^n}$ as well as

$$\frac{1}{n}\sum_{k=1}^{n} \mu(O_k^n) \le \ell(f) \le \frac{1}{n}\sum_{k=0}^{n-1} \mu(C_k^n)$$

Hence we obtain

$$\int f\, d\mu - \frac{\mu(O_0^n)}{n} \le \frac{1}{n}\sum_{k=1}^{n} \mu(O_k^n) \le \ell(f) \le \frac{1}{n}\sum_{k=0}^{n-1} \mu(C_k^n) \le \int f\, d\mu + \frac{\mu(C_0^n)}{n}$$

and letting $n \to \infty$ establishes the claim since $C_0^n = \operatorname{supp}(f)$ is compact and hence has finite measure. □

Note that this might at first sight look like a contradiction since (12.12) gives a linear functional even if μ is not regular. However, in this case the Riesz–Markov theorem merely says that there will be a corresponding regular measure which gives rise to the same integral for continuous functions. Moreover, using (12.13) one even sees that both measures agree on compact sets.

As a consequence we can also identify the dual space of $C_0(X)$ (i.e. the closure of $C_c(X)$ as a subspace of $C_b(X)$). Note that $C_0(X)$ is separable if X is locally compact and separable (Lemma B.36). Also recall that a complex measure is regular if all four positive measures in the Jordan decomposition (11.20) are. By Lemma 11.21 this is equivalent to the total variation being regular.

Theorem 12.10 (Riesz–Markov representation)**.** *Let X be a locally compact metric space. Every bounded linear functional $\ell \in C_0(X)^*$ is of the form*

$$\ell(f) = \int_X f \, d\nu \tag{12.16}$$

for some unique regular complex Borel measure ν and $\|\ell\| = |\nu|(X)$. Moreover, ℓ will be positive if and only if ν is.

If X is compact this holds for $C(X) = C_0(X)$.

Proof. First of all observe that (11.25) shows that for every regular complex measure ν equation (12.16) gives a linear functional ℓ with $\|\ell\| \le |\nu|(X)$. This functional will be positive if ν is. Moreover, we have $d\nu = h\, d|\nu|$ (Corollary 11.19) and by Theorem 10.16 we can find a sequence $h_n \in C_c(X)$ with $h_n(x) \to h(x)$ pointwise a.e. Using $\tilde{h}_n := \frac{h_n}{\max(1,|h_n|)}$ we even get such a sequence with $|\tilde{h}_n| \le 1$. Hence $\ell(\tilde{h}_n^*) = \int \tilde{h}_n^* h \, d|\nu| \to \int |h|^2 d|\nu| = |\nu|(X)$ implying $\|\ell\| \ge |\nu|(X)$.

Conversely, let ℓ be given. By the Hahn–Banach theorem we can extend ℓ to a bounded linear functional $\overline{\ell} \in B(X)^*$ which can be written as a linear combinations of positive functionals by Corollary 12.4. Hence it is no restriction to assume ℓ is positive. But for positive ℓ the previous theorem implies existence of a corresponding regular measure ν such that (12.16) holds for all $f \in C_c(X)$. Since $\overline{C_c(X)} = C_0(X)$ (12.16) holds for all $f \in C_0(X)$ by continuity. □

Example 12.3. Note that the dual space of $C_b(X)$ will in general be larger. For example, consider $C_b(\mathbb{R})$ and define $\ell(f) = \lim_{x\to\infty} f(x)$ on the subspace of functions from $C_b(\mathbb{R})$ for which this limit exists. Extend ℓ to a bounded linear functional on all of $C_b(\mathbb{R})$ using Hahn–Banach. Then ℓ restricted to $C_0(\mathbb{R})$ is zero and hence there is no associated measure such that (12.16) holds. ⋄

The above result implies that for bounded sequences, vague convergence is the same as weak-$*$ convergence in $C_0(X)^*$. One direction following from Lemma 11.41 and the other from the fact that every weak-$*$ convergent sequence is bounded.

As a consequence we can extend Helly's selection theorem. We call a sequence of complex measures ν_n vaguely convergent to a measure ν if

$$\int_X f d\nu_n \to \int_X f d\nu, \qquad f \in C_c(X). \tag{12.17}$$

This generalizes our definition for positive measures from Section 11.7. Moreover, note that in the case that the sequence is bounded, $|\nu_n|(X) \le M$, we get (12.17) for all $f \in C_0(X)$. Indeed, choose $g \in C_c(X)$ such that

$\|f-g\|_\infty < \varepsilon$ and note that $\limsup_n |\int_X f d\nu_n - \int_X f d\nu| \le \limsup_n |\int_X (f-g) d\nu_n - \int_X (f-g) d\nu| \le \varepsilon(M + |\nu|(X))$.

Theorem 12.11. *Let X be a locally compact metric space. Then every bounded sequence ν_n of regular complex measures, that is $|\nu_n|(X) \le M$, has a vaguely convergent subsequence whose limit is regular. If all ν_n are positive, every limit of a convergent subsequence is again positive.*

Proof. Let $Y = C_0(X)$. Then we can identify the space of regular complex measure $\mathcal{M}_{reg}(X)$ as the dual space Y^* by the Riesz–Markov theorem. Moreover, every bounded sequence has a weak-$*$ convergent subsequence by the Banach–Alaoglu theorem (Theorem 5.10 — if X and hence $C_0(X)$ is separable, then Lemma 4.34 will suffice) and this subsequence converges in particular vaguely.

If the measures are positive, then $\ell_n(f) = \int f \, d\nu_n \ge 0$ for every $f \ge 0$ and hence $\ell(f) = \int f \, d\nu \ge 0$ for every $f \ge 0$, where $\ell \in Y^*$ is the limit of some convergent subsequence. Hence ν is positive by the Riesz–Markov representation theorem. □

Recall once more that in the case where X is locally compact and separable, regularity will automatically hold for every Borel measure.

Problem* 12.4. *Let X be a locally compact metric space. Show that for every compact set K there is a sequence $f_n \in C_c(X)$ with $0 \le f_n \le 1$ and $f_n \downarrow \chi_K$. (Hint: Urysohn's lemma.)*

Problem 12.5. *A function*

$$F(z) := bz + a + \int_{\mathbb{R}} \frac{1+zx}{x-z} d\mu(x), \qquad z \in \mathbb{C} \setminus \mathbb{R},$$

*with $b \ge 0$, $a \in \mathbb{R}$ and μ a finite (nonnegative) measure is called a **Herglotz–Nevanlinna function**. It is easy to check that $F(z)$ is analytic (cf. Problem 9.19), satisfies $F(z^*) = F(z)^*$, and maps the upper half plane to itself (unless $b = 0$ and $\mu = 0$), that is $\operatorname{Im}(F(z)) \ge 0$ for $\operatorname{Im}(z) > 0$. In fact, it can be shown (this is the Herglotz representation theorem) that any function with these properties is of the above form.*

Show that if F_n is a sequence of Herglotz–Nevanlinna functions which converges pointwise to a function F for z in some set $U \subseteq \mathbb{C}_+$ which has a limit point in $\mathbb{C}_+$, then F is also a Herglotz–Nevanlinna function and $a_n \to a$, $b_n \to b$ and $\mu_n \to \mu$ weakly. (Hint: Note that you can get a bound on $\mu_n(\mathbb{R})$ by considering $F_n(\mathrm{i})$. Now use Helly's selection theorem (Lemma 11.44 will do) and note that by the identity theorem the limit is uniquely determined by the values of F on U. Finally recall that vague convergence of measures is just weak-$*$ convergence on $C_0(\mathbb{R})^*$ and use Lemma B.5.)

Chapter 13

Sobolev spaces

13.1. Basic properties

Let $U \subseteq \mathbb{R}^n$ be nonempty and open. Our aim is to extended the Lebesgue spaces to include derivatives. To this end we call an element $\alpha \in \mathbb{N}_0^n$ a **multi-index** and $|\alpha|$ its **order**. For $f \in C^k(U)$ and $\alpha \in \mathbb{N}_0^n$ with $|\alpha| \leq k$ we set

$$\partial_\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \quad x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}, \quad |\alpha| := \alpha_1 + \cdots + \alpha_n. \tag{13.1}$$

Then for locally integrable function $f \in L^1_{loc}(U)$ a function $h \in L^1_{loc}(U)$ satisfying

$$\int_U \varphi(x) h(x) d^n x = (-1)^{|\alpha|} \int_U (\partial_\alpha \varphi)(x) f(x) d^n x, \quad \forall \varphi \in C_c^\infty(U), \tag{13.2}$$

is called the **weak derivative** or the derivative in the sense of distributions of f. Note that by Lemma 10.21 such a function is unique if it exists. Moreover, if $f \in C^k(U)$ then integration by parts (9.63) shows that the weak derivative coincides with the usual derivative.

Example 13.1. Consider $U := \mathbb{R}$. $f(x) := |x|$, then $\partial f(x) = \operatorname{sign}(x)$ as a weak derivative. If we try to take another derivative we are lead to

$$\int_{\mathbb{R}} \varphi(x) h(x) dx = -\int_{\mathbb{R}} \varphi'(x) \operatorname{sign}(x) dx = 2\varphi(0)$$

and it is easy to see that no locally integrable function can satisfy this requirement.

In fact, in one dimension the class of weakly differentiable functions can be identified with the class of antiderivatives of integrable functions, that is, the class of absolutely continuous functions (which is discussed in detail in

Section 11.8) — see Problem 13.2. Note that in higher dimensions weakly differentiable functions might not be continuous — Problem 13.3. ⋄

Now we can define the **Sobolev space** $W^{k,p}(U)$ as the set of all functions in $L^p(U)$ which have weak derivatives up to order k in $L^p(U)$. Clearly $W^{k,p}(U)$ is a linear space since $f, g \in W^{k,p}(U)$ implies $af + bg \in W^{k,p}(U)$ for $a, b \in \mathbb{C}$ and $\partial_\alpha(af + bg) = a\partial_\alpha f + b\partial_\alpha g$ for all $|\alpha| \le k$. Moreover, for $f \in W^{k,p}(U)$ we define its norm

$$\|f\|_{k,p} := \begin{cases} \left(\sum_{|\alpha|\le k} \|\partial_\alpha f\|_p^p\right)^{1/p}, & 1 \le p < \infty, \\ \max_{|\alpha|\le k} \|\partial_\alpha f\|_\infty, & p = \infty. \end{cases} \tag{13.3}$$

We will also use the gradient $\partial u = (\partial_1 u, \dots, \partial_n u)$ and by $\|\partial u\|_p$ we will always mean $\||\partial u|\|_p$, where $|\partial u|$ denotes the Euclidean norm.

It is easy to check that with this definition $W^{k,p}$ becomes a normed linear space. Of course for $p = 2$ we have a corresponding scalar product

$$\langle f, g\rangle_{W^{k,2}} := \sum_{|\alpha|\le k} \langle \partial_\alpha f, \partial_\alpha g\rangle_{L^2} \tag{13.4}$$

and one reserves the special notation $H^k(U) := W^{k,2}(U)$ for this case. Similarly we define local versions of these spaces $W^{k,p}_{loc}(U)$ as the set of all functions in $L^p_{loc}(U)$ which have weak derivatives up to order k in $L^p_{loc}(U)$.

Theorem 13.1. *For each $k \in \mathbb{N}_0$, $1 \le p \le \infty$ the Sobolev space $W^{k,p}(U)$ is a Banach space. It is separable for $1 \le p < \infty$ and reflexive for $1 < p < \infty$.*

Proof. Let f_m be a Cauchy sequence in $W^{k,p}$. Then $\partial_\alpha f_m$ is a Cauchy sequence in L^p for every $|\alpha| \le k$. Consequently $\partial_\alpha f_m \to f_\alpha$ in L^p. Moreover, letting $m \to \infty$ in

$$\begin{aligned} \int_U \varphi f_\alpha d^n x &= \lim_{m\to\infty} \int_U \varphi(\partial_\alpha f_m) d^n x = \lim_{m\to\infty} (-1)^{|\alpha|} \int_U (\partial_\alpha \varphi) f_m d^n x \\ &= (-1)^{|\alpha|} \int_U (\partial_\alpha \varphi) f_0 d^n x, \qquad \varphi \in C_c^\infty(U), \end{aligned}$$

shows $f_0 \in W^{k,p}$ with $\partial_\alpha f_0 = f_\alpha$ for $|\alpha| \le k$. By construction $f_m \to f_0$ in $W^{k,p}$ which implies that $W^{k,p}$ is complete.

Concerning the last claim note that $W^{k,p}(U)$ can regarded a subspace of $\bigtimes_{p,|\alpha|\le k} L^p(U)$ which has the claimed properties by Lemma 10.14 and Corollary 12.2. □

Now we show that smooth functions are dense in $W^{k,p}$. A first naive approach would be to extend $f \in W^{k,p}(U)$ to all of $\mathbb{R}^n$ by setting it 0 outside U and consider $f_\varepsilon := \phi_\varepsilon * f$, where ϕ is the standard mollifier. The problem with this approach is that we generically create a non-differentiable

singularity at the boundary and hence this only works as long as we stay away from the boundary.

Lemma 13.2 (Friedrichs)**.** *Let $f \in W^{k,p}(U)$ and set $f_\varepsilon := \phi_\varepsilon * f$, where ϕ is the standard mollifier. Then for every $\varepsilon_0 > 0$ we have $f_\varepsilon \to f$ in $W^{k,p}(U_{\varepsilon_0})$ if $1 \le p < \infty$, where $U_\varepsilon := \{x \in U | \operatorname{dist}(x, \mathbb{R}^n \setminus U) > \varepsilon\}$. If $p = \infty$ we have $\partial_\alpha f_\varepsilon \to \partial_\alpha f$ a.e. for all $|\alpha| \le k$.*

Proof. Just observe that all derivatives converge in L^p for $1 \le p < \infty$ since $\partial_\alpha f_\varepsilon = (\partial_\alpha \phi_\varepsilon) * f = \phi_\varepsilon * (\partial_\alpha f)$. Here the first equality is Lemma 10.18 (ii) and the second equality only holds (by definition of the weak derivative) on U_ε since in this case $\operatorname{supp}(\phi_\varepsilon(x - .)) = B_\varepsilon(x) \subseteq U$. So if we fix $\varepsilon_0 > 0$, then $f_\varepsilon \to f$ in $W^{k,p}(U_{\varepsilon_0})$. In the case $p = \infty$ the claim follows since $L^\infty_{loc} \subseteq L^1_{loc}$ after passing to a subsequence. That selecting a subsequence is superfluous follows from Problem 10.25. □

Note that, by Problem 13.10, if $f \in W^{k,\infty}(U)$ then $\partial_\alpha f$ is locally Lipschitz continuous for all $|\alpha| \le k - 1$. Hence $\partial_\alpha f_\varepsilon \to \partial_\alpha f$ locally uniformly for all $|\alpha| \le k - 1$.

So in particular, we get convergence in $W^{k,p}(U)$ if f has compact support. To adapt this approach to work on all of U we will use a partition of unity.

Theorem 13.3 (Meyers–Serrin)**.** *Let $U \subseteq \mathbb{R}^n$ be open and $1 \le p < \infty$. Then $C^\infty(U) \cap W^{k,p}(U)$ is dense in $W^{k,p}(U)$.*

Proof. Let h_j be a smooth partition of unity as in Lemma B.31 (take any cover). Let $f \in W^{k,p}(U)$ be given and fix $\delta > 0$. Choose $\varepsilon_i > 0$ sufficiently small such that $f_j := \phi_{\varepsilon_j} * (h_j f)$ (with ϕ the standard mollifier) satisfies

$$\|f_j - h_j f\|_{W^{k,p}} \le \frac{\delta}{2^{j+1}} \quad \text{and} \quad \operatorname{supp}(f_j) \subset O_j.$$

Then $f_\delta = \sum_j f_j \in C^\infty(U)$ since our cover is locally finite. Moreover, for every relatively compact set $V \subseteq U$ we have

$$\|f_\delta - f\|_{W^{k,p}(V)} = \|\sum_j (f_j - h_j f)\|_{W^{k,p}(V)} \le \sum_j \|f_j - h_j f\|_{W^{k,p}(V)} \le \delta$$

and letting $V \nearrow U$ we get $f_\delta \in W^{k,p}(U)$ as well as $\|f_\delta - f\|_{W^{k,p}(U)} \le \delta$. □

Example 13.2. The example $f(x) := |x| \in W^{1,\infty}(-1, 1)$ shows that the theorem fails in the case $p = \infty$ since $f'(x) = \operatorname{sign}(x)$ cannot be approximated uniformly by smooth functions. ⋄

For L^p we know that smooth functions with compact support are dense. This is no longer true in general for $W^{k,p}$ since convergence of derivatives enforces that the vanishing of boundary values is preserved in the limit.

However, making this precise requires some additional effort. So for now we will just give the closure of $C_c^\infty(U)$ in $W^{k,p}(U)$ a special name $W_0^{k,p}(U)$ as well as $H_0^k(U) := W_0^{k,2}(U)$. It is easy to see that $C_c^k(U) \subseteq W_0^{k,p}(U)$ for every $1 \le p \le \infty$ and $W_c^{k,p}(U) \subseteq W_0^{k,p}(U)$ for every $1 \le p < \infty$ (mollify to get a sequence in $C_c^\infty(U)$ which converges in $W^{k,p}(U)$). Moreover, note $W_0^{k,p}(\mathbb{R}^n) = W^{k,p}(\mathbb{R}^n)$ for $1 \le p < \infty$ (Problem 13.11).

Next we collect some basic properties of weak derivatives.

Lemma 13.4. *Let $U \subseteq \mathbb{R}^n$ be open and $1 \le p \le \infty$.*

(i) *The operator $\partial_\alpha : W^{k,p}(U) \to W^{k-|\alpha|,p}(U)$ is a bounded linear map and $\partial_\beta \partial_\alpha f = \partial_\alpha \partial_\beta f = \partial_{\alpha+\beta} f$ for $f \in W^{k,p}$ and all multi-indices α, β with $|\alpha| + |\beta| \le k$.*

(ii) *We have*

$$\int_U g(\partial_\alpha f) d^n x = (-1)^{|\alpha|} \int_U (\partial_\alpha g) f d^n x, \quad g \in W_0^{k,q}(U),\ f \in W^{k,p}(U), \tag{13.5}$$

for all $|\alpha| \le k$, $\frac{1}{p} + \frac{1}{q} = 1$. This also holds for $g \in W_c^{k,q}(U)$.

(iii) *Suppose $f \in W^{1,p}(U)$ and $g \in W^{1,q}(U)$. Then $f \cdot g \in W^{1,r}(U)$, where $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$, and we have the **product rule***

$$\partial_j (f \cdot g) = (\partial_j f) g + f (\partial_j g), \qquad 1 \le j \le n. \tag{13.6}$$

The same claim holds with $q = p = r$ if $f \in W^{1,p}(U) \cap L^\infty(U)$.

(iv) *Suppose $g \in C_b^1(\mathbb{R}^m)$ and $f_1, \dots, f_m \in W_{loc}^{1,1}(U)$ are real-valued. Then $g \circ f \in W_{loc}^{1,1}(U)$ and we have the **chain rule** $\partial_j (g \circ f) = \sum_k (\partial_k g)(f) \partial_j f_k$. If in addition $f_1, \dots, f_m \in W^{1,p}(U)$ and $g(0) = 0$ or $|U| < \infty$, then $g \circ f \in W^{1,p}(U)$.*

(v) *Let $\psi : U \to V$ be a C^1 diffeomorphism such that both ψ and ψ^{-1} have bounded derivatives. Then $f \in W^{1,p}(V)$ if and only if $f \circ \psi \in W^{k,p}(U)$ and we have the **change of variables formula** $\partial_j (f \circ \psi) = \sum_k (\partial_k f)(\psi) \partial_j \psi$. Moreover, $\|f \circ \psi\|_{W^{1,p}} \le \|\partial \psi\|_\infty \|f\|_{W^{1,p}}$.*

Proof. (i) Problem 13.6. (ii) Take limits in (13.2) using Hölder's inequality. If $g \in W_c^{k,q}(U)$ only the case $q = \infty$ is of interest which follows from dominated convergence. (iii) First of all note that if $\phi, \varphi \in C_c^\infty(U)$, then $\phi\varphi \in C_c^\infty(U)$ and hence using the ordinary product rule for smooth functions and rearranging (13.2) with $\varphi \to \phi\varphi$ shows $\phi f \in W_c^{k,p}(U)$. Hence (13.5) with $g \to g\varphi \in W_c^{k,q}$ shows

$$\int_U g f (\partial_j \varphi) d^n x = -\int_U \big((\partial_j f) g + f (\partial_j g)\big) \varphi d^n x,$$

that is, the weak derivatives of $f \cdot g$ are given by the product rule and that they are in $L^r(U)$ follows from the generalized Hölder inequality (10.19). (iv) By a slight abuse of notation we will consider the vector-valued function $f = (f_1, \dots, f_m)$. Take a smooth sequence $f_n \to f$ in $W^{1,1}_{loc}(U)$ (e.g. by mollification) and let $V \subseteq U$ be relatively compact. Then $\|g(f) - g(f_n)\|_{L^1(V)} \leq \|\partial g\|_\infty \|f - f_n\|_{L^1(V)}$ by the mean value theorem. Moreover,

$$\begin{aligned}\|(\partial g)(f)\partial_j f - g(\partial)(f_n)\partial_j f_n\|_{L^1(V)} \leq & \|\partial g\|_\infty \|\partial_j f - \partial_j f_n\|_{L^1(V)} \\ & + \|((\partial g)(f) - (\partial g)(f_n))\partial_j f\|_{L^1(V)},\end{aligned}$$

where the first norm tends to zero by assumption and the second by dominated convergence after passing to a subsequence wich converges a.e. Hence by completeness of $W^{1,1}(V)$ the first part of the claim follows. The second follows from $|g(x)| \leq \|\partial g\|_\infty |x|$. (v) The weak derivative can be computed by approximation by smooth functions from the version for smooth functions as in the previous item. The claim about the norms follows from the change of variables formula for integrals. □

Finally we look at a characterization of weak derivatives in terms of difference quotients. To this end recall the translation operator $T_a f(x) = f(x - a)$ from (10.21). In addition we set

$$D^a f := \frac{T_a f - f}{|a|}.$$

and $D^\varepsilon_j = D_{\varepsilon \delta^j}$ for the difference quotient in the direction of the j'th coordinate axis.

Lemma 13.5. *Suppose $1 \leq p \leq \infty$ and $f \in W^{1,p}(U)$. Then*

$$\|T_a f - f\|_{L^p(V)} \leq |a| \|\partial f\|_{L^p(U)} \tag{13.7}$$

for every $V \subseteq U$ and all $a \in \mathbb{R}^n$ with $|a| < \operatorname{dist}(V, \partial U)$.

Conversely, let $1 < p \leq \infty$ and assume there is a constant C such that for every relatively compact set V with $\overline{V} \subseteq U$ we have $f \in L^p(V)$ and

$$\|D^\varepsilon_j f\|_{L^p(V)} \leq C \tag{13.8}$$

for all $1 \leq j \leq n$. Then $u \in W^{1,p}(V)$ with $\|\partial_j f\|_{L^p(U)} \leq C$.

Proof. By Theorem 13.3 we can assume that f is smooth without loss of generality and hence

$$|f(x + a) - f(x)| \leq |a| \int_0^1 |\partial f(x + t\, a)| dt.$$

Now integrating over V and employing Jensen's inequality further gives

$$\begin{aligned}\|T_a f - f\|^p_{L^p(V)} &\le |a|^p \int_V \left| \int_0^1 |\partial f(x + t\,a)| dt \right|^p d^n x \\ &\le |a|^p \int_V \int_0^1 |\partial f(x + t\,a)|^p dt\, d^n x \le |a|^p \|\partial f\|^p_{L^p(U)},\end{aligned}$$

for $1 \le p < \infty$. For the case $p = \infty$ see Problem 13.10.

Conversely, note that for $f \in L^p(V) \subset L^1(V)$ and $\varphi \in C^\infty_c(V)$ we have

$$\int_V f\big(D_j^\varepsilon \varphi\big) d^n x = -\int_V \big(D_j^{-\varepsilon} f\big)\varphi d^n x.$$

Moreover, by our assumption we can invoke Lemma 4.34 (with $X = L^{p'}(V)$, $\frac{1}{p} + \frac{1}{p'} = 1$) to get a sequence $\varepsilon_m \to 0$ such that $D_j^{-\varepsilon_m} \overset{*}{\rightharpoonup} g_j$. Hence taking the limit $m \to \infty$ in the above identity we obtain

$$\int_V f\big(\partial_j \varphi\big) d^n x = -\int_V g_j \varphi d^n x. \qquad \square$$

Problem 13.1. *Consider $f(x) = \sqrt{x}$, $U = (0,1)$. Compute the weak derivative. For which p is $f \in W^{1,p}(U)$?*

Problem* 13.2. *Show that f is weakly differentiable in the interval $(0,1)$ if and only if $f(x) = f(0) + \int_0^x h(t)dt$ is absolutely continuous and $f' = h$ in this case. (Hint: Lemma 10.22.)*

Problem* 13.3. *Consider $U := B_1(0) \subset \mathbb{R}^n$ and $f(x) = \tilde{f}(|x|)$ with $\tilde{f} \in C^1(0,1]$. Then $f \in W^{1,p}_{loc}(B_1(0) \setminus \{0\})$ and*

$$\partial_j f(x) = \tilde{f}'(|x|)\frac{x_j}{|x|}.$$

Show that if $\lim_{r\to 0} r^{n-1}\tilde{f}(r) = 0$ then $f \in W^{1,p}(B_1(0))$ if and only if $\tilde{f}, \tilde{f}' \in L^p((0,1), r^{n-1}dr)$.

Conclude that for $f(x) := |x|^{-\alpha}$, $\alpha > 0$, we have $f \in W^{1,p}(B_1(0))$ with

$$\partial_j f(x) = -\frac{\alpha x_j}{|x|^{\alpha+2}}$$

provided $\alpha < \frac{n-p}{p}$. (Hint: Use integration by parts on a domain which excludes $B_\varepsilon(0)$ and let $\varepsilon \to 0$.)

Problem* 13.4. *Show that for $\varphi \in C^\infty_c(\mathbb{R}^n)$ we have*

$$\varphi(0) = -\frac{1}{S_n}\int_{\mathbb{R}^n} \frac{x}{|x|^n} \cdot \partial\varphi(x) d^n x$$

Hence this weak derivative cannot be interpreted as a function. (Hint: Start with $\varphi(0) = -\int_0^\infty \big(\frac{d}{dr}\varphi(r\omega)\big)dr = -\int_0^\infty \partial\varphi(r\omega)\cdot\omega\, dr$ and integrate with respect to ω over the unit sphere S^{n-1}; cf. Lemma 9.17.)

Problem 13.5. *Suppose $V \subseteq U$ is nonempty and open. Then $f \in W^{k,p}(U)$ implies $f \in W^{k,p}(V)$.*

Problem* 13.6. *Show Lemma 13.4 (i).*

Problem* 13.7. *Suppose $f \in W^{k,p}(U)$ and $h \in C_b^k(U)$. Then $h \cdot f \in W^{k,p}(U)$ and we have* ***Leibniz' rule***

$$\partial_\alpha(h \cdot f) = \sum_{\beta \le \alpha} \binom{\alpha}{\beta} (\partial_\beta h)(\partial_{\alpha-\beta} f), \tag{13.9}$$

where $\binom{\alpha}{\beta} := \frac{\alpha!}{\beta!(\alpha-\beta)!}$, $\alpha! := \prod_{j=1}^m (\alpha_j!)$, and $\beta \le \alpha$ means $\beta_j \le \alpha_j$ for $1 \le j \le m$.

Problem 13.8. *Suppose $f \in W^{1,p}(U)$ satisfies $\partial_j f = 0$ for $1 \le j \le n$. Show that f is constant if U is connected.*

Problem 13.9. *Suppose $f \in W^{1,p}(U)$. Show that $|f| \in W^{1,p}(U)$ with*

$$\partial_j |f|(x) = \frac{\operatorname{Re}(f(x))}{|f(x)|} \partial_j \operatorname{Re}(f(x)) + \frac{\operatorname{Im}(f(x))}{|f(x)|} \partial_j \operatorname{Im}(f(x)),$$

In particular $\big|\partial_j |f|(x)\big| \le |\partial_j f(x)|$. Moreover, if f is real-valued we also have $f_\pm := \max(0, \pm f) \in W^{1,p}(U)$ with

$$\partial_j f_\pm(x) = \begin{cases} \pm \partial_j f(x), & \pm f(x) > 0, \\ 0, & \text{else}, \end{cases} \qquad \partial_j |f|(x) = \begin{cases} \partial_j f(x), & f(x) > 0, \\ -\partial_j f(x), & f(x) < 0, \\ 0, & \text{else}. \end{cases}$$

(Hint: $|f| = \lim_{\varepsilon \to 0} g_\varepsilon(\operatorname{Re}(f), \operatorname{Im}(f))$ with $g_\varepsilon(x,y) = \sqrt{x^2 + y^2 + \varepsilon^2} - \varepsilon$.)

Problem* 13.10. *Suppose that $U \subseteq \mathbb{R}^n$ is open and convex. Show that functions in $W^{1,\infty}(U)$ are Lipschitz continuous. In fact, there is a continuous embedding $W^{1,\infty}(U) \hookrightarrow C_b^{0,1}(U)$. (Hint: Start by mollifying f. Now use Problem 10.25 and the fact that a.e. point is a Lebesgue point.)*

Problem* 13.11. *Show $W_0^{k,p}(\mathbb{R}^n) = W^{k,p}(\mathbb{R}^n)$ for $1 \le p < \infty$. (Hint: Consider $f\phi_m$ with ϕ the standard mollifier.)*

Problem 13.12. *Show that the second part of Lemma 13.5 fails for $p = 1$.*

Problem 13.13. *Let $f \in L^p(U)$, $1 < p \le \infty$. Show that $f \in W^{1,p}(U)$ if and only if there exists a constant such that*

$$\left| \int_U f\big(\partial_j \varphi\big) d^n x \right| \le C \|\varphi\|_{p'}, \quad 1 \le j \le n,$$

where p' is the dual index, $\frac{1}{p} + \frac{1}{p'} = 1$.

13.2. Extension and trace operators

To proceed further we will need to be able to extend a given function beyond its original domain U. As already pointed out before, simply setting it equal to zero on $\mathbb{R}^n \setminus U$ will in general create a non-differentiable singularity along the boundary. Moreover, consider for example an annulus in $\mathbb{R}^2$ and cut it along a line, in polar coordinates $U := \{(r\cos(\varphi), r\sin(\varphi)) | 1 < r < 2,\ 0 < \varphi < 2\pi\}$. Then the function $f(x) = \varphi$ is in $W^{1,p}(U) \cap C^\infty(U)$ but, as its limits along the cut do not match, there is no way of extending it to $W^{1,p}(\mathbb{R}^2)$. Hence not every domain has the property that we can extend functions from $W^{1,p}(U)$ to $W^{1,p}(\mathbb{R}^2)$.

Of course such problems do not arise if $f \in W^{1,p}_0(U)$ since we can simply extend f to a function on $\bar{f} \in W^{1,p}(\mathbb{R}^n)$ by setting $\bar{f}(x) = 0$ for $x \in \mathbb{R}^n \setminus U$ (to see this note that if a sequence $f_n \in C^\infty_c(U)$ converges to to f in $W^{1,p}(U)$, then the corresponding sequence $\bar{f}_n \in C^\infty_c(\mathbb{R}^n)$ converges to to $\bar{f}$ in $W^{1,p}(\mathbb{R}^n)$).

We will say that a domain $U \subseteq \mathbb{R}^n$ has the **extension property** if for all $1 \le p \le \infty$ there is an extension operator $E : W^{1,p}(U) \to W^{1,p}(\mathbb{R}^n)$ such that

- E is bounded, i.e., $\|Ef\|_{W^{1,p}(\mathbb{R}^n)} \le C_{U,p}\|f\|_{W^{1,p}(U)}$ and
- $Ef\big|_U = f$.

We begin by showing that if the boundary is a hyperplane, we can do the extension by a simple reflection. To this end consider the reflection $x^\star := (x_1, \dots, x_{n-1}, -x_n)$ which is an involution on $\mathbb{R}^n$. Accordingly we define $f^\star(x) := f(x^\star)$ for functions defined on a domain U which is symmetric with respect to reflection, that is, $U^\star = U$. Also let us write $U_\pm := \{x \in U | \pm x_n > 0\}$.

Lemma 13.6. *Let $U \subseteq \mathbb{R}^n$ be symmetric with respect to reflection and $1 \le p \le \infty$. If $f \in W^{1,p}(U_+)$ then the symmetric extension $f^\star \in W^{1,p}(U)$ satisfies $\|f\|_{W^{1,p}(U)} = 2\|f\|_{W^{1,p}(U_+)}$. Moreover,*

$$(\partial_j f^\star) = \begin{cases} (\partial_j f)^\star, & 1 \le j < n, \\ \operatorname{sign}(x_n)(\partial_n f)^\star, & j = n. \end{cases} \tag{13.10}$$

Proof. It suffices to compute the weak derivatives. We start with $1 \le j < n$ and

$$\int_U u^* \partial_j \varphi d^n x = \int_{U_+} u\, \partial_j \varphi^\# d^n x,$$

where $\varphi^\#(x', x_n) = \varphi(x', x_n) + \varphi(x', -x_n)$ with $x' = (x_1, \dots, x_{n-1})$. Since $\varphi^\#$ is not compactly supported in U_+ we use a cutoff function $\eta_\varepsilon(x) = \eta(x_n/\varepsilon)$, where $\eta \in C^\infty(\mathbb{R}, [0,1])$ satisfies $\eta(r) = 0$ for $r \le \frac{1}{2}$ and $\eta(r) = 1$

for $r \geq 1$ (e.g., integrate and shift the standard mollifier to obtain such a function). Then

$$\int_U u^* \partial_j \varphi d^n x = \lim_{\varepsilon \to 0} \int_{U_+} u\, \partial_j (\eta_\varepsilon \varphi^\#) d^n x = \lim_{\varepsilon \to 0} \int_{U_+} (\partial_j u) \eta_\varepsilon \varphi^\# d^n x$$
$$= \int_{U_+} (\partial_j u) \varphi^\# d^n x = \int_U (\partial_j u)^* \varphi \, d^n x$$

for $1 \leq j < n$. For $j = n$ we proceed similarly,

$$\int_U u^* \partial_n \varphi d^n x = \int_{U_+} u\, \partial_n \varphi^\sharp d^n x,$$

where $\varphi^\sharp(x', x_n) = \varphi(x', x_n) - \varphi(x', -x_n)$. Note that $\varphi^\sharp(x', 0) = 0$ and hence $|\varphi^\sharp(x', x_n)| \leq L\, x_n$ on U_+. Using this last estimate for $\partial_n(\eta_\varepsilon \varphi^\sharp) = \partial_n(\eta_\varepsilon)\varphi^\sharp + \eta_\varepsilon \partial_n \varphi^\sharp$ we obtain as before

$$\int_U u^* \partial_n \varphi d^n x = \lim_{\varepsilon \to 0} \int_{U_+} u\, \partial_n (\eta_\varepsilon \varphi^\sharp) d^n x = \lim_{\varepsilon \to 0} \int_{U_+} (\partial_n u) \eta_\varepsilon \varphi^\# d^n x$$
$$= \int_{U_+} (\partial_n u) \varphi^\# d^n x = \int_U \operatorname{sign}(x_n) (\partial_n u)^* \varphi \, d^n x,$$

which finishes the proof. □

Corollary 13.7. *$\mathbb{R}^n_+$ has the extension property. In fact, any rectangle (not necessarily bounded) Q has the extension property. Moreover, if U is diffeomorphic to a rectangle Q with a diffeomorphism satisfying $\psi \in C^1_b(Q, U)$, $\psi^{-1} \in C^1_b(U, Q)$, then U has the extension property.*

Proof. Given a rectangle use the above lemma to extend it along every hyperplane bounding the rectangle. Finally, use a smooth cut-off function (e.g. $\psi_\varepsilon * \chi_Q$ with ε smaller than the minimal side length of Q). The last claim follows from a change of variables, Lemma 13.4 (v). □

While this already covers a large number of interesting domains, note that it fails if we look for example at the exterior of a rectangle. So our next result shows that (maybe not too surprising), that it is the boundary which will play the crucial role. To this end we recall that U is said to have a C^1 boundary if around any point $x^0 \in \partial U$ we can find a C^1 diffeomorphism ψ which straightens out the boundary (cf. Section 9.4).

Lemma 13.8. *Suppose U has a bounded C^1 boundary, then U has the extension property.*

Proof. By compactness we can find a finite number of open sets $\{U_j\}_{j=1}^m$ covering the boundary and corresponding C^1 diffeomorphism $\psi_j : U_j \to Q_j$, where Q_j is a rectangle which is symmetric with respect to reflection.

Moreover, choose an open set U_0 such that $\overline{U_0} \subset U$ and a smooth partition of unity $\{h_k\}_{k=0}^l$ subordinate to this cover (Lemma B.31). Now split $f \in W^{k,p}(U)$ according to $\sum_k f_k$, where $f_k := h_k f$. Then $h_0 f$ can be extended to $\mathbb{R}^n$ by setting it equal to 0 outside U. Moreover, f_k can be mapped to $Q_{j,+}$ using ψ_j and extended to Q_j using the symmetric extension. Note that this extension has compact support and so has the pull back $\bar{f}_k$ to U_j; in particular, it can be extended to $\mathbb{R}^n$ by setting it equal to 0 outside U_j. By construction we have $\|\bar{f}_k\|_{W^{1,p}(U_j)} \le C_j \|f_k\|_{W^{1,p}(U)} \le C_{j,k}\|f\|_{W^{1,p}(U)}$ and hence $\bar{f} = \sum_k \bar{f}$ is the required extension. □

As a first application note that by mollifying an extension we see that we can approximate by functions which are smooth up to the boundary.

Lemma 13.9. *Suppose U has the extension property, then $C_c^\infty(\mathbb{R}^n)$ is dense in $W^{1,p}(U)$ for $1 \le p < \infty$.*

Next show that functions in $W^{1,p}$ have boundary values in L^p.

Theorem 13.10. *Suppose U has a bounded C^1 boundary, then there exists a bounded* ***trace operator***

$$T : W^{1,p}(U) \to L^p(\partial U) \tag{13.11}$$

which satisfies $Tf = f\big|_{\partial U}$ for $f \in C^1(\overline{U}) \cap W^{1,p}(U)$.

Proof. In the case $p = \infty$ we observe that, since U has the extension property, we can extend f to $W^{1,\infty}(\mathbb{R}^n)$ and hence u is Lipschitz continuous by Problem 13.10. In particular it is continuous up to the boundary.

So we can focus on the case $1 \le p < \infty$. As in the proof of Lemma 13.8, using a partition of unity and straightening out the boundary, we can reduce it to the case where $f \in C_c^1(\mathbb{R}^n)$ has compact support $\operatorname{supp}(f) \subset B$ such that $\partial U \cap B \subset \partial \mathbb{R}^n_+$. Then using the Gauss–Green theorem and assuming f real-valued without loss of generality we have (cf. Problem 13.9)

$$\begin{aligned}\int_{\partial U} |f|^p d^{n-1}x &= -\int_{B_+} (|f|^p)_{x_n} d^n x = -p \int_{B_+} \operatorname{sign}(f)|f|^{p-1}(\partial_n f) d^n x \\ &\le p\|f\|_p^{p-1}\|\partial f\|_p,\end{aligned}$$

where we have used Hölders inequality in the last step. Hence the trace operator defined on $C^1(\overline{U}) \cap W^{1,p}(U)$ is bounded and since the latter set is dense, there is a unique extension to all of $W^{1,p}(U)$. □

As an application we can extend the Gauss–Green theorem and integration by parts to $W^{1,p}$ vector fields.

Lemma 13.11. *If U be a bounded C^1 domain in $\mathbb{R}^n$ and $u \in W^{1,p}(U, \mathbb{R}^n)$ is a vector field. Then the Gauss–Green formula* (9.61) *holds if the boundary*

values of u are understood as traces as in the previous theorem. Moreover, the integration by parts formula (9.63) *also holds for $f \in W^{1,p}(U)$, $f \in W^{1,q}(U)$ with $\frac{1}{p} + \frac{1}{q} = 1$.*

Proof. Since U has the extension property, we can extend u to $W^{1,\infty}(\mathbb{R}^n, \mathbb{R}^n)$ and hence u is Lipschitz continuous by Problem 13.10. Moreover, consider the mollification $u_\varepsilon = \phi_\varepsilon * u$ and apply the Gauss–Green theorem to u_ε. Now let $\varepsilon \to 0$ and observe that the right-hand side converges since $u_\varepsilon \to u$ uniformly and the left-hand side converges by dominated convergence since $\partial_j u_\varepsilon \to \partial_j u$ pointwise and $\|\partial_j u_\varepsilon\|_\infty \leq \|\partial_j u\|_\infty$. The integration by parts formula follows from the Gauss–Green theorem applied to the product fg and employing the product rule. □

Finally we identify the kernel of the trace operator.

Lemma 13.12. *If U be a bounded C^1 domain in $\mathbb{R}^n$ then the kernel of the trace operator is given by* $\operatorname{Ker}(T) = W_0^{1,p}(U)$.

Proof. Clearly $W_0^{1,p}(U) \subseteq \operatorname{Ker}(T)$. Conversely it suffices to show that $u \in \operatorname{Ker}(T)$ can be approximated by functions from $C_c^\infty(U)$. Using a partition of unity as in the proof of Lemma 13.8 we can assume $U = Q_+$, where Q is a rectangle which is symmetric with respect to reflection. Setting

$$\bar{u}(x) = \begin{cases} u(x), & x \in Q_+, \\ 0, & x \in Q_-, \end{cases}$$

integration by parts using the previous lemma shows

$$\int_Q \bar{u} \partial_j \varphi \, d^n x = \int_{Q_+} u \partial_j \varphi \, d^n x = - \int_{Q_+} (\partial_j u) \varphi \, d^n x$$

that $\bar{u} \in W^{1,p}(Q)$ with $\partial_j \bar{u}(x) = \partial_j u(x)$ for $x \in Q_+$ and $\partial_j \bar{u}(x) = 0$ else. Now consider $u_\varepsilon(x) = (\phi_\varepsilon * \bar{u})(x - \varepsilon \delta^n)$ with ϕ the standard mollifier. This is the required sequence by Lemma 10.19 and Problem 10.15. □

Problem 13.14. *Show that $f \in W_0^{k,p}(U)$ can be extended to a function $\bar{f} \in W_0^{k,p}(\mathbb{R}^n)$ by setting it equal to zero outside U. In this case the weak derivatives of $\bar{f}$ are obtained by setting the weak derivatives of f equal to zero outside U.*

Problem 13.15. *Suppose for each $x \in U$ there is an open neighborhood $V(x) \subseteq U$ such that $f \in W^{k,p}(V(x))$. Then $f \in W_{loc}^{k,p}(U)$. Moreover, if $\|f\|_{W^{k,p}(V)} \leq C$ for every relatively compact set $V \subseteq U$, then $f \in W^{k,p}(U)$.*

Problem 13.16. *Let $1 \leq p < \infty$. Show that $Tf = f\big|_{\partial U}$ for $f \in C(\overline{U}) \subseteq L^p(U)$ is unbounded (and hence has no meaningful extension to $L^p(U)$).*

13.3. Embedding theorems

The example in Problem 13.3 shows that functions in $W^{1,p}$ are not necessarily continuous (unless $n = 1$). This raises the question in what sense a function from $W^{1,p}$ is better than a function from L^p? For example, is it in L^q for some q other than p.

Theorem 13.13 (Gagliardo–Nierenberg–Sobolev)**.** *Suppose $1 \leq p < n$ and $U \subseteq \mathbb{R}^n$ is open. Then there is a continuous embedding $f \in W_0^{1,p}(U) \hookrightarrow L^q(U)$ for all $p \leq q \leq p^*$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}$. Moreover,*

$$\|f\|_{p^*} \leq \frac{p(n-1)}{(n-p)} \prod_{j=1}^{n} \|\partial_j f\|_p^{1/n} \leq \frac{p(n-1)}{n(n-p)} \sum_{j=1}^{n} \|\partial_j f\|_p. \tag{13.12}$$

Proof. It suffices to prove the case $q = p^*$ since the rest follows from interpolation (Problem 10.11). Moreover, by density it suffices to prove the inequality for $f \in C_c^\infty(\mathbb{R}^n)$. In this respect note that if you have a sequence $f_n \in C_c^\infty(U)$ which converges to some f in $W_0^{1,p}(U)$, then by (13.12) this sequence will also converge in $L^{p^*}(U)$ and by considering pointwise convergent subsequences both limits agree.

We start with the case $p = 1$ and observe

$$|f(x)| = \left| \int_{-\infty}^{x_1} \partial_1 f(r, \tilde{x}_1) dr \right| \leq \int_{-\infty}^{\infty} |\partial_1 f(r, \tilde{x}_1)| dr,$$

where we denote by $\tilde{x}_j = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)$ the vector obtained from x with the j'th component dropped. Denote by $f_1(\tilde{x}_1)$ the right-hand side of the above inequality and apply the same reasoning to the other coordinate directions to obtain

$$|f(x)|^n \leq \prod_{j=1}^{n} f_j(\tilde{x}_j).$$

Now we claim that if $f_j(\tilde{x}_j) \in L^1(\mathbb{R}^{n-1})$ then

$$\Big\| \prod_{j=1}^{n} f_j(\tilde{x}_j)^{\frac{1}{n-1}} \Big\|_{L^1(\mathbb{R}^n)} \leq \prod_{j=1}^{n} \|f_j(\tilde{x}_j)\|_{L^1(\mathbb{R}^{n-1})}^{\frac{1}{n-1}}.$$

For $n = 2$ this is just Fubini and hence we can use induction. To this end fix the last coordinate x_{n+1} and apply Hölder's inequality and the induction

hypothesis to obtain

$$\int_{\mathbb{R}^n} \prod_{j=1}^{n+1} |f_j(\tilde{x}_j)|^{\frac{1}{n}} d^n \le \|f_{n+1}^{1/n}\|_{L^n(\mathbb{R}^n)} \Big\| \prod_{j=1}^{n} |f_j(\tilde{x}_j)|^{\frac{1}{n}} \Big\|_{L^{n/(n-1)}(\mathbb{R}^n)}$$
$$= \|f_{n+1}\|_{L^1(\mathbb{R}^n)}^{1/n} \Big\| \prod_{j=1}^{n} |f_j(\tilde{x}_j)|^{\frac{1}{n-1}} \Big\|_{L^1(\mathbb{R}^n)}^{1-\frac{1}{n}} = \|f_{n+1}\|_{L^1(\mathbb{R}^n)}^{1/n} \prod_{j=1}^{n} \|f_j(\tilde{x}_j)\|_{L^1(\mathbb{R}^{n-1})}^{1/n}.$$

Now integrate this inequality with respect to the missing variable x_n and use the iterated Hölder inequality (Problem 10.9) to obtain the claim (the second inequality is just the inequality of arithmetic and geometric means).

Moreover, applying this to our situation is precisely (13.12) for the case $p = 1$. To see the general case let $f \in C_c^\infty(\mathbb{R}^n)$ and apply the case $p = 1$ to $f \to |f|^\gamma$ for $\gamma > 1$ to be determined. Then

$$\Big(\int_{\mathbb{R}^n} |f|^{\frac{\gamma n}{n-1}} d^n x \Big)^{\frac{n-1}{n}} \le \prod_{j=1}^{n} \Big(\int_{\mathbb{R}^n} \big|\partial_j |f|^\gamma\big| d^n x \Big)^{1/n} \tag{13.13}$$
$$= \gamma \prod_{j=1}^{n} \Big(\int_{\mathbb{R}^n} |f|^{\gamma-1} |\partial_j f| d^n x \Big)^{1/n} \le \gamma \||f|^{\gamma-1}\|_{p/(p-1)} \prod_{j=1}^{n} \|\partial_j f\|_p^{1/n},$$

where we have used Hölder in the last step. Now we choose $\gamma := \frac{p(n-1)}{n-p} > 1$ such that $\frac{\gamma n}{n-1} = \frac{(\gamma-1)p}{p-1} = p^*$, which gives the general case. □

Note that a simple scaling argument (Problem 13.18) shows that (13.12) can only hold for p^*. Furthermore, using an extension operator this result also extends to $W^{1,p}(U)$:

Corollary 13.14. *Suppose U has the extension property and $1 \le p < n$, then there is a continuous embedding $f \in W^{1,p}(U) \hookrightarrow L^q(U)$ for every $p \le q \le p^*$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}$.*

Note that involving the extension operator implies that we need the full $W^{1,p}$ norm to bound the L^{p*} norm. A constant functions shows that indeed an inequality involving only the derivatives on the right-hand side cannot hold on bounded domains (cf. also Theorem 13.25).

In the borderline case $p = n$ one has $p^* = \infty$, however, the example in Problem 13.19 shows that functions in $W^{1,n}$ can be unbounded if $n > 1$ (for $n = 1$ functions are bounded — see Problem 13.17). Nevertheless, we have at least the following result:

Lemma 13.15. *Suppose $p = n$ and $U \subseteq \mathbb{R}^n$ is open. Then there is a continuous embedding $f \in W_0^{1,p}(U) \hookrightarrow L^q(U)$ for every $n \le q < \infty$.*

Proof. As before it suffices to establish $\|f\|_q \le C\|f\|_{W^{1,p}}$ for $f \in C_c^\infty(\mathbb{R}^n)$. To this end we employ (13.13) with $p = n$ implying

$$\|f\|_{\gamma n/(n-1)}^\gamma \le \gamma \|f\|_{(\gamma-1)n/(n-1)}^{\gamma-1} \prod_{j=1}^n \|\partial_j f\|_n^{1/n} \le \frac{\gamma}{n} \|f\|_{(\gamma-1)n/(n-1)}^{\gamma-1} \sum_{j=1}^n \|\partial_j f\|_n.$$

Using (1.24) this gives

$$\|f\|_{\gamma n/(n-1)} \le C\Big(\|f\|_{(\gamma-1)n/(n-1)} + \sum_{j=1}^n \|\partial_j f\|_n\Big).$$

Now choosing $\gamma = n$ we get $\|f\|_{n^2/(n-1)} \le C\|f\|_{W^{1,n}}$ and by interpolation (Problem 10.11) the claim holds for $q \in [n, n\frac{n}{n-1}]$. So we can choose $\gamma = n+1$ to get the claim for $q \in [n, n+2+\frac{2}{n-1}]$ and iterating this procedure finally establishes the result. □

Corollary 13.16. *Suppose U has the extension property and $p = n$, then there is a continuous embedding $f \in W^{1,p}(U) \hookrightarrow L^q(U)$ for every $n \le q < \infty$.*

In the case $p > n$ functions from $W^{1,p}$ will be continuous (in the sense that there is a continuous representative). In fact, they will even be bounded Hölder continuous functions and hence are continuous up to the boundary (cf. Theorem 1.22 and the discussion after this theorem).

Theorem 13.17 (Morrey)**.** *Suppose $n < p \le \infty$ and $U \subseteq \mathbb{R}^n$ is open. There is a continuous embedding $f \in W_0^{1,p}(U) \hookrightarrow C_0^{0,\gamma}(\overline{U})$, where $\gamma = 1 - \frac{n}{p}$. Here $C_0^{0,\gamma}(\overline{U}) := C_b^{0,\gamma}(\overline{U}) \cap C_0(U)$ is the space of Hölder continuous functions vanishing at the boundary.*

Proof. In the case $p = \infty$ this follows from Problem 13.10. Hence we can assume $n < p < \infty$. Moreover, as before, by density we can assume $f \in C_c^\infty(\mathbb{R}^n)$.

We begin by considering a cube Q of side length r containing 0. Then, for $x \in Q$ and $\bar f = r^{-n} \int_Q f(x) d^n x$ we have

$$\bar f - f(0) = r^{-n} \int_Q \big(f(x) - f(0)\big) d^n x = r^{-n} \int_Q \int_0^1 \frac{d}{dt} f(tx) dt\, d^n x$$

and hence

$$\begin{aligned}|\bar f - f(0)| &\le r^{-n}\int_Q\int_0^1 |\partial f(tx)||x|dt\,d^nx \le r^{1-n}\int_Q\int_0^1 |\partial f(tx)|dt\,d^nx\\ &= r^{1-n}\int_0^1\int_{tQ}|\partial f(y)|\frac{d^ny}{t^n}\,dt \le r^{1-n}\int_0^1 \|\partial f\|_{L^p(tQ)}\frac{|tQ|^{1-1/p}}{t^n}\,dt\\ &\le \frac{r^\gamma}{\gamma}\|\partial f\|_{L^p(Q)}\end{aligned}$$

where we have used Hölder's inequality in the fourth step. By a translation this gives

$$|\bar f - f(x)| \le \frac{r^\gamma}{\gamma}\|\partial f\|_{L^p(Q)}$$

for any cube Q of side length r containing x and combining the corresponding estimates for two points we obtain

$$|u(x) - u(y)| \le \frac{2\|\partial f\|_{L^p(Q)}}{\gamma}|x-y|^\gamma \tag{13.14}$$

for any cube containing both x and y (note that we can chose the side length of Q to be $r = \max_{1\le j\le n}|x_j - y_j| \le |x-y|$). Since we can of course replace $L^p(Q)$ by $L^p(\mathbb{R}^n)$ we get Hölder continuity of f. Moreover, taking a cube of side length $r = 1$ containing x we get (using again Hölder)

$$|f(x)| \le |\bar f| + \frac{2\|\partial f\|_{L^p(Q)}}{\gamma} \le \|f\|_{L^p(Q)} + \frac{2\|\partial f\|_{L^p(Q)}}{\gamma} \le C\|f\|_{W^{1,p}(\mathbb{R}^n)}$$

establishing the theorem. □

Corollary 13.18. *Suppose U has the extension property and $n < p \le \infty$, then there is a continuous embedding $f \in W^{1,p}(U) \hookrightarrow C_b^{0,\gamma}(\overline{U})$, where $\gamma = 1 - \frac{n}{p}$.*

Example 13.3. The example from Problem 13.20 shows that for a domain with a cusp, functions form $W^{1,p}$ might be unbounded (and hence in particular not in $C^{1,\gamma}$) even for $p > n$. ⋄

Example 13.4. It can be shown that for $p = \infty$ this embedding is surjective (Problem 13.21). However, for $n < p < \infty$ this is not the case. To see this consider the Takagi function (Problem 11.37) which is in $C^{1,\gamma}[0,1]$ for every $\gamma < 1$ but not absolutely continuous (not even of bounded variation) and hence not in $W^{1,p}(0,1)$ for any $1 \le p \le \infty$. Note that this example immediately extends to higher dimensions by considering $f(x) = b(x_1)$ on the unit cube. ⋄

As a consequence of the proof we also get that for $n < p$ Sobolev functions are differentiable a.e.

Lemma 13.19. *Suppose $n < p \leq \infty$ and $U \subseteq \mathbb{R}^n$ is open. Then $f \in W^{1,p}_{loc}(U)$ is differentiable a.e. and the derivative equals the weak derivative.*

Proof. Since $W^{1,\infty}_{loc} \subseteq W^{1,p}_{loc}$ for any $p < \infty$ we can assume $n < p < \infty$. Let $x \in U$ be an L^p Lebesgue point (Problem 11.11) of the gradient, that is,

$$\lim_{r\to 0} \frac{1}{|Q_r(x)|} \int_{Q_r(x)} |\partial f(x) - \partial f(y)|^p d^n y \to 0,$$

where $Q_r(x)$ is a cube of side length r containing x. Now let $y \in Q_r(x)$ and $r = |y - x|$ (by shrinking the cube w.l.o.g.). Then replacing $f(y) \to f(y) - f(x) - \partial f(x) \cdot (y - x)$ in (13.14) we obtain

$$\begin{aligned}\big|f(y) - f(x) - \partial f(x) \cdot (y - x)\big| &\leq \frac{2}{\gamma}|x - y|^\gamma \left(\int_{Q_r(x)} |\partial f(x) - \partial f(z)|^p d^n z\right)^{1/p} \\ &= \frac{2}{\gamma}|x - y| \left(\frac{1}{|Q_r(x)|}\int_{Q_r(x)} |\partial f(x) - \partial f(z)|^p d^n z\right)^{1/p}\end{aligned}$$

and, since x is a L^p Lebesgue point of the gradient, the right-hand side is $o(|x - y|)$, that is, f is differentiable at x and its gradient equals its weak gradient. □

Note that since by Problem 13.10 every locally Lipschitz continuous function is locally $W^{1,\infty}$ we obtain as an immediate consequence:

Theorem 13.20 (Rademacher)**.** *Every locally Lipschitz continuous function is differentiable almost everywhere.*

So far we have only looked at first order derivatives. However, we can also cover the case of higher order derivatives by repeatedly applying the above results to the fact that $\partial_j f \in W^{k-1,p}(U)$ for $f \in W^{k,p}(U)$.

Theorem 13.21. *Suppose $U \subseteq \mathbb{R}^n$ is open and $1 \leq p \leq \infty$. There are continuous embeddings*

$$\begin{aligned} W^{k,p}_0(U) &\hookrightarrow L^q(U), \quad q \in [p, p^*_k] \text{ if } \frac{1}{p^*_k} = \frac{1}{p} - \frac{k}{n} > 0, \\ W^{k,p}_0(U) &\hookrightarrow L^q(U), \quad q \in [p, \infty) \text{ if } \frac{1}{p} = \frac{k}{n}, \\ W^{k,p}_0(U) &\hookrightarrow C^{k-l-1,\gamma}_0(\overline{U}), \quad l = \lfloor \frac{n}{p} \rfloor, \begin{cases} \gamma = 1 - \frac{n}{p} + l, & \frac{n}{p} \notin \mathbb{N}_0, \\ \gamma \in [0, 1), & \frac{n}{p} \in \mathbb{N}_0, \end{cases} \text{ if } \frac{1}{p} < \frac{k}{n}. \end{aligned}$$

If in addition $U \subseteq \mathbb{R}^n$ has the extension property, then there are continuous embeddings

$$W^{k,p}(U) \hookrightarrow L^q(U), \quad q \in [p, p_k^*] \text{ if } \frac{1}{p_k^*} = \frac{1}{p} - \frac{k}{n} > 0,$$
$$W^{k,p}(U) \hookrightarrow L^q(U), \quad q \in [p, \infty) \text{ if } \frac{1}{p} = \frac{k}{n},$$
$$W^{k,p}(U) \hookrightarrow C_b^{k-l-1,\gamma}(\overline{U}), \quad l = \lfloor \frac{n}{p} \rfloor, \begin{cases} \gamma = 1 - \frac{n}{p} + l, & \frac{n}{p} \notin \mathbb{N}_0, \\ \gamma \in [0,1), & \frac{n}{p} \in \mathbb{N}_0, \end{cases} \text{ if } \frac{1}{p} < \frac{k}{n}.$$

Proof. If $\frac{1}{p} > \frac{k}{n}$ we apply Theorem 13.13 to successively conclude $\|\partial^\alpha f\|_{L^{p_j^*}} \leq C\|f\|_{W_0^{k,p}}$ for $|\alpha| \leq k - j$ for $j = 1, \ldots, k$. If $\frac{1}{p} = \frac{k}{n}$ we proceed in the same way but use Lemma 13.15 in the last step. If $\frac{1}{p} < \frac{k}{n}$ we first apply Theorem 13.13 l times as before. If $\frac{n}{p}$ is not an integer we then apply Theorem 13.17 to conclude $\|\partial^\alpha f\|_{C_0^{0,\gamma}} \leq C\|f\|_{W_0^{k,p}}$ for $|\alpha| \leq k - l - 1$. If $\frac{n}{p}$ is an integer, we apply Theorem 13.13 $l - 1$ times and then Lemma 13.15 once to conclude $\|\partial^\alpha f\|_{L^q} \leq C\|f\|_{W_0^{k,p}}$ for any $q \in [p, \infty)$ for $|\alpha| \leq k - l$. Hence we can apply Theorem 13.17 to conclude $\|\partial^\alpha f\|_{C_0^{0,\gamma}} \leq C\|f\|_{W_0^{k,p}}$ for any $\gamma \in [0, 1)$ for $|\alpha| \leq k - l - 1$.

The second part follows analogously using the corresponding results for domains with the extension property. □

In fact, for $q \in [p, p^*)$ the embedding is even compact.

Theorem 13.22 (Rellich–Kondrachov)**.** *Suppose $U \subseteq \mathbb{R}^n$ is bounded and has the extension property. Then for $1 \leq p \leq \infty$ there are compact embeddings*

$$W^{1,p}(U) \hookrightarrow L^q(U), \quad q \in [p, p^*), \quad \frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}, \text{ if } p \leq n,$$
$$W^{1,p}(U) \hookrightarrow C(\overline{U}), \text{ if } p > n.$$

Proof. The case $p > n$ follows from $W^{1,p}(U) \hookrightarrow C_b^{0,\gamma}(\overline{U}) \hookrightarrow C(\overline{U})$, where the first embedding is continuous by Corollary 13.18 and the second is compact by the Arzelà–Ascoli theorem (Theorem B.39; see also Theorem 1.23). Next consider the case $p < n$. Let $F \subset W^{1,p}(U)$ be a bounded subset. Using an extension operator we can assume that $F \subset W^{1,p}(\mathbb{R}^n)$ with $\operatorname{supp}(f) \subseteq V$ for all $f \in F$ and some fixed set V. By the first part of Lemma 13.5 (applied on $\mathbb{R}^n$) we have

$$\|T_a f - f\|_p \leq |a| \|\partial f\|_p$$

and using the interpolation inequality from Problem 10.11 and Theorem 13.21 we have

$$\|T_a f - f\|_q \le \|T_a f - f\|_p^{1-\theta} \|T_a f - f\|_{p^*}^{\theta} \le |a|^{1-\theta} \|\partial f\|_p^{1-\theta} C^{\theta} \|f\|_{1,p}^{\theta},$$

where $\frac{1}{q} = \frac{1-\theta}{p} + \frac{\theta}{p^*}$, $\theta \in (0,1)$. Hence F is relatively compact by Theorem 10.15. In the case $p = n$, we can replace p^* by any value larger than q. □

Since for bounded U the embedding $C(\overline{U}) \hookrightarrow L^p(U)$ is continuous, we obtain:

Corollary 13.23. *Under the assumptions of the above theorem the embedding $W^{1,p}(U) \hookrightarrow L^p(U)$ is compact.*

There is also a version for $\mathbb{R}^n$.

Theorem 13.24. *Let $1 \le p \le n$. A set $F \subseteq W^{1,p}(\mathbb{R}^n)$ is relatively compact in $L^q(\mathbb{R}^n)$ for every $q \in [p, p^*)$ if F is bounded and for every $\varepsilon > 0$ there is some $r > 0$ such that $\|(1 - \chi_{B_r(0)})f\|_p < \varepsilon$ for all $f \in F$.*

Proof. Condition (i) of Theorem 10.15 is verified literally as in the previous theorem. Similarly condition (ii) follows from interpolation since $\|(1 - \chi_{B_r(0)})f\|_q \le \|(1-\chi_{B_r(0)})f\|_p^{1-\theta} \|(1-\chi_{B_r(0)})f\|_{p^*}^{\theta} \le \|(1-\chi_{B_r(0)})f\|_p^{1-\theta} \|f\|_{p^*}^{\theta}$. □

As a consequence we also can get an important inequality.

Theorem 13.25 (Poincaré inequality)**.** *Let $U \subset \mathbb{R}^n$ be an open and bounded. Then for $f \in W_0^{1,p}(U)$, $1 \le p \le \infty$, we have*

$$\|f\|_p \le C \|\partial f\|_p. \tag{13.15}$$

If in addition U is a connected subset with the extension property, then for $f \in W^{1,p}(U)$, $1 \le p \le \infty$, we have

$$\|f - (f)_U\|_p \le C \|\partial f\|_p, \tag{13.16}$$

where $(f)_U := \frac{1}{|U|} \int_U f \, d^n x$ is the average of f over U.

Proof. We begin with the second case and argue by contradiction. If the claim were wrong we could find a sequence of functions $f_m \in W^{1,p}(U)$ such that $\|f_m - (f_m)_U\|_p > m \|\partial f_m\|_p$. hence the function $g_m := \|f_m - (f_m)_U\|_p^{-1} (f_m - (f_m)_U)$ satisfies $\|g_m\|_p = 1$, $(g_m)_U = 0$, and $\|\partial g_m\|_p < \frac{1}{m}$. In particular, the sequence is bounded and by Corollary 13.23 we can assume $g_m \to g$ in $L^p(U)$ without loss of generality. Moreover, $\|g\|_p = 1$, $(g)_U = 0$, and

$$\int_U g \partial_j \varphi \, d^n x = \lim_{m\to\infty} \int_U g_m \partial_j \varphi \, d^n x = - \lim_{m\to\infty} \int_U (\partial_j g_m) \varphi \, d^n x = 0.$$

That is, $\partial_j g = 0$ and since U is connected g must be constant on U by Problem 13.8. Moreover, $(g)_U = 0$ implies $g = 0$ contradicting $\|g\|_p = 1$.

To see the first case we proceed similarly to find a sequence $g_m := \|f_m\|_p^{-1} f_m$ producing a limit such that $\|g\|_p = 1$ and $\partial_j g = 0$. Now take a ball $B := B_r(0)$ containing U such that $B \backslash U$ has positive Lebesgue measure. Observe that we can extend f_m to $\bar{f}_m \in W^{1,p}(B)$ by setting it equal to 0 outside U which will give a corresponding sequence $\bar{g}_m := \|f_m\|_p^{-1} \bar{f}_m$ and a corresponding limit $\bar{g}$. Since B is connected we again get that g is constant on B and since g vanishes on $B \setminus U$ it must vanish on all of B contradicting $\|\bar{g}\|_p = \|g\|_p = 1$. □

Example 13.5. Consider $f \in W^{1,n}(\mathbb{R}^n)$. Then note that a simple scaling shows that the constant C_r for a ball of radius r in the Poincaré inequality is given by $C_r = C_1 r$. Hence using Poincaré's and Hölder's inequalities we obtain

$$\begin{aligned}\frac{1}{|B_r(x)|}\int_{B_r(x)} |f(y) - (f)_{B_r(x)}| d^n y &\le C_1 r \int_{B_r(x)} |\partial f(y)| \frac{d^n y}{B_r(x)} \\ &\le C_1 r \left(\int_{B_r(x)} |\partial f(y)|^n \frac{d^n y}{B_r(x)} \right)^{1/n} \le \frac{C_1}{V_n^{1/n}} \|\partial f\|_n.\end{aligned}$$

Functions for which the left-hand side is bounded are called functions of **bounded mean oscillation**. ◇

Finally it is often important to know when $W^{1,p}(U)$ is an algebra: By the product rule we have $\partial_j(fg) = (\partial_j f)g + f(\partial_j g)$ and for this to be in L^p we need that f, g are bounded which follows from Morrey's inequality if $n < p$. workin a bit harder one can even show:

Theorem 13.26. *Suppose $U \subseteq \mathbb{R}^n$ is bounded and has the extension property. If $\frac{1}{p} < \frac{k}{n}$, then $W^{k,p}(U)$ is a Banach algebra with*

$$\|fg\|_{k,p} \le C \|g\|_{k,p} \|g\|_{k,p}. \tag{13.17}$$

Proof. First of all it suffices to show the inequality for the case when f and g are $C^\infty \cap W^{k,p}$. Moreover, by Leibniz' rule (Problem 13.7) it suffices to estimate $\|(\partial_\alpha f)(\partial_\beta g)\|_p$ for $|\alpha| + |\beta| \le k$. To this end we will use the generalized Hölder inequality (Problem 10.8) and hence we need to find $1 \le q_\alpha, q_\beta \le \infty$ with $\frac{1}{p} = \frac{1}{q_\alpha} + \frac{1}{q_\beta}$ such that $W^{m-|\alpha|,p} \hookrightarrow L^{q_\alpha}$ and $W^{m-|\beta|,p} \hookrightarrow L^{q_\beta}$.

Let l be the largest integer such that $\frac{1}{p} < \frac{k-l}{n}$. Then Theorem 13.21 allows us to choose $q_\alpha = \infty$, $q_\beta = p$ for $|\alpha| \le l$ and similarly $q_\alpha = p$, $q_\beta = \infty$ for $|\beta| \le l$. Otherwise, that is if $\frac{1}{p} \ge \frac{k-|\alpha|}{n}$ and $\frac{1}{p} \ge \frac{k-|\beta|}{n}$ then

Theorem 13.21 imposes the restrictions $\frac{1}{q_\alpha} \geq \frac{1}{p} - \frac{k-|\alpha|}{n}$ and $\frac{1}{q_\beta} \geq \frac{1}{p} - \frac{k-|\beta|}{n}$. Hence $\frac{1}{q_\alpha} + \frac{1}{q_\beta} \geq \frac{1}{p} - \left(\frac{k}{n} - \frac{1}{p}\right)$ and we can find the required indices. □

Problem* 13.17. *Show that* $W^{1,1}((a,b)) = AC[a,b]$ *with* $\|f\|_\infty \leq \|f\|_1 + (b-a)\|f'\|_1$. *Consequently* $W^{k,1}((a,b)) = \{f \in C^{k-1}[a,b] | f^{(k-1)} \in AC[a,b]\}$. *(Hint: Problem 13.2.)*

Problem* 13.18. *Show that the inequality* $\|f\|_q \leq C\|\partial f\|_p$ *for* $f \in W^{1,p}(\mathbb{R}^n)$ *can only hold for* $q = \frac{np}{n-p}$. *(Hint: Consider* $f_\lambda(x) = f(\lambda x)$.*)*

Problem* 13.19. *Show that* $f(x) := \log\log(1 + \frac{1}{|x|})$ *is in* $W^{1,n}(B_1(0))$ *if* $n > 1$. *(Hint: Problem 13.3.)*

Problem* 13.20. *Consider* $U := \{(x,y) \in \mathbb{R}^2 | 0 < x, y < 1,\ x^\beta < y\}$ *and* $f(x,y) := y^{-\alpha}$ *with* $\alpha, \beta > 0$. *Show* $f \in W^{1,p}(U)$ *for* $p < \frac{1+\beta}{(1+\alpha)\beta}$. *Now observe that for* $0 < \beta < 1$ *and* $\alpha < \frac{1-\beta}{2\beta}$ *we have* $2 < \frac{1+\beta}{(1+\alpha)\beta}$.

Problem* 13.21. *Show* $W^{1,\infty}(\mathbb{R}^n) = C_b^{0,1}(\mathbb{R}^n)$. *(Hint: Apply Lemma 4.34 to the differential quotient.)*

13.4. Applications to elliptic equations

The purpose of this section is to show that Sobolev spaces provide a convenient framework for treating partial differential equations by functional analytic methods. To focus on the main ideas we will start by looking at the **Laplace equation**

$$\begin{aligned} -\Delta u(x) &= f(x), & x &\in U, \\ u(x) &= 0, & x &\in \partial U, \end{aligned} \tag{13.18}$$

on a bounded domain $U \subseteq \mathbb{R}^n$ with Dirichlet boundary conditions. Here $\Delta u = \sum_{j=1}^n \partial_j^2 u$ as usual. If we regard the derivatives as weak derivatives, then our equation reads

$$\int_U (\Delta\varphi)(x) u(x) d^n x = \int_U \varphi(x) f(x) d^n x, \qquad \varphi \in C_c^\infty(U), \tag{13.19}$$

or, after an integration by parts, we can also write it in the more symmetric from

$$\sum_{j=1}^n \int_U (\partial_j \varphi)(\partial_j u) d^n x = \int_U \varphi(x) f(x) d^n x, \qquad \varphi \in C_c^\infty(U). \tag{13.20}$$

Now recall that by the Poincaré inequality (Theorem 13.25) we have a scalar product

$$\langle u, v\rangle = \sum_{j=1}^n \int_U (\partial_j u)(\partial_j v) d^m x \tag{13.21}$$

on $H^1_0(U,\mathbb{R})$ whose associated norm is equivalent to the usual one. Moreover, using the fact that $C^\infty_c(U,\mathbb{R}) \subset H^1_0(U,\mathbb{R})$ is dense we see that we can write our last form as

$$\langle v, u\rangle = \langle v, f\rangle_2, \qquad v \in H^1_0(U,\mathbb{R}), \tag{13.22}$$

where $\langle u, v\rangle_2 = \int_U u(x)v(x)d^n x$ denotes the scalar product in $L^2(U,\mathbb{R})$.

We will call a solution $u \in H^1_0(U,\mathbb{R})$ of (13.22) a **weak solution** of the Dirichlet problem (13.18). If a solution is, in addition, in $H^2(U,\mathbb{R})$, it is called a **strong solution**. In this case we can undo our integration by parts and conclude that u solves (13.18), where the derivatives are understood as weak derivatives and the boundary condition is understood in the sense of traces (at least for U with sufficiently smooth boundary; see Lemma 13.12).

Finally note that (13.22) can be understood as

$$\langle v, u\rangle = \langle Jv, f\rangle_2, \qquad v \in H^1_0(U,\mathbb{R}), \tag{13.23}$$

where $J : H^1_0(U,\mathbb{R}) \hookrightarrow L^2(U,\mathbb{R})$ is the natural embedding. Since this embedding is bounded we can use the adjoint operator to write this as

$$\langle v, u\rangle = \langle v, J^* f\rangle, \qquad v \in H^1_0(U,\mathbb{R}), \tag{13.24}$$

which shows that the weak problem (13.22) has a unique solution $u = J^* f \in H^1_0(U,\mathbb{R})$ for every $f \in L^2(U,\mathbb{R})$. Also note the estimate $\|u\| = \|J^* f\| \le \|f\|_2$ (since $\|J^*\| = \|J\| = 1$).

Now what about strong solutions? To this end we regard (13.18) as an operator equation

$$Lu = f, \tag{13.25}$$

where

$$Lu := -\Delta u, \qquad u \in \mathfrak{D}(L) := H^1_0(U,\mathbb{R}) \cap H^2(U,\mathbb{R}). \tag{13.26}$$

Then uniqueness of weak solutions implies that $\bar{L} := (JJ^*)^{-1}$ is a well-defined operator $\bar{L} : \mathfrak{D}(\bar{L}) \subset L^2(U,\mathbb{R}) \to L^2(U,\mathbb{R})$ which coincides with L when restricted to $\mathfrak{D}(L)$ (in particular, $\mathfrak{D}(L) \subseteq \mathfrak{D}(\bar{L}) = \operatorname{Ran}(JJ^*) \subset H^1_0(U,\mathbb{R}^n)$). When we also have $\mathfrak{D}(\bar{L}) \subseteq \mathfrak{D}(L)$, that is, when every weak solution is also a strong solution, is a tricky question which we will not answer here.

Instead, we point out that since the embedding J is not only continuous, but even compact by the Rellich–Kondrachov theorem (Theorem 13.22), we can apply the spectral theorem for compact operators (since JJ^* is self-adjoint Theorem 3.7 will do):

Theorem 13.27. *The operator $\bar{L}$ has a sequence of discrete real eigenvalues $0 < \lambda_0 < \lambda_1 < \cdots$ converging to ∞. The corresponding normalized eigenfunctions form an orthonormal basis for $L^2(U,\mathbb{R}^n)$.*

Observe that the lowest eigenvalue λ_0 is the optimal constant for the Poincarè inequality.

Finally we remark that our consideration extend easily to the Dirichlet problem for second order **elliptic equations** of the form

$$Lu(x) := -\sum_{i,j=1}^{n} \partial_i A_{ij}(x)\partial_j u(x) + c(x)u(x) \tag{13.27}$$

with $A_{i,j}, c \in L^\infty(U,\mathbb{R})$ with

$$a_0 = \inf_{e\in S^n, x\in U} e_i A_{ij}(x) e_j > 0, \quad c_0 = \inf_{x\in U} c(x) \geq 0. \tag{13.28}$$

As domain we choose $\mathfrak{D}(L) = \{u \in H_0^1(U,\mathbb{R}) | A_{ij}\partial_j u \in H^1(U,\mathbb{R}), 1 \leq i,j \leq n\}$. Then the elipticity condition $a_0 > 0$ ensures that the symmetric bilinear form

$$a(u,v) := \sum_{i,j=1}^{n} \int_U (A_{ij}(x)(\partial_j u(x))(\partial_i v(x)) + c(x)u(x)v(x))\, d^n x \tag{13.29}$$

gives rise to a norm which is equivalent to the usual norm on $H_0^1(U,\mathbb{R})$ and hence we can proceed as before.

Problem 13.22. *Consider the elliptic Dirichlet problem associated with*

$$Lu(x) := -\sum_{i,j=1}^{n} \partial_i A_{ij}(x)\partial_j u(x) + \sum_{j=1}^{n} b_j(x)\partial_j u(x) + c(x)u(x),$$

where $A_{i,j}, b_j, c \in L^\infty(U,\mathbb{R})$ *with*

$$a_0 = \inf_{|e|=1, x\in U} \sum_{i,j} e_i A_{ij}(x) e_j, \quad b_0 = \sup_{x\in U} |b(x)|, \quad c_0 = \inf_{x\in U} c(x).$$

Show that $Lu = f$ *has a unique weak solution in* $H_0^1(U,\mathbb{R})$ *provided* $4a_0c_0 > b_0^2$. *(Hint: Choose* $H_0^1(U,\mathbb{R})$ *equipped with* (13.21) *as underlying Hilbert space. On this Hilbert space there is a corresponding bilinear form* $a(u,v)$*, however, if* $b \neq 0$ *this form is not symmetric. To overcome this problem use the Lax–Milgram (Theorem 2.16). Again there will be a problem with the term corresponding to* b *when establishing coercivity. However, note that this* mixed *term can be controlled using the other two.)*

Chapter 14

The Fourier transform

14.1. The Fourier transform on L^1 and L^2

For $f \in L^1(\mathbb{R}^n)$ we define its **Fourier transform** via

$$\mathcal{F}(f)(p) \equiv \hat{f}(p) := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}px} f(x) d^n x. \tag{14.1}$$

Here $px = p_1 x_1 + \cdots + p_n x_n$ is the usual scalar product in $\mathbb{R}^n$ and we will use $|x| = \sqrt{x_1^2 + \cdots + x_n^2}$ for the Euclidean norm.

Lemma 14.1. *The Fourier transform is a bounded map from $L^1(\mathbb{R}^n)$ into $C_b(\mathbb{R}^n)$ satisfying*

$$\|\hat{f}\|_\infty \leq (2\pi)^{-n/2} \|f\|_1. \tag{14.2}$$

Proof. Since $|\mathrm{e}^{-\mathrm{i}px}| = 1$ the estimate (14.2) is immediate from

$$|\hat{f}(p)| \leq \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} |\mathrm{e}^{-\mathrm{i}px} f(x)| d^n x = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} |f(x)| d^n x.$$

Moreover, a straightforward application of the dominated convergence theorem shows that $\hat{f}$ is continuous. □

Note that if f is nonnegative we have equality: $\|\hat{f}\|_\infty = (2\pi)^{-n/2} \|f\|_1 = \hat{f}(0)$.

The following simple properties are left as an exercise.

Lemma 14.2. *Let $f \in L^1(\mathbb{R}^n)$. Then*

$$(f(x+a))^\wedge(p) = \mathrm{e}^{\mathrm{i}ap}\hat{f}(p), \qquad a \in \mathbb{R}^n, \tag{14.3}$$

$$(\mathrm{e}^{\mathrm{i}xa}f(x))^\wedge(p) = \hat{f}(p-a), \qquad a \in \mathbb{R}^n, \tag{14.4}$$

$$(f(\lambda x))^\wedge(p) = \frac{1}{\lambda^n}\hat{f}(\frac{p}{\lambda}), \qquad \lambda > 0, \tag{14.5}$$

$$(f(-x))^\wedge(p) = (f)^\wedge(-p). \tag{14.6}$$

Next we look at the connection with differentiation.

Lemma 14.3. *Suppose $f \in C^1(\mathbb{R}^n)$ such that $\lim_{|x|\to\infty} f(x) = 0$ and $f, \partial_j f \in L^1(\mathbb{R}^n)$ for some $1 \le j \le n$. Then*

$$(\partial_j f)^\wedge(p) = \mathrm{i}p_j\hat{f}(p). \tag{14.7}$$

Similarly, if $f(x), x_j f(x) \in L^1(\mathbb{R}^n)$ for some $1 \le j \le n$, then $\hat{f}(p)$ is differentiable with respect to p_j and

$$(x_j f(x))^\wedge(p) = \mathrm{i}\partial_j\hat{f}(p). \tag{14.8}$$

Proof. First of all, by integration by parts, we see

$$\begin{aligned}(\partial_j f)^\wedge(p) &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}px}\frac{\partial}{\partial x_j}f(x)d^n x\\ &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\left(-\frac{\partial}{\partial x_j}\mathrm{e}^{-\mathrm{i}px}\right)f(x)d^n x\\ &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\mathrm{i}p_j\mathrm{e}^{-\mathrm{i}px}f(x)d^n x = \mathrm{i}p_j\hat{f}(p).\end{aligned}$$

Similarly, the second formula follows from

$$\begin{aligned}(x_j f(x))^\wedge(p) &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n} x_j\mathrm{e}^{-\mathrm{i}px}f(x)d^n x\\ &= \frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\left(\mathrm{i}\frac{\partial}{\partial p_j}\mathrm{e}^{-\mathrm{i}px}\right)f(x)d^n x = \mathrm{i}\frac{\partial}{\partial p_j}\hat{f}(p),\end{aligned}$$

where interchanging the derivative and integral is permissible by Problem 9.15. In particular, $\hat{f}(p)$ is differentiable. □

This result immediately extends to higher derivatives. To this end let $C^\infty(\mathbb{R}^n)$ be the set of all complex-valued functions which have partial derivatives of arbitrary order. For $f \in C^\infty(\mathbb{R}^n)$ and $\alpha \in \mathbb{N}_0^n$ we set

$$\partial_\alpha f := \frac{\partial^{|\alpha|}f}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}, \quad x^\alpha := x_1^{\alpha_1}\cdots x_n^{\alpha_n}, \quad |\alpha| := \alpha_1 + \cdots + \alpha_n. \tag{14.9}$$

An element $\alpha \in \mathbb{N}_0^n$ is called a **multi-index** and $|\alpha|$ is called its **order**. We will also set $(\lambda x)^\alpha = \lambda^{|\alpha|} x^\alpha$ for $\lambda \in \mathbb{R}$. Recall the **Schwartz space**

$$\mathcal{S}(\mathbb{R}^n) := \{f \in C^\infty(\mathbb{R}^n) | \sup_x |x^\alpha (\partial_\beta f)(x)| < \infty, \forall \alpha, \beta \in \mathbb{N}_0^n\} \tag{14.10}$$

which is a subspace of $L^p(\mathbb{R}^n)$ and which is dense for $1 \le p < \infty$ (since $C_c^\infty(\mathbb{R}^n) \subset \mathcal{S}(\mathbb{R}^n)$). Together with the seminorms $\|x^\alpha(\partial_\beta f)(x)\|_\infty$ it is a Fréchet space. Note that if $f \in \mathcal{S}(\mathbb{R}^n)$, then the same is true for $x^\alpha f(x)$ and $(\partial_\alpha f)(x)$ for every multi-index α. Also, by Leibniz' rule, the product of two Schwartz functions is again a Schwartz function.

Lemma 14.4. *The Fourier transform satisfies $\mathcal{F} : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$. Furthermore, for every multi-index $\alpha \in \mathbb{N}_0^n$ and every $f \in \mathcal{S}(\mathbb{R}^n)$ we have*

$$(\partial_\alpha f)^\wedge(p) = (\mathrm{i}p)^\alpha \hat{f}(p), \qquad (x^\alpha f(x))^\wedge(p) = \mathrm{i}^{|\alpha|} \partial_\alpha \hat{f}(p). \tag{14.11}$$

Proof. The formulas are immediate from the previous lemma. To see that $\hat{f} \in \mathcal{S}(\mathbb{R}^n)$ if $f \in \mathcal{S}(\mathbb{R}^n)$, we begin with the observation that $\hat{f}$ is bounded by (14.2). But then $p^\alpha(\partial_\beta \hat{f})(p) = \mathrm{i}^{-|\alpha|-|\beta|}(\partial_\alpha x^\beta f(x))^\wedge(p)$ is bounded since $\partial_\alpha x^\beta f(x) \in \mathcal{S}(\mathbb{R}^n)$ if $f \in \mathcal{S}(\mathbb{R}^n)$. □

Hence we will sometimes write $pf(x)$ for $-\mathrm{i}\partial f(x)$, where $\partial = (\partial_1, \dots, \partial_n)$ is the **gradient**. Roughly speaking this lemma shows that the decay of a functions is related to the smoothness of its Fourier transform and the smoothness of a functions is related to the decay of its Fourier transform.

In particular, this allows us to conclude that the Fourier transform of an integrable function will vanish at ∞. Recall that we denote the space of all continuous functions $f : \mathbb{R}^n \to \mathbb{C}$ which vanish at ∞ by $C_0(\mathbb{R}^n)$.

Corollary 14.5 (Riemann-Lebesgue)**.** *The Fourier transform maps $L^1(\mathbb{R}^n)$ into $C_0(\mathbb{R}^n)$.*

Proof. First of all recall that $C_0(\mathbb{R}^n)$ equipped with the sup norm is a Banach space and that $\mathcal{S}(\mathbb{R}^n)$ is dense (Problem 1.46). By the previous lemma we have $\hat{f} \in C_0(\mathbb{R}^n)$ if $f \in \mathcal{S}(\mathbb{R}^n)$. Moreover, since $\mathcal{S}(\mathbb{R}^n)$ is dense in $L^1(\mathbb{R}^n)$, the estimate (14.2) shows that the Fourier transform extends to a continuous map from $L^1(\mathbb{R}^n)$ into $C_0(\mathbb{R}^n)$. □

Next we will turn to the inversion of the Fourier transform. As a preparation we will need the Fourier transform of a **Gaussian**.

Lemma 14.6. *We have $\mathrm{e}^{-z|x|^2/2} \in \mathcal{S}(\mathbb{R}^n)$ for $\mathrm{Re}(z) > 0$ and*

$$\mathcal{F}(\mathrm{e}^{-z|x|^2/2})(p) = \frac{1}{z^{n/2}} \mathrm{e}^{-|p|^2/(2z)}. \tag{14.12}$$

Here $z^{n/2}$ is the standard branch with branch cut along the negative real axis.

Proof. Due to the product structure of the exponential, one can treat each coordinate separately, reducing the problem to the case $n = 1$ (Problem 14.3).

Let $\phi_z(x) := \exp(-zx^2/2)$. Then $\phi_z'(x) + zx\phi_z(x) = 0$ and hence $\mathrm{i}(p\hat{\phi}_z(p) + z\hat{\phi}_z'(p)) = 0$. Thus $\hat{\phi}_z(p) = c\phi_{1/z}(p)$ and (Problem 9.23)

$$c = \hat{\phi}_z(0) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(-zx^2/2)dx = \frac{1}{\sqrt{z}}$$

at least for $z > 0$. However, since the integral is holomorphic for $\mathrm{Re}(z) > 0$ by Problem 9.19, this holds for all z with $\mathrm{Re}(z) > 0$ if we choose the branch cut of the root along the negative real axis. □

Now we can show

Theorem 14.7. *The Fourier transform is a bounded injective map from $L^1(\mathbb{R}^n)$ into $C_0(\mathbb{R}^n)$. Its inverse is given by*

$$f(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}px - \varepsilon|p|^2/2} \hat{f}(p) d^n p, \tag{14.13}$$

where the limit has to be understood in L^1. Moreover, (14.13) *holds at every Lebesgue point (cf. Theorem 11.6) and hence in particular at every point of continuity.*

Proof. Abbreviate $\phi_\varepsilon(x) := (2\pi)^{-n/2} \exp(-\varepsilon|x|^2/2)$. Then the right-hand side is given by

$$\int_{\mathbb{R}^n} \phi_\varepsilon(p) \mathrm{e}^{\mathrm{i}px} \hat{f}(p) d^n p = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \phi_\varepsilon(p) \mathrm{e}^{\mathrm{i}px} f(y) \mathrm{e}^{-\mathrm{i}py} d^n y d^n p$$

and, invoking Fubini and Lemma 14.2, we further see that this is equal to

$$= \int_{\mathbb{R}^n} (\phi_\varepsilon(p) \mathrm{e}^{\mathrm{i}px})^\wedge(y) f(y) d^n y = \int_{\mathbb{R}^n} \frac{1}{\varepsilon^{n/2}} \phi_{1/\varepsilon}(y - x) f(y) d^n y.$$

But the last integral converges to f in $L^1(\mathbb{R}^n)$ by Lemma 10.19. Moreover, it is straightforward to see that it converges at every point of continuity. The case of Lebesgue points follows from Problem 15.8. □

Of course when $\hat{f} \in L^1(\mathbb{R}^n)$, the limit is superfluous and we obtain

Corollary 14.8. *Suppose $f, \hat{f} \in L^1(\mathbb{R}^n)$. Then*

$$(\hat{f})^\vee = f, \tag{14.14}$$

where

$$\check{f}(p) := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}px} f(x) d^n x = \hat{f}(-p). \tag{14.15}$$

In particular, $\mathcal{F} : F^1(\mathbb{R}^n) \to F^1(\mathbb{R}^n)$ is a bijection, where $F^1(\mathbb{R}^n) := \{f \in L^1(\mathbb{R}^n) | \hat{f} \in L^1(\mathbb{R}^n)\}$. Moreover, $\mathcal{F} : \mathcal{S}(\mathbb{R}^n) \to \mathcal{S}(\mathbb{R}^n)$ is a bijection.

Observe that we have $F^1(\mathbb{R}^n) \subset L^1(\mathbb{R}^n) \cap C_0(\mathbb{R}^n) \subset L^p(\mathbb{R}^n)$ for any $p \in [1, \infty]$ (cf. also Problem 14.2) and choosing f continuous (14.14) will hold pointwise.

However, note that $\mathcal{F} : L^1(\mathbb{R}^n) \to C_0(\mathbb{R}^n)$ is not onto (cf. Problem 14.7). Nevertheless the inverse Fourier transform $\mathcal{F}^{-1}$ is a closed map from $\mathrm{Ran}(\mathcal{F}) \to L^1(\mathbb{R}^n)$ by Lemma 4.8.

Lemma 14.9. *Suppose $f \in F^1(\mathbb{R}^n)$. Then $f, \hat{f} \in L^2(\mathbb{R}^n)$ and*

$$\|f\|_2^2 = \|\hat{f}\|_2^2 \leq (2\pi)^{-n/2} \|f\|_1 \|\hat{f}\|_1 \tag{14.16}$$

holds.

Proof. This follows from Fubini's theorem since

$$\begin{aligned} \int_{\mathbb{R}^n} |\hat{f}(p)|^2 d^n p &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} f(x)^* \hat{f}(p) \mathrm{e}^{\mathrm{i}px} d^n p \, d^n x \\ &= \int_{\mathbb{R}^n} |f(x)|^2 d^n x \end{aligned}$$

for $f, \hat{f} \in L^1(\mathbb{R}^n)$. □

The identity $\|f\|_2 = \|\hat{f}\|_2$ is known as the **Plancherel identity**. Thus, by Theorem 1.17, we can extend $\mathcal{F}$ to all of $L^2(\mathbb{R}^n)$ by setting $\mathcal{F}(f) = \lim_{m\to\infty} \mathcal{F}(f_m)$, where f_m is an arbitrary sequence from, say, $\mathcal{S}(\mathbb{R}^n)$ converging to f in the L^2 norm.

Theorem 14.10 (Plancherel)**.** *The Fourier transform $\mathcal{F}$ extends to a unitary operator $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$.*

Proof. As already noted, $\mathcal{F}$ extends uniquely to a bounded operator on $L^2(\mathbb{R}^n)$. Since Plancherel's identity remains valid by continuity of the norm and since its range is dense, this extension is a unitary operator. □

We also note that this extension is still given by (14.1) whenever the right-hand side is integrable.

Lemma 14.11. *Let $f \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$, then* (14.1) *continues to hold, where $\mathcal{F}$ now denotes the extension of the Fourier transform from $\mathcal{S}(\mathbb{R}^n)$ to $L^2(\mathbb{R}^n)$.*

Proof. If f has compact support, then by Lemma 10.18 its mollification $\phi_\varepsilon * f \in C_c^\infty(\mathbb{R}^n)$ converges to f both in L^1 and L^2. Hence the claim holds for every f with compact support. Finally, for general $f \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ consider $f_m = f\chi_{B_m(0)}$. Then $f_m \to f$ in both $L^1(\mathbb{R}^n)$ and $L^2(\mathbb{R}^n)$ and the claim follows. □

In particular,

$$\hat{f}(p) = \lim_{m\to\infty} \frac{1}{(2\pi)^{n/2}} \int_{|x|\le m} \mathrm{e}^{-\mathrm{i}px} f(x) d^n x, \tag{14.17}$$

where the limit has to be understood in $L^2(\mathbb{R}^n)$ and can be omitted if $f \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$.

Another useful property is the convolution formula.

Lemma 14.12. *The **convolution***

$$(f * g)(x) = \int_{\mathbb{R}^n} f(y) g(x-y) d^n y = \int_{\mathbb{R}^n} f(x-y) g(y) d^n y \tag{14.18}$$

*of two functions $f, g \in L^1(\mathbb{R}^n)$ is again in $L^1(\mathbb{R}^n)$ and we have **Young's inequality***

$$\|f * g\|_1 \le \|f\|_1 \|g\|_1. \tag{14.19}$$

Moreover, its Fourier transform is given by

$$(f * g)^\wedge = (2\pi)^{n/2} \hat{f}\hat{g}. \tag{14.20}$$

Proof. The fact that $f * g$ is in L^1 together with Young's inequality follows by applying Fubini's theorem to $h(x, y) = f(x - y)g(y)$ (in fact we have shown a more general version in Lemma 10.18). For the last claim we compute

$$\begin{aligned}
(f * g)^\wedge(p) &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}px} \left(\int_{\mathbb{R}^n} f(y) g(x-y) d^n y \right) d^n x \\
&= \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}py} f(y) \left(\frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}p(x-y)} g(x-y) d^n x \right) d^n y \\
&= \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}py} f(y) \hat{g}(p) d^n y = (2\pi)^{n/2} \hat{f}(p)\hat{g}(p),
\end{aligned}$$

where we have again used Fubini's theorem. □

As a consequence we can also deal with the case of convolution on $\mathcal{S}(\mathbb{R}^n)$ as well as on $L^2(\mathbb{R}^n)$.

Corollary 14.13. *The convolution of two $\mathcal{S}(\mathbb{R}^n)$ functions as well as their product is in $\mathcal{S}(\mathbb{R}^n)$ and*

$$(f * g)^\wedge = (2\pi)^{n/2} \hat{f}\hat{g}, \qquad (fg)^\wedge = (2\pi)^{-n/2} \hat{f} * \hat{g} \tag{14.21}$$

in this case.

Proof. Clearly the product of two functions in $\mathcal{S}(\mathbb{R}^n)$ is again in $\mathcal{S}(\mathbb{R}^n)$ (show this!). Since $\mathcal{S}(\mathbb{R}^n) \subset L^1(\mathbb{R}^n)$ the previous lemma implies $(f * g)^\wedge = (2\pi)^{n/2} \hat{f}\hat{g} \in \mathcal{S}(\mathbb{R}^n)$. Moreover, since the Fourier transform is injective on $L^1(\mathbb{R}^n)$ we conclude $f * g = (2\pi)^{n/2} (\hat{f}\hat{g})^\vee \in \mathcal{S}(\mathbb{R}^n)$. Replacing f, g by $\check{f}, \check{g}$

in the last formula finally shows $\check{f} * \check{g} = (2\pi)^{n/2}(fg)^\vee$ and the claim follows by a simple change of variables using $\check{f}(p) = \hat{f}(-p)$. □

Corollary 14.14. *The convolution of two $L^2(\mathbb{R}^n)$ functions is in* $\mathrm{Ran}(\mathcal{F}) \subset C_0(\mathbb{R}^n)$ *and we have $\|f * g\|_\infty \leq \|f\|_2\|g\|_2$ as well as*

$$(fg)^\wedge = (2\pi)^{-n/2}\hat{f} * \hat{g}, \qquad (f * g)^\wedge = (2\pi)^{n/2}\hat{f}\hat{g} \tag{14.22}$$

in this case.

Proof. The inequality $\|f * g\|_\infty \leq \|f\|_2\|g\|_2$ is immediate from Cauchy–Schwarz and shows that the convolution is a continuous bilinear form from $L^2(\mathbb{R}^n)$ to $L^\infty(\mathbb{R}^n)$. Now take sequences $f_m, g_m \in \mathcal{S}(\mathbb{R}^n)$ converging to $f, g \in L^2(\mathbb{R}^n)$. Then using the previous corollary together with continuity of the Fourier transform from $L^1(\mathbb{R}^n)$ to $C_0(\mathbb{R}^n)$ and on $L^2(\mathbb{R}^n)$ we obtain

$$(fg)^\wedge = \lim_{m\to\infty}(f_m g_m)^\wedge = (2\pi)^{-n/2}\lim_{m\to\infty}\hat{f}_m * \hat{g}_m = (2\pi)^{-n/2}\hat{f} * \hat{g}.$$

Similarly,

$$(f * g)^\wedge = \lim_{m\to\infty}(f_m * g_m)^\wedge = (2\pi)^{n/2}\lim_{m\to\infty}\hat{f}_m\hat{g}_m = (2\pi)^{n/2}\hat{f}\hat{g}$$

from which that last claim follows since $\mathcal{F} : \mathrm{Ran}(\mathcal{F}) \to L^1(\mathbb{R}^n)$ is closed by Lemma 4.8. □

Finally, note that by looking at the Gaussian's $\phi_\lambda(x) = \exp(-\lambda x^2/2)$ one observes that a well centered peak transforms into a broadly spread peak and vice versa. This turns out to be a general property of the Fourier transform known as **uncertainty principle**. One quantitative way of measuring this fact is to look at

$$\|(x_j - x^0)f(x)\|_2^2 = \int_{\mathbb{R}^n}(x_j - x^0)^2|f(x)|^2 d^n x \tag{14.23}$$

which will be small if f is well concentrated around x^0 in the j'th coordinate direction.

Theorem 14.15 (Heisenberg uncertainty principle)**.** *Suppose $f \in \mathcal{S}(\mathbb{R}^n)$. Then for any $x^0, p^0 \in \mathbb{R}$ we have*

$$\|(x_j - x^0)f(x)\|_2\|(p_j - p^0)\hat{f}(p)\|_2 \geq \frac{\|f\|_2^2}{2}. \tag{14.24}$$

Proof. Replacing $f(x)$ by $\mathrm{e}^{\mathrm{i}x_j p^0}f(x + x^0 e_j)$ (where e_j is the unit vector into the j'th coordinate direction) we can assume $x^0 = p^0 = 0$ by Lemma 14.2. Using integration by parts we have

$$\|f\|_2^2 = \int_{\mathbb{R}^n}|f(x)|^2 d^n x = -\int_{\mathbb{R}^n} x_j \partial_j |f(x)|^2 d^n x = -2\mathrm{Re}\int_{\mathbb{R}^n} x_j f(x)^* \partial_j f(x) d^n x.$$

Hence, by Cauchy–Schwarz,

$$\|f\|_2^2 \le 2\|x_j f(x)\|_2 \|\partial_j f(x)\|_2 = 2\|x_j f(x)\|_2 \|p_j \hat{f}(p)\|_2$$

the claim follows. □

The name stems from quantum mechanics, where $|f(x)|^2$ is interpreted as the probability distribution for the position of a particle and $|\hat{f}(x)|^2$ is interpreted as the probability distribution for its momentum. Equation (14.24) says that the variance of both distributions cannot both be small and thus one cannot simultaneously measure position and momentum of a particle with arbitrary precision.

Another version states that f and $\hat{f}$ cannot both have compact support.

Theorem 14.16. *Suppose $f \in L^2(\mathbb{R}^n)$. If both f and $\hat{f}$ have compact support, then $f = 0$.*

Proof. Let $A, B \subset \mathbb{R}^n$ be two compact sets and consider the subspace of all functions with $\operatorname{supp}(f) \subseteq A$ and $\operatorname{supp}(\hat{f}) \subseteq B$. Then

$$f(x) = \int_{\mathbb{R}^n} K(x,y) f(y) d^n y,$$

where

$$K(x,y) := \frac{1}{(2\pi)^n} \int_B \mathrm{e}^{\mathrm{i}(x-y)p} \chi_A(y) d^n p = \frac{1}{(2\pi)^n} \hat{\chi}_B(y-x) \chi_A(y).$$

Since $K \in L^2(\mathbb{R}^n \times \mathbb{R}^n)$ the corresponding integral operator is Hilbert–Schmidt, and thus its eigenspace corresponding to the eigenvalue 1 can be at most finite dimensional.

Now if there is a nonzero f, we can find a sequence of vectors $x^n \to 0$ such the functions $f_n(x) = f(x - x^n)$ are linearly independent (look at their supports) and satisfy $\operatorname{supp}(f_n) \subseteq 2A$, $\operatorname{supp}(\hat{f}_n) \subseteq B$. But this a contradiction by the first part applied to the sets $2A$ and B. □

Problem 14.1. *Show that $\mathcal{S}(\mathbb{R}^n) \subset L^p(\mathbb{R}^n)$. (Hint: If $f \in \mathcal{S}(\mathbb{R}^n)$, then $|f(x)| \le C_m \prod_{j=1}^n (1 + x_j^2)^{-m}$ for every m.)*

Problem 14.2. *Show that $F^1(\mathbb{R}^n) \subset L^p(\mathbb{R}^n)$ with*

$$\|f\|_p \le (2\pi)^{\frac{n}{2}(1-\frac{1}{p})} \|f\|_1^{\frac{1}{p}} \|\hat{f}\|_1^{1-\frac{1}{p}}.$$

Moreover, show that $\mathcal{S}(\mathbb{R}^n) \subset F^1(\mathbb{R}^n)$ and conclude that $F^1(\mathbb{R}^n)$ is dense in $L^p(\mathbb{R}^n)$ for $p \in [1, \infty)$. (Hint: Use $x^p \le x$ for $0 \le x \le 1$ to show $\|f\|_p \le \|f\|_\infty^{1-1/p} \|f\|_1^{1/p}$.)

Problem* 14.3. *Suppose $f_j \in L^1(\mathbb{R})$, $j = 1, \dots, n$ and set $f(x) = \prod_{j=1}^n f_j(x_j)$. Show that $f \in L^1(\mathbb{R}^n)$ with $\|f\|_1 = \prod_{j=1}^n \|f_j\|_1$ and $\hat{f}(p) = \prod_{j=1}^n \hat{f}_j(p_j)$.*

Problem 14.4. *Compute the Fourier transform of the following functions $f : \mathbb{R} \to \mathbb{C}$:*

(i) $f(x) = \chi_{(-1,1)}(x)$. (ii) $f(x) = \frac{1}{x^2+k^2}$, $\operatorname{Re}(k) > 0$.

Problem 14.5. *A function $f : \mathbb{R}^n \to \mathbb{C}$ is called* ***spherically symmetric*** *if it is invariant under rotations; that is, $f(Ox) = f(x)$ for all $O \in SO(\mathbb{R}^n)$ (equivalently, f depends only on the distance to the origin $|x|$). Show that the Fourier transform of a spherically symmetric function is again spherically symmetric.*

Problem 14.6. *Suppose $f \in L^1(\mathbb{R}^n)$. If f is continuous at 0 and $\hat{f}(p) \geq 0$ then*

$$f(0) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \hat{f}(p) d^n p.$$

Use this to show the Plancherel identity for $f \in L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ by applying it to $F := f * \tilde{f}$, where $\tilde{f}(x) = f(-x)^*$.

Problem* 14.7. *Show that $\mathcal{F} : L^1(\mathbb{R}^n) \to C_0(\mathbb{R}^n)$ is not onto as follows:*

(i) *The range of $\mathcal{F}$ is dense.*

(ii) *$\mathcal{F}$ is onto if and only if it has a bounded inverse.*

(iii) *$\mathcal{F}$ has no bounded inverse.*

(Hint for (iii)*: Consider $\phi_z(x) = \exp(-zx^2/2)$ for $z = \lambda + \mathrm{i}\omega$ with $\lambda > 0$.)*

Problem 14.8 (Wiener)**.** *Suppose $f \in L^2(\mathbb{R}^n)$. Then the set $\{f(x+a) | a \in \mathbb{R}^n\}$ is total in $L^2(\mathbb{R}^n)$ if and only if $\hat{f}(p) \neq 0$ a.e. (Hint: Use Lemma 14.2 and the fact that a subspace is total if and only if its orthogonal complement is zero.)*

Problem 14.9 (Rellich)**.** *Let C be a positive constant and let $f_j : \mathbb{R}^n \to \mathbb{R}$ be two measurable functions such that $f_j(x) \geq 1$ with $\lim_{|x|\to\infty} f_j(x) = \infty$ for $j = 0, 1$. Then the set $\{f \in L^2(\mathbb{R}^n) | \, \|f_0 f\|_2 + \|f_1 \hat{f}\|_2 \leq C\}$ is compact in $L^2(\mathbb{R}^n)$. (Hint: Example 10.3.)*

Problem 14.10. *Suppose $f(x)\mathrm{e}^{k|x|} \in L^1(\mathbb{R})$ for some $k > 0$. Then $\hat{f}(p)$ has an analytic extension to the strip $|\operatorname{Im}(p)| < k$.*

Problem 14.11. *The Fourier transform of a complex measure μ is defined by*

$$\hat{\mu}(p) := \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}px} d\mu(x).$$

Show that the Fourier transform is a bounded injective map from $\mathcal{M}(\mathbb{R}^n) \to C_b(\mathbb{R}^n)$ satisfying $\|\hat{\mu}\|_\infty \leq (2\pi)^{-n/2} |\mu|(\mathbb{R}^n)$. Moreover, show $(\mu * \nu)^\wedge = (2\pi)^{n/2} \hat{\mu}\hat{\nu}$ (see Problem 10.27). (Hint: To see injectivity use Lemma 10.21 with $f = 1$ and a Schwartz function φ.)

Problem 14.12. *Consider the Fourier transform of a complex measure on* $\mathbb{R}$ *as in the previous problem. Show that*

$$\lim_{r\to\infty}\frac{1}{\sqrt{2\pi}}\int_{-r}^{r}\frac{\mathrm{e}^{\mathrm{i}bp}-\mathrm{e}^{\mathrm{i}ap}}{\mathrm{i}p}\hat{\mu}(p)dp=\frac{\mu([a,b])+\mu((a,b))}{2}.$$

(Hint: Insert the definition of $\hat{\mu}$ *and then use Fubini. To evaluate the limit you will need the Dirichlet integral from Problem 14.24).*

Problem 14.13 (Lévy continuity theorem)**.** *Let* $\mu_m(\mathbb{R}^n), \mu(\mathbb{R}^n) \le M$ *be positive measures and suppose the Fourier transforms (cf. Problem 14.11) converge pointwise*

$$\hat{\mu}_m(p)\to\hat{\mu}(p)$$

for $p\in\mathbb{R}^n$*. Then* $\mu_n\to\mu$ *vaguely. In fact, we have* (11.59) *for all* $f\in C_b(\mathbb{R}^n)$*. (Hint: Show* $\int\hat{f}d\mu=\int f(p)\hat{\mu}(p)dp$ *for* $f\in L^1(\mathbb{R}^n)$*.)*

14.2. Applications to linear partial differential equations

By virtue of Lemma 14.4 the Fourier transform can be used to map linear partial differential equations with constant coefficients to algebraic equations, thereby providing a mean of solving them. To illustrate this procedure we look at the famous **Poisson equation**, that is, given a function g, find a function f satisfying

$$-\Delta f=g. \tag{14.25}$$

For simplicity, let us start by investigating this problem in the space of Schwartz functions $\mathcal{S}(\mathbb{R}^n)$. Assuming there is a solution we can take the Fourier transform on both sides to obtain

$$|p|^2\hat{f}(p)=\hat{g}(p)\qquad\Rightarrow\qquad\hat{f}(p)=|p|^{-2}\hat{g}(p). \tag{14.26}$$

Since the right-hand side is integrable for $n\ge 3$ we obtain that our solution is necessarily given by

$$f(x)=(|p|^{-2}\hat{g}(p))^{\vee}(x). \tag{14.27}$$

In fact, this formula still works provided $g(x), |p|^{-2}\hat{g}(p)\in L^1(\mathbb{R}^n)$. Moreover, if we additionally assume $\hat{g}\in L^1(\mathbb{R}^n)$, then $|p|^2\hat{f}(p)=\hat{g}(p)\in L^1(\mathbb{R}^n)$ and Lemma 14.3 implies that $f\in C^2(\mathbb{R}^n)$ as well as that it is indeed a solution. Note that if $n\ge 3$, then $|p|^{-2}\hat{g}(p)\in L^1(\mathbb{R}^n)$ follows automatically from $g,\hat{g}\in L^1(\mathbb{R}^n)$ (show this!).

Moreover, we clearly expect that f should be given by a convolution. However, since $|p|^{-2}$ is not in $L^p(\mathbb{R}^n)$ for any p, the formulas derived so far do not apply.

Lemma 14.17. *Let $0 < \alpha < n$ and suppose $g \in L^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ as well as $|p|^{-\alpha}\hat{g}(p) \in L^1(\mathbb{R}^n)$. Then*

$$(|p|^{-\alpha}\hat{g}(p))^\vee(x) = \int_{\mathbb{R}^n} I_\alpha(|x-y|)g(y)d^n y, \tag{14.28}$$

*where the **Riesz potential** is given by*

$$I_\alpha(r) := \frac{\Gamma(\frac{n-\alpha}{2})}{2^\alpha \pi^{n/2}\Gamma(\frac{\alpha}{2})}\frac{1}{r^{n-\alpha}}. \tag{14.29}$$

Proof. Note that, while $|.|^{-\alpha}$ is not in $L^p(\mathbb{R}^n)$ for any p, our assumption $0 < \alpha < n$ ensures that the singularity at zero is integrable.

We set $\phi_t(p) = \exp(-t|p|^2/2)$ and begin with the elementary formula

$$|p|^{-\alpha} = c_\alpha \int_0^\infty \phi_t(p) t^{\alpha/2-1} dt, \quad c_\alpha = \frac{1}{2^{\alpha/2}\Gamma(\alpha/2)},$$

which follows from the definition of the gamma function (Problem 9.24) after a simple scaling. Since $|p|^{-\alpha}\hat{g}(p)$ is integrable we can use Fubini and Lemma 14.6 to obtain

$$\begin{aligned}(|p|^{-\alpha}\hat{g}(p))^\vee(x) &= \frac{c_\alpha}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}xp} \left(\int_0^\infty \phi_t(p) t^{\alpha/2-1} dt \right) \hat{g}(p) d^n p \\ &= \frac{c_\alpha}{(2\pi)^{n/2}} \int_0^\infty \left(\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}xp} \hat{\phi}_{1/t}(p)\hat{g}(p) d^n p \right) t^{(\alpha-n)/2-1} dt.\end{aligned}$$

Since $\phi, g \in L^1$ we know by Lemma 14.12 that $\hat{\phi}\hat{g} = (2\pi)^{-n/2}(\phi * g)^\wedge$ Moreover, since $\hat{\phi}\hat{g} \in L^1$ Theorem 14.7 gives us $(\hat{\phi}\hat{g})^\vee = (2\pi)^{-n/2}\phi * g$. Thus, we can make a change of variables and use Fubini once again (since $g \in L^\infty$)

$$\begin{aligned}(|p|^{-\alpha}\hat{g}(p))^\vee(x) &= \frac{c_\alpha}{(2\pi)^{n/2}} \int_0^\infty \left(\int_{\mathbb{R}^n} \phi_{1/t}(x-y) g(y) d^n y \right) t^{(\alpha-n)/2-1} dt \\ &= \frac{c_\alpha}{(2\pi)^{n/2}} \int_0^\infty \left(\int_{\mathbb{R}^n} \phi_t(x-y) g(y) d^n y \right) t^{(n-\alpha)/2-1} dt \\ &= \frac{c_\alpha}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \left(\int_0^\infty \phi_t(x-y) t^{(n-\alpha)/2-1} dt \right) g(y) d^n y \\ &= \frac{c_\alpha/c_{n-\alpha}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \frac{g(y)}{|x-y|^{n-\alpha}} d^n y\end{aligned}$$

to obtain the desired result. □

Note that the conditions of the above theorem are, for example, satisfied if $g, \hat{g} \in L^1(\mathbb{R}^n)$ which holds, for example, if $g \in \mathcal{S}(\mathbb{R}^n)$. In summary, if $g \in L^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$, $|p|^{-2}\hat{g}(p) \in L^1(\mathbb{R}^n)$ and $n \geq 3$, then

$$f = \Phi * g \tag{14.30}$$

is a classical solution of the Poisson equation, where

$$\Phi(x) := \frac{\Gamma(\frac{n}{2}-1)}{4\pi^{n/2}} \frac{1}{|x|^{n-2}}, \qquad n \geq 3, \tag{14.31}$$

is known as the **fundamental solution** of the Laplace equation.

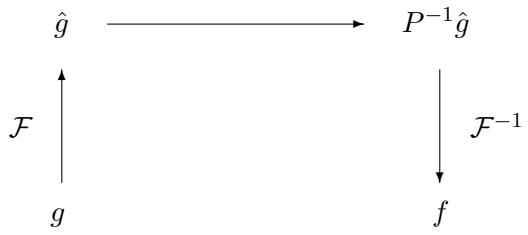
A few words about this formula are in order. First of all, our original formula in Fourier space shows that the multiplication with $|p|^{-2}$ improves the decay of $\hat{g}$ and hence, by virtue of Lemma 14.4, f should have, roughly speaking, two derivatives more than g. However, unless $\hat{g}(0)$ vanishes, multiplication with $|p|^{-2}$ will create a singularity at 0 and hence, again by Lemma 14.4, f will not inherit any decay properties from g. In fact, evaluating the above formula with $g = \chi_{B_1(0)}$ (Problem 14.14) shows that f might not decay better than Φ even for g with compact support.

Moreover, our conditions on g might not be easy to check as it will not be possible to compute $\hat{g}$ explicitly in general. So if one wants to deduce $\hat{g} \in L^1(\mathbb{R}^n)$ from properties of g, one could use Lemma 14.4 together with the Riemann–Lebesgue lemma to show that this condition holds if $g \in C^k(\mathbb{R}^n)$, $k > n-2$, such that all derivatives are integrable and all derivatives of order less than k vanish at ∞ (Problem 14.15). This seems a rather strong requirement since our solution formula will already make sense under the sole assumption $g \in L^1(\mathbb{R}^n)$. However, as the example $g = \chi_{B_1(0)}$ shows, this *solution* might not be C^2 and hence one needs to weaken the notion of a solution if one wants to include such situations. This will lead us to the concepts of weak derivatives and Sobolev spaces. As a preparation we will develop some further tools which will allow us to investigate continuity properties of the operator $\mathcal{I}_\alpha f = I_\alpha * f$ in the next section.

Before that, let us summarize the procedure in the general case. Suppose we have the following linear partial differential equations with constant coefficients:

$$P(\mathrm{i}\partial) f = g, \qquad P(\mathrm{i}\partial) = \sum_{\alpha \leq k} c_\alpha \mathrm{i}^{|\alpha|} \partial_\alpha. \tag{14.32}$$

Then the solution can be found via the procedure



and is formally given by

$$f(x) = (P(p)^{-1}\hat{g}(p))^{\vee}(x). \tag{14.33}$$

It remains to investigate the properties of the solution operator. In general, given a measurable function m one might try to define a corresponding operator via

$$A_m f := (m\hat{g})^{\vee}, \tag{14.34}$$

in which case m is known as a **Fourier multiplier**. It is said to be an L^p-multiplier if A_m can be extended to a bounded operator in $L^p(\mathbb{R}^n)$. For example, it will be an L^2 multiplier if m is bounded (in fact the converse is also true — Problem 14.16). As we have seen, in some cases A_m can be expressed as a convolution, but this is not always the case as the trivial example $m = 1$ (corresponding to the identity operator) shows.

Another famous example which can be solved in this way is the **Helmholtz equation**

$$-\Delta f + f = g. \tag{14.35}$$

As before we find that if $g(x), (1+|p|^2)^{-1}\hat{g}(p) \in L^1(\mathbb{R}^n)$ then the solution is given by

$$f(x) = ((1+|p|^2)^{-1}\hat{g}(p))^{\vee}(x). \tag{14.36}$$

Lemma 14.18. *Let $\alpha > 0$. Suppose $g \in L^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ as well as $(1+|p|^2)^{-\alpha/2}\hat{g}(p) \in L^1(\mathbb{R}^n)$. Then*

$$((1+|p|^2)^{-\alpha/2}\hat{g}(p))^{\vee}(x) = \int_{\mathbb{R}^n} J_\alpha(|x-y|)g(y)d^n y, \tag{14.37}$$

*where the **Bessel potential** is given by*

$$J_\alpha(r) := \frac{2}{(4\pi)^{n/2}\Gamma(\frac{\alpha}{2})}\left(\frac{r}{2}\right)^{-(n-\alpha)/2} K_{(n-\alpha)/2}(r), \qquad r > 0, \tag{14.38}$$

with

$$K_\nu(r) = K_{-\nu}(r) = \frac{1}{2}\left(\frac{r}{2}\right)^{\nu} \int_0^\infty \mathrm{e}^{-t-\frac{r^2}{4t}} \frac{dt}{t^{\nu+1}}, \qquad r > 0, \nu \in \mathbb{R}, \tag{14.39}$$

*the modified Bessel function of the second kind of order ν ([**31**, (10.32.10)]).*

Proof. We proceed as in the previous lemma. We set $\phi_t(p) = \exp(-t|p|^2/2)$ and begin with the elementary formula

$$\frac{\Gamma(\frac{\alpha}{2})}{(1+|p|^2)^{\alpha/2}} = \int_0^\infty t^{\alpha/2-1}\mathrm{e}^{-t(1+|p|^2)}dt.$$

Since $g, \hat{g}(p)$ are integrable we can use Fubini and Lemma 14.6 to obtain

$$\begin{aligned}(\frac{\hat{g}(p)}{(1+|p|^2)^{\alpha/2}})^\vee(x) &= \frac{\Gamma(\frac{\alpha}{2})^{-1}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}xp} \left(\int_0^\infty t^{\alpha/2-1} \mathrm{e}^{-t(1+|p|^2)} dt \right) \hat{g}(p) d^n p \\ &= \frac{\Gamma(\frac{\alpha}{2})^{-1}}{(4\pi)^{n/2}} \int_0^\infty \left(\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}xp} \hat{\phi}_{1/2t}(p) \hat{g}(p) d^n p \right) \mathrm{e}^{-t} t^{(\alpha-n)/2-1} dt \\ &= \frac{\Gamma(\frac{\alpha}{2})^{-1}}{(4\pi)^{n/2}} \int_0^\infty \left(\int_{\mathbb{R}^n} \phi_{1/2t}(x-y) g(y) d^n y \right) \mathrm{e}^{-t} t^{(\alpha-n)/2-1} dt \\ &= \frac{\Gamma(\frac{\alpha}{2})^{-1}}{(4\pi)^{n/2}} \int_{\mathbb{R}^n} \left(\int_0^\infty \phi_{1/2t}(x-y) \mathrm{e}^{-t} t^{(\alpha-n)/2-1} dt \right) g(y) d^n y\end{aligned}$$

to obtain the desired result. Using Fubini in the last step is allowed since g is bounded and $J_\alpha(|x|) \in L^1(\mathbb{R}^n)$ (Problem 14.17). □

Note that the first condition $g \in L^1(\mathbb{R}^n) \cap L^\infty(\mathbb{R}^n)$ implies $g \in L^2(\mathbb{R}^n)$ and thus the second condition $(1+|p|^2)^{-\alpha/2}\hat{g}(p) \in L^1(\mathbb{R}^n)$ will be satisfied if $\frac{n}{2} < \alpha$.

In particular, if $g, \hat{g} \in L^1(\mathbb{R}^n)$, then

$$f = J_2 * g \tag{14.40}$$

is a solution of Helmholtz equation. Note that since our multiplier $(1+|p|^2)^{-1}$ does not have a singularity near zero, the solution f will preserve (some) decay properties of g. For example, it will map Schwartz functions to Schwartz functions and thus for every $g \in \mathcal{S}(\mathbb{R}^n)$ there is a unique solution of the Helmholtz equation $f \in \mathcal{S}(\mathbb{R}^n)$. This is also reflected by the fact that the Bessel potential decays much faster than the Riesz potential. Indeed, one can show that [**31**, (10.25.3)]

$$K_\nu(r) = \sqrt{\frac{\pi}{2r}} \mathrm{e}^{-r}(1 + O(r^{-1})) \tag{14.41}$$

as $r \to \infty$. The singularity near zero is of the same type as for I_α since (see [**31**, (10.30.2) and (10.30.3)])

$$K_\nu(r) = \begin{cases} \frac{\Gamma(\nu)}{2} \left(\frac{r}{2}\right)^{-\nu} + O(r^{-\nu+2}), & \nu > 0, \\ -\log(\frac{r}{2}) + O(1), & \nu = 0, \end{cases} \tag{14.42}$$

for $r \to 0$.

Problem* 14.14. *Show that for $n = 3$ we have*

$$(\Phi * \chi_{B_1(0)})(x) = \begin{cases} \frac{1}{3|x|}, & |x| \geq 1, \\ \frac{3-|x|^2}{6}, & |x| \leq 1. \end{cases}$$

(Hint: Observe that the result depends only on $|x|$. Then choose $x = (0, 0, R)$ and evaluate the integral using spherical coordinates.)

Problem* 14.15. *Suppose $g \in C^k(\mathbb{R}^n)$ and $\partial_j^l g \in L^1(\mathbb{R}^n)$ for $j = 1, \dots, n$ and $0 \le l \le k$ as well as $\lim_{|x|\to\infty} \partial_j^l g(x) = 0$ for $j = 1, \dots, n$ and $0 \le l < k$. Then*

$$|\hat{g}(p)| \le \frac{C}{(1+|p|^2)^{k/2}}.$$

Problem* 14.16. *Show that m is an L^2 multiplier if and only if $m \in L^\infty(\mathbb{R}^n)$.*

Problem* 14.17. *Show*

$$\int_0^\infty J_\alpha(r) r^{n-1} dr = \frac{\Gamma(n/2)}{2\pi^{n/2}}, \quad \alpha > 0.$$

Conclude that

$$\|J_\alpha * g\|_p \le \|g\|_p.$$

(Hint: Fubini.)

14.3. Sobolev spaces

We have already introduced Sobolev spaces in Section 13.1. In this section we present an alternate (and in particular independent) approach to Sobolev spaces of index two (the Hilbert space case) on all of $\mathbb{R}^n$.

We begin by introducing the **Sobolev space**

$$H^r(\mathbb{R}^n) := \{f \in L^2(\mathbb{R}^n) | |p|^r \hat{f}(p) \in L^2(\mathbb{R}^n)\}. \tag{14.43}$$

The most important case is when r is an integer, however our definition makes sense for any $r \ge 0$. Moreover, note that $H^r(\mathbb{R}^n)$ becomes a Hilbert space if we introduce the scalar product

$$\langle f, g\rangle_{H^r} := \int_{\mathbb{R}^n} \hat{f}(p)^* \hat{g}(p)(1+|p|^2)^r d^n p. \tag{14.44}$$

In particular, note that by construction $\mathcal{F}$ maps $H^r(\mathbb{R}^n)$ unitarily onto $L^2(\mathbb{R}^n, (1+|p|^2)^r d^n p)$. Clearly $H^{r+1}(\mathbb{R}^n) \subset H^r(\mathbb{R}^n)$ with the embedding being continuous. Moreover, $\mathcal{S}(\mathbb{R}^n) \subset H^r(\mathbb{R}^n)$ and this subset is dense (since $\mathcal{S}(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n, (1+|p|^2)^r d^n p)$).

The motivation for the definition (14.43) stems from Lemma 14.4 which allows us to extend differentiation to a larger class. In fact, every function in $H^r(\mathbb{R}^n)$ has partial derivatives up to order $\lfloor r \rfloor$, which are defined via

$$\partial_\alpha f := ((\mathrm{i}p)^\alpha \hat{f}(p))^\vee, \qquad f \in H^r(\mathbb{R}^n),\ |\alpha| \le r. \tag{14.45}$$

By Lemma 14.4 this definition coincides with the usual one for every $f \in \mathcal{S}(\mathbb{R}^n)$.

Example 14.1. Consider $f(x) := (1-|x|)\chi_{[-1,1]}(x)$. Then $\hat{f}(p) = \sqrt{\frac{2}{\pi}} \frac{\cos(p)-1}{p^2}$ and $f \in H^1(\mathbb{R})$. The weak derivative is $f'(x) = -\operatorname{sign}(x)\chi_{[-1,1]}(x)$. ⋄

We also have

$$\begin{aligned}\int_{\mathbb{R}^n} g(x)(\partial_\alpha f)(x)d^n x &= \langle g^*, (\partial_\alpha f)\rangle = \langle \hat{g}(p)^*, (\mathrm{i}p)^\alpha \hat{f}(p)\rangle \\ &= (-1)^{|\alpha|}\langle (\mathrm{i}p)^\alpha \hat{g}(p)^*, \hat{f}(p)\rangle = (-1)^{|\alpha|}\langle \partial_\alpha g^*, f\rangle \\ &= (-1)^{|\alpha|}\int_{\mathbb{R}^n} (\partial_\alpha g)(x) f(x) d^n x, \end{aligned} \tag{14.46}$$

for $f, g \in H^r(\mathbb{R}^n)$. Furthermore, recall that a function $h \in L^1_{loc}(\mathbb{R}^n)$ satisfying

$$\int_{\mathbb{R}^n} \varphi(x)h(x)d^n x = (-1)^{|\alpha|}\int_{\mathbb{R}^n} (\partial_\alpha \varphi)(x) f(x) d^n x, \quad \forall \varphi \in C_c^\infty(\mathbb{R}^n), \tag{14.47}$$

is called the **weak derivative** or the derivative in the sense of distributions of f (by Lemma 10.21 such a function is unique if it exists). Hence, choosing $g = \varphi$ in (14.46), we see that functions in $H^r(\mathbb{R}^n)$ have weak derivatives up to order r, which are in $L^2(\mathbb{R}^n)$. Moreover, the weak derivatives coincide with the derivatives defined in (14.45). Conversely, given (14.47) with $f, h \in L^2(\mathbb{R}^n)$ we can use that $\mathcal{F}$ is unitary to conclude $\int_{\mathbb{R}^n} \hat{\varphi}(p)^* \hat{h}(p) d^n p = \int_{\mathbb{R}^n} p^\alpha \hat{\varphi}(p)\hat{f}(p) d^n p$ for all $\varphi \in C_c^\infty(\mathbb{R}^n)$. By approximation this follows for $\varphi \in H^r(\mathbb{R}^n)$ with $r \geq |\alpha|$ and hence in particular for $\hat{\varphi} \in C_c^\infty(\mathbb{R}^n)$. Consequently $p^\alpha \hat{f}(p) = \hat{h}(p)$ a.e. implying that $f \in H^r(\mathbb{R}^n)$ if all weak derivatives exist up to order r and are in $L^2(\mathbb{R}^n)$.

In this connection the following norm for $H^m(\mathbb{R}^n)$ with $m \in \mathbb{N}_0$ is more common:

$$\|f\|^2_{2,m} := \sum_{|\alpha|\leq m} \|\partial_\alpha f\|_2^2. \tag{14.48}$$

By $|p^\alpha| \leq |p|^{|\alpha|} \leq (1 + |p|^2)^{m/2}$ it follows that this norm is equivalent to (14.44).

Example 14.2. This definition of a weak derivative is tailored for the method of solving linear constant coefficient partial differential equations as outlined in Section 14.2. While Lemma 14.3 only gives us a sufficient condition on $\hat{f}$ for f to be differentiable, the weak derivatives gives us necessary and sufficient conditions. For example, we see that the Poisson equation (14.25) will have a (unique) solution $f \in H^2(\mathbb{R}^n)$ if and if $|p|^{-2}\hat{g} \in L^2(\mathbb{R}^n)$. That this is not true for all $g \in L^2(\mathbb{R}^n)$ is connected with the fact that $|p|^{-2}$ is unbounded and hence no L^2 multiplier (cf. Problem 14.16). Consequently the range of Δ when defined on $H^2(\mathbb{R}^n)$ will not be all of $L^2(\mathbb{R}^n)$ and hence the Poisson equation is not solvable within the class $H^2(\mathbb{R}^n)$ for all $g \in L^2(\mathbb{R}^n)$. Nevertheless, we get a unique weak solution under some conditions. Under which conditions this weak solution is also a classical solution can then be investigated separately.

Note that the situation is even simpler for the Helmholtz equation (14.35) since the corresponding multiplier $(1+|p|^2)^{-1}$ does map L^2 to L^2. Hence we get that the Helmholtz equation has a unique solution $f \in H^2(\mathbb{R}^n)$ if and only if $g \in L^2(\mathbb{R}^n)$. Moreover, $f \in H^{r+2}(\mathbb{R}^n)$ if and only if $g \in H^r(\mathbb{R}^n)$. $\diamond$

Of course a natural question to ask is when the weak derivatives are in fact classical derivatives. To this end observe that the Riemann–Lebesgue lemma implies that $\partial_\alpha f(x) \in C_0(\mathbb{R}^n)$ provided $p^\alpha \hat{f}(p) \in L^1(\mathbb{R}^n)$. Moreover, in this situation the derivatives will exist as classical derivatives:

Lemma 14.19. *Suppose $f \in L^1(\mathbb{R}^n)$ or $f \in L^2(\mathbb{R}^n)$ with $(1+|p|^k)\hat{f}(p) \in L^1(\mathbb{R}^n)$ for some $k \in \mathbb{N}_0$. Then $f \in C_0^k(\mathbb{R}^n)$, the set of functions with continuous partial derivatives of order k all of which vanish at ∞. Moreover,*

$$(\partial_\alpha f)^\wedge(p) = (\mathrm{i}p)^\alpha \hat{f}(p), \qquad |\alpha| \le k, \tag{14.49}$$

in this case.

Proof. We begin by observing that by Theorem 14.7

$$f(x) = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}px} \hat{f}(p) d^n p.$$

Now the claim follows as in the proof of Lemma 14.4 by differentiating the integral using Problem 9.15. □

Now we are able to prove the following embedding theorem.

Theorem 14.20 (Sobolev embedding)**.** *Suppose $r > k + \frac{n}{2}$ for some $k \in \mathbb{N}_0$. Then $H^r(\mathbb{R}^n)$ is continuously embedded into $C_0^k(\mathbb{R}^n)$ with*

$$\|\partial_\alpha f\|_\infty \le C_{n,r} \|f\|_{2,r}, \qquad |\alpha| \le k. \tag{14.50}$$

Proof. Abbreviate $\langle p \rangle = (1+|p|^2)^{1/2}$. Now use $|(\mathrm{i}p)^\alpha \hat{f}(p)| \le \langle p \rangle^{|\alpha|} |\hat{f}(p)| = \langle p \rangle^{-s} \cdot \langle p \rangle^{|\alpha|+s} |\hat{f}(p)|$. Now $\langle p \rangle^{-s} \in L^2(\mathbb{R}^n)$ if $s > \frac{n}{2}$ (use polar coordinates to compute the norm) and $\langle p \rangle^{|\alpha|+s} |\hat{f}(p)| \in L^2(\mathbb{R}^n)$ if $s + |\alpha| \le r$. Hence $\langle p \rangle^{|\alpha|} |\hat{f}(p)| \in L^1(\mathbb{R}^n)$ and the claim follows from the previous lemma. □

In fact, we can even do a bit better.

Lemma 14.21 (Morrey inequality)**.** *Suppose $f \in H^{n/2+\gamma}(\mathbb{R}^n)$ for some $\gamma \in (0,1)$. Then $f \in C_0^{0,\gamma}(\mathbb{R}^n)$, the set of functions which are Hölder continuous with exponent γ and vanish at ∞. Moreover,*

$$|f(x) - f(y)| \le C_{n,\gamma} \|\hat{f}(p)\|_{2,n/2+\gamma} |x-y|^\gamma \tag{14.51}$$

in this case.

Proof. We begin with

$$f(x+y)-f(x)=\frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\mathrm{e}^{\mathrm{i}px}(\mathrm{e}^{\mathrm{i}py}-1)\hat{f}(p)d^np$$

implying

$$|f(x+y)-f(x)|\leq\frac{1}{(2\pi)^{n/2}}\int_{\mathbb{R}^n}\frac{|\mathrm{e}^{\mathrm{i}py}-1|}{\langle p\rangle^{n/2+\gamma}}\langle p\rangle^{n/2+\gamma}|\hat{f}(p)|d^np,$$

where again $\langle p\rangle=(1+|p|^2)^{1/2}$. Hence, after applying Cauchy–Schwarz, it remains to estimate (recall (9.42))

$$\begin{aligned}\int_{\mathbb{R}^n}\frac{|\mathrm{e}^{\mathrm{i}py}-1|^2}{\langle p\rangle^{n+2\gamma}}d^np&\leq S_n\int_0^{1/|y|}\frac{(|y|r)^2}{\langle r\rangle^{n+2\gamma}}r^{n-1}dr\\&\quad+S_n\int_{1/|y|}^{\infty}\frac{4}{\langle r\rangle^{n+2\gamma}}r^{n-1}dr\\&\leq\frac{S_n}{2(1-\gamma)}|y|^{2\gamma}+\frac{S_n}{2\gamma}|y|^{2\gamma}=\frac{S_n}{2\gamma(1-\gamma)}|y|^{2\gamma},\end{aligned}$$

where $S_n=nV_n$ is the surface area of the unit sphere in $\mathbb{R}^n$. □

Using this lemma we immediately obtain:

Corollary 14.22. *Suppose $r\geq k+\gamma+\frac{n}{2}$ for some $k\in\mathbb{N}_0$ and $\gamma\in(0,1)$. Then $H^r(\mathbb{R}^n)$ is continuously embedded into $C_0^{k,\gamma}(\mathbb{R}^n)$, the set of functions in $C_0^k(\mathbb{R}^n)$ whose highest derivatives are Hölder continuous of exponent γ.*

Example 14.3. The function $f(x)=\log(|x|)$ is in $H^1(\mathbb{R}^n)$ for $n\geq 3$. In fact, the weak derivatives are given by

$$\partial_jf(x)=\frac{x_j}{|x|^2}. \tag{14.52}$$

However, observe that f is not continuous. ⋄

The last example shows that in the case $r<\frac{n}{2}$ functions in H^r are no longer necessarily continuous. In this case we at least get an embedding into some *better* L^p space:

Theorem 14.23 (Sobolev inequality)**.** *Suppose $0<r<\frac{n}{2}$. Then $H^r(\mathbb{R}^n)$ is continuously embedded into $L^p(\mathbb{R}^n)$ with $p=\frac{2n}{n-2r}$, that is,*

$$\|f\|_p\leq\tilde{C}_{n,r}\||.|^r\hat{f}(.)\|_2\leq C_{n,r}\|f\|_{2,r}. \tag{14.53}$$

Proof. We will give a prove based on the Hardy–Littlewood–Sobolev inequality to be proven in Theorem 15.10 below.

It suffices to prove the first inequality. Set $|p|^r\hat{f}(p)=\hat{g}(p)\in L^2$. Moreover, choose some sequence $f_m\in\mathcal{S}\to f\in H^r$. Then, by Lemma 14.17 $f_m=\mathcal{I}_rg_m$, and since the Hardy–Littlewood–Sobolev inequality implies that

the map $\mathcal{I}_r : L^2 \to L^p$ is continuous, we have $\|f_m\|_p = \|\mathcal{I}_r g_m\|_p \le \tilde{C}\|g_m\|_2 = \tilde{C}\|\hat{g}_m\|_2 = \tilde{C}\||p|^r \hat{f}_m(p)\|_2$ and the claim follows after taking limits. □

Problem 14.18. *Use dilations $f(x) \mapsto f(\lambda x)$, $\lambda > 0$, to show that $p = \frac{2n}{n-2r}$ is the only index for which the Sobolev inequality $\|f\|_p \le \tilde{C}_{n,r}\||p|^r \hat{f}(p)\|_2$ can hold.*

Problem 14.19. *Suppose $f \in L^2(\mathbb{R}^n)$ show that $\varepsilon^{-1}(f(x + e_j\varepsilon) - f(x)) \to g_j(x)$ in L^2 if and only if $p_j \hat{f}(p) \in L^2$, where e_j is the unit vector into the j'th coordinate direction. Moreover, show $g_j = \partial_j f$ if $f \in H^1(\mathbb{R}^n)$.*

Problem 14.20. *Suppose $f, g \in H^r(\mathbb{R}^n)$ for $r > \frac{n}{2}$. Show that $fg \in H^r(\mathbb{R}^n)$ with $\|fg\|_{2,r} \le C\|f\|_{2,r}\|g\|_{2,r}$.*

14.4. Applications to evolution equations

In this section we want to show how to apply these considerations to evolution equations. As a prototypical example we start with the Cauchy problem for the **heat equation**

$$u_t - \Delta u = 0, \qquad u(0) = g. \tag{14.54}$$

It turns out useful to view $u(t, x)$ as a function of t with values in a Banach space X. To this end we let $I \subseteq \mathbb{R}$ be some interval and denote by $C(I, X)$ the set of continuous functions from I to X. Given $t \in I$ we call $u : I \to X$ differentiable at t if the limit

$$\dot{u}(t) = \lim_{\varepsilon \to 0} \frac{u(t + \varepsilon) - u(t)}{\varepsilon} \tag{14.55}$$

exists. The set of functions $u : I \to X$ which are differentiable at all $t \in I$ and for which $\dot{u} \in C(I, X)$ is denoted by $C^1(I, X)$. As usual we set $C^{k+1}(I, X) = \{u \in C^1(I, x) | \dot{u} \in C^k(I, X)\}$. Note that if $U \in \mathscr{L}(X, Y)$ and $u \in C^1(I, X)$, then $Uu \in C^1(I, Y)$ and $\frac{d}{dt} Uu = U\dot{u}$.

A **strongly continuous operator semigroup** (also C_0-semigroup) is a family of operators $T(t) \in \mathscr{L}(X)$, $t \ge 0$, such that

(i) $T(t)g \in C([0, \infty), X)$ for every $g \in X$ (strong continuity) and
(ii) $T(0) = \mathbb{I}$, $T(t + s) = T(t)T(s)$ for every $t, s \ge 0$ (semigroup property).

Given a strongly continuous semigroup we can define its **generator** A as the linear operator

$$Af = \lim_{t \downarrow 0} \frac{1}{t}\big(T(t)f - f\big) \tag{14.56}$$

where the domain $\mathfrak{D}(A)$ is precisely the set of all $f \in X$ for which the above limit exists. The key result is that if A generates a C_0-semigroup $T(t)$,

then $u(t) := T(t)g$ will be the unique solution of the corresponding abstract Cauchy problem. More precisely we have (see Lemma 7.7):

Lemma 14.24. *Let $T(t)$ be a C_0-semigroup with generator A. If $g \in X$ with $u(t) = T(t)g \in \mathfrak{D}(A)$ for $t > 0$ then $u(t) \in C^1((0,\infty), X) \cap C([0,\infty), X)$ and $u(t)$ is the unique solution of the abstract Cauchy problem*

$$\dot{u}(t) = Au(t), \qquad u(0) = g. \tag{14.57}$$

This is, for example, the case if $g \in \mathfrak{D}(A)$ in which case we even have $u(t) \in C^1([0,\infty), X)$.

After these preparations we are ready to return to our original problem (14.54). Let $g \in L^2(\mathbb{R}^n)$ and let $u \in C^1((0,\infty), L^2(\mathbb{R}^n))$ be a solution such that $u(t) \in H^2(\mathbb{R}^n)$ for $t > 0$. Then we can take the Fourier transform to obtain

$$\hat{u}_t + |p|^2 \hat{u} = 0, \qquad \hat{u}(0) = \hat{g}. \tag{14.58}$$

Next, one verifies (Problem 14.21) that the solution (in the sense defined above) of this differential equation is given by

$$\hat{u}(t)(p) = \hat{g}(p) \mathrm{e}^{-|p|^2 t}. \tag{14.59}$$

Accordingly, the solution of our original problem is

$$u(t) = T_H(t)g, \qquad T_H(t) = \mathcal{F}^{-1} \mathrm{e}^{-|p|^2 t} \mathcal{F}. \tag{14.60}$$

Note that $T_H(t) : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is a bounded linear operator with $\|T_H(t)\| \le 1$ (since $|\mathrm{e}^{-|p|^2 t}| \le 1$). In fact, for $t > 0$ we even have $T_H(t)g \in H^r(\mathbb{R}^n)$ for any $r \ge 0$ showing that $u(t)$ is smooth even for *rough* initial functions g. In summary,

Theorem 14.25. *The family $T_H(t)$ is a C_0-semigroup whose generator is Δ, $\mathfrak{D}(\Delta) = H^2(\mathbb{R}^n)$.*

Proof. That $H^2(\mathbb{R}^n) \subseteq \mathfrak{D}(A)$ follows from Problem 14.21. Conversely, let $g \notin H^2(\mathbb{R}^n)$. Then $t^{-1}(\mathrm{e}^{-|p|^2 t} - 1) \to -|p|^2$ uniformly on every compact subset $K \subset \mathbb{R}^n$. Hence $\int_K |p|^2 |\hat{g}(p)|^2 d^n p = \int_K |Ag(x)|^2 d^n x$ which gives a contradiction as K increases. □

Next we want to derive a more explicit formula for our solution. To this end we assume $g \in L^1(\mathbb{R}^n)$ and introduce

$$\Phi_t(x) = \frac{1}{(4\pi t)^{n/2}} \mathrm{e}^{-\frac{|x|^2}{4t}}, \tag{14.61}$$

known as the **fundamental solution** of the heat equation, such that

$$\hat{u}(t) = (2\pi)^{n/2} \hat{g} \hat{\Phi}_t = (\Phi_t * g)^\wedge \tag{14.62}$$

by Lemma 14.6 and Lemma 14.12. Finally, by injectivity of the Fourier transform (Theorem 14.7) we conclude

$$u(t) = \Phi_t * g. \tag{14.63}$$

Moreover, one can check directly that (14.63) defines a solution for arbitrary $g \in L^p(\mathbb{R}^n)$.

Theorem 14.26. *Suppose $g \in L^p(\mathbb{R}^n)$, $1 \le p \le \infty$. Then (14.63) defines a solution for the heat equation which satisfies $u \in C^\infty((0,\infty) \times \mathbb{R}^n)$. The solutions has the following properties:*

(i) *If $1 \le p < \infty$, then $\lim_{t\downarrow 0} u(t) = g$ in L^p. If $p = \infty$ this holds for $g \in C_0(\mathbb{R}^n)$.*

(ii) *If $p = \infty$, then*

$$\|u(t)\|_\infty \le \|g\|_\infty. \tag{14.64}$$

If g is real-valued then so is u and

$$\inf g \le u(t) \le \sup g. \tag{14.65}$$

(iii) *(Mass conservation) If $p = 1$, then*

$$\int_{\mathbb{R}^n} u(t,x) d^n x = \int_{\mathbb{R}^n} g(x) d^n x \tag{14.66}$$

and

$$\|u(t)\|_\infty \le \frac{1}{(4\pi t)^{n/2}} \|g\|_1. \tag{14.67}$$

Proof. That $u \in C^\infty$ follows since $\Phi \in C^\infty$ from Problem 9.15. To see the remaining claims we begin by noting (by Problem 9.23)

$$\int_{\mathbb{R}^n} \Phi_t(x) d^n x = 1. \tag{14.68}$$

Now (i) follows from Lemma 10.19, (ii) is immediate, and (iii) follows from Fubini. □

Note that using Young's inequality (15.9) (to be established below) we even have

$$\|u(t)\|_\infty \le \|\Phi_t\|_q \|g\|_p = \frac{1}{q^{\frac{n}{2q}} (4\pi t)^{\frac{n}{2p}}} \|g\|_p, \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{14.69}$$

Another closely related equation is the **Schrödinger equation**

$$-\mathrm{i} u_t - \Delta u = 0, \qquad u(0) = g. \tag{14.70}$$

As before we obtain that the solution for $g \in H^2(\mathbb{R}^n)$ is given by

$$u(t) = T_S(t) g, \qquad T_S(t) = \mathcal{F}^{-1} \mathrm{e}^{-\mathrm{i}|p|^2 t} \mathcal{F}. \tag{14.71}$$

Note that $T_S(t) : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ is a unitary operator (since $|\mathrm{e}^{-\mathrm{i}|p|^2 t}| = 1$):

$$\|u(t)\|_2 = \|g\|_2. \tag{14.72}$$

However, while we have $T_H(t)g \in H^r(\mathbb{R}^n)$ whenever $g \in H^r(\mathbb{R}^n)$, unlike the heat equation, the Schrödinger equation does only preserve but not improve the regularity of the initial condition.

Theorem 14.27. *The family $T_S(t)$ is a C_0-group whose generator is $\mathrm{i}\Delta$, $\mathfrak{D}(\mathrm{i}\Delta) = H^2(\mathbb{R}^n)$.*

As in the case of the heat equation, we would like to express our solution as a convolution with the initial condition. However, now we run into the problem that $\mathrm{e}^{-\mathrm{i}|p|^2 t}$ is not integrable. To overcome this problem we consider

$$f_\varepsilon(p) = \mathrm{e}^{-(\mathrm{i}t+\varepsilon)p^2}, \quad \varepsilon > 0. \tag{14.73}$$

Then, as before we have

$$(f_\varepsilon \hat{g})^\vee(x) = \frac{1}{(4\pi(\mathrm{i}t+\varepsilon))^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{-\frac{|x-y|^2}{4(\mathrm{i}t+\varepsilon)}} g(y) d^n y \tag{14.74}$$

and hence

$$T_S(t)g(x) = \frac{1}{(4\pi\mathrm{i}t)^{n/2}} \int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\frac{|x-y|^2}{4t}} g(y) d^n y \tag{14.75}$$

for $t \neq 0$ and $g \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$. In fact, letting $\varepsilon \downarrow 0$ the left-hand side converges to $T_S(t)g$ in L^2 and the limit of the right-hand side exists pointwise by dominated convergence and its pointwise limit must thus be equal to its L^2 limit.

Using this explicit form, we can again draw some further consequences. For example, if $g \in L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n)$, then $g(t) \in C(\mathbb{R}^n)$ for $t \neq 0$ (use dominated convergence and continuity of the exponential) and satisfies

$$\|u(t)\|_\infty \leq \frac{1}{|4\pi t|^{n/2}} \|g\|_1. \tag{14.76}$$

Thus we have again spreading of wave functions in this case.

Finally we turn to the **wave equation**

$$u_{tt} - \Delta u = 0, \qquad u(0) = g, \quad u_t(0) = f. \tag{14.77}$$

This equation will fit into our framework once we transform it to a first order system with respect to time:

$$u_t = v, \quad v_t = \Delta u, \qquad u(0) = g, \quad v(0) = f. \tag{14.78}$$

After applying the Fourier transform this system reads

$$\hat{u}_t = \hat{v}, \quad \hat{v}_t = -|p|^2 \hat{u}, \qquad \hat{u}(0) = \hat{g}, \quad \hat{v}(0) = \hat{f}, \tag{14.79}$$

and the solution is given by

$$\begin{aligned}\hat{u}(t,p) &= \cos(t|p|)\hat{g}(p) + \frac{\sin(t|p|)}{|p|}\hat{f}(p),\\ \hat{v}(t,p) &= -\sin(t|p|)|p|\hat{g}(p) + \cos(t|p|)\hat{f}(p).\end{aligned} \tag{14.80}$$

Hence for $(g,f) \in H^2(\mathbb{R}^n) \oplus H^1(\mathbb{R}^n)$ our solution is given by

$$\begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = T_W(t) \begin{pmatrix} g \\ f \end{pmatrix}, \quad T_W(t) = \mathcal{F}^{-1} \begin{pmatrix} \cos(t|p|) & \frac{\sin(t|p|)}{|p|} \\ -\sin(t|p|)|p| & \cos(t|p|) \end{pmatrix} \mathcal{F}. \tag{14.81}$$

Theorem 14.28. *The family $T_W(t)$ is a C_0-semigroup whose generator is $A = \left(\begin{smallmatrix} 0 & 1 \\ \Delta & 0 \end{smallmatrix}\right)$, $\mathfrak{D}(A) = H^2(\mathbb{R}^n) \oplus H^1(\mathbb{R}^n)$.*

Note that if we use w defined via $\hat{w}(p) = |p|\hat{v}(p)$ instead of v, then

$$\begin{pmatrix} u(t) \\ w(t) \end{pmatrix} = \tilde{T}_W(t) \begin{pmatrix} g \\ h \end{pmatrix}, \quad \tilde{T}_W(t) = \mathcal{F}^{-1} \begin{pmatrix} \cos(t|p|) & \sin(t|p|) \\ -\sin(t|p|) & \cos(t|p|) \end{pmatrix} \mathcal{F}, \tag{14.82}$$

where h is defined via $\hat{h} = |p|\hat{f}$. In this case $\tilde{T}_W$ is unitary and thus

$$\|u(t)\|_2^2 + \|w(t)\|_2^2 = \|g\|_2^2 + \|h\|_2^2. \tag{14.83}$$

Note that $\|w\|_2^2 = \langle w, w\rangle = \langle \hat{w}, \hat{w}\rangle = \langle \hat{v}, |p|^2\hat{v}\rangle = -\langle v, \Delta v\rangle$.

If $n = 1$ we have $\frac{\sin(t|p|)}{|p|} \in L^2(\mathbb{R})$ and hence we can get an expression in terms of convolutions. In fact, since the inverse Fourier transform of $\frac{\sin(t|p|)}{|p|}$ is $\sqrt{\frac{\pi}{2}}\chi_{[-1,1]}(p/t)$, we obtain

$$u(t,x) = \int_{\mathbb{R}} \frac{1}{2}\chi_{[-t,t]}(x-y)f(y)dy = \frac{1}{2}\int_{x-t}^{x+t} f(y)dy$$

in the case $g = 0$. But the corresponding expression for $f = 0$ is just the time derivative of this expression and thus

$$\begin{aligned}u(t,x) &= \frac{1}{2}\frac{\partial}{\partial t}\int_{x-t}^{x+t} g(y)dy + \frac{1}{2}\int_{x-t}^{x+t} f(y)dy\\ &= \frac{g(x+t)+g(x-t)}{2} + \frac{1}{2}\int_{x-t}^{x+t} f(y)dy,\end{aligned} \tag{14.84}$$

which is known as **d'Alembert's formula**.

To obtain the corresponding formula in $n = 3$ dimensions we use the following observation

$$\frac{\partial}{\partial t}\hat{\varphi}_t(p) = \frac{\sin(t|p|)}{|p|}, \qquad \hat{\varphi}_t(p) = \frac{1-\cos(t|p|)}{|p|^2}, \tag{14.85}$$

where $\hat{\varphi}_t \in L^2(\mathbb{R}^3)$. Hence we can compute its inverse Fourier transform using

$$\varphi_t(x) = \lim_{R\to\infty} \frac{1}{(2\pi)^{3/2}} \int_{B_R(0)} \hat{\varphi}_t(p) \mathrm{e}^{\mathrm{i}px} d^3p \tag{14.86}$$

using spherical coordinates (without loss of generality we can rotate our coordinate system, such that the third coordinate direction is parallel to x)

$$\varphi_t(x) = \lim_{R\to\infty} \frac{1}{(2\pi)^{3/2}} \int_0^R \int_0^\pi \int_0^{2\pi} \frac{1-\cos(tr)}{r^2} \mathrm{e}^{\mathrm{i}r|x|\cos(\theta)} r^2 \sin(\theta) d\varphi d\theta dr. \tag{14.87}$$

Evalutaing the integrals we obtain

$$\begin{aligned}
\varphi_t(x) &= \lim_{R\to\infty} \frac{1}{\sqrt{2\pi}} \int_0^R (1-\cos(tr)) \left(\int_0^\pi \mathrm{e}^{\mathrm{i}r|x|\cos(\theta)} \sin(\theta) d\theta \right) dr \\
&= \lim_{R\to\infty} \sqrt{\frac{2}{\pi}} \int_0^R (1-\cos(tr)) \frac{\sin(r|x|)}{|x|r} dr \\
&= \lim_{R\to\infty} \frac{1}{\sqrt{2\pi}} \int_0^R \left(2\frac{\sin(r|x|)}{|x|r} + \frac{\sin(r(t-|x|))}{|x|r} - \frac{\sin(r(t+|x|))}{|x|r} \right) dr, \\
&= \lim_{R\to\infty} \frac{1}{\sqrt{2\pi}|x|} \left(2\,\mathrm{Si}(R|x|) + \mathrm{Si}(R(t-|x|)) - \mathrm{Si}(R(t+|x|)) \right),
\end{aligned} \tag{14.88}$$

where

$$\mathrm{Si}(z) = \int_0^z \frac{\sin(x)}{x} dx \tag{14.89}$$

is the **sine integral**. Using $\mathrm{Si}(-x) = -\mathrm{Si}(x)$ for $x \in \mathbb{R}$ and (Problem 14.24)

$$\lim_{x\to\infty} \mathrm{Si}(x) = \frac{\pi}{2} \tag{14.90}$$

we finally obtain (since the pointwise limit must equal the L^2 limit)

$$\varphi_t(x) = \sqrt{\frac{\pi}{2}} \frac{\chi_{[0,t]}(|x|)}{|x|}. \tag{14.91}$$

For the wave equation this implies (using Lemma 9.17)

$$\begin{aligned}
u(t,x) &= \frac{1}{4\pi} \frac{\partial}{\partial t} \int_{B_{|t|}(|x|)} \frac{1}{|x-y|} f(y) d^3y \\
&= \frac{1}{4\pi} \frac{\partial}{\partial t} \int_0^{|t|} \int_{S^2} \frac{1}{r} f(x-r\omega) r^2 d\sigma^2(\omega) dr \\
&= \frac{t}{4\pi} \int_{S^2} f(x-t\theta) d\sigma^2(\omega)
\end{aligned} \tag{14.92}$$

and thus finally

$$u(t,x) = \frac{\partial}{\partial t} \frac{t}{4\pi} \int_{S^2} g(x-t\omega) d\sigma^2(\omega) + \frac{t}{4\pi} \int_{S^2} f(x-t\omega) d\sigma^2(\omega), \tag{14.93}$$

which is known as **Kirchhoff's formula**.

Finally, to obtain a formula in $n = 2$ dimensions we use the method of descent: That is we use the fact, that our solution in two dimensions is also a solution in three dimensions which happens to be independent of the third coordinate direction. Unfortunately this does not fit within our current framework since such functions are not square integrable (unless they vanish identically). However, ignoring this fact and assuming our solution is given by Kirchhoff's formula we can simplify the integral using the fact that f does not depend on x_3. Using spherical coordinates we obtain

$$\begin{aligned}
&\frac{t}{4\pi}\int_{S^2} f(x-t\omega)d\sigma^2(\omega) = \\
&\quad = \frac{t}{2\pi}\int_0^{\pi/2}\int_0^{2\pi} f(x_1 - t\sin(\theta)\cos(\varphi), x_2 - t\sin(\theta)\sin(\varphi))\sin(\theta)d\theta d\varphi \\
&\quad \overset{\rho=\sin(\theta)}{=} \frac{t}{2\pi}\int_0^1\int_0^{2\pi} \frac{f(x_1 - t\rho\cos(\varphi), x_2 - t\rho\sin(\varphi))}{\sqrt{1-\rho^2}}\rho d\rho\, d\varphi \\
&\quad = \frac{t}{2\pi}\int_{B_1(0)} \frac{f(x-ty)}{\sqrt{1-|y|^2}}d^2y
\end{aligned}$$

which gives **Poisson's formula**

$$u(t,x) = \frac{\partial}{\partial t}\frac{t}{2\pi}\int_{B_1(0)} \frac{g(x-ty)}{\sqrt{1-|y|^2}}d^2y + \frac{t}{2\pi}\int_{B_1(0)} \frac{f(x-ty)}{\sqrt{1-|y|^2}}d^2y. \tag{14.94}$$

Problem 14.21. *Show that $u(t)$ defined in (14.59) is in $C^1((0,\infty), L^2(\mathbb{R}^n))$ and solves (14.58). (Hint: $|\mathrm{e}^{-t|p|^2} - 1| \le t|p|^2$ for $t \ge 0$.)*

Problem 14.22. *Suppose $u(t) \in C^1(I, X)$. Show that for $s, t \in I$*

$$\|u(t) - u(s)\| \le M|t-s|, \qquad M = \sup_{\tau\in[s,t]} \|\frac{du}{dt}(\tau)\|.$$

(Hint: Consider $d(\tau) = \|u(\tau) - u(s)\| - \tilde{M}(\tau - s)$ for $\tau \in [s,t]$. Suppose τ_0 is the largest τ for which the claim holds with $\tilde{M} > M$ and find a contradiction if $\tau_0 < t$.)

Problem 14.23. *Solve the **transport equation***

$$u_t + a\partial_x u = 0, \qquad u(0) = g,$$

using the Fourier transform.

Problem* 14.24. *Show the **Dirichlet integral***

$$\lim_{R\to\infty}\int_0^R \frac{\sin(x)}{x}dx = \frac{\pi}{2}.$$

Show also that the sine integral is bounded

$$|\operatorname{Si}(x)| \leq \min(x, \pi(1 + \frac{1}{2\mathrm{e}x})), \qquad x > 0.$$

(Hint: Write $\operatorname{Si}(R) = \int_0^R \int_0^\infty \sin(x)\mathrm{e}^{-xt} dt\, dx$ *and use Fubini.)*

14.5. Tempered distributions

In many situation, in particular when dealing with partial differential equations, it turns out convenient to look at generalized functions, also known as distributions.

To begin with we take a closer look at the Schwartz space $\mathcal{S}(\mathbb{R}^m)$, defined in (14.10), which already turned out to be a convenient class for the Fourier transform. For our purpose it will be crucial to have a notion of convergence in $\mathcal{S}(\mathbb{R}^m)$ and the natural choice is the topology generated by the seminorms

$$q_n(f) = \sum_{|\alpha|,|\beta| \leq n} \|x^\alpha (\partial_\beta f)(x)\|_\infty, \tag{14.95}$$

where the sum runs over all multi indices $\alpha, \beta \in \mathbb{N}_0^m$ of order less than n. Unfortunately these seminorms cannot be replaced by a single norm (and hence we do not have a Banach space) but there is at least a metric

$$d(f, g) = \sum_{n=1}^\infty \frac{1}{2^n} \frac{q_n(f-g)}{1 + q_n(f-g)} \tag{14.96}$$

and $\mathcal{S}(\mathbb{R}^m)$ is complete with respect to this metric and hence a Fréchet space:

Lemma 14.29. *The Schwartz space* $\mathcal{S}(\mathbb{R}^m)$ *together with the family of seminorms* $\{q_n\}_{n \in \mathbb{N}_0}$ *is a Fréchet space.*

Proof. It suffices to show completeness. Since a Cauchy sequence f_n is in particular a Cauchy sequence in $C^\infty(\mathbb{R}^m)$ there is a limit $f \in C^\infty(\mathbb{R}^m)$ such that all derivatives converge uniformly. Moreover, since Cauchy sequences are bounded $\|x^\alpha(\partial_\beta f_n)(x)\|_\infty \leq C_{\alpha,\beta}$ we obtain $\|x^\alpha(\partial_\beta f)(x)\|_\infty \leq C_{\alpha,\beta}$ and thus $f \in \mathcal{S}(\mathbb{R}^m)$. □

We refer to Section 5.4 for further background on Fréchet spaces. However, for our present purpose it is sufficient to observe that $f_n \to f$ if and only if $q_k(f_n - f) \to 0$ for every $k \in \mathbb{N}_0$. Moreover, (cf. Corollary 5.16) a linear map $A : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$ is continuous if and only if for every $j \in \mathbb{N}_0$ there is some $k \in \mathbb{N}_0$ and a corresponding constant C_k such that $q_j(Af) \leq C_k q_k(f)$ and a linear functional $\ell : \mathcal{S}(\mathbb{R}^m) \to \mathbb{C}$ is continuous if and only if there is some $k \in \mathbb{N}_0$ and a corresponding constant C_k such that $|\ell(f)| \leq C_k q_k(f)$.

Now the set of of all continuous linear functionals, that is the dual space $\mathcal{S}^*(\mathbb{R}^m)$, is known as the space of **tempered distributions**. To understand why this generalizes the concept of a function we begin by observing that any locally integrable function which does not grow too fast gives rise to a distribution.

Example 14.4. Let g be a locally integrable function of at most polynomial growth, that is, there is some $k \in \mathbb{N}_0$ such that $C_k := \int_{\mathbb{R}^m} |g(x)|(1 + |x|)^{-k} d^m x < \infty$. Then

$$T_g(f) := \int_{\mathbb{R}^m} g(x) f(x) d^m x$$

is a distribution. To see that T_g is continuous observe that $|T_g(f)| \leq C_k q_k(f)$. Moreover, note that by Lemma 10.21 the distribution T_g and the function g uniquely determine each other. ◇

The next question is if there are distributions which are not functions.

Example 14.5. Let $x_0 \in \mathbb{R}^m$ then

$$\delta_{x_0}(f) := f(x_0)$$

is a distribution, the **Dirac delta distribution** centered at x_0. Continuity follows from $|\delta_{x_0}(f)| \leq q_0(f)$ Formally δ_{x_0} can be written as $T_{\delta_{x_0}}$ as in the previous example where δ_{x_0} is the Dirac δ-function which satisfies $\delta_{x_0}(x) = 0$ for $x \neq x_0$ and $\delta_{x_0}(x) = \infty$ such that $\int_{\mathbb{R}^m} \delta_{x_0}(x) f(x) d^m x = f(x_0)$. This is of course nonsense as one can easily see that δ_{x_0} cannot be expressed as T_g with at locally integrable function of at most polynomial growth (show this). However, giving a precise mathematical meaning to the Dirac δ-function was one of the main motivations to develop distribution theory. ◇

Example 14.6. This example can be easily generalized: Let μ be a Borel measure on $\mathbb{R}^m$ such that $C_k := \int_{\mathbb{R}^m} (1 + |x|)^{-k} d\mu(x) < \infty$ for some k, then

$$T_\mu(f) := \int_{\mathbb{R}^m} f(x) d\mu(x)$$

is a distribution since $|T_\mu(f)| \leq C_k q_k(f)$. ◇

Example 14.7. Another interesting distribution in $\mathcal{S}^*(\mathbb{R})$ is given by

$$\big(p.v.\frac{1}{x}\big)(f) := \lim_{\varepsilon \downarrow 0} \int_{|x|>\varepsilon} \frac{f(x)}{x} dx.$$

To see that this is a distribution note that by the mean value theorem

$$\begin{aligned}|\big(p.v.\frac{1}{x}\big)(f)| &= \int_{\varepsilon<|x|<1} \frac{f(x)-f(0)}{x}dx + \int_{1<|x|} \frac{f(x)}{x}dx \\ &\le \int_{\varepsilon<|x|<1} \left|\frac{f(x)-f(0)}{x}\right| dx + \int_{1<|x|} \frac{|x\,f(x)|}{x^2}dx \\ &\le 2\sup_{|x|\le 1}|f'(x)| + 2\sup_{|x|\ge 1}|x\,f(x)|.\end{aligned}$$

This shows $|\big(p.v.\frac{1}{x}\big)(f)| \le 2q_1(f)$. ⋄

Of course, to fill distribution theory with life, we need to extend the classical operations for functions to distributions. First of all, addition and multiplication by scalars comes for free, but we can easily do more. The general principle is always the same: For any continuous linear operator $A : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$ there is a corresponding adjoint operator $A' : \mathcal{S}^*(\mathbb{R}^m) \to \mathcal{S}^*(\mathbb{R}^m)$ which extends the effect on functions (regarded as distributions of type T_g) to all distributions. We start with a simple example illustrating this procedure.

Let $h \in \mathcal{S}(\mathbb{R}^m)$, then the map $A : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$, $f \mapsto h \cdot f$ is continuous. In fact, continuity follows from the **Leibniz rule**

$$\partial_\alpha(h \cdot f) = \sum_{\beta\le\alpha} \binom{\alpha}{\beta} (\partial_\beta h)(\partial_{\alpha-\beta} f),$$

where $\binom{\alpha}{\beta} = \frac{\alpha!}{\beta!(\alpha-\beta)!}$, $\alpha! = \prod_{j=1}^m (\alpha_j!)$, and $\beta \le \alpha$ means $\beta_j \le \alpha_j$ for $1 \le j \le m$. In particular, $q_j(h \cdot f) \le C_j q_j(h) q_j(f)$ which shows that A is continuous and hence the adjoint is well defined via

$$(A'T)(f) = T(Af). \tag{14.97}$$

Now what is the effect on functions? For a distribution T_g given by an integrable function as above we clearly have

$$\begin{aligned}(A'T_g)(f) = T_g(hf) &= \int_{\mathbb{R}^m} g(x)(h(x)f(x))d^m x \\ &= \int_{\mathbb{R}^m} (g(x)h(x))f(x)d^m x = T_{gh}(f).\end{aligned} \tag{14.98}$$

So the effect of A' on functions is multiplication by h and hence A' generalizes this operation to arbitrary distributions. We will write $A'T = h \cdot T$ for notational simplicity. Note that since f can even compensate a polynomial growth, h could even be a smooth functions all whose derivatives grow at most polynomially (e.g. a polynomial):

$$C^\infty_{pg}(\mathbb{R}^m) := \{h \in C^\infty(\mathbb{R}^m) | \forall \alpha \in \mathbb{N}_0^m \; \exists C, n : \; |\partial_\alpha h(x)| \le C(1+|x|)^n\}. \tag{14.99}$$

In summary we can define

$$(h \cdot T)(f) := T(h \cdot f), \qquad h \in C^\infty_{pg}(\mathbb{R}^m). \tag{14.100}$$

Example 14.8. Let h be as above and $\delta_{x_0}(f) = f(x_0)$. Then

$$h \cdot \delta_{x_0}(f) = \delta_{x_0}(h \cdot f) = h(x_0) f(x_0)$$

and hence $h \cdot \delta_{x_0} = h(x_0)\delta_{x_0}$. ◇

Moreover, since Schwartz functions have derivatives of all orders, the same is true for tempered distributions! To this end let α be a multi-index and consider $D_\alpha : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$, $f \mapsto (-1)^{|\alpha|}\partial_\alpha f$ (the reason for the extra $(-1)^{|\alpha|}$ will become clear in a moment) which is continuous since $q_n(D_\alpha f) \le q_{n+|\alpha|}(f)$. Again we let D'_α be the corresponding adjoint operator and compute its effect on distributions given by functions g:

$$\begin{aligned}(D'_\alpha T_g)(f) &= T_g((-1)^{|\alpha|}\partial_\alpha f) = (-1)^{|\alpha|}\int_{\mathbb{R}^m} g(x)(\partial_\alpha f(x))d^m x \\ &= \int_{\mathbb{R}^m} (\partial_\alpha g(x)) f(x) d^m x = T_{\partial_\alpha g}(f),\end{aligned} \tag{14.101}$$

where we have used integration by parts in the last step which is (e.g.) permissible for $g \in C^{|\alpha|}_{pg}(\mathbb{R}^m)$ with $C^k_{pg}(\mathbb{R}^m) = \{h \in C^k(\mathbb{R}^m) | \forall |\alpha| \le k\, \exists C, n : |\partial_\alpha h(x)| \le C(1+|x|)^n\}$.

Hence for every multi-index α we define

$$(\partial_\alpha T)(f) := (-1)^{|\alpha|} T(\partial_\alpha f). \tag{14.102}$$

Example 14.9. Let α be a multi-index and $\delta_{x_0}(f) = f(x_0)$. Then

$$\partial_\alpha \delta_{x_0}(f) = (-1)^{|\alpha|}\delta_{x_0}(\partial_\alpha f) = (-1)^{|\alpha|}(\partial_\alpha f)(x_0).$$ ◇

Finally we use the same approach for the Fourier transform $\mathcal{F} : \mathcal{S}(\mathbb{R}^m) \to \mathcal{S}(\mathbb{R}^m)$, which is also continuous since $q_n(\hat{f}) \le C_n q_n(f)$ by Lemma 14.4. Since Fubini implies

$$\int_{\mathbb{R}^m} g(x)\hat{f}(x)d^m x = \int_{\mathbb{R}^m} \hat{g}(x) f(x) d^m x \tag{14.103}$$

for $g \in L^1(\mathbb{R}^m)$ (or $g \in L^2(\mathbb{R}^m)$) and $f \in \mathcal{S}(\mathbb{R}^m)$ we define the Fourier transform of a distribution to be

$$(\mathcal{F}T)(f) \equiv \hat{T}(f) := T(\hat{f}) \tag{14.104}$$

such that $\mathcal{F}T_g = T_{\hat{g}}$ for $g \in L^1(\mathbb{R}^m)$ (or $g \in L^2(\mathbb{R}^m)$).

Example 14.10. Let us compute the Fourier transform of $\delta_{x_0}(f) = f(x_0)$:

$$\hat{\delta}_{x_0}(f) = \delta_{x_0}(\hat{f}) = \hat{f}(x_0) = \frac{1}{(2\pi)^{m/2}}\int_{\mathbb{R}^m} \mathrm{e}^{-\mathrm{i}x_0 x} f(x) d^m x = T_g(f),$$

where $g(x) = (2\pi)^{-m/2}\mathrm{e}^{-\mathrm{i}x_0 x}$. ◇

Example 14.11. A slightly more involved example is the Fourier transform of $p.v.\frac{1}{x}$:

$$\begin{aligned}((p.v.\frac{1}{x}))^\wedge(f) &= \lim_{\varepsilon\downarrow 0}\int_{\varepsilon<|x|}\frac{\hat f(x)}{x}dx = \lim_{\varepsilon\downarrow 0}\int_{\varepsilon<|x|<1/\varepsilon}\int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}yx}\frac{f(y)}{x}dy\,dx\\ &= \lim_{\varepsilon\downarrow 0}\int_{\mathbb{R}}\int_{\varepsilon<|x|<1/\varepsilon}\frac{\mathrm{e}^{-\mathrm{i}yx}}{x}dx\,f(y)dy\\ &= -2\mathrm{i}\lim_{\varepsilon\downarrow 0}\int_{\mathbb{R}}\operatorname{sign}(y)\left(\int_\varepsilon^{1/\varepsilon}\frac{\sin(t)}{t}dt\right)f(y)dy\\ &= -2\mathrm{i}\lim_{\varepsilon\downarrow 0}\int_{\mathbb{R}}\operatorname{sign}(y)(\operatorname{Si}(1/\varepsilon)-\operatorname{Si}(\varepsilon))f(y)dy,\end{aligned}$$

where we have used the sine integral (14.89). Moreover, Problem 14.24 shows that we can use dominated convergence to get

$$((p.v.\frac{1}{x}))^\wedge(f) = -\mathrm{i}\pi\int_{\mathbb{R}}\operatorname{sign}(y)f(y)dy,$$

that is, $((p.v.\frac{1}{x}))^\wedge = -\mathrm{i}\pi\operatorname{sign}(y)$. ⋄

Note that since $\mathcal{F}:\mathcal{S}(\mathbb{R}^m)\to\mathcal{S}(\mathbb{R}^m)$ is a homeomorphism, so is its adjoint $\mathcal{F}':\mathcal{S}^*(\mathbb{R}^m)\to\mathcal{S}^*(\mathbb{R}^m)$. In particular, its inverse is given by

$$\check T(f) := T(\check f). \tag{14.105}$$

Moreover, all the operations for $\mathcal{F}$ carry over to $\mathcal{F}'$. For example, from Lemma 14.4 we immediately obtain

$$(\partial_\alpha T)^\wedge = (\mathrm{i}p)^\alpha\hat T, \qquad (x^\alpha T)^\wedge = \mathrm{i}^{|\alpha|}\partial_\alpha\hat T. \tag{14.106}$$

Similarly one can extend Lemma 14.2 to distributions.

Next we turn to convolutions. Since (Fubini)

$$\int_{\mathbb{R}^m}(h*g)(x)f(x)d^mx = \int_{\mathbb{R}^m} g(x)(\tilde h*f)(x)d^mx, \qquad \tilde h(x)=h(-x), \tag{14.107}$$

for integrable functions f,g,h we define

$$(h*T)(f) := T(\tilde h*f), \qquad h\in\mathcal{S}(\mathbb{R}^m), \tag{14.108}$$

which is well defined by Corollary 14.13. Moreover, Corollary 14.13 immediately implies

$$(h*T)^\wedge = (2\pi)^{n/2}\hat h\hat T, \qquad (hT)^\wedge = (2\pi)^{-n/2}\hat h*\hat T, \qquad h\in\mathcal{S}(\mathbb{R}^m). \tag{14.109}$$

Example 14.12. Note that the Dirac delta distribution acts like an identity for convolutions since

$$(h*\delta_0)(f) = \delta_0(\tilde h*f) = (\tilde h*f)(0) = \int_{\mathbb{R}^m} h(y)f(y)d^my = T_h(f). \qquad \diamond$$

In the last example the convolution is associated with a function. This turns out always to be the case.

Theorem 14.30. *For every $T \in \mathcal{S}^*(\mathbb{R}^m)$ and $h \in \mathcal{S}(\mathbb{R}^m)$ we have that $h * T$ is associated with the function*

$$h * T = T_g, \qquad g(x) := T(h(x - .)) \in C^\infty_{pg}(\mathbb{R}^m). \tag{14.110}$$

Proof. By definition $(h * T)(f) = T(\tilde{h} * f)$ and since $(\tilde{h} * f)(x) = \int h(y - x)f(y)d^m y$ the distribution T acts on $h(y - .)$ and we should be able to pull out the integral by linearity. To make this idea work let us replace the integral by a Riemann sum

$$(\tilde{h} * f)(x) = \lim_{n\to\infty} \sum_{j=1}^{n^{2m}} h(y_j^n - x)f(y_j^n)|Q_j^n|,$$

where Q_j^n is a partition of $[-\frac{n}{2}, \frac{n}{2}]^m$ into n^{2m} cubes of side length $\frac{1}{n}$ and y_j^n is the midpoint of Q_j^n. Then, if this Riemann sum converges to $h * f$ in $\mathcal{S}(\mathbb{R}^m)$, we have

$$(h * T)(f) = \lim_{n\to\infty} \sum_{j=1}^{n^{2m}} g(y_j^n)f(y_j^n)|Q_j^n|$$

and of course we expect this last limit to converge to the corresponding integral. To be able to see this we need some properties of g. Since

$$|h(z - x) - h(z - y)| \le q_1(h)|x - y|$$

by the mean value theorem and similarly

$$q_n(h(z - x) - h(z - y)) \le C_n q_{n+1}(h)|x - y|$$

we see that $x \mapsto h(. - x)$ is continuous in $\mathcal{S}(\mathbb{R}^m)$. Consequently g is continuous. Similarly, if $x = x_0 + \varepsilon e_j$ with e_j the unit vector in the j'th coordinate direction,

$$q_n\left(\frac{1}{\varepsilon}(h(. - x) - h(. - x_0)) - \partial_j h(. - x_0)\right) \le C_n q_{n+2}(h)\varepsilon$$

which shows $\partial_j g(x) = T((\partial_j h)(x - .))$. Applying this formula iteratively gives

$$\partial_\alpha g(x) = T((\partial_\alpha h)(x - .)) \tag{14.111}$$

and hence $g \in C^\infty(\mathbb{R}^m)$. Furthermore, g has at most polynomial growth since $|T(f)| \le Cq_n(f)$ implies

$$|g(x)| = |T(h(. - x))| \le Cq_n(h(. - x)) \le \tilde{C}(1 + |x|^n)q(h).$$

Combining this estimate with (14.111) even gives $g \in C^\infty_{pg}(\mathbb{R}^m)$.

In particular, since $g \cdot f \in \mathcal{S}(\mathbb{R}^m)$ the corresponding Riemann sum converges and we have $h * T = T_g$.

It remains to show that our first Riemann sum for the convolution converges in $\mathcal{S}(\mathbb{R}^m)$. It suffices to show

$$\sup_x |x|^N \left| \sum_{j=1}^{n^{2m}} h(y_j^n - x) f(y_j^n) |Q_j^n| - \int_{\mathbb{R}^m} h(y-x) f(y) d^m y \right| \to 0$$

since derivatives are automatically covered by replacing h with the corresponding derivative. The above expressions splits into two terms. The first one is

$$\sup_x |x|^N \left| \int_{|y|>n/2} h(y-x) f(y) d^m y \right| \le C q_N(h) \int_{|y|>n/2} (1+|y|^N) |f(y)| d^m y \to 0.$$

The second one is

$$\sup_x |x|^N \left| \sum_{j=1}^{n^{2m}} \int_{Q_j^n} \big(h(y_j^n - x) f(y_j^n) - h(y-x) f(y) \big) d^m y \right|$$

and the integrand can be estimated by

$$\begin{aligned} &|x|^N \big| h(y_j^n - x) f(y_j^n) - h(y-x) f(y) \big| \\ &\quad \le |x|^N \big| h(y_j^n - x) - h(y-x) \big| |f(y_j^n)| + |x|^N |h(y-x)| \big| f(y_j^n) - f(y) \big| \\ &\quad \le \big(q_{N+1}(h)(1+|y_j^n|^N) |f(y_j^n)| + q_N(h)(1+|y|^N) \big| \partial f(\tilde{y}_j^n) \big| \big) |y - y_j^n| \end{aligned}$$

and the claim follows since $|f(y)| + |\partial f(y)| \le C(1+|y|)^{-N-m-1}$. □

Example 14.13. If we take $T = \frac{1}{\pi}\big(p.v.\frac{1}{x}\big)$, then

$$(T * h)(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_{|y|>\varepsilon} \frac{h(x-y)}{y} dy = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_{|x-y|>\varepsilon} \frac{h(y)}{x-y} dy$$

which is known as the **Hilbert transform** of h. Moreover,

$$(T * h)^\wedge(p) = \sqrt{2\pi} \mathrm{i} \operatorname{sign}(p) \hat{h}(p)$$

as distributions and hence the Hilbert transform extends to a bounded operator on $L^2(\mathbb{R})$. ◇

As a consequence we get that distributions can be approximated by functions.

Theorem 14.31. *Let ϕ_ε be the standard mollifier. Then $\phi_\varepsilon * T \to T$ in $\mathcal{S}^*(\mathbb{R}^m)$.*

Proof. We need to show $\phi_\varepsilon * T(f) = T(\phi_\varepsilon * f)$ for any $f \in \mathcal{S}(\mathbb{R}^m)$. This follows from continuity since $\phi_\varepsilon * f \to f$ in $\mathcal{S}(\mathbb{R}^m)$ as can be easily seen (the derivatives follow from Lemma 10.18 (ii)). □

Note that Lemma 10.18 (i) and (ii) implies

$$\partial_\alpha(h * T) = (\partial_\alpha h) * T = h * (\partial_\alpha T). \tag{14.112}$$

When working with distributions it is also important to observe that, in contradistinction to smooth functions, they can be supported at a single point. Here the **support** $\operatorname{supp}(T)$ of a distribution is the smallest closed set for which

$$\operatorname{supp}(f) \subseteq \mathbb{R}^m \setminus V \Longrightarrow T(f) = 0.$$

An example of a distribution supported at 0 is the Dirac delta distribution δ_0 as well as all of its derivatives. It turns out that these are in fact the only examples.

Lemma 14.32. *Suppose T is a distribution supported at x_0. Then*

$$T = \sum_{|\alpha| \leq n} c_\alpha \partial_\alpha \delta_{x_0}. \tag{14.113}$$

Proof. For simplicity of notation we suppose $x_0 = 0$. First of all there is some n such that $|T(f)| \leq C q_n(f)$. Write

$$T = \sum_{|\alpha| \leq n} c_\alpha \partial_\alpha \delta_0 + \tilde{T},$$

where $c_\alpha = \frac{T(x^\alpha)}{\alpha!}$. Then $\tilde{T}$ vanishes on every polynomial of degree at most n, has support at 0, and still satisfies $|\tilde{T}(f)| \leq \tilde{C} q_n(f)$. Now let $\phi_\varepsilon(x) = \phi(\frac{x}{\varepsilon})$, where ϕ has support in $B_1(0)$ and equals 1 in a neighborhood of 0. Then $\tilde{T}(f) = \tilde{T}(g) = \tilde{T}(\phi_\varepsilon g)$, where $g(x) = f(x) - \sum_{|\alpha| \leq n} \frac{f^{(\alpha)}(0)}{\alpha!} x^\alpha$. Since $|\partial_\beta g(x)| \leq C_\beta \varepsilon^{n+1-|\beta|}$ for $x \in B_\varepsilon(0)$ Leibniz' rule implies $q_n(\phi_\varepsilon g) \leq C\varepsilon$. Hence $|\tilde{T}(f)| = |\tilde{T}(\phi_\varepsilon g)| \leq \tilde{C} q_n(\phi_\varepsilon g) \leq \hat{C}\varepsilon$ and since $\varepsilon > 0$ is arbitrary we have $|\tilde{T}(f)| = 0$, that is, $\tilde{T} = 0$. □

Example 14.14. Let us try to solve the Poisson equation in the sense of distributions. We begin with solving

$$-\Delta T = \delta_0.$$

Taking the Fourier transform we obtain

$$|p|^2 \hat{T} = (2\pi)^{-m/2}$$

and since $|p|^{-1}$ is a bounded locally integrable function in $\mathbb{R}^m$ for $m \geq 2$ the above equation will be solved by

$$\Phi := (2\pi)^{-m/2} (\frac{1}{|p|^2})^\vee.$$

Explicitly, Φ must be determined from

$$\Phi(f) = (2\pi)^{-m/2} \int_{\mathbb{R}^m} \frac{\check{f}(p)}{|p|^2} d^m p = (2\pi)^{-m/2} \int_{\mathbb{R}^m} \frac{\hat{f}(p)}{|p|^2} d^m p$$

and evaluating Lemma 14.17 at $x = 0$ we obtain for $m \geq 3$ that $\Phi = (2\pi)^{-m/2} T_{I_2}$ where I_2 is the Riesz potential. (Alternatively one can also compute the weak derivative of the Riesz potential directly, see Problems 13.3 and 13.4.) Note, that Φ is not unique since we could add a polynomial of degree two corresponding to a solution of the homogenous equation ($|p|^2 \hat{T} = 0$ implies that $\operatorname{supp}(\hat{T}) = 0$ and comparing with Lemma 14.32 shows $\hat{T} = \sum_{|\alpha| \leq 2} c_\alpha \partial_\alpha \delta_0$ and hence $T = \sum_{|\alpha| \leq 2} \tilde{c}_\alpha x^\alpha$).

Given $h \in \mathcal{S}(\mathbb{R}^m)$ we can now consider $h * \Phi$ which solves

$$-\Delta(h * \Phi) = h * (-\Delta \Phi) = h * \delta_0 = h,$$

where in the last equality we have identified h with T_h. Note that since $h * \Phi$ is associated with a function in $C^\infty_{pg}(\mathbb{R}^m)$ our distributional solution is also a classical solution. This gives the formal calculations with the Dirac delta function found in many physics textbooks a solid mathematical meaning. $\diamond$

Note that while we have been quite successful in generalizing many basic operations to distributions, our approach is limited to linear operations! In particular, it is not possible to define nonlinear operations, for example the product of two distributions within this framework. In fact, there is no associative product of two distributions extending the product of a distribution by a function from above.

Example 14.15. Consider the distributions δ_0, x, and $p.v.\frac{1}{x}$ in $\mathcal{S}^*(\mathbb{R})$. Then

$$x \cdot \delta_0 = 0, \qquad x \cdot p.v.\frac{1}{x} = 1.$$

Hence if there would be an associative product of distributions we would get $0 = (x \cdot \delta_0) \cdot p.v.\frac{1}{x} = \delta_0 \cdot (x \cdot p.v.\frac{1}{x}) = \delta_0$. $\diamond$

This is known as Schwartz' impossibility result. However, if one is content with preserving the product of functions, Colombeau algebras will do the trick.

Problem 14.25. *Compute the derivative of* $g(x) = \operatorname{sign}(x)$ *in* $\mathcal{S}^*(\mathbb{R})$.

Problem 14.26. *Let* $h \in C^\infty_{pg}(\mathbb{R}^m)$ *and* $T \in \mathcal{S}^*(\mathbb{R}^m)$. *Show*

$$\partial_\alpha(h \cdot T) = \sum_{\beta \leq \alpha} \binom{\alpha}{\beta} (\partial_\beta h)(\partial_{\alpha-\beta} T).$$

Problem 14.27. *Show that* $\operatorname{supp}(T_g) = \operatorname{supp}(g)$ *for locally integrable functions.*

Chapter 15

Interpolation

15.1. Interpolation and the Fourier transform on L^p

We will fix some measure space (X, μ) and abbreviate $L^p = L^p(X, d\mu)$ for notational simplicity. If $f \in L^{p_0} \cap L^{p_1}$ for some $p_0 < p_1$ then it is not hard to see that $f \in L^p$ for every $p \in [p_0, p_1]$ (Problem 10.11). Note that $L^{p_0} \cap L^{p_1}$ contains all integrable simple functions. Moreover, the latter functions are dense in L^p for $1 \leq p < \infty$ (for $p = \infty$ this is only true if the measure is finite — cf. Problem 10.18).

This is a first occurrence of an interpolation technique. Next we want to turn to operators. For example, we have defined the Fourier transform as an operator from $L^1 \to L^\infty$ as well as from $L^2 \to L^2$ and the question is if this can be used to extend the Fourier transform to the spaces in between.

Denote by $L^{p_0} + L^{p_1}$ the space of (equivalence classes) of measurable functions f which can be written as a sum $f = f_0 + f_1$ with $f_0 \in L^{p_0}$ and $f_1 \in L^{p_1}$ (clearly such a decomposition is not unique and different decompositions will differ by elements from $L^{p_0} \cap L^{p_1}$). Then we have

$$L^p \subseteq L^{p_0} + L^{p_1}, \qquad p_0 < p < p_1, \tag{15.1}$$

since we can always decompose a function $f \in L^p$, $1 \leq p < \infty$, as $f = f\chi_{\{x||f(x)|\leq 1\}} + f\chi_{\{x||f(x)|>1\}}$ with $f\chi_{\{x||f(x)|\leq 1\}} \in L^p \cap L^\infty$ and $f\chi_{\{x||f(x)|>1\}} \in L^1 \cap L^p$. Hence, if we have two operators $A_0 : L^{p_0} \to L^{q_0}$ and $A_1 : L^{p_1} \to L^{q_1}$ which coincide on the intersection, $A_0|_{L^{p_0} \cap L^{p_1}} = A_1|_{L^{p_0} \cap L^{p_1}}$, we can extend them by virtue of

$$A : L^{p_0} + L^{p_1} \to L^{q_0} + L^{q_1}, \qquad f_0 + f_1 \mapsto A_0 f_0 + A_1 f_1 \tag{15.2}$$

(check that A is indeed well-defined, i.e, independent of the decomposition of f into $f_0 + f_1$). In particular, this defines A on L^p for every $p \in (p_0, p_1)$

and the question is if A restricted to L^p will be a bounded operator into some L^q provided A_0 and A_1 are bounded.

To answer this question we begin with a result from complex analysis.

Theorem 15.1 (Hadamard three-lines theorem)**.** *Let S be the open strip $\{z \in \mathbb{C} | 0 < \operatorname{Re}(z) < 1\}$ and let $F : \overline{S} \to \mathbb{C}$ be continuous and bounded on $\overline{S}$ and holomorphic in S. If*

$$|F(z)| \leq \begin{cases} M_0, & \operatorname{Re}(z) = 0, \\ M_1, & \operatorname{Re}(z) = 1, \end{cases} \tag{15.3}$$

then

$$|F(z)| \leq M_0^{1-\operatorname{Re}(z)} M_1^{\operatorname{Re}(z)} \tag{15.4}$$

for every $z \in \overline{S}$.

Proof. Without loss of generality we can assume $M_0, M_1 > 0$ and after the transformation $F(z) \to M_0^{z-1} M_1^{-z} F(z)$ even $M_0 = M_1 = 1$. Now we consider the auxiliary function

$$F_n(z) = \mathrm{e}^{(z^2-1)/n} F(z)$$

which still satisfies $|F_n(z)| \leq 1$ for $\operatorname{Re}(z) = 0$ and $\operatorname{Re}(z) = 1$ since $\operatorname{Re}(z^2 - 1) \leq -\operatorname{Im}(z)^2 \leq 0$ for $z \in \overline{S}$. Moreover, by assumption $|F(z)| \leq M$ implying $|F_n(z)| \leq 1$ for $|\operatorname{Im}(z)| \geq \sqrt{\log(M)n}$. Since we also have $|F_n(z)| \leq 1$ for $|\operatorname{Im}(z)| \leq \sqrt{\log(M)n}$ by the maximum modulus principle we see $|F_n(z)| \leq 1$ for all $z \in \overline{S}$. Finally, letting $n \to \infty$ the claim follows. □

Now we are abel to show the **Riesz–Thorin interpolation theorem**

Theorem 15.2 (Riesz–Thorin)**.** *Let $(X, d\mu)$ and $(Y, d\nu)$ be σ-finite measure spaces and $1 \leq p_0, p_1, q_0, q_1 \leq \infty$. If A is a linear operator on*

$$A : L^{p_0}(X, d\mu) + L^{p_1}(X, d\mu) \to L^{q_0}(Y, d\nu) + L^{q_1}(Y, d\nu) \tag{15.5}$$

satisfying

$$\|Af\|_{q_0} \leq M_0 \|f\|_{p_0}, \qquad \|Af\|_{q_1} \leq M_1 \|f\|_{p_1}, \tag{15.6}$$

then A has continuous restrictions

$$A_\theta : L^{p_\theta}(X, d\mu) \to L^{q_\theta}(Y, d\nu), \quad \frac{1}{p_\theta} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}, \; \frac{1}{q_\theta} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1} \tag{15.7}$$

satisfying $\|A_\theta\| \leq M_0^{1-\theta} M_1^\theta$ for every $\theta \in (0, 1)$.

Proof. In the case $p_0 = p_1 = \infty$ the claim is immediate from Problem 10.11 and hence we can assume $p_\theta < \infty$ in which case the space of integrable simple

functions is dense in L^{p_θ}. We will also temporarily assume $q_\theta < \infty$. Then, by Lemma 10.6 it suffices to show

$$\left| \int (Af)(y) g(y) d\nu(y) \right| \leq M_0^{1-\theta} M_1^\theta,$$

where f, g are simple functions with $\|f\|_{p_\theta} = \|g\|_{q'_\theta} = 1$ and $\frac{1}{q_\theta} + \frac{1}{q'_\theta} = 1$.

Now choose simple functions $f(x) = \sum_j \alpha_j \chi_{A_j}(x)$, $g(x) = \sum_k \beta_k \chi_{B_k}(x)$ with $\|f\|_1 = \|g\|_1 = 1$ and set $f_z(x) = \sum_j |\alpha_j|^{1/p_z} \operatorname{sign}(\alpha_j) \chi_{A_j}(x)$, $g_z(y) = \sum_k |\beta_k|^{1-1/q_z} \operatorname{sign}(\beta_k) \chi_{B_k}(y)$ such that $\|f_z\|_{p_\theta} = \|g_z\|_{q'_\theta} = 1$ for $\theta = \operatorname{Re}(z) \in [0,1]$. Moreover, note that both functions are entire and thus the function

$$F(z) = \int (Af_z)(y) g_z d\nu(y)$$

satisfies the assumptions of the three-lines theorem. Hence we have the required estimate for integrable simple functions. Now let $f \in L^{p_\theta}$ and split it according to $f = f_0 + f_1$ with $f_0 \in L^{p_0} \cap L^{p_\theta}$ and $f_1 \in L^{p_1} \cap L^{p_\theta}$ and approximate both by integrable simple functions (cf. Problem 10.18).

It remains to consider the case $p_0 < p_1$ and $q_0 = q_1 = \infty$. In this case we can proceed as before using again Lemma 10.6 and a simple function for $g = g_z$. □

Note that the proof shows even a bit more

Corollary 15.3. *Let A be an operator defined on the space of integrable simple functions satisfying* (15.6)*. Then A has continuous extensions A_θ as in the Riesz–Thorin theorem which will agree on $L^{p_0}(X, d\mu) \cap L^{p_1}(X, d\mu)$.*

As a consequence we get two important inequalities:

Corollary 15.4 (Hausdorff–Young inequality)**.** *The Fourier transform extends to a continuous map $\mathcal{F} : L^p(\mathbb{R}^n) \to L^q(\mathbb{R}^n)$, for $1 \leq p \leq 2$, $\frac{1}{p} + \frac{1}{q} = 1$, satisfying*

$$(2\pi)^{-n/(2q)} \|\hat{f}\|_q \leq (2\pi)^{-n/(2p)} \|f\|_p. \tag{15.8}$$

We remark that the Fourier transform does not extend to a continuous map $\mathcal{F} : L^p(\mathbb{R}^n) \to L^q(\mathbb{R}^n)$, for $p > 2$ (Problem 15.1). Moreover, its range is dense for $1 < p \leq 2$ but not all of $L^q(\mathbb{R}^n)$ unless $p = q = 2$.

Corollary 15.5 (Young inequality)**.** *Let $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$ with $\frac{1}{p} + \frac{1}{q} \geq 1$. Then $f(y)g(x-y)$ is integrable with respect to y for a.e. x and the convolution satisfies $f * g \in L^r(\mathbb{R}^n)$ with*

$$\|f * g\|_r \leq \|f\|_p \|g\|_q, \tag{15.9}$$

where $\frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1$.

Proof. We consider the operator $A_g f = f * g$ which satisfies $\|A_g f\|_q \leq \|g\|_q \|f\|_1$ for every $f \in L^1$ by Lemma 10.18. Similarly, Hölder's inequality implies $\|A_g f\|_\infty \leq \|g\|_q \|f\|_{q'}$ for every $f \in L^{q'}$, where $\frac{1}{q} + \frac{1}{q'} = 1$. Hence the Riesz–Thorin theorem implies that A_g extends to an operator $A_q : L^p \to L^r$, where $\frac{1}{p} = \frac{1-\theta}{1} + \frac{\theta}{q'} = 1 - \frac{\theta}{q}$ and $\frac{1}{r} = \frac{1-\theta}{q} + \frac{\theta}{\infty} = \frac{1}{p} + \frac{1}{q} - 1$. To see that $f(y)g(x-y)$ is integrable a.e. consider $f_n(x) = \chi_{|x|\leq n}(x) \max(n, |f(x)|)$. Then the convolution $(f_n * |g|)(x)$ is finite and converges for every x by monotone convergence. Moreover, since $f_n \to |f|$ in L^p we have $f_n * |g| \to A_g f$ in L^r, which finishes the proof. □

Combining the last two corollaries we obtain:

Corollary 15.6. *Let $f \in L^p(\mathbb{R}^n)$ and $g \in L^q(\mathbb{R}^n)$ with $\frac{1}{r} = \frac{1}{p} + \frac{1}{q} - 1 \geq 0$ and $1 \leq r, p, q \leq 2$. Then*

$$(f * g)^\wedge = (2\pi)^{n/2} \hat{f} \hat{g}.$$

Proof. By Corollary 14.13 the claim holds for $f, g \in \mathcal{S}(\mathbb{R}^n)$. Now take a sequence of Schwartz functions $f_m \to f$ in L^p and a sequence of Schwartz functions $g_m \to g$ in L^q. Then the left-hand side converges in $L^{r'}$, where $\frac{1}{r'} = 2 - \frac{1}{p} - \frac{1}{q}$, by the Young and Hausdorff-Young inequalities. Similarly, the right-hand side converges in $L^{r'}$ by the generalized Hölder (Problem 10.8) and Hausdorff-Young inequalities. □

Problem* 15.1. *Show that the Fourier transform does not extend to a continuous map $\mathcal{F} : L^p(\mathbb{R}^n) \to L^q(\mathbb{R}^n)$, for $p > 2$. Use the closed graph theorem to conclude that $\mathcal{F}$ is not onto for $1 \leq p < 2$. (Hint for the case $n = 1$: Consider $\phi_z(x) = \exp(-zx^2/2)$ for $z = \lambda + \mathrm{i}\omega$ with $\lambda > 0$.)*

Problem 15.2 (Young inequality)**.** *Let $K(x, y)$ be measurable and suppose*

$$\sup_x \|K(x, .)\|_{L^r(Y, d\nu)} \leq C, \qquad \sup_y \|K(., y)\|_{L^r(X, d\mu)} \leq C.$$

where $\frac{1}{r} = \frac{1}{q} - \frac{1}{p} \geq 1$ for some $1 \leq p \leq q \leq \infty$. Then the operator $K : L^p(Y, d\nu) \to L^q(X, d\mu)$, defined by

$$(Kf)(x) = \int_Y K(x, y) f(y) d\nu(y),$$

for μ-almost every x, is bounded with $\|K\| \leq C$. (Hint: Show $\|Kf\|_\infty \leq C\|f\|_{r'}$, $\|Kf\|_r \leq C\|f\|_1$ and use interpolation.)

15.2. The Marcinkiewicz interpolation theorem

In this section we are going to look at another interpolation theorem which might be helpful in situations where the Riesz–Thorin interpolation theorem does not apply. In this respect recall, that $f(x) = \frac{1}{x}$ just fails to be integrable

over $\mathbb{R}$. To include such functions we begin by slightly weakening the L^p norms. To this end we consider the distribution function

$$E_f(r) = \mu\big(\{x \in X | \, |f(x)| > r\}\big) \tag{15.10}$$

of a measurable function $f : X \to \mathbb{C}$ with respect to μ. Note that E_f is decreasing and right continuous. Given, the distribution function we can compute the L^p norm via (Problem 9.20)

$$\|f\|_p^p = p \int_0^\infty r^{p-1} E_f(r) dr, \qquad 1 \le p < \infty. \tag{15.11}$$

In the case $p = \infty$ we have

$$\|f\|_\infty = \inf\{r \ge 0 | E_f(r) = 0\}. \tag{15.12}$$

Another relationship follows from the observation

$$\|f\|_p^p = \int_X |f|^p d\mu \ge \int_{|f|>r} r^p d\mu = r^p E_f(r) \tag{15.13}$$

which yields **Markov's inequality**

$$E_f(r) \le r^{-p} \|f\|_p^p. \tag{15.14}$$

Motivated by this we define the **weak L_p norm**

$$\|f\|_{p,w} = \sup_{r>0} r E_f(r)^{1/p}, \qquad 1 \le p < \infty, \tag{15.15}$$

and the corresponding spaces $L^{p,w}(X, d\mu)$ consist of all equivalence classes of functions which are equal a.e. for which the above *norm* is finite. Clearly the distribution function and hence the weak L^p norm depend only on the equivalence class. Despite its name the weak L_p norm turns out to be only a quasinorm (Problem 15.3). By construction we have

$$\|f\|_{p,w} \le \|f\|_p \tag{15.16}$$

and thus $L^p(X, d\mu) \subseteq L^{p,w}(X, d\mu)$. In the case $p = \infty$ we set $\|.\|_{\infty,w} = \|.\|_\infty$.

Example 15.1. Consider $f(x) = \frac{1}{x}$ in $\mathbb{R}$. Then clearly $f \notin L^1(\mathbb{R})$ but

$$E_f(r) = |\{x| \, |\frac{1}{x}| > r\}| = |\{x| \, |x| < r^{-1}\}| = \frac{2}{r}$$

shows that $f \in L^{1,w}(\mathbb{R})$ with $\|f\|_{1,w} = 2$. Slightly more general the function $f(x) = |x|^{-n/p} \notin L^p(\mathbb{R}^n)$ but $f \in L^{p,w}(\mathbb{R}^n)$. Hence $L^{p,w}(\mathbb{R}^n)$ is strictly larger than $L^p(\mathbb{R}^n)$. ⋄

Now we are ready for our interpolation result. We call an operator $T : L^p(X, d\mu) \to L^q(X, d\nu)$ **subadditive** if it satisfies

$$\|T(f+g)\|_q \le \|T(f)\|_q + \|T(g)\|_q. \tag{15.17}$$

It is said to be of **strong type** (p, q) if

$$\|T(f)\|_q \le C_{p,q} \|f\|_p \tag{15.18}$$

and of **weak type** (p,q) if

$$\|T(f)\|_{q,w} \le C_{p,q,w}\|f\|_p. \tag{15.19}$$

By (15.16) strong type (p,q) is indeed stronger than weak type (p,q) and we have $C_{p,q,w} \le C_{p,q}$.

Theorem 15.7 (Marcinkiewicz)**.** *Let $(X,d\mu)$ and $(Y,d\nu)$ measure spaces and $1 \le p_0 < p_1 \le \infty$. Let T be a subadditive operator defined for all $f \in L^p(X,d\mu)$, $p \in [p_0,p_1]$. If T is of weak type (p_0,p_0) and (p_1,p_1) then it is also of strong type (p,p) for every $p_0 < p < p_1$.*

Proof. We begin by assuming $p_1 < \infty$. Fix $f \in L^p$ as well as some number $s > 0$ and decompose $f = f_0 + f_1$ according to

$$f_0 = f\chi_{\{x\,|\,|f|>s\}} \in L^{p_0} \cap L^p, \qquad f_1 = f\chi_{\{x\,|\,|f|\le s\}} \in L^p \cap L^{p_1}.$$

Next we use (15.11),

$$\|T(f)\|_p^p = p\int_0^\infty r^{p-1}E_{T(f)}(r)dr = p2^p\int_0^\infty r^{p-1}E_{T(f)}(2r)dr$$

and observe

$$E_{T(f)}(2r) \le E_{T(f_0)}(r) + E_{T(f_1)}(r)$$

since $|T(f)| \le |T(f_0)| + |T(f_1)|$ implies $|T(f)| > 2r$ only if $|T(f_0)| > r$ or $|T(f_1)| > r$. Now using (15.14) our assumption implies

$$E_{T(f_0)}(r) \le \left(\frac{C_0\|f_0\|_{p_0}}{r}\right)^{p_0}, \qquad E_{T(f_1)}(r) \le \left(\frac{C_1\|f_1\|_{p_1}}{r}\right)^{p_1}$$

and choosing $s = r$ we obtain

$$E_{T(f)}(2r) \le \frac{C_0^{p_0}}{r^{p_0}}\int_{\{x\,|\,|f|>r\}} |f|^{p_0}d\mu + \frac{C_1^{p_1}}{r^{p_1}}\int_{\{x\,|\,|f|\le r\}} |f|^{p_1}d\mu.$$

In summary we have $\|T(f)\|_p^p \le p2^p(C_0^{p_0}I_1 + C_1^{p_1}I_2)$ with

$$\begin{aligned} I_0 &= \int_0^\infty\int_X r^{p-p_0-1}\chi_{\{(x,r)\,|\,|f(x)|>r\}}|f(x)|^{p_0}d\mu(x)\,dr \\ &= \int_X |f(x)|^{p_0}\int_0^{|f(x)|} r^{p-p_0-1}dr\,d\mu(x) = \frac{1}{p-p_0}\|f\|_p^p \end{aligned}$$

and

$$\begin{aligned} I_1 &= \int_0^\infty\int_X r^{p-p_1-1}\chi_{\{(x,r)\,|\,|f(x)|\le r\}}|f(x)|^{p_1}d\mu(x)\,dr \\ &= \int_X |f(x)|^{p_1}\int_{|f(x)|}^\infty r^{p-p_1-1}dr\,d\mu(x) = \frac{1}{p_1-p}\|f\|_p^p. \end{aligned}$$

This is the desired estimate

$$\|T(f)\|_p \le 2\big(p/(p-p_0)C_0^{p_0} + p/(p_1-p)C_1^{p_1}\big)^{1/p}\|f\|_p.$$

The case $p_1 = \infty$ is similar: Split $f \in L^{p_0}$ according to

$$f_0 = f\chi_{\{x\,|\,|f|>s/C_1\}} \in L^{p_0} \cap L^p, \qquad f_1 = f\chi_{\{x\,|\,|f|\le s/C_1\}} \in L^p \cap L^\infty$$

(if $C_1 = 0$ there is noting to prove). Then $\|T(f_1)\|_\infty \le s/C_1$ and hence $E_{T(f_1)}(s) = 0$. Thus

$$E_{T(f)}(2r) \le \frac{C_0^{p_0}}{r^{p_0}} \int_{\{x\,|\,|f|>r/C_1\}} |f|^{p_0} d\mu$$

and we can proceed as before to obtain

$$\|T(f)\|_p \le 2\big(p/(p-p_0)\big)^{1/p} C_0^{p_0/p} C_1^{1-p_0/p} \|f\|_p,$$

which is again the desired estimate. □

As with the Riesz–Thorin theorem there is also a version for operators which are of weak type (p_0, q_0) and (p_1, q_1) but the proof is slightly more involved and the above diagonal version will be sufficient for our purpose.

As a first application we will use it to investigate the **Hardy–Littlewood maximal function** defined for any locally integrable function in $\mathbb{R}^n$ via

$$\mathcal{M}(f)(x) = \sup_{r>0} \frac{1}{|B_r(x)|} \int_{B_r(x)} |f(y)| dy. \tag{15.20}$$

By the dominated convergence theorem, the integral is continuous with respect to x and consequently (Problem 8.19) $\mathcal{M}(f)$ is lower semicontinuous (and hence measurable). Moreover, its value is unchanged if we change f on sets of measure zero, so $\mathcal{M}$ is well defined for functions in $L^p(\mathbb{R}^n)$. However, it is unclear if $\mathcal{M}(f)(x)$ is finite a.e. at this point. If f is bounded we of course have the trivial estimate

$$\|\mathcal{M}(f)\|_\infty \le \|f\|_\infty. \tag{15.21}$$

Theorem 15.8 (Hardy–Littlewood maximal inequality)**.** *The maximal function is of weak type* $(1, 1)$*,*

$$E_{\mathcal{M}(f)}(r) \le \frac{3^n}{r} \|f\|_1, \tag{15.22}$$

and of strong type (p, p)*,*

$$\|\mathcal{M}(f)\|_p \le 2\left(\frac{3^n p}{p-1}\right)^{1/p} \|f\|_p, \tag{15.23}$$

for every $1 < p \le \infty$*.*

Proof. The first estimate follows literally as in the proof of Lemma 11.5 and combining this estimate with the trivial one (15.21) the Marcinkiewicz interpolation theorem yields the second. □

Using this fact, our next aim is to prove the Hardy–Littlewood–Sobolev inequality. As a preparation we show

Lemma 15.9. *Let $\phi \in L^1(\mathbb{R}^n)$ be a radial, $\phi(x) = \phi_0(|x|)$ with ϕ_0 positive and nonincreasing. Then we have the following estimate for convolutions with integrable functions:*

$$|(\phi * f)(x)| \le \|\phi\|_1 \mathcal{M}(f)(x). \tag{15.24}$$

Proof. By approximating φ_0 with simple functions of the same type, it suffices to prove that case where $\phi_0 = \sum_{j=1}^p \alpha_j \chi_{[0,r_j]}$ with $\alpha_j > 0$. Then

$$(\phi * f)(x) = \sum_j \alpha_j |B_{r_j}(0)| \frac{1}{|B_{r_j}(x)|} \int_{B_{r_j}(x)} f(y) d^n y$$

and the estimate follows upon taking absolute values and observing $\|\phi\|_1 = \sum_j \alpha_j |B_{r_j}(0)|$. □

Now we will apply this to the Riesz potential (14.29) of order α:

$$\mathcal{I}_\alpha f = I_\alpha * f. \tag{15.25}$$

Theorem 15.10 (Hardy–Littlewood–Sobolev inequality)**.** *Let $0 < \alpha < n$, $p \in (1, \frac{n}{\alpha})$, and $q = \frac{pn}{n-p\alpha} \in (\frac{n}{n-\alpha}, \infty)$ (i.e, $\frac{\alpha}{n} = \frac{1}{p} - \frac{1}{q}$). Then $\mathcal{I}_\alpha$ is of strong type (p, q),*

$$\|\mathcal{I}_\alpha f\|_q \le C_{p,\alpha,n} \|f\|_p. \tag{15.26}$$

Proof. We split the Riesz potential into two parts

$$I_\alpha = I_\alpha^0 + I_\alpha^\infty, \qquad I_\alpha^0 = I_\alpha \chi_{(0,\varepsilon)},\ I_\alpha^\infty = I_\alpha \chi_{[\varepsilon,\infty)},$$

where $\varepsilon > 0$ will be determined later. Note that $I_\alpha^0(|.|) \in L^1(\mathbb{R}^n)$ and $I_\alpha^\infty(|.|) \in L^r(\mathbb{R}^n)$ for every $r \in (\frac{n}{n-\alpha}, \infty)$. In particular, since $p' = \frac{p}{p-1} \in (\frac{n}{n-\alpha}, \infty)$, both integrals converge absolutely by the Young inequality (15.9). Next we will estimate both parts individually. Using Lemma 15.9 we obtain

$$|\mathcal{I}_\alpha^0 f(x)| \le \int_{|y|<\varepsilon} \frac{d^n y}{|y|^{n-\alpha}} \mathcal{M}(f)(x) = \frac{(n-1)V_n}{\alpha - 1} \varepsilon^n \mathcal{M}(f)(x).$$

On the other hand, using Hölder's inequality we infer

$$|\mathcal{I}_\alpha^\infty f(x)| \le \left(\int_{|y|\ge\varepsilon} \frac{d^n y}{|y|^{(n-\alpha)p'}} \right)^{1/p'} \|f\|_p = \left(\frac{(n-1)V_n}{p'(n-\alpha) - n} \right)^{1/p'} \varepsilon^{\alpha - n/p} \|f\|_p.$$

Now we choose $\varepsilon = \left(\frac{\|f\|_p}{\mathcal{M}(f)(x)} \right)^{p/n}$ such that

$$|\mathcal{I}_\alpha f(x)| \le \tilde{C} \|f\|_p^\theta \mathcal{M}(f)(x)^{1-\theta}, \qquad \theta = \frac{\alpha p}{n} \in (\frac{\alpha}{n}, 1),$$

where $\tilde{C}/2$ is the larger of the two constants in the estimates for $\mathcal{I}^0_\alpha f$ and $\mathcal{I}^\infty_\alpha f$. Taking the L^q norm in the above expression gives

$$\|\mathcal{I}_\alpha f\|_q \le \tilde{C}\|f\|_p^\theta \|\mathcal{M}(f)^{1-\theta}\|_q = \tilde{C}\|f\|_p^\theta \|\mathcal{M}(f)\|_{q(1-\theta)}^{1-\theta} = \tilde{C}\|f\|_p^\theta \|\mathcal{M}(f)\|_p^{1-\theta}$$

and the claim follows from the Hardy–Littlewood maximal inequality. □

Problem* 15.3. *Show that $E_f = 0$ if and only if $f = 0$. Moreover, show $E_{f+g}(r+s) \le E_f(r) + E_g(s)$ and $E_{\alpha f}(r) = E_f(r/|\alpha|)$ for $\alpha \neq 0$. Conclude that $L^{p,w}(X, d\mu)$ is a quasinormed space with*

$$\|f+g\|_{p,w} \le 2\big(\|f\|_{p,w} + \|g\|_{p,w}\big), \qquad \|\alpha f\|_{p,w} = |\alpha|\|f\|_{p,w}.$$

Problem 15.4. *Show $f(x) = |x|^{-n/p} \in L^{p,w}(\mathbb{R}^n)$. Compute $\|f\|_{p,w}$.*

Problem 15.5. *Show that if $f_n \to f$ in weak-L^p, then $f_n \to f$ in measure. (Hint: Problem 8.23.)*

Problem 15.6. *The right continuous generalized inverse of the distribution function is known as the **decreasing rearrangement** of f:*

$$f^*(t) = \inf\{r \ge 0 | E_f(r) \le t\}$$

(see Section 9.5 for basic properties of the generalized inverse). Note that f^* is decreasing with $f^*(0) = \|f\|_\infty$. Show that

$$E_{f^*} = E_f$$

and in particular $\|f^*\|_p = \|f\|_p$.

Problem 15.7. *Show that the maximal function of an integrable function is finite at every Lebesgue point.*

Problem* 15.8. *Let ϕ be a nonnegative nonincreasing radial function with $\|\phi\|_1 = 1$. Set $\phi_\varepsilon(x) = \varepsilon^{-n}\phi(\frac{x}{\varepsilon})$. Show that for integrable f we have $(\phi_\varepsilon * f)(x) \to f(x)$ at every Lebesgue point. (Hint: Split $\phi = \phi^\delta + \tilde{\phi}^\delta$ into a part with compact support ϕ^δ and a rest by setting $\tilde{\phi}^\delta(x) = \min(\delta, \phi(x))$. To handle the compact part use Problem 10.25. To control the contribution of the rest use Lemma 15.9.)*

Problem 15.9. *For $f \in L^1(0,1)$ define*

$$T(f)(x) = \mathrm{e}^{\mathrm{i}\arg\left(\int_0^1 f(y)dy\right)} f(x).$$

Show that T is subadditive and norm preserving. Show that T is not continuous.

Part 3

Nonlinear Functional Analysis

Chapter 16

Analysis in Banach spaces

16.1. Differentiation and integration in Banach spaces

We first review some basic facts from calculus in Banach spaces. Most facts will be similar to the situation of multivariable calculus for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. To emphasize this we will use $|.|$ for the norm in this section.

Let X and Y be two Banach spaces and let U be an open subset of X. Denote by $C(U,Y)$ the set of continuous functions from $U \subseteq X$ to Y and by $\mathscr{L}(X,Y) \subset C(X,Y)$ the Banach space of bounded linear functions equipped with the operator norm

$$\|L\| := \sup_{|u|=1} |Lu|. \tag{16.1}$$

Then a function $F : U \to Y$ is called differentiable at $x \in U$ if there exists a linear function $dF(x) \in \mathscr{L}(X,Y)$ such that

$$F(x+u) = F(x) + dF(x)\, u + o(u), \tag{16.2}$$

where o, O are the **Landau symbols**. Explicitly

$$\lim_{u\to 0} \frac{|F(x+u) - F(x) - dF(x)\, u|}{|u|} = 0. \tag{16.3}$$

The linear map $dF(x)$ is called the **Fréchet derivative** of F at x. It is uniquely defined since if $dG(x)$ were another derivative we had $(dF(x) - dG(x))u = o(u)$ implying that for every $\varepsilon > 0$ we can find a $\delta > 0$ such that $|(dF(x) - dG(x))u| \le \varepsilon |u|$ whenever $|u| \le \delta$. By homogeneity of the norm we conclude $\|dF(x) - dG(x)\| \le \varepsilon$ and since $\varepsilon > 0$ is arbitrary $dF(x) = dG(x)$.

Note that for this argument to work it is crucial that we can approach x from arbitrary directions u, which explains our requirement that U should be open.

Example 16.1. Let X be a Hilbert space and consider $F : X \to \mathbb{R}$ given by $F(x) := |x|^2$. Then

$$F(x+u) = \langle x+u, x+u\rangle = |x|^2 + 2\mathrm{Re}\langle x, u\rangle + |u|^2 = F(x) + 2\mathrm{Re}\langle x, u\rangle + o(u).$$

Hence if X is a real Hilbert space, then F is differentiable with $dF(x)u = 2\langle x, u\rangle$. However, if X is a complex Hilbert space, then F is not differentiable. In fact, in case of a complex Hilbert (or Banach) space, we obtain a version of complex differentiability which of course is much stronger than real differentiability. ◇

Example 16.2. Suppose $f \in C^1(\mathbb{R})$ with $f(0) = 0$. Let $X := \ell^p(\mathbb{N})$, then

$$F : X \to X, \qquad (x_n)_{n\in\mathbb{N}} \mapsto (f(x_n))_{n\in\mathbb{N}}$$

is differentiable for every $x \in X$ with derivative given by the multiplication operator

$$(dF(x)u)_n = f'(x_n)u_n.$$

First of all note that the mean value theorem implies $|f(t)| \le M_R|t|$ for $|t| \le R$ with $M_R := \sup_{|t|\le R} |f'(t)|$. Hence, since $\|x\|_\infty \le \|x\|_p$, we have $\|F(x)\|_p \le M_{\|x\|_\infty}\|x\|_p$ and F is well defined. This also shows that multiplication by $f'(x_n)$ is a bounded linear map. To establish differentiability we use

$$f(t+s) - f(t) - f'(t)s = s\int_0^1 \big(f'(t+s\tau) - f'(t)\big)d\tau$$

and since f' is uniformly continuous on every compact interval, we can find a $\delta > 0$ for every given $R > 0$ and $\varepsilon > 0$ such that

$$|f'(t+s) - f'(t)| < \varepsilon \quad \text{if} \quad |s| < \delta,\ |t| < R.$$

Now for $x, u \in X$ with $\|x\|_\infty < R$ and $\|u\|_\infty < \delta$ we have $|f(x_n + u_n) - f(x_n) - f'(x_n)u_n| < \varepsilon|u_n|$ and hence

$$\|F(x+u) - F(x) - dF(x)u\|_p < \varepsilon\|u\|_p$$

which establishes differentiability. Moreover, using uniform continuity of f on compact sets a similar argument shows that $dF : X \to \mathscr{L}(X, X)$ is continuous (observe that the operator norm of a multiplication operator by a sequence is the sup norm of the sequence) and hence one writes $F \in C^1(X, X)$ as usual. ◇

Differentiability implies existence of directional derivatives

$$\delta F(x, u) := \lim_{\varepsilon\to 0} \frac{F(x+\varepsilon u) - F(x)}{\varepsilon}, \qquad \varepsilon \in \mathbb{R}\setminus\{0\}, \tag{16.4}$$

which are also known as **Gâteaux derivative** or **variational derivative**. Indeed, if F is differentiable at x, then (16.2) implies

$$\delta F(x,u) = dF(x)u. \tag{16.5}$$

In particular, we call F Gâteaux differentiable at $X \in U$ if the limit on the right-hand side in (16.4) exists for all $u \in X$. However, note that Gâteaux differentiability does not imply differentiability. In fact, the Gâteaux derivative might be unbounded or it might even fail to be linear in u. Some authors require the Gâteaux derivative to be a bounded linear operator and in this case we will write $\delta F(x,u) = \delta F(x)u$ but even this additional requirement does not imply differentiability in general. Note that in any case the Gâteaux derivative is homogenous, that is, if $\delta F(x,u)$ exists, then $\delta F(x,\lambda u)$ exists for every $\lambda \in \mathbb{R}$ and

$$\delta F(x,\lambda u) = \lambda\, \delta F(x,u), \qquad \lambda \in \mathbb{R}. \tag{16.6}$$

Example 16.3. The function $F : \mathbb{R}^2 \to \mathbb{R}$ given by $F(x,y) := \frac{x^3}{x^2+y^2}$ for $(x,y) \neq 0$ and $F(0,0) = 0$ is Gâteaux differentiable at 0 with Gâteaux derivative

$$\delta F(0,(u,v)) = \lim_{\varepsilon \to 0} \frac{F(\varepsilon u, \varepsilon v)}{\varepsilon} = F(u,v),$$

which is clearly nonlinear.

The function $F : \mathbb{R}^2 \to \mathbb{R}$ given by $F(x,y) = x$ for $y = x^2$ and $F(x,0) := 0$ else is Gâteaux differentiable at 0 with Gâteaux derivative $\delta F(0) = 0$, which is clearly linear. However, F is not differentiable.

If you take a linear function $L : X \to Y$ which is unbounded, then L is everywhere Gâteaux differentiable with derivative equal to Lu, which is linear but, by construction, not bounded. ⋄

Example 16.4. Let $X := L^2(0,1)$ and consider

$$F : X \to X, \qquad x \mapsto \sin(x).$$

First of all note that by $|\sin(t)| \leq |t|$ our map is indeed from X to X and since sine is Lipschitz continuous we get the same for F: $\|F(x) - F(y)\|_2 \leq \|x - y\|_2$. Moreover, F is Gâteaux differentiable at $x = 0$ with derivative given by

$$\delta F(0) = \mathbb{I}$$

but it is not differentiable at $x = 0$.

To see that the Gâteaux derivative is the identity note that

$$\lim_{\varepsilon \to 0} \frac{\sin(\varepsilon u(t))}{\varepsilon} = u(t)$$

pointwise and hence

$$\lim_{\varepsilon\to 0}\left\|\frac{\sin(\varepsilon u(.))}{\varepsilon}-u(.)\right\|_2=0$$

by dominated convergence since $|\frac{\sin(\varepsilon u(t))}{\varepsilon}|\le|u(t)|$.

To see that F is not differentiable let

$$u_n=\pi\chi_{[0,1/n]},\qquad \|u_n\|_2=\frac{\pi}{\sqrt{n}}$$

and observe that $F(u_n)=0$, implying that

$$\frac{\|F(u_n)-u_n\|_2}{\|u_n\|_2}=1$$

does not converge to 0. Note that this problem does not occur in $X:=C[0,1]$ (Problem 16.2). ⋄

We will mainly consider Fréchet derivatives in the remainder of this chapter as it will allow a theory quite close to the usual one for multivariable functions. First of all we of course we have linearity (which is easy to check):

Lemma 16.1. *Suppose $F,G:U\to Y$ are differentiable at $x\in U$ and $\alpha,\beta\in\mathbb{C}$. Then $\alpha F+\beta G$ is differentiable at x with $d(\alpha F+\beta G)(x)=\alpha dF(x)+\beta dG(x)$. Similarly, if the Gâteaux derivatives $\delta F(x,u)$ and $\delta G(x,u)$ exist, then so does $\delta(F+G)(x,u)=\delta F(x,u)+\delta G(x,u)$.*

Next, Fréchet differentiability implies continuity:

Lemma 16.2. *Suppose $F:U\to Y$ is differentiable at $x\in U$. Then F is continuous at x. Moreover, we can find constants $M,\delta>0$ such that*

$$|F(x+u)-F(x)|\le M|u|,\qquad |u|\le\delta. \tag{16.7}$$

Proof. For every $\varepsilon>0$ we can find a $\delta>0$ such that $|F(x+u)-F(x)-dF(x)\,u|\le\varepsilon|u|$ for $|u|\le\delta$. Now choose $M=\|dF(x)\|+\varepsilon$. □

Example 16.5. Note that this lemma fails for the Gâteaux derivative as the example of an unbounded linear function shows. In fact, it already fails in $\mathbb{R}^2$ as the function $F:\mathbb{R}^2\to\mathbb{R}$ given by $F(x,y)=1$ for $y=x^2\neq 0$ and $F(x,0)=0$ else shows: It is Gâteaux differentiable at 0 with $\delta F(0)=0$ but it is not continuous since $\lim_{\varepsilon\to 0}F(\varepsilon^2,\varepsilon)=1\neq 0=F(0,0)$. ⋄

If F is differentiable for all $x\in U$ we call F **differentiable**. In this case we get a map

$$dF:\begin{array}{ccc} U & \to & \mathscr{L}(X,Y)\\ x & \mapsto & dF(x)\end{array}. \tag{16.8}$$

If $dF:U\to\mathscr{L}(X,Y)$ is continuous, we call F continuously differentiable and write $F\in C^1(U,Y)$.

If X or Y has a (finite) product structure, then the computation of the derivatives can be reduced as usual. The following facts are simple and can be shown as in the case of $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$.

Let $Y := \bigtimes_{j=1}^m Y_j$ and let $F : X \to Y$ be given by $F = (F_1, \dots, F_m)$ with $F_j : X \to Y_j$. Then $F \in C^1(X, Y)$ if and only if $F_j \in C^1(X, Y_j)$, $1 \le j \le m$, and in this case $dF = (dF_1, \dots, dF_m)$. Similarly, if $X = \bigtimes_{i=1}^n X_i$, then one can define the **partial derivative** $\partial_i F \in \mathscr{L}(X_i, Y)$, which is the derivative of F considered as a function of the i-th variable alone (the other variables being fixed). We have $dF\, u = \sum_{i=1}^n \partial_i F\, u_i$, $u = (u_1, \dots, u_n) \in X$, and $F \in C^1(X, Y)$ if and only if all partial derivatives exist and are continuous.

Example 16.6. In the case of $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, the matrix representation of dF with respect to the canonical basis in $\mathbb{R}^n$ and $\mathbb{R}^m$ is given by the partial derivatives $\partial_i F_j(x)$ and is called **Jacobi matrix** of F at x. ⋄

Given $F \in C^1(U, Y)$ we have $dF \in C(U, \mathscr{L}(X, Y))$ and we can define the second derivative (provided it exists) via

$$dF(x + v) = dF(x) + d^2F(x)v + o(v). \tag{16.9}$$

In this case $d^2F : U \to \mathscr{L}(X, \mathscr{L}(X, Y))$ which maps x to the linear map $v \mapsto d^2F(x)v$ which for fixed v is a linear map $u \mapsto (d^2F(x)v)u$. Equivalently, we could regard $d^2F(x)$ as a map $d^2F(x) : X^2 \to Y$, $(u, v) \mapsto (d^2F(x)v)u$ which is linear in both arguments. That is, $d^2F(x)$ is a bilinear map $X^2 \to Y$. The corresponding norm on $\mathscr{L}(X, \mathscr{L}(X, Y))$ explicitly spelled out reads

$$\|d^2F(x)\| = \sup_{|v|=1} \|d^2F(x)v\| = \sup_{|u|=|v|=1} \|(d^2F(x)v)u\|. \tag{16.10}$$

Example 16.7. Note that if $F \in \mathscr{L}(X, Y)$, then $dF(x) = F$ (independent of x) and $d^2F(x) = 0$. ⋄

Example 16.8. Let X be a real Hilbert space and $F(x) = |x|^2$. Then we have already seen $dF(x)u = 2\langle x, u\rangle$ and hence

$$dF(x + v)u = 2\langle x + v, u\rangle = 2\langle x, u\rangle + 2\langle v, u\rangle = dF(x)u + 2\langle v, u\rangle$$

which shows $(d^2F(x)v)u = 2\langle v, u\rangle$. ⋄

Example 16.9. Suppose $f \in C^2(\mathbb{R})$ with $f(0) = 0$ and continue Example 16.2. Then we have $F \in C^2(X, X)$ with $d^2F(x)v$ the multiplication operator by the sequence $f''(x_n)v_n$, that is,

$$((d^2F(x)v)u)_n = f''(x_n)v_n u_n.$$

Indeed, arguing in a similar fashion we can find a δ_1 such that $|f'(x_n + v_n) - f'(x_n) - f''(x_n)v_n| \le \varepsilon|v_n|$ whenever $\|x\|_\infty < R$ and $\|v\|_\infty < \delta_1$. Hence

$$\|dF(x + v) - dF(x) - d^2F(x)v\| < \varepsilon\|v\|_p$$

which shows differentiability. Moreover, since $\|d^2F(x)\| = \|f''(x)\|_\infty$ one also easily verifies that $F \in C^2(X, X)$ using uniform continuity of f'' on compact sets. ⋄

We can iterate the procedure of differentiation and write $F \in C^r(U, Y)$, $r \geq 1$, if the r-th derivative of F, d^rF (i.e., the derivative of the $(r-1)$-th derivative of F), exists and is continuous. Note that $d^rF(x)$ will be a multilinear map in r arguments as we will show below. Finally, we set $C^\infty(U, Y) = \bigcap_{r\in\mathbb{N}} C^r(U, Y)$ and, for notational convenience, $C^0(U, Y) = C(U, Y)$ and $d^0F = F$.

Example 16.10. Let X be a Banach algebra. Consider the multiplication $M : X \times X \to X$. Then

$$\partial_1 M(x, y)u = uy, \qquad \partial_2 M(x, y)u = xu$$

and hence

$$dM(x, y)(u_1, u_2) = u_1y + xu_2.$$

Consequently dM is linear in (x, y) and hence

$$(d^2M(x, y)(v_1, v_2))(u_1, u_2) = u_1v_2 + v_1u_2$$

Consequently all differentials of order higher than two will vanish and in particular $M \in C^\infty(X \times X, X)$. ⋄

If F is bijective and F, F^{-1} are both of class C^r, $r \geq 1$, then F is called a **diffeomorphism** of class C^r.

For the composition of mappings we have the usual **chain rule**.

Lemma 16.3 (Chain rule)**.** *Let $U \subseteq X$, $V \subseteq Y$ and $F \in C^r(U, V)$ and $G \in C^r(V, Z)$, $r \geq 1$. Then $G \circ F \in C^r(U, Z)$ and*

$$d(G \circ F)(x) = dG(F(x)) \circ dF(x), \qquad x \in X. \tag{16.11}$$

Proof. Fix $x \in U$, $y = F(x) \in V$ and let $u \in X$ such that $v = dF(x)u$ with $x+u \in U$ and $y+v \in V$ for $|u|$ sufficiently small. Then $F(x+u) = y+v+o(u)$ and, with $\tilde{v} = v + o(u)$,

$$G(F(x + u)) = G(y + \tilde{v}) = G(y) + dG(y)\tilde{v} + o(\tilde{v}).$$

Using $|\tilde{v}| \leq \|dF(x)\||u| + |o(u)|$ we see that $o(\tilde{v}) = o(u)$ and hence

$$G(F(x+u)) = G(y) + dG(y)v + o(u) = G(F(x)) + dG(F(x)) \circ dF(x)u + o(u)$$

as required. This establishes the case $r = 1$. The general case follows from induction. □

In particular, if $\lambda \in Y^*$ is a bounded linear functional, then $d(\lambda \circ F) = d\lambda \circ dF = \lambda \circ dF$. As an application of this result we obtain

Theorem 16.4 (Schwarz)**.** *Suppose $F \in C^2(\mathbb{R}^n, Y)$. Then*

$$\partial_i \partial_j F = \partial_j \partial_i F$$

for any $1 \leq i, j \leq n$.

Proof. First of all note that $\partial_j F(x) \in \mathscr{L}(\mathbb{R}, Y)$ and thus it can be regarded as an element of Y. Clearly the same applies to $\partial_i \partial_j F(x)$. Let $\lambda \in Y^*$ be a bounded linear functional, then $\lambda \circ F \in C^2(\mathbb{R}^2, \mathbb{R})$ and hence $\partial_i \partial_j (\lambda \circ F) = \partial_j \partial_i (\lambda \circ F)$ by the classical theorem of Schwarz. Moreover, by our remark preceding this lemma $\partial_i \partial_j (\lambda \circ F) = \partial_i \lambda (\partial_j F) = \lambda(\partial_i \partial_j F)$ and hence $\lambda(\partial_i \partial_j F) = \lambda(\partial_j \partial_i F)$ for every $\lambda \in Y^*$ implying the claim. □

Now we let $F \in C^2(X, Y)$ and look at the function $G : \mathbb{R}^2 \to Y$, $(t, s) \mapsto G(t, s) = F(x + tu + sv)$. Then one computes

$$\partial_t G(t, s)\Big|_{t=0} = dF(x + sv)u$$

and hence

$$\partial_s \partial_t G(t, s)\Big|_{(s,t)=0} = \partial_s dF(x + sv)u\Big|_{s=0} = (d^2F(x)u)v.$$

Since by the previous lemma the oder of the derivatives is irrelevant, we obtain

$$d^2F(u, v) = d^2F(v, u), \tag{16.12}$$

that is, d^2F is a symmetric bilinear form. This result easily generalizes to higher derivatives. To this end we introduce some notation first.

A function $L : \bigtimes_{j=1}^n X_j \to Y$ is called **multilinear** if it is linear with respect to each argument. It is not hard to see that L is continuous if and only if

$$\|L\| = \sup_{x : |x_1| = \cdots = |x_n| = 1} |L(x_1, \ldots, x_n)| < \infty. \tag{16.13}$$

If we take n copies of the same space, the set of multilinear functions $L : X^n \to Y$ will be denoted by $\mathscr{L}^n(X, Y)$. A multilinear function is called **symmetric** provided its value remains unchanged if any two arguments are switched. With the norm from above it is a Banach space and in fact there is a canonical isometric isomorphism between $\mathscr{L}^n(X, Y)$ and $\mathscr{L}(X, \mathscr{L}^{n-1}(X, Y))$ given by $L : (x_1, \ldots, x_n) \mapsto L(x_1, \ldots, x_n)$ maps to $x_1 \mapsto L(x_1, .)$.

Lemma 16.5. *Suppose $F \in C^r(X, Y)$. Then for every $x \in X$ we have that*

$$d^r F(x)(u_1, \ldots, u_r) = \partial_{t_1} \cdots \partial_{t_r} F(x + \sum_{i=1}^r t_i u_i)|_{t_1 = \cdots = t_r = 0}. \tag{16.14}$$

Moreover, $d^r F(x) \in \mathscr{L}^r(X, Y)$ is a bounded symmetric multilinear form.

Proof. The representation (16.14) follows using induction as before. Symmetry follows since the order of the partial derivatives can be interchanged by Lemma 16.4. □

Finally, note that to each $L \in \mathscr{L}^n(X,Y)$ we can assign its polar form $L \in C(X,Y)$ using $L(x) = L(x,\dots,x)$, $x \in X$. If L is symmetric it can be reconstructed using polarization (Problem 16.4):

$$L(u_1,\dots,u_n) = \frac{1}{n!}\partial_{t_1}\cdots\partial_{t_n} L(\sum_{i=1}^n t_i u_i). \tag{16.15}$$

We also have the following version of the **product rule**: Suppose $L \in \mathscr{L}^2(X,Y)$, then $L \in C^1(X^2,Y)$ with

$$dL(x)u = L(u_1,x_2) + L(x_1,u_2) \tag{16.16}$$

since

$$\begin{aligned} L(x_1+u_1,x_2+u_2) - L(x_1,x_2) &= L(u_1,x_2) + L(x_1,u_2) + L(u_1,u_2) \\ &= L(u_1,x_2) + L(x_1,u_2) + O(|u|^2) \end{aligned} \tag{16.17}$$

as $|L(u_1,u_2)| \le \|L\||u_1||u_2| = O(|u|^2)$. If X is a Banach algebra and $L(x_1,x_2) = x_1x_2$ we obtain the usual form of the product rule.

Next we have the following mean value theorem.

Theorem 16.6 (Mean value)**.** *Suppose $U \subseteq X$ and $F : U \to Y$ is Gâteaux differentiable at every $x \in U$. If U is convex, then*

$$|F(x) - F(y)| \le M|x-y|, \qquad M := \sup_{0\le t\le 1} |\delta F((1-t)x+ty, \frac{x-y}{|x-y|})|. \tag{16.18}$$

Conversely, (for any open U) if

$$|F(x) - F(y)| \le M|x-y|, \qquad x,y \in U, \tag{16.19}$$

then

$$\sup_{x\in U, |e|=1} |\delta F(x,e)| \le M. \tag{16.20}$$

Proof. Abbreviate $f(t) = F((1-t)x+ty)$, $0 \le t \le 1$, and hence $df(t) = \delta F((1-t)x+ty, y-x)$ implying $|df(t)| \le \tilde{M} := M|x-y|$ by (16.6). For the first part it suffices to show

$$\phi(t) = |f(t) - f(0)| - (\tilde{M}+\delta)t \le 0$$

for any $\delta > 0$. Let $t_0 = \max\{t \in [0,1] | \phi(t) \le 0\}$. If $t_0 < 1$ then

$$\begin{aligned}\phi(t_0+\varepsilon) &= |f(t_0+\varepsilon) - f(t_0) + f(t_0) - f(0)| - (\tilde{M}+\delta)(t_0+\varepsilon)\\ &\le |f(t_0+\varepsilon) - f(t_0)| - (\tilde{M}+\delta)\varepsilon + \phi(t_0)\\ &\le |df(t_0)\varepsilon + o(\varepsilon)| - (\tilde{M}+\delta)\varepsilon\\ &\le (\tilde{M} + o(1) - \tilde{M} - \delta)\varepsilon = (-\delta + o(1))\varepsilon \le 0,\end{aligned}$$

for $\varepsilon \ge 0$, small enough. Thus $t_0 = 1$.

To prove the second claim suppose we can find an $e \in X$, $|e| = 1$ such that $|\delta F(x_0, e)| = M + \delta$ for some $\delta > 0$ and hence

$$\begin{aligned}M\varepsilon &\ge |F(x_0+\varepsilon e) - F(x_0)| = |\delta F(x_0, e)\varepsilon + o(\varepsilon)|\\ &\ge (M+\delta)\varepsilon - |o(\varepsilon)| > M\varepsilon\end{aligned}$$

since we can assume $|o(\varepsilon)| < \varepsilon\delta$ for $\varepsilon > 0$ small enough, a contradiction. □

Corollary 16.7. *Suppose $U \subseteq X$ and $F \in C^1(U, Y)$. Then F is locally Lipschitz continuous.*

Proof. Fix $x \in U$ and note that by continuity there is a neighborhood U_0 of x_0 such that $\|dF(x)\| \le M$ for $x \in U_0$. Hence the claim follows from the mean value theorem. □

Note, however, that a C^1 function is in general not Lipschitz on arbitrary bounded sets since in the infinite dimensional case continuity of dF does not suffice to conclude boundedness on bounded closed sets.

Example 16.11. Let X be an infinite Hilbert space and $\{u_n\}_{n\in\mathbb{N}}$ some orthonormal set. Then the functions $F_n(x) := \max(0, 1 - 2\|x - u_n\|)$ are continuous with disjoint supports. Hence $F(x) := \sum_{n\in\mathbb{N}} nF_n(x)$ is also continuous (show this). But F is not bounded on the unit ball since $F(u_n) = n$. ◇

As an immediate consequence we obtain

Corollary 16.8. *Suppose U is a connected subset of a Banach space X. A Gâtaux differentiable mapping $F : U \to Y$ is constant if and only if $\delta F = 0$. In addition, if $F_{1,2} : U \to Y$ and $\delta F_1 = \delta F_2$, then F_1 and F_2 differ only by a constant.*

Now we turn to integration. We will only consider the case of mappings $f : I \to X$ where $I := [a, b] \subset \mathbb{R}$ is a compact interval and X is a Banach space. A function $f : I \to X$ is called **simple** if the image of f is finite, $f(I) =: \{x_i\}_{i=1}^n$, and if each inverse image $f^{-1}(x_i)$, $1 \le i \le n$ is a Borel set. The set of simple functions $S(I, X)$ forms a linear space and can be

equipped with the sup norm. The corresponding Banach space obtained after completion is called the set of **regulated functions** $R(I, X)$.

Observe that $C(I, X) \subset R(I, X)$. In fact, consider the functions $f_n := \sum_{i=0}^{n-1} f(t_i)\chi_{[t_i,t_{i+1})} \in S(I, X)$, where $t_i = a + i\frac{b-a}{n}$ and χ is the characteristic function. Since $f \in C(I, X)$ is uniformly continuous, we infer that f_n converges uniformly to f.

For $f \in S(I, X)$ we can define a linear map $\int : S(I, X) \to X$ by

$$\int_a^b f(t)dt = \sum_{i=1}^{n} x_i \lambda(f^{-1}(x_i)), \tag{16.21}$$

where λ denotes the Lebesgue measure on I. This map satisfies

$$\left| \int_a^b f(t)dt \right| \leq |f|(b-a). \tag{16.22}$$

and hence it can be extended uniquely to a linear map $\int : R(I, X) \to X$ with the same norm $(b-a)$. We even have

$$\left| \int_a^b f(t)dt \right| \leq \int_a^b |f(t)|dt \tag{16.23}$$

since this holds for simple functions by the triangle inequality and hence for all functions by approximation.

In addition, if $\lambda \in X^*$ is a continuous linear functional, then

$$\ell(\int_a^b f(t)dt) = \int_a^b \ell(f(t))dt, \qquad f \in R(I, X). \tag{16.24}$$

We will use the usual conventions $\int_{t_1}^{t_2} f(s)ds = \int_a^b \chi_{(t_1,t_2)}(s)f(s)ds$ and $\int_{t_2}^{t_1} f(s)ds = -\int_{t_1}^{t_2} f(s)ds$.

If $I \subseteq \mathbb{R}$, we have an isomorphism $\mathscr{L}(I, X) \equiv X$ and if $F : I \to X$ we will write $\dot{F}(t)$ instead of $dF(t)$ if we regard $dF(t)$ as an element of X. Note that in this case the Gâteaux and Fréchet derivatives coincide as we have

$$\dot{F}(t) = \lim_{\varepsilon \to 0} \frac{F(t+\varepsilon) - F(t)}{\varepsilon}. \tag{16.25}$$

By $C^1(I, X)$ we will denote the set of functions from $C(I, X)$ which are differentiable in the interior of I and for which the derivative has a continuous extension to $C(I, X)$.

Theorem 16.9 (Fundamental theorem of calculus)**.** *Suppose $F \in C^1(I, X)$, then*

$$F(t) = F(a) + \int_a^t \dot{F}(s)ds. \tag{16.26}$$

Conversely, if $f \in C(I, X)$, then $F(t) = \int_a^t f(s)ds \in C^1(I, X)$ and $\dot{F}(t) = f(t)$.

Proof. Let $f \in C(I,X)$ and set $G(t) := \int_a^t f(s)ds$. Then $G \in C^1(I,X)$ with $\dot{G}(t) = f(t)$ as can be seen from

$$|\int_a^{t+\varepsilon} f(s)ds - \int_a^t f(s)ds - f(t)\varepsilon| = |\int_t^{t+\varepsilon} (f(s)-f(t))ds| \le |\varepsilon| \sup_{s\in[t,t+\varepsilon]} |f(s)-f(t)|.$$

Hence if $F \in C^1(I,X)$ then $G(t) := \int_a^t (\dot{F}(s))ds$ satisfies $\dot{G} = \dot{F}$ and hence $F(t) = C + G(t)$ by Corollary 16.8. Choosing $t = a$ finally shows $F(a) = C$. □

As a simple application we obtain a generalization of the well-known fact that continuity of the directional derivatives implies continuous differentiability.

Lemma 16.10. *Suppose $F : U \subseteq X \to Y$ is Gâteaux differentiable such that the Gâteaux derivative is linear and continous, $\delta F \in C(U, \mathscr{L}(X,Y))$. Then $F \in C^1(U,Y)$ and $dF = \delta F$.*

Proof. By assumption $f(t) := F(x+tu)$ is in $C^1([0,1],Y)$ for u with sufficiently small norm. Moreover, by definition we have $\dot{f} = \delta F(x+tu)u$ and using the fundamental theorem of calculus we obtain

$$\begin{aligned} F(x+u) - F(x) = f(1) - f(0) &= \int_0^1 \dot{f}(t)dt = \int_0^1 \delta F(x+tu)u\, dt \\ &= \left(\int_0^1 \delta F(x+tu)dt\right) u, \end{aligned}$$

where the last equality follows from continuity of the integral since it clearly holds for simple functions. Consequently

$$\begin{aligned} |F(x+u) - F(x) - \delta F(x)u| &= \left|\int_0^1 \big((\delta F(x+tu) - \delta F(x))dt\big)u\right| \\ &\le \int_0^1 (\|\delta F(x+tu) - \delta F(x)\|dt)\, |u| \\ &\le \max_{t\in[0,1]} \|\delta F(x+tu) - \delta F(x)\|\, |u|. \end{aligned}$$

By the continuity assumption on δF, the right-hand side is $o(u)$ as required. □

As another consequence we obtain **Taylors theorem**.

Theorem 16.11 (Taylor)**.** *Suppose $U \subseteq X$ and $F \in C^{r+1}(U,Y)$. Then*

$$\begin{aligned} F(x+u) =& F(x) + dF(x)u + \frac{1}{2}d^2F(x)u^2 + \cdots + \frac{1}{r!}d^rF(x)u^r \\ &+ \left(\frac{1}{r!}\int_0^1 (1-t)^r d^{r+1}F(x+tu)dt\right) u^{r+1}, \end{aligned} \tag{16.27}$$

where $u^k := (u, \dots, u) \in X^k$.

Proof. As in the proof of the previous lemma, the case $r = 0$ is just the fundamental theorem of calculus applied to $f(t) := F(x + tu)$. For the induction step we use integration by parts. To this end let $f_j \in C^1([0,1], X_j)$, $L \in \mathscr{L}^2(X_1 \times X_2, Y)$ bilinear. Then the product rule (16.16) and the fundamental theorem of calculus imply

$$\int_0^1 L(\dot{f}_1(t), f_2(t))dt = L(f_1(1), f_2(1)) - L(f_1(0), f_2(0)) - \int_0^1 L(f_1(t), \dot{f}_2(t))dt.$$

Hence applying integration by parts with $L(y,t) = ty$, $f_1(t) = d^{r+1}F(x+ut)$, and $f_2(t) = \frac{(1-t)^{r+1}}{(r+1)!}$ establishes the induction step. □

Of course this also gives the Peano form for the remainder:

Corollary 16.12. *Suppose* $U \subseteq X$ *and* $F \in C^r(U, Y)$. *Then*

$$F(x+u) = F(x) + dF(x)u + \frac{1}{2}d^2F(x)u^2 + \cdots + \frac{1}{r!}d^rF(x)u^r + o(|u|^r). \quad (16.28)$$

Proof. Just estimate

$$\begin{aligned}\left|\left(\frac{1}{(r-1)!}\int_0^1 (1-t)^{r-1}d^rF(x+tu)dt - \frac{1}{r!}d^rF(x)\right)u^r\right| \\ \le \frac{|u|^r}{(r-1)!}\int_0^1 (1-t)^{r-1}\|d^rF(x+tu) - d^rF(x)\|dt \\ \le \frac{|u|^r}{r!}\sup_{0\le t\le 1}\|d^rF(x+tu) - d^rF(x)\|.\end{aligned}$$

□

Finally we remark that it is often necessary to equip $C^r(U, Y)$ with a norm. A suitable choice is

$$\|F\| = \sum_{0\le j\le r}\sup_{x\in U}\|d^jF(x)\|. \quad (16.29)$$

The set of all r times continuously differentiable functions for which this norm is finite forms a Banach space which is denoted by $C_b^r(U, Y)$.

In the definition of differentiability we have required U to be open. Of course there is no stringent reason for this and (16.3) could simply be required for all sequences from $U \setminus \{x\}$ converging to x. However, note that the derivative might not be unique in case you miss some directions (the ultimate problem occurring at an isolated point). Our requirement avoids all these issues. Moreover, there is usually another way of defining differentiability at a boundary point: By $C^r(\overline{U}, Y)$ we denote the set of all functions in $C^r(U, Y)$ all whose derivatives of order up to r have a continuous extension to $\overline{U}$. Note that if you can approach a boundary point along a half-line then

the fundamental theorem of calculus shows that the extension coincides with the Gâteaux derivative.

Problem 16.1. *Let X be a Hilbert space, $A \in \mathscr{L}(X)$ and $F(x) := \langle x, Ax\rangle$. Compute $d^n F$.*

Problem* 16.2. *Let $X := C[0,1]$ and suppose $f \in C^1(\mathbb{R})$. Show that*

$$F : C[0,1] \to C[0,1], \qquad x \mapsto f \circ x$$

is differentiable for every $x \in \ell^p(\mathbb{N})$ with derivative given by

$$(dF(x)y)(t) = f'(x(t))y(t).$$

Problem 16.3. *Let $X := \ell^2(\mathbb{N})$, $Y := \ell^1(\mathbb{N})$ and $F : X \to Y$ given by $F(x)_j := x_j^2$. Show $F \in C^\infty(X, Y)$ and compute all derivatives.*

Problem* 16.4. *Show* (16.15).

Problem 16.5. *Let X be a Banach algebra, $I \subseteq \mathbb{R}$ an open interval, and $x, y \in C^1(I, X)$. Show that $xy \in C^1(I, x)$ with $(xy)^{\cdot} = \dot{x}y + x\dot{y}$.*

Problem 16.6. *Let X be a Banach algebra and $\mathcal{G}(X)$ the group of invertible elements. Show that $I : \mathcal{G}(X) \to \mathcal{G}(X)$, $x \mapsto x^{-1}$ is differentiable with*

$$dI(x)y = -x^{-1}yx.$$

(Hint: (6.9)*)*

16.2. Minimizing functionals

Many problems in applications lead to finding the minimum (or maximum) of a given functional $F : X \to \mathbb{R}$. Since the minima of $-F$ are the maxima of F and vice versa, we will restrict our attention to minima only. Of course if $X = \mathbb{R}$ (or $\mathbb{R}^n$) we can find the local extrema by searching for the zeros of the derivative and then checking the second derivative to determine if it is a minim or maximum. In fact, by virtue of our version of Taylor's theorem (16.28) we see that F will take values above and below $F(x)$ in a vicinity of x if we can find some u such that $dF(x)u \neq 0$. Hence $dF(x) = 0$ is clearly a necessary condition for a local extremum. Moreover, if $dF(x) = 0$ we can go one step further and conclude that all values in a vicinity of x will lie above $F(x)$ provided the second derivative $d^2F(x)$ is positive in the sense that there is some $c > 0$ such that $d^2F(x)u^2 > c$ for all directions $u \in B_1(0)$. While this gives a viable solution to the problem of finding local extrema, we can easily do a bit better. To this end we look at the variations of f along lines trough x, that is, we look at the behavior of the function

$$f(t) := F(x + tu) \tag{16.30}$$

for a fixed direction $u \in B_1(0)$. Then, if F has a local extremum at x the same will be true for f and hence a necessary condition for an extremum is that the Gâteaux derivative vanishes in every direction: $\delta F(x,u) = 0$ for all unit vectors u. Similarly, a necessary condition for a local minimum at x is that f has a local minimum at 0 for all unit vectors u. For example $\delta^2 F(x,u) > 0$ for all unit vectors u. Here the higher order Gâteaux derivatives are defined as

$$\delta^n F(x,u) := \left(\frac{d}{dt}\right)^n F(x+tu)\Big|_{t=0} \tag{16.31}$$

with the derivative defined as a limit as in (16.4). That is we have the recursive definition $\delta^n F(x,u) = \lim_{\varepsilon\to 0} \varepsilon^{-1}\big(\delta^{n-1}F(x+\varepsilon u,u) - \delta^{n-1}F(x,u)\big)$. Note that if $\delta^n F(x,u)$ exists, then $\delta^n F(x,\lambda u)$ exists for every $\lambda \in \mathbb{R}$ and

$$\delta^n F(x,\lambda u) = \lambda^n \delta^n F(x,u), \qquad \lambda \in \mathbb{R}. \tag{16.32}$$

However, the condition $\delta^2 F(x,u) > 0$ for all unit vectors u is not sufficient as there are certain features you might miss when you only look at the function along rays trough a fixed point. This is demonstrated by the following example:

Example 16.12. Let $X = \mathbb{R}^2$ and consider the points $(x_n, y_n) := (\frac{1}{n}, \frac{1}{n^2})$. For each point choose a radius r_n such that the balls $B_n := B_{r_n}(x_n, y_n)$ are disjoint and lie between two parabolas $B_n \subset \{(x,y) | x \geq 0, \frac{x^2}{2} \leq y \leq 2x^2\}$. Moreover, choose a smooth nonnegative bump function $\phi(r^2)$ with support in $[-1,1]$ and maximum 1 at 0. Now consider $F(x,y) = x^2 + y^2 - 2\sum_{n\in\mathbb{N}} \rho_n \phi(\frac{(x-x_n)^2+(y-y_n)^2}{r_n^2})$, where $\rho_n = x_n^2 + y_n^2$. By construction F is smooth away from zero. Moreover, at zero F is continuous and Gâteaux differentiable of arbitrary order with $F(0,0) = 0$, $\delta F((0,0),(u,v)) = 0$, $\delta^2 F((0,0),(u,v)) = 2(u^2+v^2)$, and $\delta^k F((0,0),(u,v)) = 0$ for $k \geq 3$.

In particular, $F(ut, vt)$ has a strict local minimum at $t = 0$ for every $(u,v) \in \mathbb{R}^2 \setminus \{0\}$, but F has no local minimum at $(0,0)$ since $F(x_n, y_n) = -\rho_n$. Cleary F is not differentiable at 0. In fact, note that the Gâteaux derivatives are not continuous at 0 (the derivatives in B_n grow like r_n^{-2}). ⋄

Lemma 16.13. *Suppose $F : U \to \mathbb{R}$ has Gâteaux derivatives up to the order of two. A necessary condition for $x \in U$ to be a local extremum is that $\delta F(x,u) = 0$ and $\delta^2 F(x,u) \geq 0$ for all $u \in X$. A sufficient condition for a strict local minimum is if addition $\delta^2 F(x,u) \geq c > 0$ for all $u \in \partial B_1(0)$ and $\delta^2 F$ is continuous at x uniformly with respect to $u \in \partial B_1(0)$.*

Proof. The necessary conditions have already been established. To see the sufficient conditions note that the assumptions on $\delta^2 F$ imply that there is some $\varepsilon > 0$ such that $\delta^2 F(y,u) \geq \frac{c}{2}$ for all $y \in B_\varepsilon(x)$ and all $u \in \partial B_1(0)$.

Equivalently, $\delta^2 F(y,u) \geq \frac{c}{2}|u|^2$ for all $y \in B_\varepsilon(x)$ and all $u \in X$. Hence applying Taylor's theorem to $f(t)$ using $\ddot{f}(t) = \delta^2 F(x+tu,u)$ gives

$$F(x+u) = f(1) = f(0) + \int_0^1 (1-s)\ddot{f}(s)ds \geq F(x) + \frac{c}{4}|u|^2$$

for $u \in B_\varepsilon(0)$. □

Note that if $F \in C^2(U,\mathbb{R})$ then $\delta^2 F(x,u) = d^2F(x)u^2$ and we obtain

Corollary 16.14. *Suppose $F \in C^2(U,\mathbb{R})$. A sufficient condition for $x \in U$ to be a strict local minimum is $dF(x) = 0$ and $d^2F(x)u^2 \geq c|u|^2$ for all $u \in X$.*

Proof. Observe that by $|\delta^2 F(x,u) - \delta^2 F(y,u)| \leq \|d^2F(x) - d^2F(y)\||u|^2$ the continuity requirement from the previous lemma is satisfied. □

Example 16.13. If X is a Hilbert space, then the symmetric bilinear form d^2F has a corresponding self-adjoint operator $A \in \mathscr{L}(X)$ such that $d^2F(u,v) = \langle u, Av\rangle$ and the condition $d^2F(x)u^2 \geq c|u|^2$ is equivalent to the spectral condition $\sigma(A) \subset [c,\infty)$. In the finite dimensional case A is of course the Jacobi matrix and the spectral conditions says that all eigenvalues must be positive. ◇

Example 16.14. Let $X := \ell^2(\mathbb{N})$ and consider

$$F(x) := \sum_{n\in\mathbb{N}} \left(\frac{x_n^2}{2n^2} - x_n^4 \right).$$

Then $F \in C^2(X,\mathbb{R})$ with $dF(x)u = \sum_{n\in\mathbb{N}}(\frac{x_n}{n^2} + 4x_n^3)u_n$ and $d^2F(x)(u,v) = \sum_{n\in\mathbb{N}}(\frac{1}{n^2} + 12x_n^2)v_n u_n$. In particular, $F(0) = 0$, $dF(0) = 0$ and and $d^2F(0)u^2 = \sum_n n^{-2}u_n^2 > 0$ for $u \neq 0$. However, $F(\delta^m/m) < 0$ shows that 0 is no local minimum. So the condition $d^2F(x)u^2 > 0$ is not sufficient in infinite dimensions. It is however, sufficient in finite dimensions since compactness of the unit ball leads to the stronger condition $d^2F(x,u) \geq c > 0$ for all $u \in \partial B_1(0)$. ◇

Example 16.15. Consider a classical particle whose location at time t is given by $q(t)$. Then the **least action principle** states that, if the particle moves from $q(a)$ to $q(b)$, the path of the particle will make the action functional

$$S(q) := \int_a^b L(t,q(t),\dot{q}(t))dt$$

stationary, that is

$$\delta S(q) = 0.$$

Here $L : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the **Lagrangian** of the system. The name suggests that the action should attain a minimum, but this is not always the case and hence it is also referred to as **stationary action principle**.

More precisely, let $L \in C^2(\mathbb{R}^{2n+1}, \mathbb{R})$ and in order to incorporate the requirement that the initial and end points are fixed, we take $X = \{x \in C^2([a,b], \mathbb{R}^n) | x(a) = x(b) = 0\}$ and consider

$$q(t) := q(a) + \frac{t-a}{b-a} q(b) + x(t), \qquad x \in X.$$

Hence we want to compute the Gâteaux derivative of $F(x) := S(q)$, where x and q are related as above with $q(a)$, $q(b)$ fixed. Then

$$\begin{aligned} \delta F(x,u) &= \frac{d}{dt}\Big|_{t=0} \int_a^b L(s, q(s) + t\,u(s), \dot q(s) + t\,\dot u(s)) ds \\ &= \int_a^b \big(L_q(s, q(s), \dot q(s)) u(s) + L_{\dot q}(s, q(s), \dot q(s)) \dot u(s)\big) ds \\ &= \int_a^b \big(L_q(s, q(s), \dot q(s)) u(s) - \partial_s L_{\dot q}(s, q(s), \dot q(s))\big) u(s) ds, \end{aligned}$$

where we have used integration by parts (including the boundary conditions) to obtain the last equality. Here L_q, $L_{\dot q}$ are the gradients with respect to q, $\dot q$, respectively, and products are understood as scalar products in $\mathbb{R}^n$.

If we want this to vanish for all $u \in X$ we obtain the corresponding **Euler–Lagrange equation**

$$\partial_s L_{\dot q}(s, q(s), \dot q(s)) = L_q(s, q(s), \dot q(s)).$$

For example, for a classical particle of mass $m > 0$ moving in a conservative force field described by a potential $V \in C^1(\mathbb{R}^n, \mathbb{R})$ the Lagrangian is given by the difference between kinetic and potential energy

$$L(t, q, \dot q) := \frac{m}{2} \dot q^2 - V(q)$$

and the Euler–Lagrange equations read

$$m \ddot q = -V_q(q),$$

which are just Newton's equation of motion. ⋄

Finally we note that the situation simplifies a lot when F is convex. Our first observation is that a local minimum is automatically a global one.

Lemma 16.15. *Suppose $C \subseteq X$ is convex and $F : C \to \mathbb{R}$ is convex. Every local minimum is a global minimum. Moreover, if F is strictly convex then the minimum is unique.*

Proof. Suppose x is a local minimum and $F(y) < F(x)$. Then $F(\lambda y + (1-\lambda)x) \le \lambda F(y) + (1-\lambda) F(x) < F(x)$ for $\lambda \in (0,1)$ contradicts the fact that x

is a local minimum. If x, y are two global minima, then $F(\lambda y + (1-\lambda)x) < F(y) = F(x)$ yielding a contradiction unless $x = y$. □

Moreover, to find the global minimum it suffices to find a point where the Gâteaux derivative vanishes.

Lemma 16.16. *Suppose $C \subseteq X$ is convex and $F : C \to \mathbb{R}$ is convex. If the Gâteaux derivative exists at an interior point $x \in C$ and satisfies $\delta F(x, u) = 0$ for all $u \in X$, then x is a global minimum.*

Proof. By assumption $f(t) := F(x + tu)$ is a convex function defined on an interval containing 0 with $f'(0) = 0$. If y is another point we can choose $u = y - x$ and Lemma 10.2 (iii) implies $F(y) = f(1) \geq f(0) = F(x)$. □

As in the one-dimensional case, convexity can be read off from the second derivative.

Lemma 16.17. *Suppose $C \subseteq X$ is open and convex and $F : C \to \mathbb{R}$ has Gâteaux derivatives up to order two. Then F is convex if and only if $\delta^2 F(x, u) \geq 0$ for all $x \in C$ and $u \in X$. Moreover, F is strictly convex if $\delta^2 F(x, u) > 0$ for all $x \in C$ and $u \in X \setminus \{0\}$.*

Proof. We consider $f(t) := F(x + tu)$ as before such that $f'(t) = \delta F(x + tu, u)$, $f''(t) = \delta^2 F(x + tu, u)$. Moreover, note that f is (strictly) convex for all $x \in C$ and $u \in X \setminus \{0\}$ if and only if F is (strictly) convex. Indeed, if F is (strictly) convex so is f as is easy to check. To see the converse note

$$F(\lambda y + (1-\lambda)x) = f(\lambda) \leq \lambda f(1) - (1-\lambda) f(0) = \lambda F(y) - (1-\lambda) F(x)$$

with strict inequality if f is strictly convex. The rest follows from Problem 10.5. □

There is also a version using only first derivatives plus the concept of a monotone operator. A map $F : U \subseteq X \to X^*$ is **monotone** if

$$(F(x) - F(y))(x - y) \geq 0, \qquad x, y \in U.$$

It is called **strictly monotone** if we have strict inequality for $x \neq y$. Monotone operators will be the topic of Chapter 19.

Lemma 16.18. *Suppose $C \subseteq X$ is open and convex and $F : C \to \mathbb{R}$ has Gâteaux derivatives $\delta F(x) \in X^*$ for every $x \in C$. Then F is (strictly) convex if and only if δF is (strictly) monotone.*

Proof. Note that by assumption $\delta F : C \to X^*$ and the claim follows as in the previous lemma from Problem 10.5 since $f'(t) = \delta F(x+tu)u$ which shows that δF is (strictly) monotone if and only if f' is (strictly) increasing. □

Example 16.16. The length of a curve $q : [a, b] \to \mathbb{R}^n$ is given by (cf. Problem 11.43)

$$\int_a^b |q'(s)|ds.$$

Of course we know that the shortest curve between two given points q_0 and q_1 is a straight line. Notwithstanding that this is evident, defining the length as the total variation (cf. again Problem 11.43), let us show this by seeking the minimum of the following functional

$$F(x) := \int_a^b |q'(s)|ds, \qquad q(t) = x(t) + q_0 + \frac{t-a}{b-a}(q_1 - q_0)$$

for $x \in X := \{x \in C^1([a, b], \mathbb{R}^n) | x(a) = x(b) = 0\}$. Unfortunately our integrand will not be differentiable unless $|\dot{q}| \geq c$. However, since the absolute value is convex, so is F and it will suffice to search for a local minimum within the convex open set $C := \{x \in X || \dot{x}| < \frac{|q_1 - q_0|}{2(b-a)}\}$. We compute

$$\delta F(x, u) = \int_a^b \frac{q'(s)u'(s)}{|q'(s)|} ds$$

which shows (Lemma 10.22) that q' must be constant. Hence the local minimum in C is indeed a straight line and this must also be a global minimum in X. However, since the length of a curve is independent of its parametrization, this minimum is not unique! ◇

Example 16.17. Let us try to find a curve $y(x)$ from $y(0) = 0$ to $y(x_1) = y_1$ which minimizes

$$F(y) := \int_0^{x_1} \sqrt{\frac{1 + y'(x)^2}{x}} dx.$$

Note that since the function $t \mapsto \sqrt{1 + t^2}$ is convex, we obtain that F is convex. Hence it suffices to find a zero of

$$\delta F(y, u) = \int_0^{x_1} \frac{y'(x)u'(x)}{\sqrt{x(1 + y'(x)^2)}} dx,$$

which shows (Lemma 10.22) that $\frac{y'}{\sqrt{x(1+y'^2)}} = C^{-1/2}$ is constant or equivalently

$$y'(x) = \sqrt{\frac{x}{C - x}}$$

and hence

$$y(x) = C \arctan\left(\sqrt{\frac{x}{C - x}}\right) - \sqrt{x(C - x)}.$$

The constant C has to be chosen such that $y(x_1)$ matches the given value y_1. Note that $C \mapsto y(x_1)$ decreases from $\frac{\pi x_1}{2}$ to 0 and hence there will be a unique $C > x_1$ for $0 < y_1 < \frac{\pi x_1}{2}$. ◇

Problem 16.7. *Consider the least action principle for a classical one-dimensional particle. Show that*

$$\delta^2 F(x,u) = \int_a^b \big(m\,\dot u(s)^2 - V''(q(s))u(s)^2\big)ds.$$

Moreover, show that we have indeed a minimum if $V'' \le 0$.

16.3. Contraction principles

Let X be a Banach space. A fixed point of a mapping $F : C \subseteq X \to C$ is an element $x \in C$ such that $F(x) = x$. Moreover, F is called a contraction if there is a contraction constant $\theta \in [0, 1)$ such that

$$|F(x) - F(\tilde x)| \le \theta |x - \tilde x|, \qquad x, \tilde x \in C. \tag{16.33}$$

Note that a contraction is continuous. We also recall the notation $F^n(x) = F(F^{n-1}(x))$, $F^0(x) = x$.

Theorem 16.19 (Contraction principle)**.** *Let C be a nonempty closed subset of a Banach space X and let $F : C \to C$ be a contraction, then F has a unique fixed point $\overline{x} \in C$ such that*

$$|F^n(x) - \overline{x}| \le \frac{\theta^n}{1-\theta}|F(x) - x|, \qquad x \in C. \tag{16.34}$$

Proof. If $x = F(x)$ and $\tilde x = F(\tilde x)$, then $|x - \tilde x| = |F(x) - F(\tilde x)| \le \theta |x - \tilde x|$ shows that there can be at most one fixed point.

Concerning existence, fix $x_0 \in C$ and consider the sequence $x_n = F^n(x_0)$. We have

$$|x_{n+1} - x_n| \le \theta |x_n - x_{n-1}| \le \cdots \le \theta^n |x_1 - x_0|$$

and hence by the triangle inequality (for $n > m$)

$$\begin{aligned} |x_n - x_m| &\le \sum_{j=m+1}^{n} |x_j - x_{j-1}| \le \theta^m \sum_{j=0}^{n-m-1} \theta^j |x_1 - x_0| \\ &\le \frac{\theta^m}{1-\theta}|x_1 - x_0|. \end{aligned} \tag{16.35}$$

Thus x_n is Cauchy and tends to a limit $\overline{x}$. Moreover,

$$|F(\overline{x}) - \overline{x}| = \lim_{n\to\infty} |x_{n+1} - x_n| = 0$$

shows that $\overline{x}$ is a fixed point and the estimate (16.34) follows after taking the limit $m \to \infty$ in (16.35). □

Note that we can replace θ^n by any other summable sequence θ_n (Problem 16.9):

Theorem 16.20 (Weissinger)**.** *Let C be a nonempty closed subset of a Banach space X. Suppose $F : C \to C$ satisfies*

$$|F^n(x) - F^n(y)| \le \theta_n |x - y|, \qquad x, y \in C, \tag{16.36}$$

with $\sum_{n=1}^{\infty} \theta_n < \infty$. Then F has a unique fixed point $\overline{x}$ such that

$$|F^n(x) - \overline{x}| \le \left(\sum_{j=n}^{\infty} \theta_j \right) |F(x) - x|, \qquad x \in C. \tag{16.37}$$

Next, we want to investigate how fixed points of contractions vary with respect to a parameter. Let X, Y be Banach spaces, $U \subseteq X$, $V \subseteq Y$ be open and consider $F : \overline{U} \times V \to U$. The mapping F is called a uniform contraction if there is a $\theta \in [0, 1)$ such that

$$|F(x, y) - F(\tilde{x}, y)| \le \theta |x - \tilde{x}|, \quad x, \tilde{x} \in \overline{U},\ y \in V, \tag{16.38}$$

that us, the contraction constant θ is independent of y.

Theorem 16.21 (Uniform contraction principle)**.** *Let U, V be nonempty open subsets of Banach spaces X, Y, respectively. Let $F : \overline{U} \times V \to U$ be a uniform contraction and denote by $\overline{x}(y) \in U$ the unique fixed point of $F(., y)$. If $F \in C^r(U \times V, U)$, $r \ge 0$, then $\overline{x}(.) \in C^r(V, U)$.*

Proof. Let us first show that $\overline{x}(y)$ is continuous. From

$$\begin{aligned} |\overline{x}(y + v) - \overline{x}(y)| &= |F(\overline{x}(y + v), y + v) - F(\overline{x}(y), y + v) \\ &\quad + F(\overline{x}(y), y + v) - F(\overline{x}(y), y)| \\ &\le \theta |\overline{x}(y + v) - \overline{x}(y)| + |F(\overline{x}(y), y + v) - F(\overline{x}(y), y)| \end{aligned} \tag{16.39}$$

we infer

$$|\overline{x}(y + v) - \overline{x}(y)| \le \frac{1}{1 - \theta} |F(\overline{x}(y), y + v) - F(\overline{x}(y), y)| \tag{16.40}$$

and hence $\overline{x}(y) \in C(V, U)$. Now let $r := 1$ and let us formally differentiate $\overline{x}(y) = F(\overline{x}(y), y)$ with respect to y,

$$d\,\overline{x}(y) = \partial_x F(\overline{x}(y), y) d\,\overline{x}(y) + \partial_y F(\overline{x}(y), y). \tag{16.41}$$

Considering this as a fixed point equation $T(x', y) = x'$, where $T(., y) : \mathscr{L}(Y, X) \to \mathscr{L}(Y, X)$, $x' \mapsto \partial_x F(\overline{x}(y), y)x' + \partial_y F(\overline{x}(y), y)$ is a uniform contraction since we have $\|\partial_x F(\overline{x}(y), y)\| \le \theta$ by Theorem 16.6. Hence we get a unique continuous solution $\overline{x}'(y)$. It remains to show

$$\overline{x}(y + v) - \overline{x}(y) - \overline{x}'(y)v = o(v). \tag{16.42}$$

Let us abbreviate $u := \overline{x}(y+v) - \overline{x}(y)$, then using (16.41) and the fixed point property of $\overline{x}(y)$ we see

$$\begin{aligned}(1 - \partial_x F(\overline{x}(y), y))(u - \overline{x}'(y)v) &= \\ &= F(\overline{x}(y) + u, y + v) - F(\overline{x}(y), y) - \partial_x F(\overline{x}(y), y)u - \partial_y F(\overline{x}(y), y)v \\ &= o(u) + o(v) \end{aligned} \tag{16.43}$$

since $F \in C^1(U \times V, U)$ by assumption. Moreover, $\|(1 - \partial_x F(\overline{x}(y), y))^{-1}\| \leq (1-\theta)^{-1}$ and $u = O(v)$ (by (16.40)) implying $u - \overline{x}'(y)v = o(v)$ as desired.

Finally, suppose that the result holds for some $r - 1 \geq 1$. Thus, if F is C^r, then $\overline{x}(y)$ is at least C^{r-1} and the fact that $d\,\overline{x}(y)$ satisfies (16.41) implies $\overline{x}(y) \in C^r(V, U)$. □

As an important consequence we obtain the implicit function theorem.

Theorem 16.22 (Implicit function)**.** *Let X, Y, and Z be Banach spaces and let U, V be open subsets of X, Y, respectively. Let $F \in C^r(U \times V, Z)$, $r \geq 0$, and fix $(x_0, y_0) \in U \times V$. Suppose $\partial_x F \in C(U \times V, \mathscr{L}(X, Z))$ exists (if $r = 0$) and $\partial_x F(x_0, y_0) \in \mathscr{L}(X, Z)$ is an isomorphism. Then there exists an open neighborhood $U_1 \times V_1 \subseteq U \times V$ of (x_0, y_0) such that for each $y \in V_1$ there exists a unique point $(\xi(y), y) \in U_1 \times V_1$ satisfying $F(\xi(y), y) = F(x_0, y_0)$. Moreover, ξ is in $C^r(V_1, Z)$ and fulfills (for $r \geq 1$)*

$$d\xi(y) = -(\partial_x F(\xi(y), y))^{-1} \circ \partial_y F(\xi(y), y). \tag{16.44}$$

Proof. Using the shift $F \to F - F(x_0, y_0)$ we can assume $F(x_0, y_0) = 0$. Next, the fixed points of $G(x, y) = x - (\partial_x F(x_0, y_0))^{-1} F(x, y)$ are the solutions of $F(x, y) = 0$. The function G has the same smoothness properties as F and since $\partial_x G(x_0, y_0) = 0$, we can find balls U_1 and V_1 around x_0 and y_0 such that $\|\partial_x G(x, y)\| \leq \theta < 1$ for $(x, y) \in U_1 \times V_1$. Thus by the mean value theorem (Theorem 16.6) $G(., y)$ is a uniform contraction on U_1 for $y \in V_1$. Moreover, choosing the radius of V_1 sufficiently small such that $|G(x_0, y) - G(x_0, y_0)| < (1 - \theta)r$ for $y \in V_1$, where r is the radius of U_1, we get

$$|G(x, y) - x_0| = |G(x, y) - G(x_0, y_0)| \leq \theta |x - x_0| + (1 - \theta)r < r$$

for $(x, y) \in \overline{U_1} \times V_1$, that is, $G : \overline{U_1} \times V_1 \to U_1$. The rest follows from the uniform contraction principle. Formula (16.44) follows from differentiating $F(\xi(y), y) = 0$ using the chain rule. □

Note that our proof is constructive, since it shows that the solution $\xi(y)$ can be obtained by iterating $x - (\partial_x F(x_0, y_0))^{-1}(F(x, y) - F(x_0, y_0))$.

Moreover, as a corollary of the implicit function theorem we also obtain the inverse function theorem.

Theorem 16.23 (Inverse function)**.** *Suppose $F \in C^r(U,Y)$, $r \geq 1$, $U \subseteq X$, and let $dF(x_0)$ be an isomorphism for some $x_0 \in U$. Then there are neighborhoods U_1, V_1 of x_0, $F(x_0)$, respectively, such that $F \in C^r(U_1,V_1)$ is a diffeomorphism.*

Proof. Apply the implicit function theorem to $G(x,y) = y - F(x)$. □

Example 16.18. It is important to emphasize that invertibility of dF on all of U does not imply injectivity on U as the following example in $X := \mathbb{R}^2$ shows: $F(x,y) = (\mathrm{e}^{2x} - y^2 + 3, 4\mathrm{e}^{2x}y - y^3)$. ⋄

Example 16.19. Let X be a Banach algebra and $\mathcal{G}(X)$ the group of invertible elements. We have seen that multiplication is $C^\infty(X \times X, X)$ and hence taking the inverse is also $C^\infty(\mathcal{G}(X), \mathcal{G}(X))$. Consequently, $\mathcal{G}(X)$ is an (in general infinite-dimensional) **Lie group**. ⋄

Further applications will be given in the next section.

Problem 16.8. *Derive Newton's method for finding the zeros of a twice continuously differentiable function $f(x)$,*

$$x_{n+1} = F(x_n), \qquad F(x) = x - \frac{f(x)}{f'(x)},$$

from the contraction principle by showing that if $\overline{x}$ is a zero with $f'(\overline{x}) \neq 0$, then there is a corresponding closed interval C around $\overline{x}$ such that the assumptions of Theorem 16.19 are satisfied.

Problem* 16.9. *Prove Theorem 16.20. Moreover, suppose $F : C \to C$ and that F^n is a contraction. Show that the fixed point of F^n is also one of F. Hence Theorem 16.20 (except for the estimate) can also be considered as a special case of Theorem 16.19 since the assumption implies that F^n is a contraction for n sufficiently large.*

16.4. Ordinary differential equations

As a first application of the implicit function theorem, we prove (local) existence and uniqueness for solutions of ordinary differential equations in Banach spaces. Let X be a Banach space, $U \subseteq X$ a (nonempty) open subset, and $I \subseteq \mathbb{R}$ a compact interval. Denote by $C(I,U)$ the Banach space of bounded continuous functions equipped with the sup norm.

The following lemma, known as **omega lemma**, will be needed in the proof of the next theorem.

Lemma 16.24. *Suppose $I \subseteq \mathbb{R}$ is a compact interval and $f \in C^r(U,Y)$. Then $f_* \in C^r(C(I,U), C(I,Y))$, where*

$$(f_*x)(t) := f(x(t)). \tag{16.45}$$

Proof. Fix $x_0 \in C(I,U)$ and $\varepsilon > 0$. For each $t \in I$ we have a $\delta(t) > 0$ such that $\overline{B_{2\delta(t)}}(x_0(t)) \subset U$ and $|f(x) - f(x_0(t))| \le \varepsilon/2$ for all x with $|x - x_0(t)| \le 2\delta(t)$. The balls $B_{\delta(t)}(x_0(t))$, $t \in I$, cover the set $\{x_0(t)\}_{t\in I}$ and since I is compact, there is a finite subcover $B_{\delta(t_j)}(x_0(t_j))$, $1 \le j \le n$. Let $\|x - x_0\| \le \delta := \min_{1\le j\le n} \delta(t_j)$. Then for each $t \in I$ there is a t_j such that $|x_0(t) - x_0(t_j)| \le \delta(t_j)$ and hence $|f(x(t)) - f(x_0(t))| \le |f(x(t)) - f(x_0(t_j))| + |f(x_0(t_j)) - f(x_0(t))| \le \varepsilon$ since $|x(t) - x_0(t_j)| \le |x(t) - x_0(t)| + |x_0(t) - x_0(t_j)| \le 2\delta(t_j)$. This settles the case $r = 0$.

Next let us turn to $r = 1$. We claim that df_* is given by $(df_*(x_0)x)(t) := df(x_0(t))x(t)$. To show this we use Taylor's theorem (cf. the proof of Corollary 16.12) to conclude that

$$|f(x_0(t)+x)-f(x_0(t))-df(x_0(t))x| \le |x| \sup_{0\le s\le 1} \|df(x_0(t)+sx)-df(x_0(t))\|.$$

By the first part $(df)_*$ is continuous and hence for a given ε we can find a corresponding δ such that $|x(t) - y(t)| \le \delta$ implies $\|df(x(t)) - df(y(t))\| \le \varepsilon$ and hence $\|df(x_0(t) + sx) - df(x_0(t))\| \le \varepsilon$ for $|x_0(t) + sx - x_0(t)| \le |x| \le \delta$. But this shows differentiability of f_* as required. and it remains to show that df_* is continuous. To see this we use the linear map

$$\begin{array}{rccl} \lambda: & C(I,\mathscr{L}(X,Y)) & \to & \mathscr{L}(C(I,X),C(I,Y)) \ , \\ & T & \mapsto & T_* \end{array}$$

where $(T_*x)(t) := T(t)x(t)$. Since we have

$$\|T_*x\| = \sup_{t\in I} |T(t)x(t)| \le \sup_{t\in I} \|T(t)\||x(t)| \le \|T\|\|x\|,$$

we infer $|\lambda| \le 1$ and hence λ is continuous. Now observe $df_* = \lambda \circ (df)_*$.

The general case $r > 1$ follows from induction. □

Now we come to our existence and uniqueness result for the initial value problem in Banach spaces.

Theorem 16.25. *Let I be an open interval, U an open subset of a Banach space X and Λ an open subset of another Banach space. Suppose $F \in C^r(I \times U \times \Lambda, X)$, $r \ge 1$, then the initial value problem*

$$\dot{x} = F(t,x,\lambda), \qquad x(t_0) = x_0, \quad (t_0,x_0,\lambda) \in I \times U \times \Lambda, \tag{16.46}$$

has a unique solution $x(t,t_0,x_0,\lambda) \in C^r(I_1 \times I_2 \times U_1 \times \Lambda_1, X)$, where $I_{1,2}$, U_1, and Λ_1 are open subsets of I, U, and Λ, respectively. The sets I_2, U_1, and Λ_1 can be chosen to contain any point $t_0 \in I$, $x_0 \in U$, and $\lambda_0 \in \Lambda$, respectively.

Proof. Adding t and λ to the dependent variables x, that is considering $(\tau,x,\lambda) \in \mathbb{R} \times X \times \Lambda$ and augmenting the differential equation according to

$(\dot\tau, \dot x, \dot\lambda) = (1, F(\tau, x, \lambda), 0)$, we can assume that F is independent of t and λ. Moreover, by a translation we can even assume $t_0 = 0$.

Our goal is to invoke the implicit function theorem. In order to do this we introduce an additional parameter $\varepsilon \in \mathbb{R}$ and consider

$$\dot x = \varepsilon F(x_0 + x), \quad x \in D^1 := \{x \in C^1([-1,1], B_\delta(0)) | x(0) = 0\}, \quad (16.47)$$

such that we know the solution for $\varepsilon = 0$. The implicit function theorem will show that solutions still exist as long as ε remains small. At first sight this doesn't seem to be good enough for us since our original problem corresponds to $\varepsilon = 1$. But since ε corresponds to a scaling $t \to \varepsilon t$, the solution for one $\varepsilon > 0$ suffices. Now let us turn to the details.

Our problem (16.47) is equivalent to looking for zeros of the function

$$\begin{aligned} G: \quad D^1 \times U_0 \times \mathbb{R} &\to C([-1,1], X), \\ (x, x_0, \varepsilon) &\mapsto \dot x - \varepsilon F(x_0 + x), \end{aligned} \quad (16.48)$$

where U_0 is a neighborhood of x_0 and δ sufficiently small such that $U_0 + B_\delta(0) \subseteq U$. Lemma 16.24 ensures that this function is C^1. Now fix x_0, then $G(0, x_0, 0) = 0$ and $\partial_x G(0, x_0, 0) = T$, where $Tx := \dot x$. Since $(T^{-1}x)(t) = \int_0^t x(s)ds$ we can apply the implicit function theorem to conclude that there is a unique solution $x(x_0, \varepsilon) \in C^1(U_1 \times (-\varepsilon_0, \varepsilon_0), D^1) \hookrightarrow C^1([-1,1] \times U_1 \times (-\varepsilon_0, \varepsilon_0), X)$. In particular, the map $(t, x_0) \mapsto x_0 + x(x_0, \varepsilon)(t/\varepsilon)$ is in $C^1((-\varepsilon, \varepsilon) \times U_1, X)$. Hence it is the desired solution of our original problem. This settles the case $r = 1$.

For $r > 1$ we use induction. Suppose $F \in C^{r+1}$ and let $x(t, x_0)$ be the solution which is at least C^r. Moreover, $y(t, x_0) := \partial_{x_0} x(t, x_0)$ satisfies

$$\dot y = \partial_x F(x(t, x_0))y, \qquad y(0) = \mathbb{I},$$

and hence $y(t, x_0) \in C^r$. Moreover, the differential equation shows $\partial_t x(t, x_0) = F(x(t, x_0)) \in C^r$ which shows $x(t, x_0) \in C^{r+1}$. □

Example 16.20. The simplest example is a linear equation

$$\dot x = Ax, \qquad x(0) = x_0,$$

where $A \in \mathscr{L}(X)$. Then it is easy to verify that the solution is given by

$$x(t) = \exp(tA)x_0,$$

where

$$\exp(tA) := \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k.$$

It is easy to check that the last series converges absolutely (cf. also Problem 1.35) and solves the differential equation (Problem 16.10). ⋄

Example 16.21. The classical example $\dot{x} = x^2$, $x(0) = x_0$, in $X := \mathbb{R}$ with solution

$$x(t) = \frac{x_0}{1 - x_0 t}, \qquad t \in \begin{cases} (-\infty, \frac{1}{x_0}), & x_0 > 0, \\ \mathbb{R}, & x_0 = 0, \\ (\frac{1}{x_0}, \infty), & x_0 < 0. \end{cases}$$

shows that solutions might not exist for all $t \in \mathbb{R}$ even though the differential equation is defined for all $t \in \mathbb{R}$. ⋄

This raises the question about the maximal interval on which a solution of the initial value problem (16.46) can be defined.

Suppose that solutions of the initial value problem (16.46) exist locally and are unique (as guaranteed by Theorem 16.25). Let ϕ_1, ϕ_2 be two solutions of (16.46) defined on the open intervals I_1, I_2, respectively. Let $I := I_1 \cap I_2 = (T_-, T_+)$ and let (t_-, t_+) be the maximal open interval on which both solutions coincide. I claim that $(t_-, t_+) = (T_-, T_+)$. In fact, if $t_+ < T_+$, both solutions would also coincide at t_+ by continuity. Next, considering the initial value problem with initial condition $x(t_+) = \phi_1(t_+) = \phi_2(t_+)$ shows that both solutions coincide in a neighborhood of t_+ by local uniqueness. This contradicts maximality of t_+ and hence $t_+ = T_+$. Similarly, $t_- = T_-$.

Moreover, we get a solution

$$\phi(t) := \begin{cases} \phi_1(t), & t \in I_1, \\ \phi_2(t), & t \in I_2, \end{cases} \tag{16.49}$$

defined on $I_1 \cup I_2$. In fact, this even extends to an arbitrary number of solutions and in this way we get a (unique) solution defined on some maximal interval.

Theorem 16.26. *Suppose the initial value problem* (16.46) *has a unique local solution (e.g. the conditions of Theorem 16.25 are satisfied). Then there exists a unique maximal solution defined on some maximal interval* $I_{(t_0,x_0)} = (T_-(t_0, x_0), T_+(t_0, x_0))$.

Proof. Let $\mathcal{S}$ be the set of all solutions ϕ of (16.46) which are defined on an open interval I_ϕ. Let $\mathcal{I} := \bigcup_{\phi \in \mathcal{S}} I_\phi$, which is again open. Moreover, if $t_1 > t_0 \in \mathcal{I}$, then $t_1 \in I_\phi$ for some ϕ and thus $[t_0, t_1] \subseteq I_\phi \subseteq \mathcal{I}$. Similarly for $t_1 < t_0$ and thus $\mathcal{I}$ is an open interval containing t_0. In particular, it is of the form $\mathcal{I} = (T_-, T_+)$. Now define $\phi_{max}(t)$ on $\mathcal{I}$ by $\phi_{max}(t) := \phi(t)$ for some $\phi \in \mathcal{S}$ with $t \in I_\phi$. By our above considerations any two ϕ will give the same value, and thus $\phi_{max}(t)$ is well-defined. Moreover, for every $t_1 > t_0$ there is some $\phi \in \mathcal{S}$ such that $t_1 \in I_\phi$ and $\phi_{max}(t) = \phi(t)$ for $t \in (t_0 - \varepsilon, t_1 + \varepsilon)$ which shows that ϕ_{max} is a solution. By construction there cannot be a solution defined on a larger interval. □

The solution found in the previous theorem is called the **maximal solution**. A solution defined for all $t \in \mathbb{R}$ is called a **global solution**. Clearly every global solution is maximal.

The next result gives a simple criterion for a solution to be global.

Lemma 16.27. *Suppose $F \in C^1(\mathbb{R} \times X, X)$ and let $x(t)$ be a maximal solution of the initial value problem* (16.46)*. Suppose $|F(t, x(t))|$ is bounded on finite t-intervals. Then $x(t)$ is a global solution.*

Proof. Let (T_-, T_+) be the domain of $x(t)$ and suppose $T_+ < \infty$. Then $|F(t, x(t))| \le C$ for $t \in (t_0, T_+)$ and for $t_0 s < t < T_+$ we have

$$|x(t) - x(s)| \le \int_s^t |\dot{x}(\tau)| d\tau = \int_s^t |F(\tau, x(\tau))| d\tau \le C|t - s|.$$

Thus $x(t_n)$ is Cauchy whenever t_n is and hence $\lim_{t \to T_+} x(t) = x_+$ exists. Now let $y(t)$ be the solution satisfying the initial condition $y(T_+) = x_+$. Then

$$\tilde{x}(t) = \begin{cases} x(t), & t < T_+, \\ y(t), & t \ge T_+, \end{cases}$$

is a larger solution contradicting maximality of T_+. □

Example 16.22. Finally, we want to to apply this to a famous example, the so-called **FPU lattices** (after Enrico Fermi, John Pasta, and Stanislaw Ulam who investigated such systems numerically). This is a simple model of a linear chain of particles coupled via nearest neighbor interactions. Let us assume for simplicity that all particles are identical and that the interaction is described by a potential $V \in C^2(\mathbb{R})$. Then the equation of motions are given by

$$\ddot{q}_n(t) = V'(q_{n+1} - q_n) - V'(q_n - q_{n-1}), \qquad n \in \mathbb{Z},$$

where $q_n(t) \in \mathbb{R}$ denotes the position of the n'th particle at time $t \in \mathbb{R}$ and the particle index n runs trough all integers. If the potential is quadratic, $V(r) = \frac{k}{2} r^2$, then we get the discrete linear wave equation

$$\ddot{q}_n(t) = k\big(q_{n+1}(t) - 2q_n(t) + q_{n-1}(t)\big).$$

If we use the fact that the Jacobi operator $Aq_n = q_{n+1} - 2q_n + q_{n-1}$ is a bounded operator in $X = \ell^p(\mathbb{Z})$ we can easily solve this system as in the case of ordinary differential equations. In fact, if $q^0 = q(0)$ and $p^0 = \dot{q}(0)$ are the initial conditions then one can easily check (cf. Problem 16.10) that the solution is given by

$$q(t) = \cos(tA^{1/2})q^0 + \frac{\sin(tA^{1/2})}{A^{1/2}} p^0.$$

In the Hilbert space case $p = 2$ these functions of our operator A could be defined via the spectral theorem but here we just use the more direct definition

$$\cos(tA^{1/2}) := \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!} A^k, \qquad \frac{\sin(tA^{1/2})}{A^{1/2}} := \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k+1)!} A^k.$$

In the general case an explicit solution is no longer possible but we are still able to show global existence under appropriate conditions. To this end we will assume that V has a global minimum at 0 and hence looks like $V(r) = V(0) + \frac{k}{2} r^2 + o(r^2)$. As $V(0)$ does not enter our differential equation we will assume $V(0) = 0$ without loss of generality. Moreover, we will also introduce $p_n := \dot{q}_n$ to have a first order system

$$\dot{q}_n = p_n, \qquad \dot{p}_n = V'(q_{n+1} - q_n) - V'(q_n - q_{n-1}).$$

Since $V' \in C^1(\mathbb{R})$ with $V'(0) = 0$ it gives rise to a C^1 map on $\ell^p(\mathbb{N})$ (see Example 16.2). Since the same is true for shifts, the chain rule implies that the right-hand side of our system is a C^1 map and hence Theorem 16.25 gives us existence of a local solution. To get global solutions we will need a bound on solutions. This will follow from the fact that the energy of the system

$$H(p, q) := \sum_{n \in \mathbb{Z}} \left(\frac{p_n^2}{2} + V(q_{n+1} - q_n) \right)$$

is conserved. To ensure that the above sum is finite we will choose $X := \ell^2(\mathbb{Z}) \oplus \ell^2(\mathbb{Z})$ as our underlying Banach (in this case even Hilbert) space. Recall that since we assume V to have a minimum at 0 we have $|V(r)| \leq C_R r^2$ for $|r| < R$ and hence $H(p, q) < \infty$ for $(p, q) \in X$. Under these assumptions it is easy to check that $H \in C^1(X, \mathbb{R})$ and that

$$\begin{aligned}
\frac{d}{dt} H(p(t), q(t)) &= \sum_{n \in \mathbb{Z}} \Big(\dot{p}_n(t) p_n(t) + V'(q_{n+1}(t) - q_n(t)) \big(\dot{q}_{n+1}(t) - \dot{q}_n(t) \big) \Big) \\
&= \sum_{n \in \mathbb{Z}} \Big(\big(V'(q_{n+1} - q_n) - V'(q_n - q_{n-1}) \big) p_n(t) \\
&\qquad + V'(q_{n+1}(t) - q_n(t)) \big(p_{n+1}(t) - p_n(t) \big) \Big) \\
&= \sum_{n \in \mathbb{Z}} \Big(- V'(q_n - q_{n-1}) p_n(t) + V'(q_{n+1}(t) - q_n(t)) p_{n+1}(t) \Big) \\
&= 0
\end{aligned}$$

provided $(p(t), q(t))$ solves our equation. Consequently, since $V \geq 0$,

$$\|p(t)\|_2 \leq 2H(p(t), q(t)) = 2H(p(0), q(0)).$$

Moreover, $q_n(t) = q_n(0) + \int_0^t p_n(s)ds$ (note that since the ℓ^2 norm is stronger than the ℓ^∞ norm, $q_n(t)$ is differentiable for fixed n) implies

$$\|q(t)\|_2 \le \|q(0)\|_2 + \int_0^t \|p_n(s)\|_2 ds \le \|q(0)\|_2 + 2H(p(0), q(0))t.$$

So Lemma 16.27 ensures that solutions are global in X. Of course every solution from X is also a solution from $Y = \ell^p(\mathbb{Z}) \oplus \ell^p(\mathbb{Z})$ for all $p \ge 2$ (since the $\|.\|_2$ norm is stronger than the $\|.\|_p$ norm for $p \ge 2$).

Examples include the original FPU β-model $V_\beta(r) := \frac{1}{2}r^2 + \frac{\beta}{4}r^4$, $\beta > 0$, and the famous Toda lattice $V(r) := \mathrm{e}^{-r} + r - 1$. ⋄

Example 16.23. Consider the **discrete nonlinear Schrödinger equation (dNLS)**

$$\mathrm{i}\dot{u}(t) = Hu(t) \pm |u(t)|^{2p}u(t), \quad t \in \mathbb{R},$$

where $Hu_n = u_{n+1} + u_{n-1} + q_n u_n$ is the Jacobi operator, in $X = \ell^2(\mathbb{Z})$. Here $q \in \ell^\infty(\mathbb{Z})$ is a real-valued sequence corresponding to an external potential and $q = 0$ (or $q = -2$, depending on your preferences) is the free Jacobi operator. Clearly the right-hand side is C^1 for $p \ge 0$ and hence there is a unique local solution by Theorem 16.25. Moreover, for a solution we have

$$\frac{d}{dt}\|u(t)\|_2 = 2\mathrm{Re}\langle \dot{u}(t), u(t)\rangle = 2\mathrm{Im}\left(\langle Hu, u\rangle \pm \langle |u(t)|^{2p}u(t), u(t)\rangle\right) = 0$$

and hence the dNLS has a unique global norm preserving solution $u \in C^1(\mathbb{R}, \ell^2(\mathbb{Z}))$. Note that this in fact works for any self-adjoint $H \in \mathscr{L}(X)$. ⋄

It should be mentioned that the above theory does not suffice to cover partial differential equations. In fact, if we replace the difference operator by a differential operator we run into the problem that differentiation is not a continuous process!

Problem* 16.10. *Let*

$$f(z) := \sum_{j=0}^{\infty} f_j z^j, \qquad |z| < R,$$

be a convergent power series with convergence radius $R > 0$. Suppose X is a Banach space and $A \in \mathscr{L}(X)$ is a bounded operator with $\|A\| < R$. Show that

$$f(tA) := \sum_{j=0}^{\infty} f_j t^j A^j$$

is in $C^\infty(I, \mathscr{L}(X))$, $I = (-R\|A\|^{-1}, R\|A\|^{-1})$ and

$$\frac{d^n}{dt^n} f(tA) = A^n f^{(n)}(tA), \qquad n \in \mathbb{N}_0.$$

(Compare also Problem 1.35.)

Problem 16.11. *Consider the FPU α-model $V_\alpha(r) := \frac{1}{2}r^2 + \frac{\alpha}{3}r^3$. Show that solutions satisfying $\|q_{n+1}(0) - q_n(0)\|_\infty < \frac{1}{|\alpha|}$ and $H(p(0), q(0)) < \frac{1}{6\alpha^2}$ are global in $X := \ell^2(\mathbb{Z}) \oplus \ell^2(\mathbb{Z})$. (Hint: Of course local solutions follow from our considerations above. Moreover, note that $V(r)$ has a maximum at $r = -\frac{1}{\alpha}$. Now use conservation of energy to conclude that the solution cannot escape the region $|r| < \frac{1}{|\alpha|}$.)*

16.5. Bifurcation theory

One of the most basic tasks is finding the zeros of a given function $F \in C^k(U, Y)$, where $U \subseteq X$ and X, Y are Banach spaces. Frequently the equation will depend on a parameter $\mu \in \mathbb{R}$ (of course we could also consider the case where the the parameter is again in some Banach space, but we will only consider this basic case here). That is, we are looking for solutions $x(\mu)$ of the equation

$$F(\mu, x) = 0, \tag{16.50}$$

where $F \in C^k(I \times U, Y)$ for some suitable $k \in \mathbb{N}$ and some open interval $I \subseteq \mathbb{R}$. Moreover, we are interested in the case of values $\mu_0 \in I$, where there is a change in the number of solutions (i.e. where new solutions appear or old solutions disappear as μ increases). Such points $\mu_0 \in I$ will be called **bifurcation points**. Clearly this cannot happen at a point where the implicit function theorem is applicable and hence a necessary condition for a bifurcation point is that

$$\partial_x F(\mu_0, x_0) \tag{16.51}$$

is not invertible at some point x_0 satisfying the equation $F(\mu_0, x_0) = 0$.

Example 16.24. Consider $f(\mu, x) = x^2 - \mu$ in $C^\infty(\mathbb{R}\times\mathbb{R}, \mathbb{R})$. Then $f(\mu, x) = x^2 - \mu = 0$, $\partial_x f(\mu, x) = 2x = 0$ shows that $\mu_0 = 0$ is the only possible bifurcation point. Since there are no solutions for $\mu < 0$ and two solutions $x_0(\mu) = \pm\sqrt{\mu}$ for $\mu > 0$, we see that $\mu_0 = 0$ is a bifurcation point. Consider $f(\mu, x) = x^3 - \mu$ in $C^\infty(\mathbb{R} \times \mathbb{R}, \mathbb{R})$. Then $f(\mu, x) = x^3 - \mu = 0$, $\partial_x f(\mu, x) = 3x^3 = 0$ shows that again $\mu_0 = 0$ is the only possible bifurcation point. However, this time there is only one solution $x(\mu) = \operatorname{sign}(\mu)|\mu|^{1/3}$ for all $\mu \in \mathbb{R}$ and hence there is no bifurcation occurring at $\mu_0 = 0$. ⋄

So the derivative $\partial_x F$ tells us where to look for bifurcation points while further derivatives can be used to determine what kind of bifurcation (if any) occurs. Here we want to show how this can be done in the infinite dimensional case.

Suppose we have an abstract problem $F(\mu, x) = 0$ with $\mu \in \mathbb{R}$ and $x \in X$ some Banach space. We assume that $F \in C^1(\mathbb{R} \times X, X)$ and that there is a trivial solution $x = 0$, that is, $F(\mu, 0) = 0$.

The first step is to split off the trivial solution and reduce it effectively to a finite-dimensional problem. To this end we assume that we have found a point $\mu_0 \in \mathbb{R}$ such that the derivative $A := \partial_x F(\mu_0, 0)$ is not invertible. Moreover, we will assume that A is a Fredholm operator such that there exists (cf. Section 6.6) continuous linear projections

$$P : X = \operatorname{Ker}(A) \dot{+} X_0 \to \operatorname{Ker}(A), \qquad Q : X = X_1 \dot{+} \operatorname{Ran}(A) \to X_1. \tag{16.52}$$

Now split our equation into a system of two equations according to the above splitting of the underlying Banach space:

$$F(\mu, x) = 0 \qquad \Leftrightarrow \qquad F_1(\mu, u, v) = 0,\ F_2(\mu, u, v) = 0, \tag{16.53}$$

where $x = u + v$ with $u = Px \in \operatorname{Ker}(A)$, $v = (1-P)x \in X_0$ and $F_1(\mu, u, v) = Q\, F(\mu, u + v)$, $F_2(\mu, u, v) = (1 - Q)F(\mu, u + v)$.

Since P, Q are bounded this system is still C^1 and the derivatives are given by (recall the block structure of A from (6.75))

$$\begin{aligned} &\partial_u F_1(\mu_0, 0, 0) = 0, \qquad &&\partial_v F_1(\mu_0, 0, 0) = 0, \\ &\partial_u F_2(\mu_0, 0, 0) = 0, \qquad &&\partial_v F_2(\mu_0, 0, 0) = A_0. \end{aligned} \tag{16.54}$$

Moreover, since A_0 is an isomorphism, the implicit function theorem tells us that we can (locally) solve F_2 for v. That is, there exists a neighborhood U of $(\mu_0, 0) \in \mathbb{R} \times \operatorname{Ker}(A)$ and a unique function $\psi \in C^1(U, X_0)$ such that

$$F_2(\mu, u, \psi(\mu, u)) = 0, \qquad (\mu, u) \in U. \tag{16.55}$$

In particular, by the uniqueness part we have $\psi(\mu, 0) = 0$. Moreover, $\partial_u \psi(\mu_0, 0) = -A_0^{-1} \partial_u F_2(\mu_0, 0, 0) = 0$.

Plugging this into the first equation reduces to original system to the finite dimensional system

$$\tilde{F}_1(\mu, u) = F_1(\mu, u, \psi(\mu, u)) = 0. \tag{16.56}$$

Of course the chain rule tells us that $\tilde{F} \in C^1$. Moreover, we still have $\tilde{F}_1(\mu, 0) = F_1(\mu, 0, \psi(\mu, 0)) = QF(\mu, 0) = 0$ as well as

$$\partial_u \tilde{F}_1(\mu_0, 0) = \partial_u F_1(\mu_0, 0, 0) + \partial_v F_1(\mu_0, 0, 0) \partial_u \psi(\mu_0, 0) = 0. \tag{16.57}$$

This is known as **Lyapunov–Schmidt reduction**.

Now that we have reduced the problem to a finite-dimensional system, it remains to find conditions such that the finite dimensional system has a nontrivial solution. For simplicity we make the requirement

$$\dim \operatorname{Ker}(A) = \dim \operatorname{Coker}(A) = 1 \tag{16.58}$$

such that we actually have a problem in $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Explicitly, let u_0 span $\operatorname{Ker}(A)$ and let u_1 span X_1. Then we can write

$$\tilde{F}_1(\mu, \lambda u_0) = f(\mu, \lambda) u_1, \tag{16.59}$$

where $f \in C^1(V,\mathbb{R})$ with $V = \{(\mu,\lambda)|(\mu,\lambda u_0) \in U\} \subseteq \mathbb{R}^2$ a neighborhood of $(\mu_0,0)$. Of course we still have $f(\mu,0) = 0$ for $(\mu,0) \in V$ as well as

$$\partial_\lambda f(\mu_0,0)u_1 = \partial_u \tilde{F}_1(\mu_0,0)u_0 = 0. \tag{16.60}$$

It remains to investigate f. To split off the trivial solution it suggests itself to write

$$f(\mu,\lambda) = \lambda\, g(\mu,\lambda) \tag{16.61}$$

We already have

$$g(\mu_0,0) = \partial_\lambda f(\mu_0,0) = 0 \tag{16.62}$$

and hence if

$$0 \neq \partial_\mu g(\mu_0,0) = \partial_\mu\partial_\lambda f(\mu_0,0) \neq 0 \tag{16.63}$$

the implicit function theorem implies existence of a function $\mu(\lambda)$ with $\mu(0) = \mu_0$ and $g(\mu(\lambda),\lambda) = 0$. Moreover, $\mu'(0) = -\frac{\partial_\lambda g(\mu_0,0)}{\partial_\mu g(\mu_0,0)} = -\frac{\partial_\lambda^2 f(\mu_0,0)}{2\partial_\mu\partial_\lambda f(\mu_0,0)}$.

Of course this last condition is a bit problematic since up to this point we only have $f \in C^1$ and hence $g \in C^0$. However, if we change our original assumption to $F \in C^2$ we get $f \in C^2$ and thus $g \in C^1$.

So all we need to do is to trace back our definitions and compute

$$\begin{aligned}\partial_\lambda^2 f(\mu_0,0)u_1 &= \partial_\lambda^2 \tilde{F}_1(\mu,\lambda u_0)\Big|_{\lambda=0} = \partial_\lambda^2 F_1(\mu,\lambda u_0,\psi(\mu,\lambda u_0))\Big|_{\lambda=0}\\ &= \partial_u^2 F_1(\mu_0,0,0)(u_0,u_0) = Q\partial_x^2 F(\mu_0,0)(u_0,u_0)\end{aligned}$$

(recall $\partial_u\psi(\mu_0,0) = 0$) and

$$\begin{aligned}\partial_\mu\partial_\lambda f(\mu_0,0)u_1 &= \partial_\mu\partial_\lambda \tilde{F}_1(\mu,\lambda u_0)\Big|_{\lambda=0,\mu=\mu_0}\\ &= \partial_\mu\partial_\lambda F_1(\mu,\lambda u_0,\psi(\mu,\lambda u_0))\Big|_{\lambda=0,\mu=\mu_0}\\ &= Q\partial_\mu\partial_x F(\mu_0,0)u_0.\end{aligned}$$

Theorem 16.28 (Crandall–Rabinowitz)**.** *Let $F \in C^2(\mathbb{R}\times X,X)$ with $F(\mu,x) = 0$ for all $\mu \in \mathbb{R}$. Suppose that for some $\mu_0 \in \mathbb{R}$ we have that $\partial_x F(\mu_0,0)$ is a Fredholm operator of index zero with a one-dimensional kernel spanned by $u_0 \in X$. Then, if*

$$\partial_\mu\partial_x F(\mu_0,0)u_0 \notin \operatorname{Ran}(\partial_x F(\mu_0,0)) \tag{16.64}$$

there are open neighborhoods $I \subseteq \mathbb{R}$ of 0, $J \subseteq \mathbb{R}$ of μ_0, and $U \subseteq X$ of 0 plus corresponding functions $\mu \in C^1(I,J)$ and $\psi \in C^1(J\times U,X_0)$ such that every nontrivial solution of $F(\mu,x) = 0$ in a neighborhood of $(\mu_0,0)$ is given by

$$x(\lambda) = \lambda u_0 + \psi(\mu(\lambda),\lambda u_0). \tag{16.65}$$

Moreover,

$$\mu(\lambda) = \mu_0 - \frac{\ell_1(\partial_x^2 F(\mu_0,0)(u_0,u_0))}{2\ell_1(\partial_\mu\partial_x F(\mu_0,0)u_0)}\lambda + o(\lambda), \quad x(\lambda) = \lambda u_0 + o(\lambda). \tag{16.66}$$

where ℓ_1 is any nontrivial functional which vanishes on $\operatorname{Ran}(\partial_x F(\mu_0, 0))$.

Note that if $Q\partial_x^2 F(\mu_0, 0)(u_0, u_0) \neq 0$ we could have also solved for λ obtaining a function $\lambda(\mu)$ with $\lambda(\mu_0) = 0$. However, in this case it is not obvious that $\lambda(\mu) \neq 0$ for $\mu \neq \mu_0$, and hence that we get a nontrivial solution, unless we also require $Q\partial_\mu \partial_x F(\mu_0, 0)u_0 \neq 0$ which brings us back to our previous condition. If both conditions are met, then $\mu'(0) \neq 0$ and there is a unique nontrivial solution $x(\mu)$ which crosses the trivial solution non transversally at μ_0. This is known as **transcritical bifurcation**. If $\mu'(0) = 0$ but $\mu''(0) \neq 0$ (assuming this derivative exists), then two solutions will branch off (either for $\mu > \mu_0$ or for $\mu < \mu_0$ depending on the sign of the second derivative). This is known as a **pitchfork bifurcation** and is typical in case of an odd function $F(\mu, -x) = -F(\mu, x)$.

Example 16.25. Now we can establish existence of a stationary solution of the dNLS of the form

$$u_n(t) = \mathrm{e}^{-\mathrm{i}t\omega}\phi_n(\omega)$$

Plugging this ansatz into the dNLS we get the stationary dNLS

$$H\phi - \omega\phi \pm |\phi|^{2p}\phi = 0.$$

Of course we always have the trivial solution $\phi = 0$.

Applying our analysis to

$$F(\omega, \phi) = (H - \omega)\phi \pm |\phi|^{2p}\phi, \qquad p > \frac{1}{2},$$

we have

$$\partial_\phi F(\omega, \phi)u = (H - \omega)u \pm (2p + 1)|\phi|^{2p}u$$

and in particular $\partial_\phi F(\omega, 0) = H - \omega$ and hence ω must be an eigenvalue of H. In fact, if ω_0 is a discrete eigenvalue, then self-adjointness implies that $H - \omega_0$ is Fredholm of index zero. Moreover, if there are two eigenfunction u and v, then one checks that the Wronskian $W(u, v) = u(n)v(n+1) - u(n+1)v(n)$ is constant. But square summability implies that the Wronskian must vanish and hence u and v must be linearly dependent (note that a solution of $Hu = \omega_0 u$ vanishing at two consecutive points must vanish everywhere). Hence eigenvalues are always simple for our Jacobi operator H. Finally, if u_0 is the eigenfunction corresponding to ω_0 we have

$$\partial_\omega \partial_\phi F(\omega_0, 0)u_0 = -u_0 \notin \operatorname{Ran}(H - \omega_0) = \operatorname{Ker}(H - \omega_0)^\perp$$

and the Crandall–Rabinowitz theorem ensures existence of a stationary solution ϕ for ω in a neighborhood of ω_0. Note that

$$\partial_\phi^2 F(\omega, \phi)(u, v) = \pm 2p(2p + 1)|\phi|^{2p-1} \operatorname{sign}(\phi)uv$$

and hence $\partial_\phi^2 F(\omega, 0) = 0$. This is of course not surprising and related to the symmetry $F(\omega, -\phi) = -F(\omega, \phi)$ which implies that zeros branch off in symmetric pairs.

Of course this leaves the problem of finding an discrete eigenvalue open. One can show that for the free operator H_0 (with $q = 0$) the spectrum is $\sigma(H_0) = [-2, 2]$ and that there are no eigenvalues (in fact, the discrete Fourier transform will map H_0 to a multiplication operator in $L^2[-\pi, \pi]$). If $q \in c_0(\mathbb{Z})$, then the corresponding multiplication operator is compact and $\sigma_{ess}(H) = \sigma(H_0)$ by Weyl's theorem (Theorem 6.36). Hence every point in $\sigma(H) \setminus [-2, 2]$ will be an isolated eigenvalue. ⋄

Problem 16.12. *Show that if* $F(\mu, -x) = -F(\mu, x)$, *then* $\psi(\mu, -u) = -\psi(\mu, u)$ *and* $\mu(-\lambda) = \mu(\lambda)$.

Chapter 17

The Brouwer mapping degree

17.1. Introduction

Many applications lead to the problem of finding zeros of a mapping $f : U \subseteq X \to X$, where X is some (real) Banach space. That is, we are interested in the solutions of

$$f(x) = 0, \quad x \in U. \tag{17.1}$$

In most cases it turns out that this is too much to ask for, since determining the zeros analytically is in general impossible.

Hence one has to ask some weaker questions and hope to find answers for them. One such question would be "Are there any solutions, respectively, how many are there?". Luckily, these questions allow some progress.

To see how, lets consider the case $f \in \mathcal{H}(\mathbb{C})$, where $\mathcal{H}(U)$ denotes the set of **holomorphic functions** on a domain $U \subset \mathbb{C}$. Recall the concept of the **winding number** from complex analysis. The winding number of a path $\gamma : [0, 1] \to \mathbb{C} \setminus \{z_0\}$ around a point $z_0 \in \mathbb{C}$ is defined by

$$n(\gamma, z_0) := \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{dz}{z - z_0} \in \mathbb{Z}. \tag{17.2}$$

It gives the number of times γ encircles z_0 taking orientation into account. That is, encirclings in opposite directions are counted with opposite signs.

In particular, if we pick $f \in \mathcal{H}(\mathbb{C})$ one computes (assuming $0 \notin f(\gamma)$)

$$n(f(\gamma), 0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{f'(z)}{f(z)} dz = \sum_k n(\gamma, z_k)\alpha_k, \tag{17.3}$$

where z_k denote the zeros of f and α_k their respective multiplicity. Moreover, if γ is a Jordan curve encircling a simply connected domain $U \subset \mathbb{C}$, then $n(\gamma, z_k) = 0$ if $z_k \notin \overline{U}$ and $n(\gamma, z_k) = 1$ if $z_k \in U$. Hence $n(f(\gamma), 0)$ counts the number of zeros inside U.

However, this result is useless unless we have an efficient way of computing $n(f(\gamma), 0)$ (which does not involve the knowledge of the zeros z_k). This is our next task.

Now, lets recall how one would compute complex integrals along complicated paths. Clearly, one would use homotopy invariance and look for a simpler path along which the integral can be computed and which is homotopic to the original one. In particular, if $f : \gamma \to \mathbb{C}\setminus\{0\}$ and $g : \gamma \to \mathbb{C}\setminus\{0\}$ are homotopic, we have $n(f(\gamma), 0) = n(g(\gamma), 0)$ (which is known as Rouché's theorem).

More explicitly, we need to find a mapping g for which $n(g(\gamma), 0)$ can be computed and a **homotopy** $H : [0, 1]\times\gamma \to \mathbb{C}\setminus\{0\}$ such that $H(0, z) = f(z)$ and $H(1, z) = g(z)$ for $z \in \gamma$. For example, how many zeros of $f(z) = \frac{1}{2}z^6 + z - \frac{1}{3}$ lie inside the unit circle? Consider $g(z) = z$, then $H(t, z) = (1 - t)f(z) + t\, g(z)$ is the required homotopy since $|f(z) - g(z)| < |g(z)|$, $|z| = 1$, implying $H(t, z) \neq 0$ on $[0, 1] \times \gamma$. Hence $f(z)$ has one zero inside the unit circle.

Summarizing, given a (sufficiently smooth) domain U with enclosing Jordan curve ∂U, we have defined a degree $\deg(f, U, z_0) = n(f(\partial U), z_0) = n(f(\partial U) - z_0, 0) \in \mathbb{Z}$ which counts the number of solutions of $f(z) = z_0$ inside U. The invariance of this degree with respect to certain deformations of f allowed us to explicitly compute $\deg(f, U, z_0)$ even in nontrivial cases.

Our ultimate goal is to extend this approach to continuous functions $f : \mathbb{R}^n \to \mathbb{R}^n$. However, such a generalization runs into several problems. First of all, it is unclear how one should define the multiplicity of a zero. But even more severe is the fact, that the number of zeros is unstable with respect to small perturbations. For example, consider $f_\varepsilon : [-1, 2] \to \mathbb{R}$, $x \mapsto x^2 - \varepsilon$. Then f_ε has no zeros for $\varepsilon < 0$, one zero for $\varepsilon = 0$, two zeros for $0 < \varepsilon \leq 1$, one for $1 < \varepsilon \leq \sqrt{2}$, and none for $\varepsilon > \sqrt{2}$. This shows the following facts.

(i) Zeros with $f' \neq 0$ are stable under small perturbations.

(ii) The number of zeros can change if two zeros with opposite sign change (i.e., opposite signs of f') run into each other.

(iii) The number of zeros can change if a zero drops over the boundary.

Hence we see that we cannot expect too much from our degree. In addition, since it is unclear how it should be defined, we will first require some basic

properties a degree should have and then we will look for functions satisfying these properties.

17.2. Definition of the mapping degree and the determinant formula

To begin with, let us introduce some useful notation. Throughout this section U will be a bounded open subset of $\mathbb{R}^n$. For $f \in C^1(U, \mathbb{R}^n)$ the Jacobi matrix of f at $x \in U$ is $df(x) = (\partial_{x_i} f_j(x))_{1 \le i,j \le n}$ and the Jacobi determinant of f at $x \in U$ is

$$J_f(x) := \det df(x). \tag{17.4}$$

The set of **regular values** is

$$\mathrm{RV}(f) := \{y \in \mathbb{R}^n | \forall x \in f^{-1}(y) : J_f(x) \neq 0\}. \tag{17.5}$$

Its complement $\mathrm{CV}(f) := \mathbb{R}^n \setminus \mathrm{RV}(f)$ is called the set of **critical values**. We will also need the spaces

$$\bar{C}^k(U, \mathbb{R}^n) := C^k(U, \mathbb{R}^n) \cap C(\overline{U}, \mathbb{R}^n) \tag{17.6}$$

and regard them as subspaces of the Banach space $C(\overline{U}, \mathbb{R}^n)$ (cf. Section 1.8 or 16.1). Note that $\bar{C}^\infty(U, \mathbb{R}^n)$ is dense in $C(\overline{U}, \mathbb{R}^n)$. To see this you can either apply Stone–Weierstraß or use the Tietze extension theorem to extend $f \in C(\overline{U}, \mathbb{R}^n)$ to all of $\mathbb{R}^n$ and then mollify. If you use mollification and $f \in \bar{C}^k(U, \mathbb{R}^n)$ then all derivatives up to order k will converge uniformly on compact subsets of U. Finally , for $y \in \mathbb{R}^n$ we set

$$\bar{C}^k_y(\overline{U}, \mathbb{R}^n) := \{f \in \bar{C}^k(U, \mathbb{R}^n) | y \notin f(\partial U)\} \tag{17.7}$$

and $\bar{C}_y(U, \mathbb{R}^n) := \bar{C}^0_y(U, \mathbb{R}^n)$.

Note that, since U is bounded, ∂U is compact and so is $f(\partial U)$ if $f \in C(\overline{U}, \mathbb{R}^n)$. In particular,

$$\mathrm{dist}(y, f(\partial U)) = \min_{x \in \partial U} |y - f(x)| \tag{17.8}$$

is positive for $f \in \bar{C}_y(U, \mathbb{R}^n)$ and thus $\bar{C}_y(U, \mathbb{R}^n)$ is an open subset of $C(\overline{U}, \mathbb{R}^n)$.

Now that these things are out of the way, we come to the formulation of the requirements for our degree.

A function deg which assigns each $f \in \bar{C}_y(U, \mathbb{R}^n)$, $y \in \mathbb{R}^n$, a real number $\deg(f, U, y)$ will be called degree if it satisfies the following conditions.

(D1). $\deg(f, U, y) = \deg(f - y, U, 0)$ (*translation invariance*).

(D2). $\deg(\mathbb{I}, U, y) = 1$ if $y \in U$ (*normalization*).

(D3). If $U_{1,2}$ are open, disjoint subsets of U such that $y \notin f(\overline{U} \setminus (U_1 \cup U_2))$, then $\deg(f, U, y) = \deg(f, U_1, y) + \deg(f, U_2, y)$ (*additivity*).

(D4). If $H(t) = (1-t)f + tg \in \bar{C}_y(U, \mathbb{R}^n)$, $t \in [0,1]$, then $\deg(f, U, y) = \deg(g, U, y)$ (*homotopy invariance*).

Before we draw some first conclusions form this definition, let us discuss the properties (D1)–(D4) first. (D1) is natural since $\deg(f, U, y)$ should have something to do with the solutions of $f(x) = y$, $x \in U$, which is the same as the solutions of $f(x) - y = 0$, $x \in U$. $(D2)$ is a normalization since any multiple of deg would also satisfy the other requirements. (D3) is also quite natural since it requires deg to be additive with respect to components. In addition, it implies that sets where $f \neq y$ do not contribute. (D4) is not that natural since it already rules out the case where deg is the cardinality of $f^{-1}(\{y\})$. On the other hand it will give us the ability to compute $\deg(f, U, y)$ in several cases.

Theorem 17.1. *Suppose* deg *satisfies (D1)–(D4) and let* $f, g \in \bar{C}_y(U, \mathbb{R}^n)$, *then the following statements hold.*

(i). *We have* $\deg(f, \emptyset, y) = 0$. *Moreover, if* U_i, $1 \leq i \leq N$, *are disjoint open subsets of* U *such that* $y \notin f(\overline{U} \setminus \bigcup_{i=1}^N U_i)$, *then* $\deg(f, U, y) = \sum_{i=1}^N \deg(f, U_i, y)$.

(ii). *If* $y \notin f(U)$, *then* $\deg(f, U, y) = 0$ *(but not the other way round). Equivalently, if* $\deg(f, U, y) \neq 0$, *then* $y \in f(U)$.

(iii). *If* $|f(x) - g(x)| < |f(x) - y|$, $x \in \partial U$, *then* $\deg(f, U, y) = \deg(g, U, y)$. *In particular, this is true if* $f(x) = g(x)$ *for* $x \in \partial U$.

Proof. For the first part of (i) use (D3) with $U_1 = U$ and $U_2 = \emptyset$. For the second part use $U_2 = \emptyset$ in (D3) if $i = 1$ and the rest follows from induction. For (ii) use $i = 1$ and $U_1 = \emptyset$ in (ii). For (iii) note that $H(t, x) = (1-t)f(x) + t\, g(x)$ satisfies $|H(t, x) - y| \geq \operatorname{dist}(y, f(\partial U)) - |f(x) - g(x)|$ for x on the boundary. □

Item (iii) is a version of **Rouché's theorem** for our degree. Next we show that (D4) implies several at first sight stronger looking facts.

Theorem 17.2. *We have that* $\deg(., U, y)$ *and* $\deg(f, U, .)$ *are both continuous. In fact, we even have*

(i). $\deg(., U, y)$ *is constant on each component of* $\bar{C}_y(U, \mathbb{R}^n)$.

(ii). $\deg(f, U, .)$ *is constant on each component of* $\mathbb{R}^n \setminus f(\partial U)$.

Moreover, if $H : [0,1] \times \overline{U} \to \mathbb{R}^n$ *and* $y : [0,1] \to \mathbb{R}^n$ *are both continuous such that* $H(t) \in \bar{C}_{y(t)}(U, \mathbb{R}^n)$, $t \in [0,1]$, *then* $\deg(H(0), U, y(0)) = \deg(H(1), U, y(1))$.

Proof. For (i) let C be a component of $\bar{C}_y(U, \mathbb{R}^n)$ and let $d_0 \in \deg(C, U, y)$. It suffices to show that $\deg(., U, y)$ is locally constant. But if $|g - f| <$

$\operatorname{dist}(y, f(\partial U))$, then $\deg(f, U, y) = \deg(g, U, y)$ by (D4) since $|H(t) - y| \geq |f - y| - |g - f| > 0$, $H(t) = (1-t)f + t\,g$. The proof of (ii) is similar. For the remaining part observe, that if $H : [0,1] \times \overline{U} \to \mathbb{R}^n$, $(t, x) \mapsto H(t, x)$, is continuous, then so is $H : [0,1] \to C(\overline{U}, \mathbb{R}^n)$, $t \mapsto H(t)$, since $\overline{U}$ is compact. Hence, if in addition $H(t) \in \bar{C}_y(U, \mathbb{R}^n)$, then $\deg(H(t), U, y)$ is independent of t and if $y = y(t)$ we can use $\deg(H(0), U, y(0)) = \deg(H(t) - y(t), U, 0) = \deg(H(1), U, y(1))$. □

In this context note that a Banach space X is locally path-connected and hence the components of any open subset are open (in the topology of X) and path-connected (see Lemma B.32 (vi)).

Moreover, note that this result also shows why $\deg(f, U, y)$ cannot be defined meaningful for $y \in f(\partial D)$. Indeed, approaching y from within different components of $\mathbb{R}^n \setminus f(\partial U)$ will result in different limits in general!

Now let us try to compute deg using its properties. If you are not interested in how to derive the determinant formula for the degree from its properties you can of course take it as a definition and skip to the next section.

Let's start with a simple case and suppose $f \in \bar{C}^1_y(U, \mathbb{R}^n)$ and $y \notin \operatorname{CV}(f)$. Without restriction we consider $y = 0$. In addition, we avoid the trivial case $f^{-1}(\{0\}) = \emptyset$. Since the points of $f^{-1}(\{0\})$ inside U are isolated (use $J_f(x) \neq 0$ and the inverse function theorem) they can only cluster at the boundary ∂U. But this is also impossible since f would equal 0 at the limit point on the boundary by continuity. Hence $f^{-1}(\{0\}) = \{x^i\}_{i=1}^N$. Picking sufficiently small neighborhoods $U(x^i)$ around x^i we consequently get

$$\deg(f, U, 0) = \sum_{i=1}^{N} \deg(f, U(x^i), 0). \tag{17.9}$$

It suffices to consider one of the zeros, say x^1. Moreover, we can even assume $x^1 = 0$ and $U(x^1) = B_\delta(0)$. Next we replace f by its linear approximation around 0. By the definition of the derivative we have

$$f(x) = df(0)x + |x| r(x), \qquad r \in C(B_\delta(0), \mathbb{R}^n), \quad r(0) = 0. \tag{17.10}$$

Now consider the homotopy $H(t, x) = df(0)x + (1-t)|x| r(x)$. In order to conclude $\deg(f, B_\delta(0), 0) = \deg(df(0), B_\delta(0), 0)$ we need to show $0 \notin H(t, \partial B_\delta(0))$. Since $J_f(0) \neq 0$ we can find a constant λ such that $|df(0)x| \geq \lambda |x|$ and since $r(0) = 0$ we can decrease δ such that $|r| < \lambda$. This implies $|H(t, x)| \geq ||df(0)x| - (1-t)|x||r(x)|| \geq \lambda\delta - \delta|r| > 0$ for $x \in \partial B_\delta(0)$ as desired.

In summary we have

$$\deg(f, U, 0) = \sum_{i=1}^{N} \deg(df(x^i), B_\delta(0), 0) \tag{17.11}$$

and it remains to compute the degree of a nonsingular matrix. To this end we need the following lemma.

Lemma 17.3. *Two nonsingular matrices $M_{1,2} \in \mathrm{GL}(n)$ are homotopic in $\mathrm{GL}(n)$ if and only if* $\operatorname{sign}\det M_1 = \operatorname{sign}\det M_2$.

Proof. We will show that any given nonsingular matrix M is homotopic to $\operatorname{diag}(\operatorname{sign}\det M, 1, \dots, 1)$, where $\operatorname{diag}(m_1, \dots, m_n)$ denotes a diagonal matrix with diagonal entries m_i.

In fact, note that adding one row to another and multiplying a row by a positive constant can be realized by continuous deformations such that all intermediate matrices are nonsingular. Hence we can reduce M to a diagonal matrix $\operatorname{diag}(m_1, \dots, m_n)$ with $(m_i)^2 = 1$. Next,

$$\begin{pmatrix} \pm\cos(\pi t) & \mp\sin(\pi t) \\ \sin(\pi t) & \cos(\pi t) \end{pmatrix},$$

shows that $\operatorname{diag}(\pm 1, 1)$ and $\operatorname{diag}(\mp 1, -1)$ are homotopic. Now we apply this result to all two by two subblocks as follows. For each i starting from n and going down to 2 transform the subblock $\operatorname{diag}(m_{i-1}, m_i)$ into $\operatorname{diag}(1, 1)$ respectively $\operatorname{diag}(-1, 1)$. The result is the desired form for M.

To conclude the proof note that a continuous deformation within $\mathrm{GL}(n)$ cannot change the sign of the determinant since otherwise the determinant would have to vanish somewhere in between (i.e., we would leave $\mathrm{GL}(n)$). □

Using this lemma we can now show the main result of this section.

Theorem 17.4. *Suppose $f \in \bar{C}^1_y(U, \mathbb{R}^n)$ and $y \notin \mathrm{CV}(f)$, then a degree satisfying (D1)–(D4) satisfies*

$$\deg(f, U, y) = \sum_{x \in f^{-1}(\{y\})} \operatorname{sign}\, J_f(x), \tag{17.12}$$

where the sum is finite and we agree to set $\sum_{x\in\emptyset} = 0$.

Proof. By the previous lemma we obtain

$$\deg(df(0), B_\delta(0), 0) = \deg(\operatorname{diag}(\operatorname{sign} J_f(0), 1, \dots, 1), B_\delta(0), 0)$$

since $\det M \neq 0$ is equivalent to $Mx \neq 0$ for $x \in \partial B_\delta(0)$. Hence it remains to show $\deg(M_\pm, B_\delta(0), 0) = \pm 1$, where $M_\pm := \operatorname{diag}(\pm 1, 1, \dots, 1)$. For M_+ this is true by (D2) and for M_- we note that we can replace $B_\delta(0)$ by any neighborhood U of 0. Now abbreviate $U_1 := \{x \in \mathbb{R}^n \,||x_i| < 1, 1 \le i \le n\}$, $U_2 :=$

$\{x \in \mathbb{R}^n | 1 < x_1 < 3, |x_i| < 1, 2 \le i \le n\}$, $U := \{x \in \mathbb{R}^n | -1 < x_1 < 3, |x_i| < 1, 2 \le i \le n\}$, and $g(r) = 2 - |r-1|$, $h(r) = 1 - r^2$. Consider the two functions $f_1(x) = (1 - g(x_1)h(x_2)\cdots h(x_n), x_2, \ldots, x_n)$ and $f_2(x) = (1, x_2, \ldots, x_n)$. Clearly $f_1^{-1}(\{0\}) = \{x^1, x^2\}$ with $x^1 = 0$, $x^2 = (2, 0, \ldots, 0)$ and $f_2^{-1}(0) = \emptyset$. Since $f_1(x) = f_2(x)$ for $x \in \partial U$ we infer $\deg(f_1, U, 0) = \deg(f_2, U, 0) = 0$. Moreover, we have $\deg(f_1, U, 0) = \deg(f_1, U_1, 0) + \deg(f_1, U_2, 0)$ and hence $\deg(M_-, U_1, 0) = \deg(df_1(x^1), U_1, 0) = \deg(f_1, U_1, 0) = -\deg(f_1, U_2, 0) = -\deg(df_1(x^2), U_1, 0) = -\deg(\mathbb{I}, U_1, 0) = -1$ as claimed. □

Up to this point we have only shown that a degree (provided there is one at all) necessarily satisfies (17.12). Once we have shown that regular values are dense, it will follow that the degree is uniquely determined by (17.12) since the remaining values follow from point (iii) of Theorem 17.1. On the other hand, we don't even know whether a degree exists since it is unclear whether (17.12) satisfies (D4). Hence we need to show that (17.12) can be extended to $f \in \bar{C}_y(U, \mathbb{R}^n)$ and that this extension satisfies our requirements (D1)–(D4).

17.3. Extension of the determinant formula

Our present objective is to show that the determinant formula (17.12) can be extended to all $f \in \bar{C}_y(U, \mathbb{R}^n)$. As a preparation we prove that the set of regular values is dense as a warm up. This is a consequence of a special case of Sard's theorem which says that $\mathrm{CV}(f)$ has zero measure.

Lemma 17.5 (Sard)**.** *Suppose $f \in C^1(U, \mathbb{R}^n)$, then the Lebesgue measure of $\mathrm{CV}(f)$ is zero.*

Proof. Since the claim is easy for linear mappings our strategy is as follows. We divide U into sufficiently small subsets. Then we replace f by its linear approximation in each subset and estimate the error.

Let $\mathrm{CP}(f) := \{x \in U | J_f(x) = 0\}$ be the set of critical points of f. We first pass to cubes which are easier to divide. Let $\{Q_i\}_{i \in \mathbb{N}}$ be a countable cover for U consisting of open cubes such that $\overline{Q_i} \subset U$. Then it suffices to prove that $f(\mathrm{CP}(f) \cap Q_i)$ has zero measure since $\mathrm{CV}(f) = f(\mathrm{CP}(f)) = \bigcup_i f(\mathrm{CP}(f) \cap Q_i)$ (the Q_i's are a cover).

Let Q be anyone of these cubes and denote by ρ the length of its edges. Fix $\varepsilon > 0$ and divide Q into N^n cubes Q_i of length ρ/N. These cubes don't have to be open and hence we can assume that they cover Q. Since $df(x)$ is uniformly continuous on $\overline{Q}$ we can find an N (independent of i) such that

$$|f(x) - f(\tilde{x}) - df(\tilde{x})(x - \tilde{x})| \le \int_0^1 |df(\tilde{x} + t(x - \tilde{x})) - df(\tilde{x})||\tilde{x} - x| dt \le \frac{\varepsilon\rho}{N} \tag{17.13}$$

for $\tilde{x}, x \in Q_i$. Now pick a Q_i which contains a critical point $\tilde{x}_i \in \mathrm{CP}(f)$. Without restriction we assume $\tilde{x}_i = 0$, $f(\tilde{x}_i) = 0$ and set $M := df(\tilde{x}_i)$. By $\det M = 0$ there is an orthonormal basis $\{b^i\}_{1 \le i \le n}$ of $\mathbb{R}^n$ such that b^n is orthogonal to the image of M. In addition,

$$Q_i \subseteq \{\sum_{i=1}^{n} \lambda_i b^i | \sqrt{\sum_{i=1}^{n} |\lambda_i|^2} \le \sqrt{n}\frac{\rho}{N}\}$$

and hence there is a constant (again independent of i) such that

$$MQ_i \subseteq \{\sum_{i=1}^{n-1} \lambda_i b^i | \, |\lambda_i| \le C\sqrt{n}\frac{\rho}{N}\}$$

(e.g., $C := \max_{x \in \overline{Q}} |df(x)|$). Next, by our estimate (17.13) we even have

$$f(Q_i) \subseteq \{\sum_{i=1}^{n} \lambda_i b^i | \, |\lambda_i| \le (C+\varepsilon)\sqrt{n}\frac{\rho}{N}, |\lambda_n| \le \varepsilon\sqrt{n}\frac{\rho}{N}\}$$

and hence the measure of $f(Q_i)$ is smaller than $\frac{\tilde{C}\varepsilon}{N^n}$. Since there are at most N^n such Q_i's, we see that the measure of $f(\mathrm{CP}(f) \cap Q)$ is smaller than $\tilde{C}\varepsilon$. □

By (ii) of Theorem 17.2, $\deg(f, U, y)$ should be constant on each component of $\mathbb{R}^n \setminus f(\partial U)$. Unfortunately, if we connect y and a nearby regular value $\tilde{y}$ by a path, then there might be some critical values in between.

Example 17.1. The function $f(x) := x^2 \sin(\frac{\pi}{2x})$ is in $\bar{C}^1_0([-1,1], \mathbb{R})$. It has 0 as a critical value and the critical values accumulate at 0. ◇

To overcome this problem we need a definition for deg which works for critical values as well. Let us try to look for an integral representation. Formally (17.12) can be written as $\deg(f, U, y) = \int_U \delta_y(f(x)) J_f(x) d^n x$, where $\delta_y(.)$ is the Dirac distribution at y. But since we don't want to mess with distributions, we replace $\delta_y(.)$ by $\phi_\varepsilon(. - y)$, where $\{\phi_\varepsilon\}_{\varepsilon > 0}$ is a family of functions such that ϕ_ε is supported on the ball $B_\varepsilon(0)$ of radius ε around 0 and satisfies $\int_{\mathbb{R}^n} \phi_\varepsilon(x) d^n x = 1$.

Lemma 17.6 (Heinz)**.** *Let $f \in \bar{C}^1_y(U, \mathbb{R}^n)$, $y \notin \mathrm{CV}(f)$. Then*

$$\deg(f, U, y) = \int_U \phi_\varepsilon(f(x) - y) J_f(x) d^n x \tag{17.14}$$

for all positive ε smaller than a certain ε_0 depending on f and y. Moreover, $\mathrm{supp}(\phi_\varepsilon(f(.) - y)) \subset U$ for $\varepsilon < \mathrm{dist}(y, f(\partial U))$.

Proof. If $f^{-1}(\{y\}) = \emptyset$, we can set $\varepsilon_0 = \mathrm{dist}(y, f(\overline{U}))$, implying $\phi_\varepsilon(f(x) - y) = 0$ for $x \in \overline{U}$.

If $f^{-1}(\{y\}) = \{x^i\}_{1\le i\le N}$, we can find an $\varepsilon_0 > 0$ such that $f^{-1}(B_{\varepsilon_0}(y))$ is a union of disjoint neighborhoods $U(x^i)$ of x^i by the inverse function theorem. Moreover, after possibly decreasing ε_0 we can assume that $f|_{U(x^i)}$ is a bijection and that $J_f(x)$ is nonzero on $U(x^i)$. Again $\phi_\varepsilon(f(x)-y) = 0$ for $x \in \overline{U} \setminus \bigcup_{i=1}^N U(x^i)$ and hence

$$\begin{aligned}\int_U \phi_\varepsilon(f(x)-y)J_f(x)d^n x &= \sum_{i=1}^N \int_{U(x^i)} \phi_\varepsilon(f(x)-y)J_f(x)d^n x \\ &= \sum_{i=1}^N \operatorname{sign}(J_f(x^i)) \int_{B_{\varepsilon_0}(0)} \phi_\varepsilon(\tilde{x})d^n\tilde{x} = \deg(f,U,y),\end{aligned}$$

where we have used the change of variables $\tilde{x} = f(x) - y$ in the second step. □

Our new integral representation makes sense even for critical values. But since ε_0 depends on f and y, continuity is not clear. This will be tackled next.

The key idea is to rewrite $\deg(f,U,y^2) - \deg(f,U,y^1)$ as an integral over a divergence supported in U and then apply the Gauss–Green theorem. For this purpose the following result will be used.

Lemma 17.7. *Suppose $f \in C^2(U,\mathbb{R}^n)$ and $u \in C^1(\mathbb{R}^n,\mathbb{R}^n)$, then*

$$(\operatorname{div} u)(f)J_f = \operatorname{div} D_f(u), \tag{17.15}$$

where $D_f(u)_j$ is the determinant of the matrix obtained from df by replacing the j-th column by $u(f)$.

Proof. We compute

$$\operatorname{div} D_f(u) = \sum_{j=1}^n \partial_{x_j} D_f(u)_j = \sum_{j,k=1}^n D_f(u)_{j,k},$$

where $D_f(u)_{j,k}$ is the determinant of the matrix obtained from the matrix associated with $D_f(u)_j$ by applying ∂_{x_j} to the k-th column. Since $\partial_{x_j}\partial_{x_k} f = \partial_{x_k}\partial_{x_j} f$ we infer $D_f(u)_{j,k} = -D_f(u)_{k,j}$, $j \neq k$, by exchanging the k-th and the j-th column. Hence

$$\operatorname{div} D_f(u) = \sum_{i=1}^n D_f(u)_{i,i}.$$

Now let $J_f^{(i,j)}(x)$ denote the (i,j) cofactor of $df(x)$ and recall the cofactor expansion of the determinant $\sum_{i=1}^n J_f^{(i,j)}\partial_{x_i} f_k = \delta_{j,k} J_f$. Using this to expand

the determinant $D_f(u)_{i,i}$ along the i-th column shows

$$\begin{aligned}\operatorname{div} D_f(u) &= \sum_{i,j=1}^n J_f^{(i,j)} \partial_{x_i} u_j(f) = \sum_{i,j=1}^n J_f^{(i,j)} \sum_{k=1}^n (\partial_{x_k} u_j)(f) \partial_{x_i} f_k \\ &= \sum_{j,k=1}^n (\partial_{x_k} u_j)(f) \sum_{i=1}^n J_f^{(i,j)} \partial_{x_i} f_k = \sum_{j=1}^n (\partial_{x_j} u_j)(f) J_f\end{aligned}$$

as required. □

Now we can prove

Theorem 17.8. *There is a unique degree* deg *satisfying (D1)–(D4). Moreover,* $\deg(.,U,y) : \bar{C}_y(U,\mathbb{R}^n) \to \mathbb{Z}$ *is constant on each component and given* $f \in \bar{C}_y(U,\mathbb{R}^n)$ *we have*

$$\deg(f,U,y) = \sum_{x \in \tilde{f}^{-1}(y)} \operatorname{sign} J_{\tilde{f}}(x) \tag{17.16}$$

where $\tilde{f} \in \bar{C}^1_y(U,\mathbb{R}^n)$ *is in the same component of* $\bar{C}_y(U,\mathbb{R}^n)$*, say* $|f-\tilde{f}| < \operatorname{dist}(y, f(\partial U))$*, such that* $y \in \operatorname{RV}(\tilde{f})$*.*

Proof. We will first show that our integral formula works in fact for all $\varepsilon < \rho := \operatorname{dist}(y, f(\partial U))$. For this we will make some additional assumptions: Let $f \in \bar{C}^2(U,\mathbb{R}^n)$ and choose a family of functions $\phi_\varepsilon \in C^\infty((0,\infty))$ with $\operatorname{supp}(\phi_\varepsilon) \subset (0,\varepsilon)$ such that $S_n \int_0^\varepsilon \phi(r) r^{n-1} dr = 1$. Consider

$$I_\varepsilon(f,U,y) := \int_U \phi_\varepsilon(|f(x)-y|) J_f(x) d^n x.$$

Then $I := I_{\varepsilon_1} - I_{\varepsilon_2}$ will be of the same form but with ϕ_ε replaced by $\varphi := \phi_{\varepsilon_1} - \phi_{\varepsilon_2}$, where $\varphi \in C^\infty((0,\infty))$ with $\operatorname{supp}(\varphi) \subset (0,\rho)$ and $\int_0^\rho \varphi(r) r^{n-1} dr = 0$. To show that $I = 0$ we will use our previous lemma with u chosen such that $\operatorname{div}(u(x)) = \varphi(|x|)$. To this end we make the ansatz $u(x) = \psi(|x|)x$ such that $\operatorname{div}(u(x)) = |x|\psi'(|x|) + n\psi(|x|)$. Our requirement now leads to an ordinary differential equation whose solution is

$$\psi(r) = \frac{1}{r^n} \int_0^r s^{n-1} \varphi(s) ds.$$

Moreover, one checks $\psi \in C^\infty((0,\infty))$ with $\operatorname{supp}(\psi) \subset (0,\rho)$. Thus our lemma shows

$$I = \int_U \operatorname{div} D_{f-y}(u) d^n x$$

and since the integrand vanishes in a neighborhood of ∂U we can extend it to all of $\mathbb{R}^n$ by setting it zero outside U and choose a cube $Q \supset U$. Then elementary coordinatewise integration gives $I = \int_Q \operatorname{div} D_{f-y}(u) d^n x = 0$.

Now fix $\delta < \rho$ and look at $I_\varepsilon(f+g,U,y)$ for $g \in B_\delta(f) \cap \bar{C}^2(U,\mathbb{R}^n) \subset C(\overline{U},\mathbb{R}^n)$ and $\varepsilon < \rho - \delta < \operatorname{dist}((f+g)(\partial U), y)$ fixed. Then $g \mapsto I_\varepsilon(f+g,U,y)$ is continuous and it is integer valued (since it is equal to our determinant formula) one a dense set. Consequently it must be constant and we can extend it to a function $\bar{I}_\varepsilon$ on all of $B_\delta(f)$. Note that by mollifying $f \in \bar{C}^1(U,\mathbb{R}^n)$ we get a sequence of smooth functions for which both f and df converge uniformly on compact subsets of U and hence I_ε converges, such that for such f we still have $\bar{I}_\varepsilon(f,U,y) = I_\varepsilon(f,U,y)$.

Now we set

$$\deg(f,U,y) := I_\varepsilon(\tilde{f},U,y),$$

where $\tilde{f} \in \bar{C}^1(U,\mathbb{R}^n)$ with $\varepsilon < \rho$ and $|\tilde{f} - f| < \rho - \varepsilon$. Then (D1) holds since it holds for I_ε, (D2) holds since I_ε extends the determinant formula, (D3) holds since the integrand of I_ε vanishes on $\overline{U} \setminus (U_1 \cup U_2)$, and (D4) holds since we can choose $\varepsilon < \max_{t\in[0,1]} \operatorname{dist}(H(t)(\partial U), y)$ such that $I_\varepsilon(H(t),U,y)$ is continuous and hence constant for $t \in [0,1]$. □

To conclude this section, let us give a few simple examples illustrating the use of the Brouwer degree.

Example 17.2. First, let's investigate the zeros of

$$f(x_1,x_2) := (x_1 - 2x_2 + \cos(x_1+x_2), x_2 + 2x_1 + \sin(x_1+x_2)).$$

Denote the linear part by

$$g(x_1,x_2) := (x_1 - 2x_2, x_2 + 2x_1).$$

Then we have $|g(x)| = \sqrt{5}|x|$ and $|f(x) - g(x)| = 1$ and hence $h(t) = (1-t)g + t\,f = g + t(f-g)$ satisfies $|h(t)| \geq |g| - t|f-g| > 0$ for $|x| > 1/\sqrt{5}$ implying

$$\deg(f, B_r(0), 0) = \deg(g, B_r(0), 0) = 1, \qquad r > 1/\sqrt{5}.$$

Moreover, since $J_f(x) = 5 + 3\cos(x_1+x_2) + \sin(x_1+x_2) > 1$ the determinant formula (17.12) for the degree implies that $f(x) = 0$ has a unique solution in $\mathbb{R}^2$. This solution even has to lie on the circle $|x| = 1/\sqrt{5}$ since $f(x) = 0$ implies $1 = |f(x) - g(x)| = |g(x)| = \sqrt{5}|x|$. ⋄

Next let us prove the following result which implies the **hairy ball (or hedgehog) theorem**.

Theorem 17.9. *Suppose U is open, bounded and contains the origin and let $f : \partial U \to \mathbb{R}^n \setminus \{0\}$ be continuous. If n is odd, then there exists a $x \in \partial U$ and a $\lambda \neq 0$ such that $f(x) = \lambda x$.*

Proof. By Theorem 18.10 we can assume $f \in C(\overline{U},\mathbb{R}^n)$ and since n is odd we have $\deg(-\mathbb{I},U,0) = -1$. Now if $\deg(f,U,0) \neq -1$, then $H(t,x) = (1-t)f(x) - tx$ must have a zero $(t_0,x_0) \in (0,1) \times \partial U$ and hence $f(x_0) =$

$\frac{t_0}{1-t_0}x_0$. Otherwise, if $\deg(f,U,0)=-1$ we can apply the same argument to $H(t,x)=(1-t)f(x)+tx$. □

In particular, this result implies that a continuous tangent vector field on the unit sphere $f: S^{n-1}\to\mathbb{R}^n$ (with $f(x)x=0$ for all $x\in S^{n-1}$) must vanish somewhere if n is odd. Or, for $n=3$, you cannot smoothly comb a hedgehog without leaving a bald spot or making a parting. It is however possible to comb the hair smoothly on a torus and that is why the magnetic containers in nuclear fusion are toroidal.

Example 17.3. The result fails in even dimensions as the example $n=2$, $U=B_1(0)$, $f(x_1,x_2)=(-x_2,x_1)$ shows. ⋄

Another simple consequence is the fact that a vector field on $\mathbb{R}^n$, which points outwards (or inwards) on a sphere, must vanish somewhere inside the sphere (Problem 17.2).

One more useful observation is that odd functions have odd degree:

Theorem 17.10 (Borsuk)**.** *Let $0\in U\subseteq\mathbb{R}^n$ be open, bounded and symmetric with respect to the origin (i.e, $U=-U$). Let $f\in\bar C_0(U)$ be odd (i.e., $f(-x)=-f(x)$). Then* $\deg(f,U,0)$ *is odd.*

Proof. If $f\in\bar C^1_0(U)$ and $0\in\mathrm{RV}(f)$, then the claim is straightforward since

$$\deg(f,U,0)=\operatorname{sign}J_f(0)+\sum_{x\in f^{-1}(0)\setminus\{0\}}\operatorname{sign}J_f(x),$$

where the sum is even since for every $x\in f^{-1}(0)\setminus\{0\}$ we also have $-x\in f^{-1}(0)\setminus\{0\}$ as well as $J_f(x)=J_f(-x)$.

Hence we need to reduce the general case to this one. Clearly if $f\in\bar C_0(U)$ we can choose an approximating $f_0\in\bar C^1_0(U)$ and replacing f_0 by its odd part $\frac12(f_0(x)-f_0(-x))$ we can assume f_0 to be odd. Moreover, if $J_{f_0}(0)=0$ we can replace f_0 by $f_0(x)+\delta x$ such that 0 is regular. However, if we choose a nearby regular value y and consider $f_0(x)-y$ we have the problem that constant functions are even. Hence we will try the next best thing and perturb by a function which is constant in all except one direction. To this end we choose an odd function $\varphi\in C^1(\mathbb{R})$ such that $\varphi'(0)=0$ (since we don't want to alter the behavior at 0) and $\varphi(t)\neq0$ for $t\neq0$. Now we consider $f_1(x)=f_0(x)-\varphi(x_1)y^1$ and note

$$df_1(x)=df_0(x)-d\varphi(x_1)y^1=df_0(x)-d\varphi(x_1)\frac{f_0(x)}{\varphi(x_1)}=\varphi(x_1)d\big(\frac{f_0(x)}{\varphi(x_1)}\big)$$

for every $x\in U_1:=\{x\in U|x_1\neq0\}$ with $f_1(x)=0$. Hence if y^1 is chosen such that $y^1\in\mathrm{RV}(h_1)$, where $h_1: U_1\to\mathbb{R}^n$, $x\mapsto\frac{f_0(x)}{\varphi(x_1)}$, then 0 will be a regular value for f_1 when restricted to $V_1:=U_1$. Now we repeat this

procedure and consider $f_2(x) = f_1(x) - \varphi(x_2)y^2$ with $y^2 \in \mathrm{RV}(h_2)$ as before. Then every point $x \in V_2 := U_1 \cup U_2$ with $f_2(x) = 0$ either satisfies $x_2 \neq 0$ and thus is regular by our choice of y^2 or satisfies $x_2 = 0$ and thus is regular since it is in V_1 where $f_2 = f_1$. After n steps we reach $V_n = U \setminus \{0\}$ and f_n is the approximation we are looking for. □

At first sight the obvious conclusion that an odd function has a zero does not seem too spectacular since the fact that f is odd already implies $f(0) = 0$. However, the result gets more interesting upon observing that it suffices when the boundary values are odd. Moreover, local constancy of the degree implies that f does not only attain 0 but also any y in a neighborhood of 0. The next two important consequences are based on this observation:

Theorem 17.11 (Borsuk–Ulam)**.** *Let $0 \in U \subseteq \mathbb{R}^n$ be open, bounded and symmetric with respect to the origin. Let $f \in C(\partial U, \mathbb{R}^m)$ with $m < n$. Then there is some $x \in \partial U$ with $f(x) = f(-x)$.*

Proof. Consider $g(x) = f(x) - f(-x)$ and extend it to a continuous odd function $\overline{U} \to \mathbb{R}^n$ (extend the domain by Tietze and then take the odd part, finally fill up the missing coordinates by setting them equal to 0). If g does not vanish on ∂U, we get that $\deg(g, U, y) = \deg(g, U, 0) \neq 0$ for y in a neighborhood of 0 and thus the image of g contains a neighborhood of 0 (in $\mathbb{R}^n$), which contradicts the fact that the image is in $\mathbb{R}^m \times \{0\} \subset \mathbb{R}^n$. □

This theorem is often illustrated by the fact that there are always two opposite points on the earth which have the same weather (in the sense that they have the same temperature and the same pressure). In a similar manner one can also derive the **invariance of domain theorem**.

Theorem 17.12 (Brouwer)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}^n$ be continuous and locally injective. Then $f(U)$ is also open.*

Proof. It suffices to show that every point $x \in U$ contains a neighborhood $B_r(x)$ such that the image $f(B_r(x))$ contains a ball centered at $f(x)$. By simple translations we can assume $x = 0$ as well as $f(x) = 0$. Now choose r sufficiently small such that f restricted to $\bar{B}_r(0)$ is injective and consider $H(t, x) := f(\frac{1}{1+t}x) - f(-\frac{t}{1+t}x)$ for $t \in [0, 1]$ and $x \in \bar{B}_r(0)$. Moreover, if $H(t, x) = 0$ then by injectivity $\frac{1}{1+t}x = -\frac{t}{1+t}x$, that is, $x = 0$. Thus $\deg(f, B_r(0), 0) = \deg(H(1), B_r(0), 0) \neq 0$ since $H(1) = f(\frac{1}{2}x) - f(-\frac{t}{2}x)$ is odd. But then we also have $\deg(f, B_r(0), y) \neq 0$ for $y \in B_\varepsilon(0)$ and thus $B_\varepsilon(0) \subseteq f(B_r(0))$. □

An easy consequence worth while noting is the topological invariance of dimension:

Corollary 17.13. *If $m < n$ and U is a nonempty open subset of $\mathbb{R}^n$, then there is no continuous injective mapping from U to $\mathbb{R}^m$.*

Proof. Suppose there where such a map and extend it to a map from U to $\mathbb{R}^n$ by setting the additional coordinates equal to zero. The resulting map contradicts the invariance of domain theorem. □

In particular, $\mathbb{R}^m$ and $\mathbb{R}^n$ are not homeomorphic for $m \neq n$.

Problem 17.1. *Suppose $D = (a, b) \subset \mathbb{R}^1$. Show*

$$\deg(f, (a, b), y) = \frac{1}{2}\big(\operatorname{sign}(f(b) - y) - \operatorname{sign}(f(a) - y)\big).$$

In particular, our degree reduces to the intermediate value theorem in this case.

Problem* 17.2. *Suppose $f : \bar{B}_r(0) \to \mathbb{R}^n$ is continuous and satisfies*

$$f(x)x > 0, \qquad |x| = r.$$

Then $f(x)$ vanishes somewhere inside $B_r(0)$.

Problem 17.3. *Show that in Borsuk's theorem the condition f is odd can be replaced by $f(x) \neq tf(-x)$ for all $x \in \partial U$ and $t \in (0, 1]$. Note that this condition will hold if $\operatorname{sign}(f(x)) \neq \operatorname{sign}(f(-x))$, $x \in \partial U$ (where $\operatorname{sign}(f(x)) := \frac{f(x)}{|f(x)|}$).*

17.4. The Brouwer fixed point theorem

Now we can show that the famous Brouwer fixed point theorem is a simple consequence of the properties of our degree.

Theorem 17.14 (Brouwer fixed point)**.** *Let K be a topological space homeomorphic to a compact, convex subset of $\mathbb{R}^n$ and let $f \in C(K, K)$, then f has at least one fixed point.*

Proof. Clearly we can assume $K \subset \mathbb{R}^n$ since homeomorphisms preserve fixed points. Now lets assume $K = \bar{B}_r(0)$. If there is a fixed point on the boundary $\partial B_r(0))$ we are done. Otherwise $H(t, x) = x - t\, f(x)$ satisfies $0 \notin H(t, \partial B_r(0))$ since $|H(t, x)| \geq |x| - t|f(x)| \geq (1 - t)r > 0$, $0 \leq t < 1$. And the claim follows from $\deg(x - f(x), B_r(0), 0) = \deg(x, B_r(0), 0) = 1$.

Now let K be convex. Then $K \subseteq B_\rho(0)$ and, by the Hilbert projection theorem (Theorem 2.11) (or alternatively by the Tietze extension theorem or its variant Theorem 18.10 below), we can find a continuous retraction $R : \mathbb{R}^n \to K$ (i.e., $R(x) = x$ for $x \in K$) and consider $\tilde{f} = f \circ R \in C(\bar{B}_\rho(0), \bar{B}_\rho(0))$. By our previous analysis, there is a fixed point $x = \tilde{f}(x) \in \operatorname{conv}(f(K)) \subseteq K$. □

Note that any compact, convex subset of a finite dimensional Banach space (complex or real) is isomorphic to a compact, convex subset of $\mathbb{R}^n$ since linear transformations preserve both properties. In addition, observe that all assumptions are needed. For example, the map $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x+1$, has no fixed point ($\mathbb{R}$ is homeomorphic to a bounded set but not to a compact one). The same is true for the map $f : \partial B_1(0) \to \partial B_1(0)$, $x \mapsto -x$ ($\partial B_1(0) \subset \mathbb{R}^n$ is simply connected for $n \geq 3$ but not homeomorphic to a convex set).

As an easy example of how to use the Brouwer fixed point theorem we show the famous **Perron–Frobenius theorem**.

Theorem 17.15 (Perron–Frobenius)**.** *Let A be an $n \times n$ matrix all whose entries are nonnegative and there is an m such the entries of A^m are all positive. Then A has a positive eigenvalue and the corresponding eigenvector can be chosen to have positive components.*

Proof. We equip $\mathbb{R}^n$ with the norm $|x|_1 := \sum_{j=1}^n |x_j|$ and set $\Delta := \{x \in \mathbb{R}^n | x_j \geq 0, |x|_1 = 1\}$. For $x \in \Delta$ we have $Ax \neq 0$ (since $A^m x \neq 0$) and hence

$$f : \Delta \to \Delta, \quad x \mapsto \frac{Ax}{|Ax|_1}$$

has a fixed point x_0 by the Brouwer fixed point theorem. Then $Ax_0 = |Ax_0|_1 x_0$ and x_0 has positive components since $A^m x_0 = |Ax_0|_1^m x_0$ has. □

Let me remark that the Brouwer fixed point theorem is equivalent to the fact that there is no continuous retraction $R : \overline{B_1(0)} \to \partial B_1(0)$ (with $R(x) = x$ for $x \in \partial B_1(0)$) from the unit ball to the unit sphere in $\mathbb{R}^n$.

In fact, if R would be such a retraction, $-R$ would have a fixed point $x_0 \in \partial B_1(0)$ by Brouwer's theorem. But then $x_0 = -R(x_0) = -x_0$ which is impossible. Conversely, if a continuous function $f : \overline{B_1(0)} \to \overline{B_1(0)}$ has no fixed point we can define a retraction $R(x) = f(x) + t(x)(x - f(x))$, where $t(x) \geq 0$ is chosen such that $|R(x)|^2 = 1$ (i.e., $R(x)$ lies on the intersection of the line spanned by x, $f(x)$ with the unit sphere).

Using this equivalence the Brouwer fixed point theorem can also be derived easily by showing that the homology groups of the unit ball $\overline{B_1(0)}$ and its boundary (the unit sphere) differ (see, e.g., [**35**] for details).

17.5. Kakutani's fixed point theorem and applications to game theory

In this section we want to apply Brouwer's fixed point theorem to show the existence of Nash equilibria for n-person games. As a preparation we extend Brouwer's fixed point theorem to set-valued functions.

Denote by $\mathrm{CS}(K)$ the set of all nonempty convex subsets of K.

Theorem 17.16 (Kakutani)**.** *Suppose K is a compact convex subset of $\mathbb{R}^n$ and $f : K \to \mathrm{CS}(K)$. If the set*

$$\Gamma := \{(x, y) | y \in f(x)\} \subseteq K^2 \tag{17.17}$$

is closed, then there is a point $x \in K$ such that $x \in f(x)$.

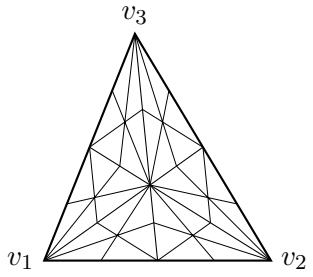
Proof. Our strategy is to apply Brouwer's theorem, hence we need a function related to f. For this purpose it is convenient to assume that K is a simplex

$$K = \mathrm{conv}(v_1, \ldots, v_m), \qquad m \leq n,$$

where v_i are the vertices. Recall that each point $x \in K$ can be uniquely represented by its barycentric coordinates $\lambda_i(x)$ (i.e., $\lambda_i \geq 0$, $\sum_{i=1}^m \lambda_i(x) = 1$ and $x = \sum_{i=1}^m \lambda_i v_i$). Now if we pick $y_i \in f(v_i)$ we could set

$$f^1(x) = \sum_{i=1}^m \lambda_i(x) y_i.$$

By construction, $f^1 \in C(K, K)$ and there is a fixed point x^1. But unless x^1 is one of the vertices, this doesn't help us too much. So lets choose a better function as follows. Consider the k-th barycentric subdivision, that is, for every permutation $v_{\sigma_1}, \ldots, v_{\sigma_m}$ of the vertices you consider the simplex $\mathrm{conv}(v_{\sigma_1}, \frac{1}{2}(v_{\sigma_1} + v_{\sigma_2}), \ldots, \frac{1}{m}(v_{\sigma_1} + \cdots + v_{\sigma_m}))$. This gives you $m!$ smaller simplices (note that the maximal distance between vertices of the subsimplices decreases by a factor $\frac{m-1}{m}$ during the subdivision) whose union is the simplex you have started with. Now repeat this construction k times.



For each vertex v_i in this subdivision pick an element $y_i \in f(v_i)$. Now define $f^k(v_i) = y_i$ and extend f^k to the interior of each subsimplex as before. Hence $f^k \in C(K, K)$ and there is a fixed point x^k in one of the subsimplices. Denote this subsimplex be $\mathrm{conv}(v_1^k, \ldots, v_m^k)$ such that

$$x^k = \sum_{i=1}^m \lambda_i^k v_i^k = \sum_{i=1}^m \lambda_i^k y_i^k, \quad y_i^k = f^k(v_i^k). \tag{17.18}$$

Since $(x^k, \lambda_1^k, \ldots, \lambda_m^k, y_1^k, \ldots, y_m^k) \in K \times [0,1]^m \times K^m$ we can assume that this sequence converges to some limit $(x^0, \lambda_1^0, \ldots, \lambda_m^0, y_1^0, \ldots, y_m^0)$ after passing to a subsequence. Since the subsimplices shrink to a point, this implies $v_i^k \to x^0$

and hence $y_i^0 \in f(x^0)$ since $(v_i^k, y_i^k) \in \Gamma \to (v_i^0, y_i^0) \in \Gamma$ by the closedness assumption. Now (17.18) tells us

$$x^0 = \sum_{i=1}^{m} \lambda_i^0 y_i^0 \in f(x^0)$$

since $f(x^0)$ is convex and the claim holds if K is a simplex.

If K is not a simplex, we can pick a simplex S containing K and proceed as in the proof of the Brouwer theorem. □

If $f(x)$ contains precisely one point for all x, then Kakutani's theorem reduces to the Brouwer's theorem (show that the closedness of Γ is equivalent to continuity of f).

Now we want to see how this applies to game theory.

An **n-person game** consists of n players who have m_i possible actions to choose from. The set of all possible actions for the i-th player will be denoted by $\Phi_i = \{1, \dots, m_i\}$. An element $\varphi_i \in \Phi_i$ is also called a pure strategy for reasons to become clear in a moment. Once all players have chosen their move φ_i, the payoff for each player is given by the **payoff** function

$$R_i(\varphi) \in \mathbb{R}, \qquad \varphi = (\varphi_1, \dots, \varphi_n) \in \Phi = \bigtimes_{i=1}^{n} \Phi_i \tag{17.19}$$

of the i-th player. We will consider the case where the game is repeated a large number of times and where in each step the players choose their action according to a fixed strategy. Here a **strategy** s_i for the i-th player is a probability distribution on Φ_i, that is, $s_i = (s_i^1, \dots, s_i^{m_i})$ such that $s_i^k \geq 0$ and $\sum_{k=1}^{m_i} s_i^k = 1$. The set of all possible strategies for the i-th player is denoted by S_i. The number s_i^k is the probability for the k-th pure strategy to be chosen. Consequently, if $s = (s_1, \dots, s_n) \in S = \bigtimes_{i=1}^{n} S_i$ is a collection of strategies, then the probability that a given collection of pure strategies gets chosen is

$$s(\varphi) = \prod_{i=1}^{n} s_i(\varphi), \qquad s_i(\varphi) = s_i^{k_i},\ \varphi = (k_1, \dots, k_n) \in \Phi \tag{17.20}$$

(assuming all players make their choice independently) and the expected payoff for player i is

$$R_i(s) = \sum_{\varphi \in \Phi} s(\varphi) R_i(\varphi). \tag{17.21}$$

By construction, $R_i : S \to \mathbb{R}$ is polynomial and hence in particular continuous.

The question is of course, what is an optimal strategy for a player? If the other strategies are known, a **best reply** of player i against s would be

a strategy $\overline{s_i}$ satisfying

$$R_i(s \setminus \overline{s}_i) = \max_{\tilde{s}_i \in S_i} R_i(s \setminus \tilde{s}_i) \tag{17.22}$$

Here $s \setminus \tilde{s}_i$ denotes the strategy combination obtained from s by replacing s_i by $\tilde{s}_i$. The set of all best replies against s for the i-th player is denoted by $B_i(s)$. Since

$$R_i(s) = \sum_{k=1}^{m_i} s_i^k R_i(s/k) \tag{17.23}$$

we have $\overline{s}_i \in B_i(s)$ if and only if $\overline{s}_i^k = 0$ whenever $R_i(s\backslash k) < \max_{1 \le l \le m_i} R_i(s\backslash l)$. In particular, since there are no restrictions on the other entries, $B_i(s)$ is a nonempty convex set.

Let $s, \overline{s} \in S$, we call $\overline{s}$ a best reply against s if $\overline{s}_i$ is a best reply against s for all i. The set of all best replies against s is $B(s) = \bigtimes_{i=1}^n B_i(s)$.

A strategy combination $\overline{s} \in S$ is a **Nash equilibrium** for the game if it is a best reply against itself, that is,

$$\overline{s} \in B(\overline{s}). \tag{17.24}$$

Or, put differently, $\overline{s}$ is a Nash equilibrium if no player can increase his payoff by changing his strategy as long as all others stick to their respective strategies. In addition, if a player sticks to his equilibrium strategy, he is assured that his payoff will not decrease no matter what the others do.

To illustrate these concepts, let us consider the famous *prisoner's dilemma*. Here we have two players which can choose to defect or to cooperate. The payoff is symmetric for both players and given by the following diagram

R_1	d_2	c_2
d_1	0	2
c_1	-1	1

R_2	d_2	c_2
d_1	0	-1
c_1	2	1

(17.25)

where c_i or d_i means that player i cooperates or defects, respectively. You should think of two prisoners who are offered a reduced sentence if they testify against the other.

It is easy to see that the (pure) strategy pair (d_1, d_2) is the only Nash equilibrium for this game and that the expected payoff is 0 for both players. Of course, both players could get the payoff 1 if they both agree to cooperate. But if one would break this agreement in order to increase his payoff, the other one would get less. Hence it might be safer to defect.

Now that we have seen that Nash equilibria are a useful concept, we want to know when such an equilibrium exists. Luckily we have the following result.

Theorem 17.17 (Nash)**.** *Every n-person game has at least one Nash equilibrium.*

Proof. The definition of a Nash equilibrium begs us to apply Kakutani's theorem to the set-valued function $s \mapsto B(s)$. First of all, S is compact and convex and so are the sets $B(s)$. Next, observe that the closedness condition of Kakutani's theorem is satisfied since if $s^m \in S$ and $\overline{s}^m \in B(s^n)$ both converge to s and $\overline{s}$, respectively, then (17.22) for s^m, $\overline{s}^m$

$$R_i(s^m \setminus \tilde{s}_i) \leq R_i(s^m \setminus \overline{s}_i^m), \qquad \tilde{s}_i \in S_i,\ 1 \leq i \leq n,$$

implies (17.22) for the limits s, $\overline{s}$

$$R_i(s \setminus \tilde{s}_i) \leq R_i(s \setminus \overline{s}_i), \qquad \tilde{s}_i \in S_i,\ 1 \leq i \leq n,$$

by continuity of $R_i(s)$. □

17.6. Further properties of the degree

We now prove some additional properties of the mapping degree. The first one will relate the degree in $\mathbb{R}^n$ with the degree in $\mathbb{R}^m$. It will be needed later on to extend the definition of degree to infinite dimensional spaces. By virtue of the canonical embedding $\mathbb{R}^m \hookrightarrow \mathbb{R}^m \times \{0\} \subset \mathbb{R}^n$ we can consider $\mathbb{R}^m$ as a subspace of $\mathbb{R}^n$. We can project $\mathbb{R}^n$ to $\mathbb{R}^m$ by setting the last $n-m$ coordinates equal to zero.

Theorem 17.18 (Reduction property)**.** *Let $U \subseteq \mathbb{R}^n$ be open and bounded, $f \in C(\overline{U}, \mathbb{R}^m)$ and $y \in \mathbb{R}^m \setminus (\mathbb{I} + f)(\partial U)$, then*

$$\deg(\mathbb{I} + f, U, y) = \deg(\mathbb{I} + f_m, U_m, y), \tag{17.26}$$

where $f_m = f|_{U_m}$, where U_m is the projection of U to $\mathbb{R}^m$.

Proof. After perturbing f a little, we can assume $f \in C^1(U, \mathbb{R}^m)$ without loss of generality. Let $x \in (\mathbb{I} + f)^{-1}(\{y\})$, then $x = y - f(x) \in \mathbb{R}^m$ implies $(\mathbb{I} + f)^{-1}(\{y\}) = (\mathbb{I} + f_m)^{-1}(\{y\})$. Moreover,

$$\begin{aligned} J_{\mathbb{I}+f}(x) &= \det(\mathbb{I} + df)(x) = \det \begin{pmatrix} \delta_{ij} + \partial_j f_i(x) & \partial_j f_j(x) \\ 0 & \delta_{ij} \end{pmatrix} \\ &= \det(\delta_{ij} + \partial_j f_i) = J_{\mathbb{I}+f_m}(x) \end{aligned}$$

So if $y \in \mathrm{RV}(\mathbb{I}+f_m)$ we immediately get $\deg(\mathbb{I}+f, U, y) = \deg(\mathbb{I}+f_m, U_m, y)$ as desired. Otherwise, if $y \in \mathrm{CV}(\mathbb{I} + f_m)$ we can choose some $\tilde{y} \in \mathrm{RV}(\mathbb{I} + f_m)$ (and hence also $\tilde{y} \in \mathrm{RV}(\mathbb{I} + f)$ by the first part) with $|y - \tilde{y}| < \min(\mathrm{dist}(y, f(\partial U)), \mathrm{dist}(y, f_m(\partial U_m)))$ and the claim follows from the regular case. □

Let $U \subseteq \mathbb{R}^n$ and $f \in C(\overline{U}, \mathbb{R}^n)$ be as usual. By Theorem 17.2 we know that $\deg(f, U, y)$ is the same for every y in a connected component of $\mathbb{R}^n \setminus f(\partial U)$. Since $\mathbb{R}^n \setminus f(\partial U)$ is open and locally path connected, these components are open. We will denote these components by G_j and write $\deg(f, U, G_j) := \deg(f, U, y)$ if $y \in G_j$. In this context observe that since $f(\partial U)$ is compact any unbounded component (there will be two for $n = 1$ and one for $n > 1$) will have degree zero.

Theorem 17.19 (Product formula)**.** *Let $U \subseteq \mathbb{R}^n$ be a bounded and open set and denote by G_j the connected components of $\mathbb{R}^n \setminus f(\partial U)$. If $g \circ f \in \bar{C}_y(U, \mathbb{R}^n)$, then*

$$\deg(g \circ f, U, y) = \sum_j \deg(f, U, G_j) \deg(g, G_j, y), \tag{17.27}$$

where only finitely many terms in the sum are nonzero (and in particular, summands corresponding to unbounded components are considered to be zero).

Proof. Since $y \notin (g \circ f)(\partial U)$ we have $g^{-1}(\{y\}) \cap f(\partial U) = \emptyset$, that is, $g^{-1}(\{y\}) \subset \bigcap_j G_j$. Moreover, since $f(\overline{U})$ is compact, we can find an $r > 0$ such that $f(\overline{U}) \subseteq B_r(0)$. Moreover, since $g^{-1}(\{y\})$ is closed, $g^{-1}(\{y\}) \cap B_r(0)$ is compact and hence can be covered by finitely many components, say $g^{-1}(\{y\}) \subset \bigcap_{j=1}^m G_j$. In particular, the others will be either unbounded or have $\deg(f, G_k, y) = 0$ and hence only finitely many terms in the above sum are nonzero.

We begin by computing $\deg(g \circ f, U, y)$ in the case where $f, g \in C^1$ and $y \notin \mathrm{CV}(g \circ f)$. Since $d(g \circ f)(x) = dg(f(x)) \circ df(x)$ the claim is a straightforward calculation

$$\begin{aligned}
\deg(g \circ f, U, y) &= \sum_{x \in (g \circ f)^{-1}(\{y\})} \operatorname{sign}(J_{g \circ f}(x)) \\
&= \sum_{x \in (g \circ f)^{-1}(\{y\})} \operatorname{sign}(J_g(f(x))) \operatorname{sign}(J_f(x)) \\
&= \sum_{z \in g^{-1}(\{y\})} \operatorname{sign}(J_g(z)) \sum_{x \in f^{-1}(\{z\}))} \operatorname{sign}(J_f(x)) \\
&= \sum_{z \in g^{-1}(\{y\})} \operatorname{sign}(J_g(z)) \deg(f, U, z)
\end{aligned}$$

and, using our cover $\{G_j\}_{j=1}^m$,

$$\begin{aligned}\deg(g\circ f,U,y) &= \sum_{j=1}^m \sum_{z\in g^{-1}(\{y\})\cap G_j} \operatorname{sign}(J_g(z))\deg(f,U,z)\\ &= \sum_{j=1}^m \deg(f,U,G_j) \sum_{z\in g^{-1}(\{y\})\cap G_j} \operatorname{sign}(J_g(z))\\ &= \sum_{j=1}^m \deg(f,U,G_j)\deg(g,G_j,y).\end{aligned}$$

Moreover, this formula still holds for $y\in \operatorname{CV}(g\circ f)$ and for $g\in C$ by construction of the Brouwer degree. However, the case $f\in C$ will need a closer investigation since the components G_j depend on f. To overcome this problem we will introduce the sets

$$L_l := \{z\in\mathbb{R}^n\setminus f(\partial U)|\deg(f,U,z)=l\}.$$

Observe that L_l, $l\neq 0$, must be a union of some sets from $\{G_j\}_{j=1}^m$, that is, $L_l=\bigcup_{k=1}^{m_l} G_{j_k^l}$ and $\bigcup_{l\neq 0} L_l=\bigcup_{j=1}^m G_j$.

Now choose $\tilde f\in C^1$ such that $|f(x)-\tilde f(x)|<2^{-1}\operatorname{dist}(g^{-1}(\{y\}),f(\partial U))$ for $x\in\overline{U}$ and define $\tilde G_j$, $\tilde L_l$ accordingly. Then we have $L_l\cap g^{-1}(\{y\})=\tilde L_l\cap g^{-1}(\{y\})$ by Theorem 17.1 (iii) and hence $\deg(g,\tilde L_l,y)=\deg(g,L_l,y)$ by Theorem 17.1 (i) implying

$$\begin{aligned}\deg(g\circ f,U,y)=\deg(g\circ\tilde f,U,y) &= \sum_{j=1}^{\tilde m}\deg(\tilde f,U,\tilde G_j)\deg(g,\tilde G_j,y)\\ &= \sum_{l\neq 0} l\deg(g,\tilde L_l,y)=\sum_{l\neq 0} l\deg(g,L_l,y)\\ &= \sum_{l\neq 0}\sum_{k=1}^{m_l} l\deg(g,G_{j_k^l},y)=\sum_{j=1}^m\deg(f,U,G_j)\deg(g,G_j,y)\end{aligned}$$

which proves the claim. □

17.7. The Jordan curve theorem

In this section we want to show how the product formula (17.27) for the Brouwer degree can be used to prove the famous **Jordan curve theorem** which states that a homeomorphic image of the circle dissects $\mathbb{R}^2$ into two components (which necessarily have the image of the circle as common boundary). In fact, we will even prove a slightly more general result.

Theorem 17.20. *Let $C_j\subset\mathbb{R}^n$, $j=1,2$, be homeomorphic compact sets. Then $\mathbb{R}^n\setminus C_1$ and $\mathbb{R}^n\setminus C_2$ have the same number of connected components.*

Proof. Denote the components of $\mathbb{R}^n \setminus C_1$ by H_j and those of $\mathbb{R}^n \setminus C_2$ by K_j. Since our sets are closed these components are open. Moreover, $\partial H_j \subseteq C_1$ since a sequence from H_j cannot converge to a (necessarily interior) point of H_k for some $k \neq j$. Let $h : C_1 \to C_2$ be a homeomorphism with inverse $k : C_2 \to C_1$. By Theorem 18.10 we can extend both to $\mathbb{R}^n$. Then Theorem 17.1 (iii) and the product formula imply

$$1 = \deg(k \circ h, H_j, y) = \sum_l \deg(h, H_j, G_l) \deg(k, G_l, y)$$

for any $y \in H_j$, where G_l are the components of $\mathbb{R}^n \setminus h(\partial H_j)$. Now we have

$$\bigcup_i K_i = \mathbb{R}^n \setminus C_2 \subseteq \mathbb{R}^n \setminus h(\partial H_j) = \bigcup_l G_l$$

and hence for every i we have $K_i \subseteq G_l$ for some l since components are connected. Let $N_l := \{i | K_i \subseteq G_l\}$ and observe that we have $\deg(k, G_l, y) = \sum_{i \in N_l} \deg(k, K_i, y)$ by Theorem 17.1 (i) since $k^{-1}(\{y\}) \subseteq \bigcup_j K_j$ and of course $\deg(h, H_j, K_i) = \deg(h, H_j, G_l)$ for every $i \in N_l$. Therefore,

$$1 = \sum_l \sum_{i \in N_l} \deg(h, H_j, K_i) \deg(k, K_i, y) = \sum_i \deg(h, H_j, K_i) \deg(k, K_i, H_j)$$

By reversing the role of C_1 and C_2, the same formula holds with H_j and K_i interchanged.

Hence

$$\sum_i 1 = \sum_i \sum_j \deg(h, H_j, K_i) \deg(k, K_i, H_j) = \sum_j 1$$

shows that if either the number of components of $\mathbb{R}^n \setminus C_1$ or the number of components of $\mathbb{R}^n \setminus C_2$ is finite, then so is the other and both are equal. Otherwise there is nothing to prove. □

Chapter 18

The Leray–Schauder mapping degree

18.1. The mapping degree on finite dimensional Banach spaces

The objective of this section is to extend the mapping degree from $\mathbb{R}^n$ to general Banach spaces. Naturally, we will first consider the finite dimensional case.

Let X be a (real) Banach space of dimension n and let ϕ be any isomorphism between X and $\mathbb{R}^n$. Then, for $f \in \bar{C}_y(U, X)$, $U \subset X$ open, $y \in X$, we can define

$$\deg(f, U, y) := \deg(\phi \circ f \circ \phi^{-1}, \phi(U), \phi(y)) \tag{18.1}$$

provided this definition is independent of the isomorphism chosen. To see this let ψ be a second isomorphism. Then $A = \psi \circ \phi^{-1} \in \mathrm{GL}(n)$. Abbreviate $f^* = \phi \circ f \circ \phi^{-1}$, $y^* = \phi(y)$ and pick $\tilde{f}^* \in \bar{C}^1_y(\phi(U), \mathbb{R}^n)$ in the same component of $\bar{C}_y(\phi(U), \mathbb{R}^n)$ as f^* such that $y^* \in \mathrm{RV}(f^*)$. Then $A \circ \tilde{f}^* \circ A^{-1} \in \bar{C}^1_y(\psi(U), \mathbb{R}^n)$ is the same component of $\bar{C}_y(\psi(U), \mathbb{R}^n)$ as $A \circ f^* \circ A^{-1} = \psi \circ f \circ \psi^{-1}$ (since A is also a homeomorphism) and

$$J_{A \circ \tilde{f}^* \circ A^{-1}}(Ay^*) = \det(A) J_{\tilde{f}^*}(y^*) \det(A^{-1}) = J_{\tilde{f}^*}(y^*) \tag{18.2}$$

by the chain rule. Thus we have $\deg(\psi \circ f \circ \psi^{-1}, \psi(U), \psi(y)) = \deg(\phi \circ f \circ \phi^{-1}, \phi(U), \phi(y))$ and our definition is independent of the basis chosen. In addition, it inherits all properties from the mapping degree in $\mathbb{R}^n$. Note also that the reduction property holds if $\mathbb{R}^m$ is replaced by an arbitrary subspace X_1 since we can always choose $\phi : X \to \mathbb{R}^n$ such that $\phi(X_1) = \mathbb{R}^m$.

Our next aim is to tackle the infinite dimensional case. The following example due to Kakutani shows that the Brouwer fixed point theorem (and hence also the Brouwer degree) does not generalize to infinite dimensions directly.

Example 18.1. Let X be the Hilbert space $\ell^2(\mathbb{N})$ and let R be the right shift given by $Rx := (0, x_1, x_2, \dots)$. Define $f : \bar{B}_1(0) \to \bar{B}_1(0)$, $x \mapsto \sqrt{1-\|x\|^2}\delta_1 + Rx = (\sqrt{1-\|x\|^2}, x_1, x_2, \dots)$. Then a short calculation shows $\|f(x)\|^2 = (1-\|x\|^2) + \|x\|^2 = 1$ and any fixed point must satisfy $\|x\| = 1$, $x_1 = \sqrt{1-\|x\|^2} = 0$ and $x_{j+1} = x_j$, $j \in \mathbb{N}$ giving the contradiction $x_j = 0$, $j \in \mathbb{N}$. ⋄

However, by the reduction property we expect that the degree should hold for functions of the type $\mathbb{I} + F$, where F has finite dimensional range. In fact, it should work for functions which can be approximated by such functions. Hence as a preparation we will investigate this class of functions.

18.2. Compact maps

Let X, Y be Banach spaces and $U \subseteq X$. A map $F : U \subset X \to Y$ is called **finite dimensional** if its range is finite dimensional. In addition, it is called **compact** if it is continuous and maps bounded sets into relatively compact ones. The set of all compact maps is denoted by $\mathcal{C}(U, Y)$ and the set of all compact, finite dimensional maps is denoted by $\mathcal{F}(U, Y)$. Both sets are normed linear spaces and we have $\mathcal{C}(U, Y) \subseteq C_b(U, Y)$ if U is bounded (recall that compact sets are automatically bounded).

If K is compact, then $\mathcal{C}(K, Y) = C(K, Y)$ (since the continuous image of a compact set is compact) and if $\dim(Y) < \infty$, then $\mathcal{F}(U, Y) = \mathcal{C}(U, Y)$. In particular, if $U \subset \mathbb{R}^n$ is bounded, then $\mathcal{F}(\overline{U}, \mathbb{R}^n) = \mathcal{C}(\overline{U}, \mathbb{R}^n) = C(\overline{U}, \mathbb{R}^n)$.

Example 18.2. Note that for nonliner functions it is important to include continuity in the definition of compactness. Indeed, if X is a Hilbert space with an orthonormal basis $\{u_j\}_{j\in\mathbb{N}}$, then

$$\phi(x) = \begin{cases} j(1-2|x-x_j|), & x \in B_{1/2}(x_j), \\ 0, & \text{else}, \end{cases}$$

is in $C(B_1(0), \mathbb{R})$ but not bounded. Hence $F(x) = \phi(x)x_1$ is one-dimensional but not compact. Choosing $\phi(x) = 1$ for $x \in B_{1/2}(0)$ and $\phi(x) = 0$ else gives a map F which maps bounded sets to relatively compact ones but which is not continuous. ⋄

Now let us collect some results needed in the sequel.

Lemma 18.1. *If $K \subset X$ is compact, then for every $\varepsilon > 0$ there is a finite dimensional subspace $X_\varepsilon \subseteq X$ and a continuous map $P_\varepsilon : K \to X_\varepsilon$ such that $|P_\varepsilon(x) - x| \leq \varepsilon$ for all $x \in K$.*

Proof. Pick $\{x_i\}_{i=1}^n \subseteq K$ such that $\bigcup_{i=1}^n B_\varepsilon(x_i)$ covers K. Let $\{\phi_i\}_{i=1}^n$ be a partition of unity (restricted to K) subordinate to $\{B_\varepsilon(x_i)\}_{i=1}^n$, that is, $\phi_i \in C(K, [0,1])$ with $\operatorname{supp}(\phi_i) \subset B_\varepsilon(x_i)$ and $\sum_{i=1}^n \phi_i(x) = 1$, $x \in K$. Set

$$P_\varepsilon(x) = \sum_{i=1}^n \phi_i(x) x_i,$$

then

$$|P_\varepsilon(x) - x| = |\sum_{i=1}^n \phi_i(x) x - \sum_{i=1}^n \phi_i(x) x_i| \leq \sum_{i=1}^n \phi_i(x) |x - x_i| \leq \varepsilon. \qquad \square$$

This lemma enables us to prove the following important result.

Theorem 18.2. *Let U be bounded, then the closure of $\mathcal{F}(U,Y)$ in $C_b(U,Y)$ is $\mathcal{C}(U,Y)$.*

Proof. Suppose $F_N \in \mathcal{C}(U,Y)$ converges to F. If $F \notin \mathcal{C}(U,Y)$ then we can find a sequence $x_n \in U$ such that $|F(x_n) - F(x_m)| \geq \rho > 0$ for $n \neq m$. If N is so large that $|F - F_N| \leq \rho/4$, then

$$\begin{aligned} |F_N(x_n) - F_N(x_m)| &\geq |F(x_n) - F(x_m)| - |F_N(x_n) - F(x_n)| \\ &\quad - |F_N(x_m) - F(x_m)| \\ &\geq \rho - 2\frac{\rho}{4} = \frac{\rho}{2} \end{aligned}$$

This contradiction shows $\overline{\mathcal{F}(U,Y)} \subseteq \mathcal{C}(U,Y)$. Conversely, let $F \in \mathcal{C}(U,Y)$, set $K := \overline{F(U)}$ and choose P_ε according to Lemma 18.1. Then $F_\varepsilon = P_\varepsilon \circ F \in \mathcal{F}(U,Y)$ converges to F. Hence $\mathcal{C}(U,Y) \subseteq \overline{\mathcal{F}(U,Y)}$ and we are done. $\square$

Finally, let us show some interesting properties of mappings $\mathbb{I} + F$, where $F \in \mathcal{C}(U,Y)$.

Lemma 18.3. *Let $\overline{U} \subseteq X$ be bounded and closed. Suppose $F \in \mathcal{C}(\overline{U}, X)$, then $\mathbb{I} + F$ is **proper** (i.e., inverse images of compact sets are compact) and maps closed subsets to closed subsets.*

Proof. Let $A \subseteq \overline{U}$ be closed and suppose $y_n = (\mathbb{I} + F)(x_n) \in (\mathbb{I} + F)(A)$ converges to some point y. Since $y_n - x_n = F(x_n) \in F(U)$ we can assume that $y_n - x_n \to z$ after passing to a subsequence and hence $x_n \to x = y - z \in A$. Since $y = x + F(x) \in (\mathbb{I} + F)(A)$, $(\mathbb{I} + F)(A)$ is closed.

Next, let $\overline{U}$ be closed and $K \subset X$ be compact. Let $\{x_n\} \subseteq (\mathbb{I}+F)^{-1}(K)$. Then $y_n := x_n + F(x_n) \in K$ and we can pass to a subsequence $y_{n_m} =$

$x_{n_m} + F(x_{n_m})$ such that $y_{n_m} \to y$. As before this implies $x_{n_m} \to x$ and thus $(\mathbb{I} + F)^{-1}(K)$ is compact. □

Finally note that if $F \in \mathcal{C}(\overline{U}, Y)$ and $G \in C(Y, Z)$, then $G \circ F \in \mathcal{C}(\overline{U}, Z)$ and similarly, if $G \in C_b(\overline{V}, \overline{U})$, then $F \circ G \in \mathcal{C}(\overline{V}, Y)$.

Now we are all set for the definition of the Leray–Schauder degree, that is, for the extension of our degree to infinite dimensional Banach spaces.

18.3. The Leray–Schauder mapping degree

For an open set $U \subset X$ we set

$$\bar{\mathcal{C}}_y(U, X) := \{F \in \mathcal{C}(\overline{U}, X) | y \notin (\mathbb{I} + F)(\partial U)\} \tag{18.3}$$

and $\bar{\mathcal{F}}_y(U, X) := \{F \in \mathcal{F}(\overline{U}, X) | y \notin (\mathbb{I} + F)(\partial U)\}$. Note that for $F \in \bar{\mathcal{C}}_y(U, X)$ we have $\operatorname{dist}(y, (\mathbb{I} + F)(\partial U)) > 0$ since $\mathbb{I} + F$ maps closed sets to closed sets (cf. Problem B.22).

Abbreviate $\rho := \operatorname{dist}(y, (\mathbb{I} + F)(\partial U))$ and pick $F_1 \in \mathcal{F}(\overline{U}, X)$ such that $|F - F_1| < \rho$ implying $F_1 \in \bar{\mathcal{F}}_y(U, X)$. Next, let X_1 be a finite dimensional subspace of X such that $F_1(U) \subset X_1$, $y \in X_1$ and set $U_1 := U \cap X_1$. Then we have $F_1 \in \bar{\mathcal{F}}_y(U_1, X_1)$ and might define

$$\deg(\mathbb{I} + F, U, y) := \deg(\mathbb{I} + F_1, U_1, y) \tag{18.4}$$

provided we show that this definition is independent of F_1 and X_1 (as above). Pick another map $F_2 \in \mathcal{F}(\overline{U}, X)$ such that $|F - F_2| < \rho$ and let X_2 be a corresponding finite dimensional subspace as above. Consider $X_0 := X_1 + X_2$, $U_0 = U \cap X_0$, then $F_i \in \bar{\mathcal{F}}_y(U_0, X_0)$, $i = 1, 2$, and

$$\deg(\mathbb{I} + F_i, U_0, y) = \deg(\mathbb{I} + F_i, U_i, y), \quad i = 1, 2, \tag{18.5}$$

by the reduction property. Moreover, set $H(t) = \mathbb{I} + (1-t)F_1 + t\,F_2$ implying $H(t) \in \bar{\mathcal{C}}_y(U_0, X_0)$, $t \in [0, 1]$, since $|H(t) - (\mathbb{I} + F)| < \rho$ for $t \in [0, 1]$. Hence homotopy invariance

$$\deg(\mathbb{I} + F_1, U_0, y) = \deg(\mathbb{I} + F_2, U_0, y) \tag{18.6}$$

shows that (18.4) is independent of F_1, X_1.

Theorem 18.4. *Let U be a bounded open subset of a (real) Banach space X and let $F \in \bar{\mathcal{C}}_y(U, X)$, $y \in X$. Then the following hold true.*

(i). $\deg(\mathbb{I} + F, U, y) = \deg(\mathbb{I} + F - y, U, 0)$.

(ii). $\deg(\mathbb{I}, U, y) = 1$ *if* $y \in U$.

(iii). *If $U_{1,2}$ are open, disjoint subsets of U such that $y \notin f(\overline{U} \setminus (U_1 \cup U_2))$, then* $\deg(\mathbb{I} + F, U, y) = \deg(\mathbb{I} + F, U_1, y) + \deg(\mathbb{I} + F, U_2, y)$.

(iv). *If* $H : [0,1] \times \overline{U} \to X$ *and* $y : [0,1] \to X$ *are both continuous such that* $H(t) \in \bar{\mathcal{C}}_{y(t)}(U,X)$, $t \in [0,1]$, *then* $\deg(\mathbb{I} + H(0), U, y(0)) = \deg(\mathbb{I} + H(1), U, y(1))$.

Proof. Except for (iv) all statements follow easily from the definition of the degree and the corresponding property for the degree in finite dimensional spaces. Considering $H(t,x) - y(t)$, we can assume $y(t) = 0$ by (i). Since $H([0,1], \partial U)$ is compact, we have $\rho = \operatorname{dist}(y, H([0,1], \partial U)) > 0$. By Theorem 18.2 we can pick $H_1 \in \mathcal{F}([0,1] \times U, X)$ such that $|H(t) - H_1(t)| < \rho$, $t \in [0,1]$. This implies $\deg(\mathbb{I} + H(t), U, 0) = \deg(\mathbb{I} + H_1(t), U, 0)$ and the rest follows from Theorem 17.2. □

In addition, Theorem 17.1 and Theorem 17.2 hold for the new situation as well (no changes are needed in the proofs).

Theorem 18.5. *Let* $F, G \in \bar{\mathcal{C}}_y(U,X)$, *then the following statements hold.*

(i). *We have* $\deg(\mathbb{I} + F, \emptyset, y) = 0$. *Moreover, if* U_i, $1 \le i \le N$, *are disjoint open subsets of* U *such that* $y \notin (\mathbb{I} + F)(\overline{U} \setminus \bigcup_{i=1}^N U_i)$, *then* $\deg(\mathbb{I} + F, U, y) = \sum_{i=1}^N \deg(\mathbb{I} + F, U_i, y)$.

(ii). *If* $y \notin (\mathbb{I} + F)(U)$, *then* $\deg(\mathbb{I} + F, U, y) = 0$ *(but not the other way round). Equivalently, if* $\deg(\mathbb{I} + F, U, y) \neq 0$, *then* $y \in (\mathbb{I} + F)(U)$.

(iii). *If* $|F(x) - G(x)| < \operatorname{dist}(y, (\mathbb{I} + F)(\partial U))$, $x \in \partial U$, *then* $\deg(\mathbb{I} + F, U, y) = \deg(\mathbb{I} + G, U, y)$. *In particular, this is true if* $F(x) = G(x)$ *for* $x \in \partial U$.

(iv). $\deg(\mathbb{I} + ., U, y)$ *is constant on each component of* $\bar{\mathcal{C}}_y(U,X)$.

(v). $\deg(\mathbb{I} + F, U, .)$ *is constant on each component of* $X \setminus (\mathbb{I} + F)(\partial U)$.

Note that it is easy to generalize Borsuk's theorem.

Theorem 18.6 (Borsuk)**.** *Let* $U \subseteq X$ *be open, bounded and symmetric with respect to the origin (i.e,* $U = -U$*). Let* $F \in \bar{\mathcal{C}}_0(U)$ *be odd (i.e.,* $F(-x) = -F(x)$*). Then* $\deg(\mathbb{I} + F, U, 0)$ *is odd.*

Proof. Choose F_1 and U_1 as in the definition of the degree. Then U_1 is symmetric and F_1 can be chosen to be odd by replacing it by its odd part. Hence the claim follows from the finite dimensional version. □

In the same way as in the finite dimensional case we also obtain the **invariance of domain theorem**.

Theorem 18.7 (Brouwer)**.** *Let* $U \subseteq X$ *be open and let* $F \in \mathcal{C}(U,x)$ *be compact with* $\mathbb{I} + F$ *locally injective. Then* $\mathbb{I} + F$ *is also open.*

18.4. The Leray–Schauder principle and the Schauder fixed point theorem

As a first consequence we note the Leray–Schauder principle which says that a priori estimates yield existence.

Theorem 18.8 (Schaefer fixed point or Leray–Schauder principle)**.** *Suppose $F \in \mathcal{C}(X,X)$ and any solution x of $x = tF(x)$, $t \in [0,1]$ satisfies the a priori bound $|x| \le M$ for some $M > 0$, then F has a fixed point.*

Proof. Pick $\rho > M$ and observe $\deg(\mathbb{I} - F, B_\rho(0), 0) = \deg(\mathbb{I}, B_\rho(0), 0) = 1$ using the compact homotopy $H(t,x) := -tF(x)$. Here $H(t) \in \bar{\mathcal{C}}_0(B_\rho(0), X)$ due to the a priori bound. □

Now we can extend the Brouwer fixed point theorem to infinite dimensional spaces as well.

Theorem 18.9 (Schauder fixed point)**.** *Let K be a closed, convex, and bounded subset of a Banach space X. If $F \in \mathcal{C}(K,K)$, then F has at least one fixed point. The result remains valid if K is only homeomorphic to a closed, convex, and bounded subset.*

Proof. Since K is bounded, there is a $\rho > 0$ such that $K \subseteq B_\rho(0)$. By Theorem 18.10 below we can find a continuous retraction $R : X \to K$ (i.e., $R(x) = x$ for $x \in K$) and consider $\tilde{F} = F \circ R \in \mathcal{C}(\overline{B_\rho(0)}, \overline{B_\rho(0)})$. Now either $t\tilde{F}$ has a fixed point on the boundary ∂U or the compact homotopy $H(t,x) := -t\tilde{F}(x)$ satisfies $0 \notin (\mathbb{I} - t\tilde{F})(\partial U)$ and thus $\deg(\mathbb{I} - \tilde{F}, B_\rho(0), 0) = \deg(\mathbb{I}, B_\rho(0), 0) = 1$. Hence there is a point $x_0 = \tilde{F}(x_0) \in K$. Since $\tilde{F}(x_0) = F(x_0)$ for $x_0 \in K$ we are done. □

It remains to prove the following variant of the **Tietze extension theorem** needed in the proof.

Theorem 18.10. *Let X be a metric space, Y a normed space and let K be a closed subset of X. Then $F \in C(K,Y)$ has a continuous extension $\bar{F} \in C(X,Y)$ such that $\bar{F}(X) \subseteq \operatorname{conv}(F(K))$.*

Proof. Consider the open cover $\{B_{\rho(x)}(x)\}_{x \in X \setminus K}$ for $X \setminus K$, where $\rho(x) = \operatorname{dist}(x,K)/2$. Choose a (locally finite) partition of unity $\{\phi_\lambda\}_{\lambda \in \Lambda}$ subordinate to this cover (cf. Lemma B.30) and set

$$\bar{F}(x) := \sum_{\lambda \in \Lambda} \phi_\lambda(x) F(x_\lambda) \text{ for } x \in X \setminus K,$$

where $x_\lambda \in K$ satisfies $\operatorname{dist}(x_\lambda, \operatorname{supp}\phi_\lambda) \le 2\operatorname{dist}(K, \operatorname{supp}\phi_\lambda)$. By construction, $\bar{F}$ is continuous except for possibly at the boundary of K. Fix $x_0 \in \partial K$, $\varepsilon > 0$ and choose $\delta > 0$ such that $|F(x) - F(x_0)| \le \varepsilon$ for all

$x \in K$ with $|x - x_0| < 4\delta$. We will show that $|\bar{F}(x) - F(x_0)| \leq \varepsilon$ for all $x \in X$ with $|x - x_0| < \delta$. Suppose $x \notin K$, then $|\bar{F}(x) - F(x_0)| \leq \sum_{\lambda\in\Lambda} \phi_\lambda(x)|F(x_\lambda) - F(x_0)|$. By our construction, x_λ should be close to x for all λ with $x \in \operatorname{supp}\phi_\lambda$ since x is close to K. In fact, if $x \in \operatorname{supp}\phi_\lambda$ we have

$$\begin{aligned} |x - x_\lambda| &\leq \operatorname{dist}(x_\lambda, \operatorname{supp}\phi_\lambda) + \operatorname{diam}(\operatorname{supp}\phi_\lambda) \\ &\leq 2\operatorname{dist}(K, \operatorname{supp}\phi_\lambda) + \operatorname{diam}(\operatorname{supp}\phi_\lambda), \end{aligned}$$

where $\operatorname{diam}(\operatorname{supp}\phi_\lambda) := \sup_{x,y\in\operatorname{supp}\phi_\lambda} |x-y|$. Since our partition of unity is subordinate to the cover $\{B_{\rho(x)}(x)\}_{x\in X\setminus K}$ we can find a $\tilde{x} \in X \setminus K$ such that $\operatorname{supp}\phi_\lambda \subset B_{\rho(\tilde{x})}(\tilde{x})$ and hence $\operatorname{diam}(\operatorname{supp}\phi_\lambda) \leq 2\rho(\tilde{x}) \leq \operatorname{dist}(K, B_{\rho(\tilde{x})}(\tilde{x})) \leq \operatorname{dist}(K, \operatorname{supp}\phi_\lambda)$. Putting it all together implies that we have $|x - x_\lambda| \leq 3\operatorname{dist}(K, \operatorname{supp}\phi_\lambda) \leq 3|x_0 - x|$ whenever $x \in \operatorname{supp}\phi_\lambda$ and thus

$$|x_0 - x_\lambda| \leq |x_0 - x| + |x - x_\lambda| \leq 4|x_0 - x| \leq 4\delta$$

as expected. By our choice of δ we have $|F(x_\lambda) - F(x_0)| \leq \varepsilon$ for all λ with $\phi_\lambda(x) \neq 0$. Hence $|F(x) - F(x_0)| \leq \varepsilon$ whenever $|x - x_0| \leq \delta$ and we are done. □

Example 18.3. Consider the nonlinear integral equation

$$x = F(x), \qquad F(x)(t) := \int_0^1 \mathrm{e}^{-ts}\cos(\lambda x(s))ds$$

in $X := C[0,1]$ with $\lambda > 0$. Then one checks that $F \in C(X, X)$ since

$$\begin{aligned} |F(x)(t) - F(y)(t)| &\leq \int_0^1 \mathrm{e}^{-ts}|\cos(\lambda x(s)) - \cos(\lambda y(s))|ds \\ &\leq \int_0^1 \mathrm{e}^{-ts}\lambda|x(s) - y(s)|ds \leq \lambda\|x - y\|_\infty. \end{aligned}$$

In particular, for $\lambda < 1$ we have a contraction and the contraction principle gives us existence of a unique fixed point. Moreover, proceeding similarly, one obtains estimates for the norm of $F(x)$ and its derivative:

$$\|F(x)\|_\infty \leq 1, \qquad \|F(x)'\|_\infty \leq 1.$$

Hence the Arzelà–Ascoli theorem (Theorem B.39) implies that the image of F is a compact subset of the unit ball and hence $F \in \mathcal{C}(\bar{B}_1(0), \bar{B}_1(0))$. Thus the Schauder fixed point theorem guarantees a fixed point for all $\lambda > 0$. ⋄

Finally, let us prove another fixed point theorem which covers several others as special cases.

Theorem 18.11. *Let $U \subset X$ be open and bounded and let $F \in \mathcal{C}(\overline{U}, X)$. Suppose there is an $x_0 \in U$ such that*

$$F(x) - x_0 \neq \alpha(x - x_0), \qquad x \in \partial U,\ \alpha \in (1, \infty). \tag{18.7}$$

Then F has a fixed point.

Proof. Consider $H(t,x) := x - x_0 - t(F(x) - x_0)$, then we have $H(t,x) \neq 0$ for $x \in \partial U$ and $t \in [0,1]$ by assumption. If $H(1,x) = 0$ for some $x \in \partial U$, then x is a fixed point and we are done. Otherwise we have $\deg(\mathbb{I} - F, U, 0) = \deg(\mathbb{I} - x_0, U, 0) = \deg(\mathbb{I}, U, x_0) = 1$ and hence F has a fixed point. □

Now we come to the anticipated corollaries.

Corollary 18.12. *Let $F \in \mathcal{C}(\bar{B}_\rho(0), X)$. Then F has a fixed point if one of the following conditions holds.*

(i) *$F(\partial B_\rho(0)) \subseteq \bar{B}_\rho(0)$ (Rothe).*

(ii) *$|F(x) - x|^2 \geq |F(x)|^2 - |x|^2$ for $x \in \partial B_\rho(0)$ (Altman).*

(iii) *X is a Hilbert space and $\langle F(x), x\rangle \leq |x|^2$ for $x \in \partial B_\rho(0)$ (Krasnosel'skii).*

Proof. Our strategy is to verify (18.7) with $x_0 = 0$. (i). $F(\partial B_\rho(0)) \subseteq \bar{B}_\rho(0)$ and $F(x) = \alpha x$ for $|x| = \rho$ implies $|\alpha|\rho \leq \rho$ and hence (18.7) holds. (ii). $F(x) = \alpha x$ for $|x| = \rho$ implies $(\alpha - 1)^2\rho^2 \geq (\alpha^2 - 1)\rho^2$ and hence $\alpha \leq 1$. (iii). Special case of (ii) since $|F(x) - x|^2 = |F(x)|^2 - 2\langle F(x), x\rangle + |x|^2$. □

18.5. Applications to integral and differential equations

In this section we want to show how our results can be applied to integral and differential equations. To be able to apply our results we will need to know that certain integral operators are compact.

Lemma 18.13. *Suppose $I = [a,b] \subset \mathbb{R}$ and $f \in C(I \times I \times \mathbb{R}^n, \mathbb{R}^n)$, $\tau \in C(I, I)$, then*

$$\begin{aligned} F: \; C(I,\mathbb{R}^n) &\to C(I,\mathbb{R}^n) \\ x(t) &\mapsto F(x)(t) = \textstyle\int_a^{\tau(t)} f(t,s,x(s))ds \end{aligned} \tag{18.8}$$

is compact.

Proof. We first need to prove that F is continuous. Fix $x_0 \in C(I,\mathbb{R}^n)$ and $\varepsilon > 0$. Set $\rho := \|x_0\|_\infty + 1$ and abbreviate $\bar{B} = \bar{B}_\rho(0) \subset \mathbb{R}^n$. The function f is uniformly continuous on $Q := I \times I \times \bar{B}$ since Q is compact. Hence for $\varepsilon_1 := \varepsilon/(b-a)$ we can find a $\delta \in (0,1]$ such that $|f(t,s,x) - f(t,s,y)| \leq \varepsilon_1$ for $|x - y| < \delta$. But this implies

$$\begin{aligned} \|F(x) - F(x_0)\|_\infty &\leq \sup_{t\in I} \int_a^{\tau(t)} |f(t,s,x(s)) - f(t,s,x_0(s))|ds \\ &\leq \sup_{t\in I} (b-a)\varepsilon_1 = \varepsilon, \end{aligned}$$

for $\|x - x_0\|_\infty < \delta$. In other words, F is continuous. Next we note that if $U \subset C(I, \mathbb{R}^n)$ is bounded, say $U \subset \bar{B}_\rho(0)$, then

$$\|F(x)\|_\infty \le \sup_{x\in U} \left| \int_a^{\tau(t)} f(t, s, x(s))ds \right| \le (b - a)M, \quad x \in U,$$

where $M := \max |f(I, I, \bar{B})|$. Moreover, the family $F(U)$ is equicontinuous. Fix ε and $\varepsilon_1 := \varepsilon/(2(b - a))$, $\varepsilon_2 := \varepsilon/(2M)$. Since f and τ are uniformly continuous on $I \times I \times \bar{B}$ and I, respectively, we can find a $\delta > 0$ such that $|f(t, s, x) - f(t_0, s, x)| \le \varepsilon_1$ and $|\tau(t) - \tau(t_0)| \le \varepsilon_2$ for $|t - t_0| < \delta$. Hence we infer for $|t - t_0| < \delta$

$$\begin{aligned} |F(x)(t) - F(x)(t_0)| &= \left| \int_a^{\tau(t)} f(t, s, x(s))ds - \int_a^{\tau(t_0)} f(t_0, s, x(s))ds \right| \\ &\le \int_a^{\tau(t_0)} |f(t, s, x(s)) - f(t_0, s, x(s))|ds + \left| \int_{\tau(t_0)}^{\tau(t)} |f(t, s, x(s))|ds \right| \\ &\le (b - a)\varepsilon_1 + \varepsilon_2 M = \varepsilon. \end{aligned}$$

This implies that $F(U)$ is relatively compact by the Arzelà–Ascoli theorem (Theorem B.39). Thus F is compact. □

As a first application we use this result to show existence of solutions to integral equations.

Theorem 18.14. *Let F be as in the previous lemma. Then the integral equation*

$$x - \lambda F(x) = y, \qquad \lambda \in \mathbb{R},\ y \in C(I, \mathbb{R}^n) \tag{18.9}$$

has at least one solution $x \in C(I, \mathbb{R}^n)$ if $|\lambda| \le \frac{\rho}{(b-a)M(\rho)}$, where $M(\rho) = \max_{(s,t,x)\in I\times I\times \bar{B}_\rho(0)} |f(s, t, x - y(s))|$ and $\rho > 0$ is arbitrary.

Proof. Note that, by our assumption on λ, $\lambda F + y$ maps $\bar{B}_\rho(y)$ into itself. Now apply the Schauder fixed point theorem. □

This result immediately gives the Peano theorem for ordinary differential equations.

Theorem 18.15 (Peano)**.** *Consider the initial value problem*

$$\dot{x} = f(t, x), \quad x(t_0) = x_0, \tag{18.10}$$

where $f \in C(I\times U, \mathbb{R}^n)$ and $I \subset \mathbb{R}$ is an interval containing t_0. Then (18.10) *has at least one local solution $x \in C^1([t_0-\varepsilon, t_0+\varepsilon], \mathbb{R}^n)$, $\varepsilon > 0$. For example, any ε satisfying $\varepsilon M(\varepsilon, \rho) \le \rho$, $\rho > 0$ with $M(\varepsilon, \rho) := \max |f([t_0 - \varepsilon, t_0 + \varepsilon] \times \bar{B}_\rho(x_0))|$ works. In addition, if $M(\varepsilon, \rho) \le \tilde{M}(\varepsilon)(1 + \rho)$, then there exists a global solution.*

Proof. For notational simplicity we make the shift $t \to t - t_0$, $x \to x - x_0$, $f(t, x) \to f(t + t_0, x + t_0)$ and assume $t_0 = 0$, $x_0 = 0$. In addition, it suffices to consider $t \geq 0$ since $t \to -t$ amounts to $f \to -f$.

Now observe, that (18.10) is equivalent to

$$x(t) - \int_0^t f(s, x(s))ds = 0, \qquad x \in C([0, \varepsilon], \mathbb{R}^n)$$

and the first part follows from our previous theorem. To show the second, fix $\varepsilon > 0$ and assume $M(\varepsilon, \rho) \leq \tilde{M}(\varepsilon)(1 + \rho)$. Then

$$|x(t)| \leq \int_0^t |f(s, x(s))|ds \leq \tilde{M}(\varepsilon) \int_0^t (1 + |x(s)|)ds$$

implies $|x(t)| \leq \exp(\tilde{M}(\varepsilon)\varepsilon)$ by Gronwall's inequality (Problem 18.1). Hence we have an a priori bound which implies existence by the Leary–Schauder principle. Since ε was arbitrary we are done. □

As another example we look at the stationary Navier–Stokes equation. Our goal is to use the Leray–Schauder principle to prove an existence and uniqueness result for solutions.

Let U ($\neq \emptyset$) be an open, bounded, and connected subset of $\mathbb{R}^3$. We assume that U is filled with an incompressible fluid described by its velocity field $v(t, x) \in \mathbb{R}^3$ and its pressure $p(t, x) \in \mathbb{R}$ at time $t \in \mathbb{R}$ and at a point $x \in U$. The requirement that the fluid is incompressible implies $\nabla \cdot v = 0$ (here we use a dot to emphasize a scalar product in $\mathbb{R}^3$), which follows from the Gauss theorem since the flux trough any closed surface must be zero. Moreover, the outer force density (force per volume) will be denoted by $K(x) \in \mathbb{R}^3$ and is assumed to be known (e.g. gravity).

Then the **Navier–Stokes equation** governing the motion of the fluid reads

$$\rho \partial_t v = \eta \Delta v - \rho(v \cdot \nabla)v - \nabla p + K, \tag{18.11}$$

where $\eta > 0$ is the viscosity constant and $\rho > 0$ is the density of the fluid. In addition to the incompressibility condition $\nabla v = 0$ we also require the boundary condition $v|_{\partial U} = 0$, which follows from experimental observations.

In what follows we will only consider the stationary Navier–Stokes equation

$$0 = \eta \Delta v - \rho(v \cdot \nabla)v - \nabla p + K. \tag{18.12}$$

Our first step is to switch to a weak formulation and rewrite this equation in integral form, which is more suitable for our further analysis. We pick as underlying Hilbert space $H_0^1(U, \mathbb{R}^3)$ with scalar product

$$\langle u, v \rangle = \sum_{i,j=1}^{3} \int_U (\partial_j u_i)(\partial_j v_i)dx. \tag{18.13}$$

Recall that by the Poincaré inequality (Theorem 13.25) the corresponding norm is equivalent to the usual one. In order to take care of the incompressibility condition we will choose

$$\mathcal{X} := \{v \in H^1_0(U, \mathbb{R}^3) | \nabla \cdot v = 0\}. \tag{18.14}$$

as our configuration space (check that this is a closed subspace of $H^1_0(U, \mathbb{R}^3)$).

Now we multiply (18.12) by $w \in \mathcal{X}$ and integrate over U

$$\begin{aligned}\int_U \Big(\eta \Delta v - \rho(v \cdot \nabla)v - K\Big) \cdot w \, d^3x &= \int_U (\nabla p) \cdot w \, d^3x \\ &= \int_U p(\nabla w) d^3x = 0,\end{aligned} \tag{18.15}$$

where we have used integration by parts (Lemma 13.4 (iii)) to conclude that the pressure term drops out of our picture. Using further integration by parts we finally arrive at the weak formulation of the stationary Navier–Stokes equation

$$\eta \langle v, w \rangle - a(v, v, w) - \int_U K \cdot w \, d^3x = 0, \quad \text{for all } w \in \mathcal{X}, \tag{18.16}$$

where

$$a(u, v, w) := \sum_{j,k=1}^{3} \int_U u_k v_j (\partial_k w_j) \, d^3x. \tag{18.17}$$

In other words, (18.16) represents a necessary solubility condition for the Navier–Stokes equations and a solution of (18.16) will also be called a **weak solution**. If we can show that a weak solution is in $H^2(U, \mathbb{R}^3)$, then we can undo the integration by parts and obtain again (18.15). Since the integral on the left-hand side vanishes for all $w \in \mathcal{X}$, one can conclude that the expression in parenthesis must be the gradient of some function $p \in L^2(U)$ and hence one recovers the original equation. In particular, note that p follows from v up to a constant if U is connected.

For later use we note

$$\begin{aligned}a(v, v, v) &= \sum_{j,k} \int_U v_k v_j (\partial_k v_j) \, d^3x = \frac{1}{2} \sum_{j,k} \int_U v_k \partial_k (v_j v_j) \, d^3x \\ &= -\frac{1}{2} \sum_{j,k} \int_U (v_j v_j) \partial_k v_k \, d^3x = 0, \qquad v \in \mathcal{X}.\end{aligned} \tag{18.18}$$

We proceed by studying (18.16). Let $K \in L^2(U, \mathbb{R}^3)$, then $\int_U K \cdot w \, d^3x$ is a bounded linear functional on $\mathcal{X}$ and hence there is a $\tilde{K} \in \mathcal{X}$ such that

$$\int_U K \cdot w \, d^3x = \langle \tilde{K}, w \rangle, \qquad w \in \mathcal{X}. \tag{18.19}$$

Moreover, applying the Cauchy–Schwarz inequality twice to each summand in $a(u,v,w)$ we see

$$\begin{aligned}|a(u,v,w)| &\le \sum_{j,k} \Big(\int_U (u_k v_j)^2 dx \Big)^{1/2} \Big(\int_U (\partial_k w_j)^2 dx \Big)^{1/2} \\ &\le \|w\| \sum_{j,k} \Big(\int_U (u_k)^4 dx \Big)^{1/4} \Big(\int_U (v_j)^4 dx \Big)^{1/4} = \|u\|_4 \|v\|_4 \|w\|. \end{aligned} \tag{18.20}$$

Since by the Gagliardo–Nierenberg–Sobolev inequality (Theorem 13.13) there is a continous embedding $L^4(U,\mathbb{R}^3) \hookrightarrow H^1(U,\mathbb{R}^3)$ (which is in this context also known as Ladyzhenskaya inequality), the map $a(u,v,.)$ is a bounded linear functional in $\mathcal{X}$ whenever $u,v \in \mathcal{X}$, and hence there is an element $B(u,v) \in \mathcal{X}$ such that

$$a(u,v,w) = \langle B(u,v), w\rangle, \qquad w \in \mathcal{X}. \tag{18.21}$$

In addition, the map $B : \mathcal{X}^2 \to \mathcal{X}$ is bilinear and bounded $\|B(u,v)\| \le \|u\|_4 \|v\|_4$. In summary we obtain

$$\langle \eta v - B(v,v) - \tilde{K}, w\rangle = 0, \qquad w \in \mathcal{X}, \tag{18.22}$$

and hence

$$\eta v - B(v,v) = \tilde{K}. \tag{18.23}$$

So in order to apply the theory from our previous chapter, we choose the Banach space $Y := L^4(U,\mathbb{R}^3)$ such that $\mathcal{X} \hookrightarrow Y$ is compact by the Rellich–Kondrachov theorem (Theorem 13.22).

Motivated by this analysis we formulate the following theorem which implies existence of weak solutions and uniqueness for sufficiently small outer forces.

Theorem 18.16. *Let $\mathcal{X}$ be a Hilbert space, Y a Banach space, and suppose there is a compact embedding $\mathcal{X} \hookrightarrow Y$. In particular, $\|u\|_Y \le \beta\|u\|$. Let $a : \mathcal{X}^3 \to \mathbb{R}$ be a multilinear form such that*

$$|a(u,v,w)| \le \alpha \|u\|_Y \|v\|_Y \|w\| \tag{18.24}$$

and $a(v,v,v) = 0$. Then for any $\tilde{K} \in \mathcal{X}$, $\eta > 0$ we have a solution $v \in \mathcal{X}$ to the problem

$$\eta\langle v,w\rangle - a(v,v,w) = \langle \tilde{K}, w\rangle, \qquad w \in \mathcal{X}. \tag{18.25}$$

Moreover, if $2\alpha\beta\|\tilde{K}\| < \eta^2$ this solution is unique.

Proof. It is no loss to set $\eta = 1$. Arguing as before we see that our equation is equivalent to

$$v - B(v,v) + \tilde{K} = 0,$$

where our assumption (18.24) implies

$$\|B(u,v)\| \le \alpha\|u\|_Y\|v\|_Y \le \alpha\beta^2\|u\|\|v\|$$

Here the second equality follows since the embedding $\mathcal{X} \hookrightarrow Y$ is continuous.

Abbreviate $F(v) = B(v,v)$. Observe that F is locally Lipschitz continuous since if $\|u\|, \|v\| \le \rho$ we have

$$\begin{aligned}\|F(u) - F(v)\| &= \|B(u-v,u) - B(v,u-v)\| \le 2\alpha\beta\rho\|u-v\|_Y \\ &\le 2\alpha\beta^2\rho\|u-v\|.\end{aligned}$$

Moreover, let v_n be a bounded sequence in $\mathcal{X}$. After passing to a subsequence we can assume that v_n is Cauchy in Y and hence $F(v_n)$ is Cauchy in $\mathcal{X}$ by $\|F(u) - F(v)\| \le 2\alpha\,\rho\|u-v\|_Y$. Thus $F : \mathcal{X} \to \mathcal{X}$ is compact.

Hence all we need to apply the Leray–Schauder principle is an a priori estimate. Suppose v solves $v = tF(v) + t\tilde{K}$, $t \in [0,1]$, then

$$\langle v, v\rangle = t\,a(v,v,v) + t\langle\tilde{K}, v\rangle = t\langle\tilde{K}, v\rangle.$$

Hence $\|v\| \le \|\tilde{K}\|$ is the desired estimate and the Leray–Schauder principle yields existence of a solution.

Now suppose there are two solutions v_i, $i = 1, 2$. By our estimate they satisfy $\|v_i\| \le \|\tilde{K}\|$ and hence $\|v_1 - v_2\| = \|F(v_1) - F(v_2)\| \le 2\alpha\beta^2\,\|\tilde{K}\|\|v_1 - v_2\|$ which is a contradiction if $2\alpha\beta^2\,\|\tilde{K}\| < 1$. □

Problem* 18.1 (Gronwall's inequality)**.** *Let $\alpha \ge 0$ and $\beta, \varphi : [0,T] \to [0,\infty)$ be integrable functions satisfying*

$$\varphi(t) \le \alpha + \int_0^t \beta(s)\varphi(s)ds.$$

Then $\varphi(t) \le \alpha e^{\int_0^t \beta(s)ds}$. (Hint: Differentiate $\log\left(\alpha + \int_0^t \beta(s)\varphi(s)ds\right)$.)

Chapter 19

Monotone maps

19.1. Monotone maps

The Leray–Schauder theory can only be applied to compact perturbations of the identity. If F is not compact, we need different tools. In this section we briefly present another class of maps, namely monotone ones, which allow some progress.

If $F : \mathbb{R} \to \mathbb{R}$ is continuous and we want $F(x) = y$ to have a unique solution for every $y \in \mathbb{R}$, then f should clearly be strictly monotone increasing (or decreasing) and satisfy $\lim_{x\to\pm\infty} F(x) = \pm\infty$. Rewriting these conditions slightly such that they make sense for vector valued functions the analogous result holds.

Lemma 19.1. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies*

$$\lim_{|x|\to\infty} \frac{F(x)x}{|x|} = \infty. \tag{19.1}$$

Then the equation

$$F(x) = y \tag{19.2}$$

has a solution for every $y \in \mathbb{R}^n$. If F is strictly monotone

$$(F(x) - F(y))(x - y) > 0, \qquad x \neq y, \tag{19.3}$$

then this solution is unique.

Proof. Our first assumption implies that $G(x) = F(x) - y$ satisfies $G(x)x = F(x)x - yx > 0$ for $|x|$ sufficiently large. Hence the first claim follows from Problem 17.2. The second claim is trivial. □

Now we want to generalize this result to infinite dimensional spaces. Throughout this chapter, $\mathfrak{H}$ will be a real Hilbert space with scalar product $\langle .,.. \rangle$. A map $F : \mathfrak{H} \to \mathfrak{H}$ is called **monotone** if

$$\langle F(x) - F(y), x - y \rangle \geq 0, \qquad x, y \in \mathfrak{H}, \tag{19.4}$$

strictly monotone if

$$\langle F(x) - F(y), x - y \rangle > 0, \qquad x \neq y \in \mathfrak{H}, \tag{19.5}$$

and finally **strongly monotone** if there is a constant $C > 0$ such that

$$\langle F(x) - F(y), x - y \rangle \geq C\|x - y\|^2, \qquad x, y \in \mathfrak{H}. \tag{19.6}$$

Note that the same definitions can be made for a Banach space X and mappings $F : X \to X^*$.

Observe that if F is strongly monotone, then it is **coercive**

$$\lim_{\|x\| \to \infty} \frac{\langle F(x), x \rangle}{\|x\|} = \infty. \tag{19.7}$$

(Just take $y = 0$ in the definition of strong monotonicity.) Hence the following result is not surprising.

Theorem 19.2 (Zarantonello)**.** *Suppose $F \in C(\mathfrak{H}, \mathfrak{H})$ is (globally) Lipschitz continuous and strongly monotone. Then, for each $y \in \mathfrak{H}$ the equation*

$$F(x) = y \tag{19.8}$$

has a unique solution $x(y) \in \mathfrak{H}$ which depends continuously on y.

Proof. Set

$$G(x) := x - t(F(x) - y), \qquad t > 0,$$

then $F(x) = y$ is equivalent to the fixed point equation

$$G(x) = x.$$

It remains to show that G is a contraction. We compute

$$\begin{aligned} \|G(x) - G(\tilde{x})\|^2 &= \|x - \tilde{x}\|^2 - 2t\langle F(x) - F(\tilde{x}), x - \tilde{x} \rangle + t^2 \|F(x) - F(\tilde{x})\|^2 \\ &\leq (1 - 2\frac{C}{L}(Lt) + (Lt)^2)\|x - \tilde{x}\|^2, \end{aligned}$$

where L is a Lipschitz constant for F (i.e., $\|F(x) - F(\tilde{x})\| \leq L\|x - \tilde{x}\|$). Thus, if $t \in (0, \frac{2C}{L})$, G is a uniform contraction and the rest follows from the uniform contraction principle. □

Again observe that our proof is constructive. In fact, the best choice for t is clearly $t = \frac{C}{L^2}$ such that the contraction constant $\theta = 1 - (\frac{C}{L})^2$ is minimal. Then the sequence

$$x_{n+1} = x_n - \frac{C}{L^2}(F(x_n) - y), \quad x_0 = y, \tag{19.9}$$

converges to the solution.

Example 19.1. Let $A \in \mathscr{L}(\mathfrak{H})$ and consider $F(x) = Ax$. Then the condition

$$\langle Ax, x\rangle \geq C\|x\|^2$$

implies that A has a bounded inverse $A^{-1} : \mathfrak{H} \to \mathfrak{H}$ with $\|A^{-1}\| \leq C^{-1}$ (cf. Problem 1.33). ⋄

19.2. The nonlinear Lax–Milgram theorem

As a consequence of the last theorem we obtain a nonlinear version of the Lax–Milgram theorem. We want to investigate the following problem:

$$a(x, y) = b(y), \qquad \text{for all } y \in \mathfrak{H}, \tag{19.10}$$

where $a : \mathfrak{H}^2 \to \mathbb{R}$ and $b : \mathfrak{H} \to \mathbb{R}$. For this equation the following result holds.

Theorem 19.3 (Nonlinear Lax–Milgram theorem)**.** *Suppose $b \in \mathscr{L}(\mathfrak{H}, \mathbb{R})$ and $a(x, .) \in \mathscr{L}(\mathfrak{H}, \mathbb{R})$, $x \in \mathfrak{H}$, are linear functionals such that there are positive constants L and C such that for all $x, y, z \in \mathfrak{H}$ we have*

$$a(x, x - y) - a(y, x - y) \geq C\|x - y\|^2 \tag{19.11}$$

and

$$|a(x, z) - a(y, z)| \leq L\|z\|\|x - y\|. \tag{19.12}$$

Then there is a unique $x \in \mathfrak{H}$ such that (19.10) *holds.*

Proof. By the Riesz lemma (Theorem 2.10) there are elements $F(x) \in \mathfrak{H}$ and $z \in \mathfrak{H}$ such that $a(x, y) = b(y)$ is equivalent to $\langle F(x) - z, y\rangle = 0$, $y \in \mathfrak{H}$, and hence to

$$F(x) = z.$$

By (19.11) the map F is strongly monotone. Moreover, by (19.12) we infer

$$\|F(x) - F(y)\| = \sup_{\tilde{x}\in\mathfrak{H}, \|\tilde{x}\|=1} |\langle F(x) - F(y), \tilde{x}\rangle| \leq L\|x - y\|$$

that F is Lipschitz continuous. Now apply Theorem 19.2. □

The special case where $a \in \mathscr{L}^2(\mathfrak{H}, \mathbb{R})$ is a bounded bilinear form which is strongly coercive, that is,

$$a(x, x) \geq C\|x\|^2, \qquad x \in \mathfrak{H}, \tag{19.13}$$

is usually known as (linear) Lax–Milgram theorem (Theorem 2.16).

Example 19.2. Note that if $b = 0$ we could replace $c(x)u(x)$ in (13.27) by $F(x, u(x))$, where $F : U \times \mathbb{R}^n \to \mathbb{R}$ such that $|F(x, u_1) - F(x, u_2)| \leq L|u_1 - u_2|$ and $(F(x, u_1) - F(x, u_2))(u_1 - u_2) \geq 0$. ⋄

19.3. The main theorem of monotone maps

Now we return to the investigation of $F(x) = y$ and weaken the conditions of Theorem 19.2. We will assume that $\mathfrak{H}$ is a separable Hilbert space and that $F : \mathfrak{H} \to \mathfrak{H}$ is a continuous, coercive monotone map. In fact, it suffices to assume that F is **demicontinuous**

$$\lim_{n\to\infty} \langle F(x_n), y\rangle = \langle F(x), y\rangle, \qquad \text{for all } y \in \mathfrak{H} \tag{19.14}$$

whenever $x_n \to x$.

The idea is as follows: Start with a finite dimensional subspace $\mathfrak{H}_n \subset \mathfrak{H}$ and project the equation $F(x) = y$ to $\mathfrak{H}_n$ resulting in an equation

$$F_n(x_n) = y_n, \qquad x_n, y_n \in \mathfrak{H}_n. \tag{19.15}$$

More precisely, let P_n be the (linear) projection onto $\mathfrak{H}_n$ and set $F_n(x_n) = P_nF(x_n)$, $y_n = P_ny$ (verify that F_n is continuous and monotone!).

Now Lemma 19.1 ensures that there exists a solution x_n. Now chose the subspaces $\mathfrak{H}_n$ such that $\mathfrak{H}_n \to \mathfrak{H}$ (i.e., $\mathfrak{H}_n \subset \mathfrak{H}_{n+1}$ and $\bigcup_{n=1}^{\infty} \mathfrak{H}_n$ is dense). Then our hope is that x_n converges to a solution x.

This approach is quite common when solving equations in infinite dimensional spaces and is known as **Galerkin approximation**. It can often be used for numerical computations and the right choice of the spaces $\mathfrak{H}_n$ will have a significant impact on the quality of the approximation.

So how should we show that x_n converges? First of all observe that our construction of x_n shows that x_n lies in some ball with radius R_n, which is chosen such that

$$\langle F_n(x), x\rangle > \|y_n\|\|x\|, \qquad \|x\| \geq R_n,\ x \in \mathfrak{H}_n. \tag{19.16}$$

Since $\langle F_n(x), x\rangle = \langle P_nF(x), x\rangle = \langle F(x), P_nx\rangle = \langle F(x), x\rangle$ for $x \in \mathfrak{H}_n$ we can drop all n's to obtain a constant R (depending on $\|y\|$) which works for all n. So the sequence x_n is uniformly bounded

$$\|x_n\| \leq R. \tag{19.17}$$

Now by Theorem 4.30 there exists a weakly convergent subsequence. That is, after dropping some terms, we can assume that there is some x such that $x_n \rightharpoonup x$, that is,

$$\langle x_n, z\rangle \to \langle x, z\rangle, \qquad \text{for every } z \in \mathfrak{H}. \tag{19.18}$$

And it remains to show that x is indeed a solution. This follows from

Lemma 19.4. *Suppose $F : \mathfrak{H} \to \mathfrak{H}$ is demicontinuous and monotone, then*

$$\langle y - F(z), x - z\rangle \geq 0 \quad \textit{for every } z \in \mathfrak{H} \tag{19.19}$$

implies $F(x) = y$.

Proof. Choose $z = x \pm tw$, then $\mp\langle y - F(x \pm tw), w\rangle \geq 0$ and by continuity $\mp\langle y - F(x), w\rangle \geq 0$. Thus $\langle y - F(x), w\rangle = 0$ for every w implying $y - F(x) = 0$. □

Now we can show

Theorem 19.5 (Browder–Minty)**.** *Let $\mathfrak{H}$ be a separable Hilbert space. Suppose $F : \mathfrak{H} \to \mathfrak{H}$ is demicontinuous, coercive and monotone. Then the equation*

$$F(x) = y \tag{19.20}$$

has a solution for every $y \in \mathfrak{H}$. If F is strictly monotone then this solution is unique.

Proof. We have $\langle y - F(z), x_n - z\rangle = \langle y_n - F_n(z), x_n - z\rangle \geq 0$ for $z \in \mathfrak{H}_n$. Taking the limit implies $\langle y - F(z), x - z\rangle \geq 0$ for every $z \in \mathfrak{H}_\infty = \bigcup_{n=1}^{\infty} \mathfrak{H}_n$. Since $\mathfrak{H}_\infty$ is dense, $\langle y - F(z), x - z\rangle \geq 0$ for every $z \in \mathfrak{H}$ by continuity and hence $F(x) = y$ by our lemma. □

Note that in the infinite dimensional case we need monotonicity even to show existence. Moreover, this result can be further generalized in two more ways. First of all, the Hilbert space $\mathfrak{H}$ can be replaced by a reflexive Banach space X if $F : X \to X^*$. The proof is similar. Secondly, it suffices if

$$t \mapsto \langle F(x + ty), z\rangle \tag{19.21}$$

is continuous for $t \in [0, 1]$ and all $x, y, z \in \mathfrak{H}$, since this condition together with monotonicity can be shown to imply demicontinuity.

Appendix A

Some set theory

At the beginning of the 20th century Russel showed with his famous paradox that naive set theory can lead into contradictions. Hence it was replaced by **axiomatic set theory**, more specific we will take the **Zermelo–Fraenkel set theory** (**ZF**), which assumes existence of some sets (like the empty set and the integers) and defines what operations are allowed. Somewhat informally (i.e. without writing them using the symbolism of first order logic) they can be stated as follows:

- **Axiom of empty set**. There is a set $\emptyset$ which contains no elements.
- **Axiom of extensionality**. Two sets A and B are equal $A = B$ if they contain the same elements. If a set A contains all elements from a set B, it is called a subset $A \subseteq B$. In particular $A \subseteq B$ and $B \subseteq A$ if and only if $A = B$.

The last axiom implies that the empty set is unique and that any set which is not equal to the empty set has at least one element.

- **Axiom of pairing**. If A and B are sets, then there exists a set $\{A, B\}$ which contains A and B as elements. One writes $\{A, A\} = \{A\}$. By the axiom of extensionality we have $\{A, B\} = \{B, A\}$.
- **Axiom of union**. Given a set $\mathcal{F}$ whose elements are again sets, there is a set $A = \bigcup \mathcal{F}$ containing every element that is a member of some member of $\mathcal{F}$. In particular, given two sets A, B there exists a set $A \cup B = \bigcup\{A, B\}$ consisting of the elements of both sets. Note that this definition ensures that the union is commutative $A \cup B = B \cup A$ and associative $(A \cup B) \cup C = A \cup (B \cup C)$. Note also $\bigcup\{A\} = A$.

- **Axiom schema of specification**. Given a set A and a logical statement $\phi(x)$ depending on $x \in A$ we can form the set $B = \{x \in A | \phi(x)\}$ of all elements from A obeying ϕ. For example, given two sets A and B we can define their intersection as $A \cap B = \{x \in A \cup B | (x \in A) \wedge (x \in B)\}$ and their complement as $A \setminus B = \{x \in A | x \notin B\}$. Or the intersection of a family of sets $\mathcal{F}$ as $\bigcap \mathcal{F} = \{x \in \bigcup \mathcal{F} | \forall F \in \mathcal{F} : x \in F\}$.
- **Axiom of power set**. For any set A, there is a **power set** $\mathfrak{P}(A)$ that contains every subset of A.

From these axioms one can define ordered pairs as $(x, y) = \{\{x\}, \{x, y\}\}$ and the Cartesian product as $A \times B = \{z \in \mathfrak{P}(A \cup \mathfrak{P}(A \cup B)) | \exists x \in A, y \in B : z = (x, y)\}$. Functions $f : A \to B$ are defined as single valued relations, that is $f \subseteq A \times B$ such that $(x, y) \in f$ and $(x, \tilde{y}) \in f$ implies $y = \tilde{y}$.

- **Axiom schema of replacement**. For every function f the image of a set A is again a set $B = \{f(x) | x \in A\}$.

So far the previous axioms were concerned with ensuring that the usual set operations required in mathematics yield again sets. In particular, we can start constructing sets with any given finite number of elements starting from the empty set: $\emptyset$ (no elements), $\{\emptyset\}$ (one element), $\{\emptyset, \{\emptyset\}\}$ (two elements), etc. However, while existence of infinite sets (like e.g. the integers) might seem *obvious* at this point, it cannot be deduced from the axioms we have so far. Hence it has to be added as well.

- **Axiom of infinity**. There exists a set A which contains the empty set and for every element $x \in A$ we also have $x \cup \{x\} \in A$. The smallest such set $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \dots\}$ can be identified with the integers via $0 = \emptyset$, $1 = \{\emptyset\}$, $2 = \{\emptyset, \{\emptyset\}\}$, $\dots$

Now we finally have the integers and thus everything we need to start constructing the rational, real, and complex numbers in the usual way. Hence we only add one more axiom to exclude some pathological objects which will lead to contradictions.

- **Axiom of Regularity**. Every nonempty set A contains an element x with $x \cap A = \emptyset$. This excludes for example the possibility that a set contains itself as an element (apply the axiom to $\{A\}$). Similarly, we can only have $A \in B$ or $B \in A$ but not both (apply it to the set $\{A, B\}$).

Hence a set is something which can be constructed from the above axioms. Of course this raises the question if these axioms are consistent but as has been shown by Gödel this question cannot be answered: If ZF contains a statement of its own consistency then ZF is inconsistent. In fact, the

same holds for any other sufficiently rich (such that one can do basic math) system of axioms. In particular, it also holds for ZFC defined below. So we have to live with the fact that someday someone might come and prove that ZFC is inconsistent.

Starting from ZF one can develop basic analysis (including the construction of the real numbers). However, it turns out that several fundamental results require yet another construction for their proof:

Given an index set A and for every $\alpha \in A$ some set M_α the product $\bigtimes_{\alpha\in A} M_\alpha$ is defined to be the set of all functions $\varphi : A \to \bigcup_{\alpha\in A} M_\alpha$ which assign each element $\alpha \in A$ some element $m_\alpha \in M_\alpha$. If all sets M_α are nonempty it seems quite reasonable that there should be such a *choice function* which chooses an element from M_α for every $\alpha \in A$. However, no matter how obvious this might seem, it cannot be deduced from the ZF axioms alone and hence has to be added:

- **Axiom of Choice**: Given an index set A and nonempty sets $\{M_\alpha\}_{\alpha\in A}$ their product $\bigtimes_{\alpha\in A} M_\alpha$ is nonempty.

ZF augmented by the axiom of choice is known as **ZFC** and we accept it as the fundament upon which our functional analytic house is built.

Note that the axiom of choice is not only used to ensure that infinite products are nonempty but also in many proofs! For example, suppose you start with a set M_1 and recursively construct some sets M_n such that in every step you have a nonempty set. Then the axiom of choice guarantees the existence of a sequence $x = (x_n)_{n\in\mathbb{N}}$ with $x_n \in M_n$.

The axiom of choice has many important consequences (many of which are in fact equivalent to the axiom of choice and it is hence a matter of taste which one to choose as axiom).

A **partial order** is a binary relation "$\preceq$" over a set $\mathcal{P}$ such that for all $A, B, C \in \mathcal{P}$:

- $A \preceq A$ (reflexivity),
- if $A \preceq B$ and $B \preceq A$ then $A = B$ (antisymmetry),
- if $A \preceq B$ and $B \preceq C$ then $A \preceq C$ (transitivity).

It is custom to write $A \prec B$ if $A \preceq B$ and $A \neq B$.

Example A.1. Let $\mathcal{P}(X)$ be the collections of all subsets of a set X. Then $\mathcal{P}$ is partially ordered by inclusion $\subseteq$. ⋄

It is important to emphasize that two elements of $\mathcal{P}$ need not be comparable, that is, in general neither $A \preceq B$ nor $B \preceq A$ might hold. However, if any two elements are comparable, $\mathcal{P}$ will be called **totally ordered**. A set with a total order is called **well-ordered** if every nonempty subset has

a **least element**, that is some $A \in \mathcal{P}$ with $A \preceq B$ for every $B \in \mathcal{P}$. Note that the least element is unique by antisymmetry.

Example A.2. $\mathbb{R}$ with $\leq$ is totally ordered and $\mathbb{N}$ with $\leq$ is well-ordered. $\diamond$

On every well-ordered set we have the

Theorem A.1 (Induction principle)**.** *Let K be well ordered and let $S(k)$ be a statement for arbitrary $k \in K$. Then, if $A(l)$ true for all $l \prec k$ implies $A(k)$ true, then $A(k)$ is true for all $k \in K$.*

Proof. Otherwise the set of all k for which $A(k)$ is false had a least element k_0. But by our choice of k_0, $A(l)$ holds for all $l \prec k_0$ and thus for k_0 contradicting our assumption. $\square$

The induction principle also shows that in a well-ordered set functions f can be defined recursively, that is, by a function φ which computes the value of $f(k)$ from the values $f(l)$ for all $l \prec k$. Indeed, the induction principle implies that on the set $M_k = \{l \in K | l \prec k\}$ there is at most one such function f_k. Since k is arbitrary, f is unique. In case of the integers existence of f_k is also clear provided $f(1)$ is given. In general, one can prove existence provided f_k is given for some k but we will not need this.

If $\mathcal{P}$ is partially ordered, then every totally ordered subset is also called a **chain**. If $\mathcal{Q} \subseteq \mathcal{P}$, then an element $M \in \mathcal{P}$ satisfying $A \preceq M$ for all $A \in \mathcal{Q}$ is called an **upper bound**.

Example A.3. Let $\mathcal{P}(X)$ as before. Then a collection of subsets $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{P}(X)$ satisfying $A_n \subseteq A_{n+1}$ is a chain. The set $\bigcup_n A_n$ is an upper bound. $\diamond$

An element $M \in \mathcal{P}$ for which $M \preceq A$ for some $A \in \mathcal{P}$ is only possible if $M = A$ is called a **maximal element**.

Theorem A.2 (Zorn's lemma)**.** *Every partially ordered set in which every chain has an upper bound contains at least one maximal element.*

Proof. Suppose it were false. Then to every chain C we can assign an element $m(C)$ such that $m(C) \succ x$ for all $x \in C$ (here we use the axiom of choice). We call a chain C distinguished if it is well-ordered and if for every segment $C_x = \{y \in C | y \prec x\}$ we have $m(C_x) = x$. We will also regard C as a segment of itself.

Then (since for the least element of C we have $C_x = \emptyset$) every distinguished chain must start like $m(\emptyset) \prec m(m(\emptyset)) \prec \cdots$ and given two segments C, D we expect that always one must be a segment of the other.

So let us first prove this claim. Suppose D is not a segment of C. Then we need to show $C = D_z$ for some z. We start by showing that $x \in C$ implies $x \in D$ and $C_x = D_x$. To see this suppose it were wrong and let

x be the least $x \in C$ for which it fails. Then $y \in K_x$ implies $y \in L$ and hence $K_x \subset L$. Then, since $C_x \neq D$ by assumption, we can find a least $z \in D \setminus C_x$. In fact we must even have $z \succ C_x$ since otherwise we could find a $y \in C_x$ such that $x \succ y \succ z$. But then, using that it holds for y, $y \in D$ and $C_y = D_y$ so we get the contradiction $z \in D_y = C_y \subset C_x$. So $z \succ C_x$ and thus also $C_x = D_z$ which in turn shows $x = m(C_x) = m(D_z) = z$ and proves that $x \in C$ implies $x \in D$ and $C_x = D_x$. In particular $C \subset D$ and as before $C = D_z$ for the least $z \in D \setminus C$. This proves the claim.

Now using this claim we see that we can take the union over all distinguished chains to get a maximal distinguished chain C_{max}. But then we could add $m(C_{max}) \notin C_{max}$ to C_{max} to get a larger distinguished chain $C_{max} \cup \{m(C_{max})\}$ contradicting maximality. □

We will also frequently use the **cardinality** of sets: Two sets A and B have the same cardinality, written as $|A| = |B|$, if there is a bijection $\varphi : A \to B$. We write $|A| \leq |B|$ if there is an injective map $\varphi : A \to B$. Note that $|A| \leq |B|$ and $|B| \leq |C|$ implies $|A| \leq |C|$. A set A is called infinite if $|A| \geq |\mathbb{N}|$, countable if $|A| \leq |\mathbb{N}|$, and countably infinite if $|A| = |\mathbb{N}|$.

Theorem A.3 (Schröder–Bernstein)**.** $|A| \leq |B|$ *and* $|B| \leq |A|$ *implies* $|A| = |B|$.

Proof. Suppose $\varphi : A \to B$ and $\psi : B \to A$ are two injective maps. Now consider sequences x_n defined recursively via $x_{2n+1} = \varphi(x_{2n})$ and $x_{2n+1} = \psi(x_{2n})$. Given a start value $x_0 \in A$ the sequence is uniquely defined but might terminate at a negative integer since our maps are not surjective. In any case, if an element appears in two sequences, the elements to the left and to the right must also be equal (use induction) and hence the two sequences differ only by an index shift. So the ranges of such sequences form a partition for $A \dot\cup B$ and it suffices to find a bijection between elements in one partition. If the sequence stops at an element in A we can take φ. If the sequence stops at an element in B we can take ψ^{-1}. If the sequence is doubly infinite either of the previous choices will do. □

Theorem A.4 (Zermelo)**.** *Either* $|A| \leq |B|$ *or* $|B| \leq |A|$.

Proof. Consider the set of all bijective functions $\varphi_\alpha : A_\alpha \to B$ with $A_\alpha \subseteq A$. Then we can define a partial ordering via $\varphi_\alpha \preceq \varphi_\beta$ if $A_\alpha \subseteq A_\beta$ and $\varphi_\beta|_{A_\alpha} = \varphi_\alpha$. Then every chain has an upper bound (the unique function defined on the union of all domains) and by Zorn's lemma there is a maximal element φ_m. For φ_m we have either $A_m = A$ or $\varphi_m(A_m) = B$ since otherwise there is some $x \in A \setminus A_m$ and some $y \in B \setminus f(A_m)$ which could be used to extend φ_m to $A_m \cup \{x\}$ by setting $\varphi(x) = y$. But if $A_m = A$ we have $|A| \leq |B|$ and if $\varphi_m(A_m) = B$ we have $|B| \leq |A|$. □

The cardinality of the power set $\mathfrak{P}(A)$ is strictly larger than the cardinality of A.

Theorem A.5 (Cantor)**.** $|A| < |\mathfrak{P}(A)|$.

Proof. Suppose there were a bijection $\varphi : A \to \mathfrak{P}(A)$. Then, for $B = \{x \in A | x \notin \varphi(x)\}$ there must be some y such that $B = \varphi(y)$. But $y \in B$ if and only if $y \notin \varphi(y) = B$, a contradiction. □

This innocent looking result also caused some grief when announced by Cantor as it clearly gives a contradiction when applied to the *set of all sets* (which is fortunately not a legal object in ZFC).

The following result and its corollary will be used to determine the cardinality of unions and products.

Lemma A.6. *Any infinite set can be written as a disjoint union of countably infinite sets.*

Proof. Consider collections of disjoint countably infinite subsets. Such collections can be partially ordered by inclusion and hence there is a maximal collection by Zorn's lemma. If the union of such a maximal collection falls short of the whole set the complement must be finite. Since this finite reminder can be added to a set of the collection we are done. □

Corollary A.7. *Any infinite set can be written as a disjoint union of two disjoint subsets having the same cardinality as the original set.*

Proof. By the lemma we can write $A = \bigcup A_\alpha$, where all A_α are countably infinite. Now split $A_\alpha = B_\alpha \cup C_\alpha$ into two disjoint countably infinite sets (map A bijective to $\mathbb{N}$ and the split into even and odd elements). Then the desired splitting is $A = B \cup C$ with $B = \bigcup B_\alpha$ and $C = \bigcup C_\alpha$. □

Theorem A.8. *Suppose A or B is infinite. Then $|A \cup B| = \max\{|A|, |B|\}$.*

Proof. Suppose A is infinite and $|B| \leq |A|$. Then $|A| \leq |A \cup B| \leq |A \uplus B| \leq |A \uplus A| = |A|$ by the previous corollary. Here $\uplus$ denotes the disjoint union. □

A standard theorem proven in every introductory course is that $\mathbb{N} \times \mathbb{N}$ is countable. The generalization of this result is also true.

Theorem A.9. *Suppose A is infinite and $B \neq \emptyset$. Then $|A \times B| = \max\{|A|, |B|\}$.*

Proof. Without loss of generality we can assume $|B| \leq |A|$ (otherwise exchange both sets). Then $|A| \leq |A \times B| \leq |A \times A|$ and it suffices to show $|A \times A| = |A|$.

We proceed as before and consider the set of all bijective functions $\varphi_\alpha : A_\alpha \to A_\alpha \times A_\alpha$ with $A_\alpha \subseteq A$ with the same partial ordering as before. By Zorn's lemma there is a maximal element φ_m. Let A_m be its domain and let $A'_m = A \setminus A_m$. We claim that $|A'_m| < |A_m|$. If not, A'_m had a subset A''_m with the same cardinality of A_m and hence we had a bijection from $A''_m \to A''_m \times A''_m$ which could be used to extend φ. So $|A'_m| < |A_m|$ and thus $|A| = |A_m \cup A'_m| = |A_m|$. Since we have shown $|A_m \times A_m| = |A_m|$ the claim follows. □

Example A.4. Note that for $A = \mathbb{N}$ we have $|\mathfrak{P}(\mathbb{N})| = |\mathbb{R}|$. Indeed, since $|\mathbb{R}| = |\mathbb{Z} \times [0,1)| = |[0,1)|$ it suffices to show $|\mathfrak{P}(\mathbb{N})| = |[0,1)|$. To this end note that $\mathfrak{P}(\mathbb{N})$ can be identified with the set of all sequences with values in $\{0,1\}$ (the value of the sequence at a point tells us wether it is in the corresponding subset). Now every point in $[0,1)$ can be mapped to such a sequence via its binary expansion. This map is injective but not surjective since a point can have different binary expansions: $|[0,1)| \leq |\mathfrak{P}(\mathbb{N})|$. Conversely, given a sequence $a_n \in \{0,1\}$ we can map it to the number $\sum_{n=1}^{\infty} a_n 4^{-n}$. Since this map is again injective (note that we avoid expansions which are eventually 1) we get $|\mathfrak{P}(\mathbb{N})| \leq |[0,1)|$. ⋄

Hence we have

$$|\mathbb{N}| < |\mathfrak{P}(\mathbb{N})| = |\mathbb{R}| \tag{A.1}$$

and the **continuum hypothesis** states that there are no sets whose cardinality lie in between. It was shown by Gödel and Cohen that it, as well as its negation, is consistent with ZFC and hence cannot be decided within this framework.

Problem A.1. *Show that Zorn's lemma implies the axiom of choice. (Hint: Consider the set of all partial choice functions defined on a subset.)*

Problem A.2. *Show $|\mathbb{R}^{\mathbb{N}}| = |\mathbb{R}|$. (Hint: Without loss we can replace $\mathbb{R}$ by $(0,1)$ and identify each $x \in (0,1)$ with its decimal expansion. Now the digits in a given sequence are indexed by two countable parameters.)*

Appendix B

Metric and topological spaces

This chapter collects some basic facts from metric and topological spaces as a reference for the main text. I presume that you are familiar with most of these topics from your calculus course. As a general reference I can warmly recommend Kelley's classical book [**24**] or the nice book by Kaplansky [**22**]. As always such a brief compilation introduces a zoo of properties. While sometimes the connection between these properties are straightforward, othertimes they might be quite tricky. So if at some point you are wondering if there exists an infinite multi-variable sub-polynormal Woffle which does not satisfy the lower regular Q-property, start searching in the book by Steen and Seebach [**41**].

B.1. Basics

One of the key concepts in analysis is convergence. To define convergence requires the notion of distance. Motivated by the Euclidean distance one is lead to the following definition:

A **metric space** is a space X together with a distance function $d : X \times X \to [0, \infty)$ such that for arbitrary points $x, y, z \in X$ we have

(i) $d(x, y) = 0$ if and only if $x = y$,

(ii) $d(x, y) = d(y, x)$,

(iii) $d(x, z) \leq d(x, y) + d(y, z)$ (**triangle inequality**).

If (i) does not hold, d is called a **pseudometric**. As a straightforward consequence we record the **inverse triangle inequality** (Problem B.1)

$$|d(x,y) - d(z,y)| \leq d(x,z). \tag{B.1}$$

Example B.1. The role model for a metric space is of course Euclidean space $\mathbb{R}^n$ (or $\mathbb{C}^n$) together with $d(x,y) := (\sum_{k=1}^n |x_k - y_k|^2)^{1/2}$. ⋄

Several notions from $\mathbb{R}^n$ carry over to metric spaces in a straightforward way. The set

$$B_r(x) := \{y \in X | d(x,y) < r\} \tag{B.2}$$

is called an **open ball** around x with radius $r > 0$. We will write $B_r^X(x)$ in case we want to emphasize the corresponding space. A point x of some set $U \subseteq X$ is called an **interior point** of U if U contains some ball around x. If x is an interior point of U, then U is also called a **neighborhood** of x. A point x is called a **limit point** of U (also **accumulation** or **cluster point**) if for any open ball $B_r(x)$, there exists at least one point in $B_r(x) \cap U$ distinct from x. Note that a limit point x need not lie in U, but U must contain points arbitrarily close to x. A point x is called an **isolated point** of U if there exists a neighborhood of x not containing any other points of U. A set which consists only of isolated points is called a **discrete set**. If any neighborhood of x contains at least one point in U and at least one point not in U, then x is called a **boundary point** of U. The set of all boundary points of U is called the boundary of U and denoted by ∂U.

Example B.2. Consider $\mathbb{R}$ with the usual metric and let $U := (-1,1)$. Then every point $x \in U$ is an interior point of U. The points $[-1,1]$ are limit points of U, and the points $\{-1,+1\}$ are boundary points of U.

Let $U := \mathbb{Q}$, the set of rational numbers. Then U has no interior points and $\partial U = \mathbb{R}$. ⋄

A set all of whose points are interior points is called **open**. The family of open sets $\mathcal{O}$ satisfies the properties

(i) $\emptyset, X \in \mathcal{O}$,

(ii) $O_1, O_2 \in \mathcal{O}$ implies $O_1 \cap O_2 \in \mathcal{O}$,

(iii) $\{O_\alpha\} \subseteq \mathcal{O}$ implies $\bigcup_\alpha O_\alpha \in \mathcal{O}$.

That is, $\mathcal{O}$ is closed under finite intersections and arbitrary unions. Indeed, (i) is obvious, (ii) follows since the intersection of two open balls centered at x is again an open ball centered at x (explicitly $B_{r_1}(x) \cap B_{r_2}(x) = B_{\min(r_1,r2)}(x)$), and (iii) follows since every ball contained in one of the sets is also contained in the union.

Now it turns out that for defining convergence, a distance is slightly more than what is actually needed. In fact, it suffices to know when a point

is in the neighborhood of another point. And if we adapt the definition of a neighborhood by requiring it to contain an open set around x, then we see that it suffices to know when a set is open. This motivates the following definition:

A space X together with a family of sets $\mathcal{O}$, the open sets, satisfying (i)–(iii), is called a **topological space**. The notions of interior point, limit point, and neighborhood carry over to topological spaces if we replace open ball around x by open set containing x.

There are usually different choices for the topology. Two not too interesting examples are the **trivial topology** $\mathcal{O} = \{\emptyset, X\}$ and the **discrete topology** $\mathcal{O} = \mathfrak{P}(X)$ (the power set of X). Given two topologies $\mathcal{O}_1$ and $\mathcal{O}_2$ on X, $\mathcal{O}_1$ is called **weaker** (or **coarser**) than $\mathcal{O}_2$ if $\mathcal{O}_1 \subseteq \mathcal{O}_2$. Conversely, $\mathcal{O}_1$ is called **stronger** (or **finer**) than $\mathcal{O}_2$ if $\mathcal{O}_2 \subseteq \mathcal{O}_1$.

Example B.3. Note that different metrics can give rise to the same topology. For example, we can equip $\mathbb{R}^n$ (or $\mathbb{C}^n$) with the Euclidean distance $d(x, y)$ as before or we could also use

$$\tilde{d}(x, y) := \sum_{k=1}^{n} |x_k - y_k|. \tag{B.3}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} |x_k| \le \sqrt{\sum_{k=1}^{n} |x_k|^2} \le \sum_{k=1}^{n} |x_k| \tag{B.4}$$

shows $B_{r/\sqrt{n}}(x) \subseteq \tilde{B}_r(x) \subseteq B_r(x)$, where B, $\tilde{B}$ are balls computed using d, $\tilde{d}$, respectively. In particular, both distances will lead to the same notion of convergence. ⋄

Example B.4. We can always replace a metric d by the bounded metric

$$\tilde{d}(x, y) := \frac{d(x, y)}{1 + d(x, y)} \tag{B.5}$$

without changing the topology (since the family of open balls does not change: $B_\delta(x) = \tilde{B}_{\delta/(1+\delta)}(x)$). To see that $\tilde{d}$ is again a metric, observe that $f(r) = \frac{r}{1+r}$ is monotone as well as concave and hence subadditive, $f(r + s) \le f(r) + f(s)$ (cf. Problem B.3). ⋄

Every subspace Y of a topological space X becomes a topological space of its own if we call $O \subseteq Y$ open if there is some open set $\tilde{O} \subseteq X$ such that $O = \tilde{O} \cap Y$. This natural topology $\mathcal{O} \cap Y$ is known as the **relative topology** (also **subspace**, **trace** or **induced topology**).

Example B.5. The set $(0, 1] \subseteq \mathbb{R}$ is not open in the topology of $X := \mathbb{R}$, but it is open in the relative topology when considered as a subset of $Y := [-1, 1]$. ⋄

A family of open sets $\mathcal{B} \subseteq \mathcal{O}$ is called a **base** for the topology if for each x and each neighborhood $U(x)$, there is some set $O \in \mathcal{B}$ with $x \in O \subseteq U(x)$. Since an open set O is a neighborhood of every one of its points, it can be written as $O = \bigcup_{O \supseteq \tilde{O} \in \mathcal{B}} \tilde{O}$ and we have

Lemma B.1. *A family of open sets $\mathcal{B} \subseteq \mathcal{O}$ is a base for the topology if and only if every open set can be written as a union of elements from $\mathcal{B}$.*

Proof. To see the converse let x and $U(x)$ be given. Then $U(x)$ contains an open set O containing x which can be written as a union of elements from $\mathcal{B}$. One of the elements in this union must contain x and this is the set we are looking for. □

A family of open sets $\mathcal{B} \subseteq \mathcal{O}$ is called a **subbase** for the topology if every open set can be written as a union of finite intersections of elements from $\mathcal{B}$.

Example B.6. The intervals form a base for the topology on $\mathbb{R}$. Slightly more general, the open balls are a base for the topology in a metric space. Intervals of the form (α, ∞) or $(-\infty, \alpha)$ with $\alpha \in \mathbb{R}$ are a subbase for topology of $\mathbb{R}$. ⋄

Note that a subbase $\mathcal{B}$ generates the topology in the sense that the corresponding topology is the coarsest topology for which all of the sets from $\mathcal{B}$ are open.

Example B.7. The **extended real numbers** $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ have a topology generated by the extended intervals of the form $(\alpha, \infty]$ or $[-\infty, \alpha)$ with $\alpha \in \mathbb{R}$. Note that the map $f(x) := \frac{x}{1+|x|}$ maps $\bar{\mathbb{R}} \to [-1, 1]$. This map becomes isometric if we choose $d(x, y) = |f(x) - f(y)|$ as a metric on $\bar{\mathbb{R}}$. It is not hard to verify that this metric generates our topology and hence we can think of $\bar{\mathbb{R}}$ as $[-1, 1]$. ⋄

There is also a local version of the previous notions. A **neighborhood base** for a point x is a collection of neighborhoods $\mathcal{B}(x)$ of x such that for each neighborhood $U(x)$, there is some set $B \in \mathcal{B}(x)$ with $B \subseteq U(x)$. Note that the sets in a neighborhood base are not required to be open.

If every point has a countable neighborhood base, then X is called **first countable**. If there exists a countable base, then X is called **second countable**. Note that a second countable space is in particular first countable since for every base $\mathcal{B}$ the subset $\mathcal{B}(x) := \{O \in \mathcal{B} | x \in O\}$ is a neighborhood base for x.

Example B.8. In a metric space the open balls $\{B_{1/m}(x)\}_{m \in \mathbb{N}}$ are a neighborhood base for x. Hence every metric space is first countable. Taking the union over all x, we obtain a base. In the case of $\mathbb{R}^n$ (or $\mathbb{C}^n$) it even suffices

to take balls with rational center, and hence $\mathbb{R}^n$ (as well as $\mathbb{C}^n$) is second countable. ⋄

Given two topologies on X their intersection will again be a topology on X. In fact, the intersection of an arbitrary collection of topologies is again a topology and hence given a collection $\mathcal{M}$ of subsets of X we can define the topology generated by $\mathcal{M}$ as the smallest topology (i.e., the intersection of all topologies) containing $\mathcal{M}$. Note that if $\mathcal{M}$ is closed under finite intersections and $\emptyset, X \in \mathcal{M}$, then it will be a base for the topology generated by $\mathcal{M}$ (Problem B.8).

Given two bases we can use them to check if the corresponding topologies are equal.

Lemma B.2. *Let $\mathcal{O}_j$, $j = 1, 2$ be two topologies for X with corresponding bases $\mathcal{B}_j$. Then $\mathcal{O}_1 \subseteq \mathcal{O}_2$ if and only if for every $x \in X$ and every $B_1 \in \mathcal{B}_1$ with $x \in B_1$ there is some $B_2 \in \mathcal{B}_2$ such that $x \in B_2 \subseteq B_1$.*

Proof. Suppose $\mathcal{O}_1 \subseteq \mathcal{O}_2$, then $B_1 \in \mathcal{O}_2$ and there is a corresponding B_2 by the very definition of a base. Conversely, let $O_1 \in \mathcal{O}_1$ and pick some $x \in O_1$. Then there is some $B_1 \in \mathcal{B}_1$ with $x \in B_1 \subseteq O_1$ and by assumption some $B_2 \in \mathcal{B}_2$ such that $x \in B_2 \subseteq B_1 \subseteq O_1$ which shows that x is an interior point with respect to $\mathcal{O}_2$. Since x was arbitrary we conclude $O_1 \in \mathcal{O}_2$. □

The next definition will ensure that limits are unique: A topological space is called a **Hausdorff space** if for any two different points there are always two disjoint neighborhoods.

Example B.9. Any metric space is a Hausdorff space: Given two different points x and y, the balls $B_{d/2}(x)$ and $B_{d/2}(y)$, where $d = d(x, y) > 0$, are disjoint neighborhoods. A pseudometric space will in general not be Hausdorff since two points of distance 0 cannot be separated by open balls. ⋄

The complement of an open set is called a **closed set**. It follows from **De Morgan's laws**

$$X \setminus \big(\bigcup_\alpha U_\alpha\big) = \bigcap_\alpha (X \setminus U_\alpha), \quad X \setminus \big(\bigcap_\alpha U_\alpha\big) = \bigcup_\alpha (X \setminus U_\alpha) \tag{B.6}$$

that the family of closed sets $\mathcal{C}$ satisfies

(i) $\emptyset, X \in \mathcal{C}$,

(ii) $C_1, C_2 \in \mathcal{C}$ implies $C_1 \cup C_2 \in \mathcal{C}$,

(iii) $\{C_\alpha\} \subseteq \mathcal{C}$ implies $\bigcap_\alpha C_\alpha \in \mathcal{C}$.

That is, closed sets are closed under finite unions and arbitrary intersections.

The smallest closed set containing a given set U is called the **closure**

$$\overline{U} := \bigcap_{C\in\mathcal{C},U\subseteq C} C, \tag{B.7}$$

and the largest open set contained in a given set U is called the **interior**

$$U^\circ := \bigcup_{O\in\mathcal{O},O\subseteq U} O. \tag{B.8}$$

It is not hard to see that the closure satisfies the following axioms (**Kuratowski closure axioms**):

(i) $\overline{\emptyset} = \emptyset$,
(ii) $U \subset \overline{U}$,
(iii) $\overline{\overline{U}} = \overline{U}$,
(iv) $\overline{U \cup V} = \overline{U} \cup \overline{V}$.

In fact, one can show that these axioms can equivalently be used to define the topology by observing that the closed sets are precisely those which satisfy $\overline{U} = U$.

Lemma B.3. *Let X be a topological space. Then the interior of U is the set of all interior points of U, and the closure of U is the union of U with all limit points of U. Moreover, $\partial U = \overline{U} \setminus U^\circ$.*

Proof. The first claim is straightforward. For the second claim observe that by Problem B.6 we have that $\overline{U} = (X \setminus (X \setminus U)^\circ)$, that is, the closure is the set of all points which are not interior points of the complement. That is, $x \notin \overline{U}$ iff there is some open set O containing x with $O \subseteq X \setminus U$. Hence, $x \in \overline{U}$ iff for all open sets O containing x we have $O \not\subseteq X \setminus U$, that is, $O \cap U \neq \emptyset$. Hence, $x \in \overline{U}$ iff $x \in U$ or if x is a limit point of U. The last claim is left as Problem B.7. □

Example B.10. For any $x \in X$ the **closed ball**

$$\bar{B}_r(x) := \{y \in X | d(x, y) \leq r\} \tag{B.9}$$

is a closed set (check that its complement is open). But in general we have only

$$\overline{B_r(x)} \subseteq \bar{B}_r(x) \tag{B.10}$$

since an isolated point y with $d(x, y) = r$ will not be a limit point. In $\mathbb{R}^n$ (or $\mathbb{C}^n$) we have of course equality. ⋄

Problem B.1. *Show that $|d(x, y) - d(z, y)| \leq d(x, z)$.*

Problem B.2. *Show the **quadrangle inequality** $|d(x, y) - d(x', y')| \leq d(x, x') + d(y, y')$.*

Problem B.3. *Show that if $f : [0, \infty) \to \mathbb{R}$ is concave, $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$ for $\lambda \in [0, 1]$, and satisfies $f(0) = 0$, then it is subadditive, $f(x+y) \leq f(x)+f(y)$. If in addition f is increasing and d is a pseudometric, then so is $f(d)$. (Hint: Begin by showing $f(\lambda x) \geq \lambda f(x)$.)*

Problem B.4. *Show De Morgan's laws.*

Problem B.5. *Show that the closure satisfies the Kuratowski closure axioms.*

Problem B.6. *Show that the closure and interior operators are dual in the sense that*

$$X \setminus \overline{U} = (X \setminus U)^\circ \qquad \text{and} \qquad X \setminus U^\circ = \overline{(X \setminus U)}.$$

In particular, the closure is the set of all points which are not interior points of the complement. (Hint: De Morgan's laws.)

Problem B.7. *Show that the boundary of U is given by $\partial U = \overline{U} \setminus U^\circ$.*

Problem B.8. *Suppose $\mathcal{M}$ is a collection of sets closed under finite intersections containing $\emptyset$ and X. Then the topology generated by $\mathcal{M}$ is given by $\mathcal{O}(\mathcal{M}) := \{\bigcup_\alpha M_\alpha | M_\alpha \in \mathcal{M}\}$.*

B.2. Convergence and completeness

A sequence $(x_n)_{n=1}^\infty \in X^{\mathbb{N}}$ is said to **converge** to some point $x \in X$ if $\lim_{n\to\infty} d(x, x_n) = 0$. We write $\lim_{n\to\infty} x_n = x$ or $x_n \to x$ as usual in this case. Clearly the **limit** x is unique if it exists (this is not true for a pseudometric). We will also frequently identify the sequence with its values x_n for simplicity of notation.

Note that convergent sequences are bounded. Here a set $U \subseteq X$ is called **bounded** if it is contained within a ball, that is, $U \subseteq B_r(x)$ for some $x \in X$ and $r > 0$.

Note that convergence can also be equivalently formulated in topological terms: A sequence $(x_n)_{n=1}^\infty$ converges to x if and only if for every neighborhood $U(x)$ of x there is some $N \in \mathbb{N}$ such that $x_n \in U(x)$ for $n \geq N$. In a Hausdorff space the limit is unique. However, sequences usually do not suffice to describe a topology and, in general, definitions in terms of sequences are weaker (see the example below). This could be avoided by using generalized sequences, so-called nets, where the index set $\mathbb{N}$ is replaced by arbitrary directed sets. We will not pursue this here.

Example B.11. For example, we can call a set U **sequentially closed** if every convergent sequence from U also has its limit in U. If U is closed, then every point in the complement is an inner point of the complement, thus no sequence from U can converge to such a point. Hence every closed

set is sequentially closed. In a metric space (or more generally in a first countable space) we can find a sequence for every limit point x by choosing a point (different from x) from every set in a neighborhood base. Hence the converse is also true in this case. ⋄

Note that the argument from the previous example shows that in a first countable space sequentially closed is the same as closed. In particular, in this case the family of closed sets is uniquely determined by the convergent sequences:

Lemma B.4. *Two first countable topologies agree if and only if they give rise to the same convergent sequences.*

Of course every subsequence of a convergent sequence will converge to the same limit and we have the following converse:

Lemma B.5. *Let X be a topological space, $(x_n)_{n=1}^{\infty} \in X^{\mathbb{N}}$ a sequence and $x \in X$. If every subsequence has a further subsequence which converges to x, then x_n converges to x.*

Proof. We argue by contradiction: If $x_n \not\to x$ we can find a neighborhood $U(x)$ and a subsequence $x_{n_k} \notin U(x)$. But then no subsequence of x_{n_k} can converge to x. □

This innocent observation is often useful to establish convergence in situations where one knows that the limit of a subsequence solves a given problem together with uniqueness of solutions for this problem. It can also be used to show that a notion of convergence does not stem from a topology (cf. Probelm 9.6).

In summary: A metric induces a natural topology and a topology induces a natural notion of convergence. However, a notion of convergence might not stem form a topology (or different topologies might give rise to the same notion of convergence) and a topology might not stem from a metric.

A sequence $(x_n)_{n=1}^{\infty} \in X^{\mathbb{N}}$ is called a **Cauchy sequence** if for every $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that

$$d(x_n, x_m) \le \varepsilon, \qquad n, m \ge N. \tag{B.11}$$

Every convergent sequence is a Cauchy sequence. If the converse is also true, that is, if every Cauchy sequence has a limit, then X is called **complete**. It is easy to see that a Cauchy sequence converges if and only if it has a convergent subsequence.

Example B.12. Both $\mathbb{R}^n$ and $\mathbb{C}^n$ are complete metric spaces. ⋄

Example B.13. The metric

$$d(x, y) := |\arctan(x) - \arctan(y)| \tag{B.12}$$

gives rise to the standard topology on $\mathbb{R}$ (since arctan is bi-Lipschitz on every compact interval). However, $x_n = n$ is a Cauchy sequence with respect to this metric but not with respect to the usual metric. Moreover, any sequence with $x_n \to \infty$ or $x_n \to -\infty$ will be Cauchy with respect to this metric and hence (show this) for the completion of $\mathbb{R}$ precisely the two new points $-\infty$ and $+\infty$ have to be added (cf. Example B.7). $\diamond$

As noted before, in a metric space x is a limit point of U if and only if there exists a sequence $(x_n)_{n=1}^{\infty} \subseteq U \setminus \{x\}$ with $\lim_{n\to\infty} x_n = x$. Hence U is closed if and only if for every convergent sequence the limit is in U. In particular,

Lemma B.6. *A subset of a complete metric space is again a complete metric space if and only if it is closed.*

A set $U \subseteq X$ is called **dense** if its closure is all of X, that is, if $\overline{U} = X$. A space is called **separable** if it contains a countable dense set.

Lemma B.7. *A metric space is separable if and only if it is second countable as a topological space.*

Proof. From every dense set we get a countable base by considering open balls with rational radii and centers in the dense set. Conversely, from every countable base we obtain a dense set by choosing an element from each set in the base. $\square$

Lemma B.8. *Let X be a separable metric space. Every subset Y of X is again separable.*

Proof. Let $A = \{x_n\}_{n\in\mathbb{N}}$ be a dense set in X. The only problem is that $A \cap Y$ might contain no elements at all. However, some elements of A must be at least arbitrarily close to this intersection: Let $J \subseteq \mathbb{N}^2$ be the set of all pairs (n, m) for which $B_{1/m}(x_n) \cap Y \neq \emptyset$ and choose some $y_{n,m} \in B_{1/m}(x_n) \cap Y$ for all $(n, m) \in J$. Then $B = \{y_{n,m}\}_{(n,m)\in J} \subseteq Y$ is countable. To see that B is dense, choose $y \in Y$. Then there is some sequence x_{n_k} with $d(x_{n_k}, y) < 1/k$. Hence $(n_k, k) \in J$ and $d(y_{n_k,k}, y) \leq d(y_{n_k,k}, x_{n_k}) + d(x_{n_k}, y) \leq 2/k \to 0$. $\square$

If X is an (incomplete) metric space, consider the set of all Cauchy sequences $\mathcal{X}$ from X. Call two Cauchy sequences equivalent if their difference converges to zero and denote by $\bar{X}$ the set of all equivalence classes. Moreover, the quadrangle inequality (Problem B.2) shows that if $x = (x_n)_{n\in\mathbb{N}}$ and $y = (y_n)_{n\in\mathbb{N}}$ are Cauchy sequences, so is $d(x_n, y_n)$ and hence we can define a metric on $\bar{X}$ via

$$d_{\bar{X}}([x], [y]) = \lim_{n\to\infty} d_X(x_n, y_n). \tag{B.13}$$

Indeed, it is straightforward to check that $d_{\bar{X}}$ is well defined (independent of the representative) and inherits all properties of a metric from d_X. Moreover, $d_{\bar{X}}$ agrees with d_X whenever both Cauchy sequences converge in X.

Theorem B.9. *The space $\bar{X}$ is a metric space containing X as a dense subspace if we identify $x \in X$ with the equivalence class of all sequences converging to x. Moreover, this embedding is isometric.*

Proof. The map $J : X \to \bar{X}$, $x_0 \mapsto [(x_0, x_0, \dots)]$ is an isometric embedding (i.e., it is injective and preserves the metric). Moreover, for a Cauchy sequence $x = (x_n)_{n\in\mathbb{N}}$ the sequence $J(x_n)$ converges to x and hence $J(X)$ is dense in $\bar{X}$. It remains to show that $\bar{X}$ is complete. Let $\xi_n = [(x_{n,j})_{j\in\mathbb{N}}]$ be a Cauchy sequence in $\bar{X}$. Without loss of generality (by dropping terms) we can choose the representatives $x_{n,j}$ such that $d(x_{n,j}, x_{n,k}) \le \frac{1}{n}$ for $j, k \ge n$. Then it is not hard to see that $\xi = [(x_{j,j})_{j\in\mathbb{N}}]$ is its limit. □

Let me remark that the completion $\bar{X}$ is unique. More precisely, suppose $\tilde{X}$ is another complete metric space which contains X as a dense subset such that the embedding $\tilde{J} : X \hookrightarrow \tilde{X}$ is isometric. Then $I = \tilde{J} \circ J^{-1} : J(X) \to \tilde{J}(X)$ has a unique isometric extension $\bar{I} : \bar{X} \to \tilde{X}$ (compare Theorem B.38 below). In particular, it is no restriction to assume that a metric space is complete.

Problem B.9. *Let $U \subseteq V$ be subsets of a metric space X. Show that if U is dense in V and V is dense in X, then U is dense in X.*

Problem B.10. *Let X be a metric space and denote by $B(X)$ the set of all bounded functions $X \to \mathbb{C}$. Introduce the metric*

$$d(f, g) = \sup_{x\in X} |f(x) - g(x)|.$$

Show that $B(X)$ is complete.

Problem B.11. *Let X be a metric space and $B(X)$ as in the previous problem. Consider the embedding $J : X \hookrightarrow B(X)$ defind via*

$$y \mapsto J(x)(y) = d(x, y) - d(x_0, y)$$

for some fixed $x_0 \in X$. Show that this embedding is isometric. Hence $\overline{J(X)}$ is another (equivalent) completion of X.

B.3. Functions

Next, we come to functions $f : X \to Y$, $x \mapsto f(x)$. We use the usual conventions $f(U) := \{f(x)|x \in U\}$ for $U \subseteq X$ and $f^{-1}(V) := \{x|f(x) \in V\}$ for $V \subseteq Y$. Note

$$U \subseteq f^{-1}(f(U)), \qquad f(f^{-1}(V)) \subseteq V. \tag{B.14}$$

The set $\mathrm{Ran}(f) := f(X)$ is called the **range** of f, and X is called the **domain** of f. A function f is called **injective** or **one-to-one** if for each $y \in Y$ there is at most one $x \in X$ with $f(x) = y$ (i.e., $f^{-1}(\{y\})$ contains at most one point) and **surjective** or **onto** if $\mathrm{Ran}(f) = Y$. A function f which is both injective and surjective is called **bijective**.

Recall that we always have

$$\begin{aligned} f^{-1}(\bigcup_\alpha V_\alpha) &= \bigcup_\alpha f^{-1}(V_\alpha), \qquad f^{-1}(\bigcap_\alpha V_\alpha) = \bigcap_\alpha f^{-1}(V_\alpha), \\ f^{-1}(Y \setminus V) &= X \setminus f^{-1}(V) \end{aligned} \tag{B.15}$$

as well as

$$\begin{aligned} f(\bigcup_\alpha U_\alpha) &= \bigcup_\alpha f(U_\alpha), \qquad f(\bigcap_\alpha U_\alpha) \subseteq \bigcap_\alpha f(U_\alpha), \\ f(X) \setminus f(U) &\subseteq f(X \setminus U) \end{aligned} \tag{B.16}$$

with equality if f is injective.

A function f between metric spaces X and Y is called continuous at a point $x \in X$ if for every $\varepsilon > 0$ we can find a $\delta > 0$ such that

$$d_Y(f(x), f(y)) \le \varepsilon \quad \text{if} \quad d_X(x, y) < \delta. \tag{B.17}$$

If f is continuous at every point, it is called **continuous**. In the case $d_Y(f(x), f(y)) = d_X(x, y)$ we call f **isometric** and every isometry is of course continuous.

Lemma B.10. *Let X, Y be metric spaces. The following are equivalent:*

(i) *f is continuous at x (i.e.,* (B.17) *holds).*

(ii) *$f(x_n) \to f(x)$ whenever $x_n \to x$.*

(iii) *For every neighborhood V of $f(x)$ the preimage $f^{-1}(V)$ is a neighborhood of x.*

Proof. (i) $\Rightarrow$ (ii) is obvious. (ii) $\Rightarrow$ (iii): If (iii) does not hold, there is a neighborhood V of $f(x)$ such that $B_\delta(x) \not\subseteq f^{-1}(V)$ for every δ. Hence we can choose a sequence $x_n \in B_{1/n}(x)$ such that $f(x_n) \notin f^{-1}(V)$. Thus $x_n \to x$ but $f(x_n) \not\to f(x)$. (iii) $\Rightarrow$ (i): Choose $V = B_\varepsilon(f(x))$ and observe that by (iii), $B_\delta(x) \subseteq f^{-1}(V)$ for some δ. □

In a general topological space we use (iii) as the definition of continuity and (ii) is called **sequential continuity**. Then continuity will imply sequential continuity but the converse will not be true unless we assume (e.g.) that X is first countable (Problem B.12).

In particular, (iii) implies that f is continuous if and only if the preimage of every open set is again open (equivalently, the inverse image of every closed set is closed). Note that by (B.15) it suffices to check this for sets in

a subbase. If the image of every open set is open, then f is called **open**. A bijection f is called a **homeomorphism** if both f and its inverse f^{-1} are continuous. Note that if f is a bijection, then f^{-1} is continuous if and only if f is open. Two topological spaces are called **homeomorphic** if there is a homeomorphism between them.

The **support** of a function $f : X \to \mathbb{C}^n$ is the closure of all points x for which $f(x)$ does not vanish; that is,

$$\operatorname{supp}(f) := \overline{\{x \in X | f(x) \neq 0\}}. \tag{B.18}$$

Problem B.12. *Let X, Y be topological spaces. Show that if $f : X \to Y$ is continuous at $x \in X$ then it is also sequential continuous. Show that the converse holds if X is first countable.*

Problem B.13. *Let $f : X \to Y$ be continuous. Then $f(\overline{A}) \subseteq \overline{f(A)}$.*

Problem B.14. *Let X, Y be topological spaces and let $f : X \to Y$ be continuous. Show that if X is separable, then so is $f(Y)$.*

Problem B.15. *Let X be a topological space and $f : X \to \mathbb{R}$. Let $x_0 \in X$ and let $\mathcal{B}(x_0)$ be a neighborhood base for x_0. Define*

$$\liminf_{x\to x_0} f(x) := \sup_{U\in\mathcal{B}(x_0)} \inf_{U(x_0)} f, \qquad \limsup_{x\to x_0} f(x) := \inf_{U\in\mathcal{B}(x_0)} \sup_{U(x_0)} f.$$

Show that both are independent of the neighborhood base and satisfy

(i) $\liminf_{x\to x_0}(-f(x)) = -\limsup_{x\to x_0} f(x)$.

(ii) $\liminf_{x\to x_0}(\alpha f(x)) = \alpha \liminf_{x\to x_0} f(x)$, $\alpha \geq 0$.

(iii) $\liminf_{x\to x_0}(f(x) + g(x)) \geq \liminf_{x\to x_0} f(x) + \liminf_{x\to x_0} g(x)$.

Moreover, show that

$$\liminf_{n\to\infty} f(x_n) \geq \liminf_{x\to x_0} f(x), \qquad \limsup_{n\to\infty} f(x_n) \leq \limsup_{x\to x_0} f(x)$$

for every sequence $x_n \to x_0$ and there exists a sequence attaining equality if X is a metric space.

B.4. Product topologies

If X and Y are metric spaces, then $X \times Y$ together with

$$d((x_1, y_1), (x_2, y_2)) := d_X(x_1, x_2) + d_Y(y_1, y_2) \tag{B.19}$$

is a metric space. A sequence (x_n, y_n) converges to (x, y) if and only if $x_n \to x$ and $y_n \to y$. In particular, the projections onto the first $(x, y) \mapsto x$, respectively, onto the second $(x, y) \mapsto y$, coordinate are continuous. Moreover, if X and Y are complete, so is $X \times Y$.

In particular, by the inverse triangle inequality (B.1),

$$|d(x_n, y_n) - d(x, y)| \le d(x_n, x) + d(y_n, y), \tag{B.20}$$

we see that $d : X \times X \to \mathbb{R}$ is continuous.

Example B.14. If we consider $\mathbb{R} \times \mathbb{R}$, we do not get the Euclidean distance of $\mathbb{R}^2$ unless we modify (B.19) as follows:

$$\tilde{d}((x_1, y_1), (x_2, y_2)) := \sqrt{d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2}. \tag{B.21}$$

As noted in our previous example, the topology (and thus also convergence/continuity) is independent of this choice. ⋄

If X and Y are just topological spaces, the **product topology** is defined by calling $O \subseteq X \times Y$ open if for every point $(x, y) \in O$ there are open neighborhoods U of x and V of y such that $U \times V \subseteq O$. In other words, the products of open sets form a base of the product topology. Again the projections onto the first and second component are continuous. In the case of metric spaces this clearly agrees with the topology defined via the product metric (B.19). There is also another way of constructing the product topology, namely, as the weakest topology which makes the projections continuous. In fact, this topology must contain all sets which are inverse images of open sets $U \subseteq X$, that is all sets of the form $U \times Y$ as well as all inverse images of open sets $V \subseteq Y$, that is all sets of the form $X \times V$. Adding finite intersections we obtain all sets of the form $U \times V$ and hence the same base as before. In particular, a sequence (x_n, y_n) will converge if and only if both components converge.

Note that the product topology immediately extends to the product of an arbitrary number of spaces $X := \bigtimes_{\alpha \in A} X_\alpha$ by defining it as the weakest topology which makes all projections $\pi_\alpha : X \to X_\alpha$ continuous.

Example B.15. Let X be a topological space and A an index set. Then $X^A = \bigtimes_A X$ is the set of all functions $x : A \to X$ and a neighborhood base at x are sets of functions which coincide with x at a given finite number of points. Convergence with respect to the product topology corresponds to pointwise convergence (note that the projection π_α is the point evaluation at α: $\pi_\alpha(x) = x(\alpha)$). If A is uncountable (and X is not equipped with the trivial topology), then there is no countable neighborhood base (if there were such a base, it would involve only a countable number of points, now choose a point from the complement ...). In particular, there is no corresponding metric even if X has one. Moreover, this topology cannot be characterized with sequences alone. For example, let $X = \{0, 1\}$ (with the discrete topology) and $A = \mathbb{R}$. Then the set $F = \{x | x^{-1}(1) \text{ is countable}\}$ is sequentially closed but its closure is all of $\{0, 1\}^{\mathbb{R}}$ (every set from our neighborhood base contains an element which vanishes except at finitely many points). ⋄

In fact this is a special case of a more general construction which is often used. Let $\{f_\alpha\}_{\alpha\in A}$ be a collection of functions $f_\alpha : X \to Y_\alpha$, where Y_α are some topological spaces. Then we can equip X with the weakest topology (known as the **initial topology**) which makes all f_α continuous. That is, we take the topology generated by sets of the forms $f_\alpha^{-1}(O_\alpha)$, where $O_\alpha \subseteq Y_\alpha$ is open, known as open **cylinders**. Finite intersections of such sets, known as open **cylinder sets**, are hence a base for the topology and a sequence x_n will converge to x if and only if $f_\alpha(x_n) \to f_\alpha(x)$ for all $\alpha \in A$. In particular, if the collection is countable, then X will be first (or second) countable if all Y_α are.

The initial topology has the following characteristic property:

Lemma B.11. *Let X have the initial topology from a collection of functions $\{f_\alpha : X \to Y_\alpha\}_{\alpha\in A}$ and let Z be another topological space. A function $f : Z \to X$ is continuous (at z) if and only if $f_\alpha \circ f$ is continuous (at z) for all $\alpha \in A$.*

Proof. If f is continuous at z, then so is the composition $f_\alpha \circ f$. Conversely, let $U \subseteq X$ be a neighborhood of $f(z)$. Then $\bigcap_{j=1}^n f_{\alpha_j}^{-1}(O_{\alpha_j}) \subseteq U$ for some α_j and some open neighborhoods O_{α_j} of $f_{\alpha_j}(f(z))$. But then $f^{-1}(U)$ contains the neighborhood $f^{-1}(\bigcap_{j=1}^n f_{\alpha_j}^{-1}(O_{\alpha_j})) = \bigcap_{j=1}^n (f_{\alpha_j} \circ f)^{-1}(O_{\alpha_j})$ of z. □

If all Y_α are Hausdorff and if the collection $\{f_\alpha\}_{\alpha\in A}$ **separates points**, that is for every $x \neq y$ there is some α with $f_\alpha(x) \neq f_\alpha(y)$, then X will again be Hausdorff. Indeed for $x \neq y$ choose α such that $f_\alpha(x) \neq f_\alpha(y)$ and let U_α, V_α be two disjoint neighborhoods separating $f_\alpha(x), f_\alpha(y)$. Then $f_\alpha^{-1}(U_\alpha), f_\alpha^{-1}(V_\alpha)$ are two disjoint neighborhoods separating x, y. In particular, $X = \bigtimes_{\alpha\in A} X_\alpha$ is Hausdorff if all X_α are.

Note that a similar construction works in the other direction. Let $\{f_\alpha\}_{\alpha\in A}$ be a collection of functions $f_\alpha : X_\alpha \to Y$, where X_α are some topological spaces. Then we can equip Y with the strongest topology (known as the **final topology**) which makes all f_α continuous. That is, we take as open sets those for which $f_\alpha^{-1}(O)$ is open for all $\alpha \in A$.

Example B.16. Let $\sim$ be an equivalence relation on X with equivalence classes $[x] = \{y \in X | x \sim y\}$. Then the **quotient topology** on the set of equivalence classes $X/\!\sim$ is the final topology of the projection map $\pi : X \to X/\!\sim$. ◇

Lemma B.12. *Let Y have the final topology from a collection of functions $\{f_\alpha : X_\alpha \to Y\}_{\alpha\in A}$ and let Z be another topological space. A function $f : Y \to Z$ is continuous if and only if $f \circ f_\alpha$ is continuous for all $\alpha \in A$.*

Proof. If f is continuous, then so is the composition $f \circ f_\alpha$. Conversely, let $V \subseteq Z$ be open. Then $f \circ f_\alpha$ implies $(f \circ f_\alpha)^{-1}(V) = f_\alpha^{-1}(f^{-1}(V))$ open for all α and hence $f^{-1}(V)$ open. □

Problem B.16. *Let $X = \bigtimes_{\alpha \in A} X_\alpha$ with the product topology. Show that the projection maps are open.*

Problem B.17 (Gluing lemma)**.** *Suppose X, Y are topological spaces and $f_\alpha : A_\alpha \to Y$ are continuous functions defined on $A_\alpha \subseteq X$. Suppose $f_\alpha = f_\beta$ on $A_\alpha \cap A_\beta$ such that $f : A := \bigcup_\alpha A_\alpha \to Y$ is well defined by $f(x) = f_\alpha(x)$ if $x \in A_\alpha$. Show that f is continuous if either all sets A_α are open or if the collection A_α is finite and all are closed.*

Problem B.18. *Let $\{(X_j, d_j)\}_{j \in \mathbb{N}}$ be a sequence of metric spaces. Show that*

$$d(x,y) = \sum_{j \in \mathbb{N}} \frac{1}{2^j} \frac{d_j(x_j, y_j)}{1 + d_j(x_j, y_j)} \quad \text{or} \quad d(x,y) = \max_{j \in \mathbb{N}} \frac{1}{2^j} \frac{d_j(x_j, y_j)}{1 + d_j(x_j, y_j)}$$

is a metric on $X = \bigtimes_{n \in \mathbb{N}} X_n$ which generates the product topology. Show that X is complete if all X_n are.

B.5. Compactness

A **cover** of a set $Y \subseteq X$ is a family of sets $\{U_\alpha\}$ such that $Y \subseteq \bigcup_\alpha U_\alpha$. A cover is called open if all U_α are open. Any subset of $\{U_\alpha\}$ which still covers Y is called a **subcover**.

Lemma B.13 (Lindelöf)**.** *If X is second countable, then every open cover has a countable subcover.*

Proof. Let $\{U_\alpha\}$ be an open cover for Y, and let $\mathcal{B}$ be a countable base. Since every U_α can be written as a union of elements from $\mathcal{B}$, the set of all $B \in \mathcal{B}$ which satisfy $B \subseteq U_\alpha$ for some α form a countable open cover for Y. Moreover, for every B_n in this set we can find an α_n such that $B_n \subseteq U_{\alpha_n}$. By construction, $\{U_{\alpha_n}\}$ is a countable subcover. □

A **refinement** $\{V_\beta\}$ of a cover $\{U_\alpha\}$ is a cover such that for every β there is some α with $V_\beta \subseteq U_\alpha$. A cover is called **locally finite** if every point has a neighborhood that intersects only finitely many sets in the cover.

Lemma B.14 (Stone)**.** *In a metric space every countable open cover has a locally finite open refinement.*

Proof. Denote the cover by $\{O_j\}_{j\in\mathbb{N}}$ and introduce the sets

$$\hat{O}_{j,n} := \bigcup_{x\in A_{j,n}} B_{2^{-n}}(x), \text{ where}$$

$$A_{j,n} := \{x \in O_j \setminus (O_1 \cup \cdots \cup O_{j-1}) | x \notin \bigcup_{k\in\mathbb{N},1\le l<n} \hat{O}_{k,l} \text{ and } B_{3\cdot 2^{-n}}(x) \subseteq O_j\}.$$

Then, by construction, $\hat{O}_{j,n}$ is open, $\hat{O}_{j,n} \subseteq O_j$, and it is a cover since for every x there is a smallest j such that $x \in O_j$ and a smallest n such that $B_{3\cdot 2^{-n}}(x) \subseteq O_j$ implying $x \in \hat{O}_{k,l}$ for some $l \le n$.

To show that $\hat{O}_{j,n}$ is locally finite fix some x and let j be the smallest integer such that $x \in \hat{O}_{j,n}$ for some n. Moreover, choose m such that $B_{2^{-m}}(x) \subseteq \hat{O}_{j,n}$. It suffices to show that:
(i) If $i \ge n + m$ then $B_{2^{-n-m}}(x)$ is disjoint from $\hat{O}_{k,i}$ for all k.
(ii) If $i < n + m$ then $B_{2^{-n-m}}(x)$ intersects $\hat{O}_{k,i}$ for at most one k.

To show (i) observe that since $i > n$ every ball $B_{2^{-i}}(y)$ used in the definition of $\hat{O}_{k,i}$ has its center outside of $\hat{O}_{j,n}$. Hence $d(x, y) \ge 2^{-m}$ and $B_{2^{-n-m}}(x) \cap B_{2^{-i}}(x) = \emptyset$ since $i \ge m + 1$ as well as $n + m \ge m + 1$.

To show (ii) let $y \in \hat{O}_{j,i}$ and $z \in \hat{O}_{k,i}$ with $j < k$. We will show $d(y, z) > 2^{-n-m+1}$. There are points r and s such that $y \in B_{2^{-i}}(r) \subseteq \hat{O}_{j,i}$ and $z \in B_{2^{-i}}(s) \subseteq \hat{O}_{k,i}$. Then by definition $B_{3\cdot 2^{-i}}(r) \subseteq O_j$ but $s \notin O_j$. So $d(r, s) \ge 3 \cdot 2^{-i}$ and $d(y, z) > 2^{-i} \ge 2^{-n-m+1}$. □

A subset $K \subset X$ is called **compact** if every open cover of K has a finite subcover. A set is called **relatively compact** if its closure is compact.

Lemma B.15. *A topological space is compact if and only if it has the **finite intersection property**: The intersection of a family of closed sets is empty if and only if the intersection of some finite subfamily is empty.*

Proof. By taking complements, to every family of open sets there corresponds a family of closed sets and vice versa. Moreover, the open sets are a cover if and only if the corresponding closed sets have empty intersection. □

Lemma B.16. *Let X be a topological space.*

(i) *The continuous image of a compact set is compact.*
(ii) *Every closed subset of a compact set is compact.*
(iii) *If X is Hausdorff, every compact set is closed.*
(iv) *The finite union of compact sets is compact.*
(v) *If X is Hausdorff, any intersection of compact sets is compact.*

Proof. (i) Observe that if $\{O_\alpha\}$ is an open cover for $f(Y)$, then $\{f^{-1}(O_\alpha)\}$ is one for Y.

(ii) Let $\{O_\alpha\}$ be an open cover for the closed subset Y (in the induced topology). Then there are open sets $\tilde{O}_\alpha$ with $O_\alpha = \tilde{O}_\alpha \cap Y$ and $\{\tilde{O}_\alpha\} \cup \{X \setminus Y\}$ is an open cover for X which has a finite subcover. This subcover induces a finite subcover for Y.

(iii) Let $Y \subseteq X$ be compact. We show that $X \setminus Y$ is open. Fix $x \in X \setminus Y$ (if $Y = X$ there is nothing to do). By the definition of Hausdorff, for every $y \in Y$ there are disjoint neighborhoods $V(y)$ of y and $U_y(x)$ of x. By compactness of Y, there are $y_1, \ldots, y_n$ such that the $V(y_j)$ cover Y. But then $\bigcap_{j=1}^n U_{y_j}(x)$ is a neighborhood of x which does not intersect Y.

(iv) Note that a cover of the union is a cover for each individual set and the union of the individual subcovers is the subcover we are looking for.

(v) Follows from (ii) and (iii) since an intersection of closed sets is closed. □

As a consequence we obtain a simple criterion when a continuous function is a homeomorphism.

Corollary B.17. *Let X and Y be topological spaces with X compact and Y Hausdorff. Then every continuous bijection $f : X \to Y$ is a homeomorphism.*

Proof. It suffices to show that f maps closed sets to closed sets. By (ii) every closed set is compact, by (i) its image is also compact, and by (iii) it is also closed. □

Concerning products of compact sets we have

Theorem B.18 (Tychonoff)**.** *The product $\bigtimes_{\alpha\in A} K_\alpha$ of an arbitrary collection of compact topological spaces $\{K_\alpha\}_{\alpha\in A}$ is compact with respect to the product topology.*

Proof. We say that a family $\mathcal{F}$ of closed subsets of K has the finite intersection property if the intersection of every finite subfamily has nonempty intersection. The collection of all such families which contain $\mathcal{F}$ is partially ordered by inclusion and every chain has an upper bound (the union of all sets in the chain). Hence, by Zorn's lemma, there is a maximal family $\mathcal{F}_M$ (note that this family is closed under finite intersections).

Denote by $\pi_\alpha : K \to K_\alpha$ the projection onto the α component. Then the closed sets $\{\overline{\pi_\alpha(F)}\}_{F\in\mathcal{F}_M}$ also have the finite intersection property and since K_α is compact, there is some $x_\alpha \in \bigcap_{F\in\mathcal{F}_M} \overline{\pi_\alpha(F)}$. Consequently, if F_α is a closed neighborhood of x_α, then $\pi_\alpha^{-1}(F_\alpha) \in \mathcal{F}_M$ (otherwise there would

be some $F \in \mathcal{F}_M$ with $F \cap \pi_\alpha^{-1}(F_\alpha) = \emptyset$ contradicting $F_\alpha \cap \pi_\alpha(F) \neq \emptyset$). Furthermore, for every finite subset $A_0 \subseteq A$ we have $\bigcap_{\alpha\in A_0} \pi_\alpha^{-1}(F_\alpha) \in \mathcal{F}_M$ and so every neighborhood of $x = (x_\alpha)_{\alpha\in A}$ intersects F. Since F is closed, $x \in F$ and hence $x \in \bigcap_{\mathcal{F}_M} F$. □

A subset $K \subset X$ is called **sequentially compact** if every sequence from K has a convergent subsequence whose limit is in K. In a metric space, compact and sequentially compact are equivalent.

Lemma B.19. *Let X be a metric space. Then a subset is compact if and only if it is sequentially compact.*

Proof. Without loss of generality we can assume the subset to be all of X. Suppose X is compact and let x_n be a sequence which has no convergent subsequence. Then $K := \{x_n\}$ has no limit points and is hence compact by Lemma B.16 (ii). For every n there is a ball $B_{\varepsilon_n}(x_n)$ which contains only finitely many elements of K. However, finitely many suffice to cover K, a contradiction.

Conversely, suppose X is sequentially compact and let $\{O_\alpha\}$ be some open cover which has no finite subcover. For every $x \in X$ we can choose some $\alpha(x)$ such that if $B_r(x)$ is the largest ball contained in $O_{\alpha(x)}$, then either $r \geq 1$ or there is no β with $B_{2r}(x) \subset O_\beta$ (show that this is possible). Now choose a sequence x_n such that $x_n \notin \bigcup_{m<n} O_{\alpha(x_m)}$. Note that by construction the distance $d = d(x_m, x_n)$ to every successor of x_m is either larger than 1 or the ball $B_{2d}(x_m)$ will not fit into any of the O_α.

Now let y be the limit of some convergent subsequence and fix some $r \in (0, 1)$ such that $B_r(y) \subseteq O_{\alpha(y)}$. Then this subsequence must eventually be in $B_{r/5}(y)$, but this is impossible since if $d := d(x_{n_1}, x_{n_2})$ is the distance between two consecutive elements of this subsequence, then $B_{2d}(x_{n_1})$ cannot fit into $O_{\alpha(y)}$ by construction whereas on the other hand $B_{2d}(x_{n_1}) \subseteq B_{4r/5}(a) \subseteq O_{\alpha(y)}$. □

If we drop the requirement that the limit must be in K, we obtain relatively compact sets:

Corollary B.20. *Let X be a metric space and $K \subset X$. Then K is relatively compact if and only if every sequence from K has a convergent subsequence (the limit must not be in K).*

Proof. For any sequence $x_n \in \overline{K}$ we can find a nearby sequence $y_n \in K$ with $x_n - y_n \to 0$. If we can find a convergent subsequence of y_n then the corresponding subsequence of x_n will also converge (to the same limit) and $\overline{K}$ is (sequentially) compact in this case. The converse is trivial. □

As another simple consequence observe that

Corollary B.21. *A compact metric space X is complete and separable.*

Proof. Completeness is immediate from the previous lemma. To see that X is separable note that, by compactness, for every $n \in \mathbb{N}$ there is a finite set $S_n \subseteq X$ such that the balls $\{B_{1/n}(x)\}_{x \in S_n}$ cover X. Then $\bigcup_{n \in \mathbb{N}} S_n$ is a countable dense set. □

In a metric space, a set is called **bounded** if it is contained inside some ball. Clearly the union of two bounded sets is bounded. Moreover, compact sets are always bounded since the can be covered by finitely many balls. In $\mathbb{R}^n$ (or $\mathbb{C}^n$) the converse also holds.

Theorem B.22 (Heine–Borel)**.** *In $\mathbb{R}^n$ (or $\mathbb{C}^n$) a set is compact if and only if it is bounded and closed.*

Proof. By Lemma B.16 (ii), (iii), and Tychonoff's theorem it suffices to show that a closed interval in $I \subseteq \mathbb{R}$ is compact. Moreover, by Lemma B.19, it suffices to show that every sequence in $I = [a, b]$ has a convergent subsequence. Let x_n be our sequence and divide $I = [a, \frac{a+b}{2}] \cup [\frac{a+b}{2}, b]$. Then at least one of these two intervals, call it I_1, contains infinitely many elements of our sequence. Let $y_1 = x_{n_1}$ be the first one. Subdivide I_1 and pick $y_2 = x_{n_2}$, with $n_2 > n_1$ as before. Proceeding like this, we obtain a Cauchy sequence y_n (note that by construction $I_{n+1} \subseteq I_n$ and hence $|y_n - y_m| \leq \frac{b-a}{2n}$ for $m \geq n$). □

By Lemma B.19 this is equivalent to

Theorem B.23 (Bolzano–Weierstraß)**.** *Every bounded infinite subset of $\mathbb{R}^n$ (or $\mathbb{C}^n$) has at least one limit point.*

Combining Theorem B.22 with Lemma B.16 (i) we also obtain the **extreme value theorem**.

Theorem B.24 (Weierstraß)**.** *Let X be compact. Every continuous function $f : X \to \mathbb{R}$ attains its maximum and minimum.*

A metric space X for which the Heine–Borel theorem holds is called **proper**. Lemma B.16 (ii) shows that X is proper if and only if every closed ball is compact. Note that a proper metric space must be complete (since every Cauchy sequence is bounded). A topological space is called **locally compact** if every point has a compact neighborhood. Clearly a proper metric space is locally compact. A topological space is called **σ-compact**, if it can be written as a countable union of compact sets. Again a proper space is σ-compact.

Lemma B.25. *For a metric space X the following are equivalent:*

(i) *X is separable and locally compact.*

(ii) *X contains a countable base consisting of relatively compact sets.*

(iii) *X is locally compact and σ-compact.*

(iv) *X can be written as the union of an increasing sequence U_n of relatively compact open sets satisfying $\overline{U_n} \subseteq U_{n+1}$ for all n.*

Proof. (i) $\Rightarrow$ (ii): Let $\{x_n\}$ be a dense set. Then the balls $B_{n,m} = B_{1/m}(x_n)$ form a base. Moreover, for every n there is some m_n such that $B_{n,m}$ is relatively compact for $m \le m_n$. Since those balls are still a base we are done. (ii) $\Rightarrow$ (iii): Take the union over the closures of all sets in the base. (iii) $\Rightarrow$ (vi): Let $X = \bigcup_n K_n$ with K_n compact. Without loss $K_n \subseteq K_{n+1}$. For a given compact set K we can find a relatively compact open set $V(K)$ such that $K \subseteq V(K)$ (cover K by relatively compact open balls and choose a finite subcover). Now define $U_n = V(\overline{U_n})$. (vi) $\Rightarrow$ (i): Each of the sets $\overline{U_n}$ has a countable dense subset by Corollary B.21. The union gives a countable dense set for X. Since every $x \in U_n$ for some n, X is also locally compact. □

Note that since a sequence as in (iv) is a cover for X, every compact set is contained in U_n for n sufficiently large.

Example B.17. $X = (0, 1)$ with the usual metric is locally compact and σ-compact but not proper. ◇

Example B.18. Consider $\ell^2(\mathbb{N})$ with the standard basis δ^j. Let $X_j := \{\lambda\delta^j | \lambda \in [0, 1]\}$ and note that the metric on X_j inherited from $\ell^2(\mathbb{Z})$ is the same as the usual metric from $\mathbb{R}$. Then $X := \bigcup_{j\in\mathbb{N}} X_j$ is a complete separable σ-compact space, which is not locally compact. In fact, consider a ball of radius ε around zero. Then $(\varepsilon/2)\delta^j \in B_\varepsilon(0)$ is a bounded sequence which has no convergent subsequence since $d((\varepsilon/2)\delta^j, (\varepsilon/2)\delta^k) = \varepsilon/\sqrt{2}$ for $k \ne j$. ◇

However, under the assumptions of the previous lemma we can always switch to a new metric which generates the same topology and for which X is proper. To this end recall that a function $f : X \to Y$ between topological spaces is called **proper** if the inverse image of a compact set is again compact. Now given a proper function (Problem B.25) there is a new metric with the claimed properties (Problem B.21).

A subset U of a complete metric space X is called **totally bounded** if for every $\varepsilon > 0$ it can be covered with a finite number of balls of radius ε. We will call such a cover an ε-cover. Clearly every totally bounded set is bounded.

Example B.19. Of course in $\mathbb{R}^n$ the totally bounded sets are precisely the bounded sets. This is in fact true for every proper metric space since the closure of a bounded set is compact and hence has a finite cover. $\diamond$

Lemma B.26. *Let X be a complete metric space. Then a set is relatively compact if and only it is totally bounded.*

Proof. Without loss of generality we can assume our set to be closed. Clearly a compact set K is closed and totally bounded (consider the cover by all balls of radius ε with center in the set and choose a finite subcover). Conversely, we will show that K is sequentially compact. So start with $\varepsilon_1 = 1$ and choose a finite cover of balls with radius ε_1. One of these balls contains an infinite number of elements x_n^1 from our sequence x_n. Choose $\varepsilon_2 = \frac{1}{2}$ and repeat the process with the sequence x_n^1. The resulting diagonal sequence x_n^n gives a subsequence which is Cauchy and hence converges by completeness. $\square$

Problem B.19 (One point compactification)**.** *Suppose X is a locally compact Hausdorff space which is not compact. Introduce a new point ∞, set $\hat{X} = X \cup \{\infty\}$ and make it into a topological space by calling $O \subseteq \hat{X}$ open if either $\infty \notin O$ and O is open in X or if $\infty \in O$ and $\hat{X} \setminus O$ is compact. Show that $\hat{X}$ is a compact Hausdorff space which contains X as a dense subset.*

Problem B.20. *Show that every open set $O \subseteq \mathbb{R}$ can be written as a countable union of disjoint intervals. (Hint: Consider the set $\{I_\alpha\}$ of all maximal open subintervals of O; that is, $I_\alpha \subseteq O$ and there is no other subinterval of O which contains I_α.*

Problem B.21. *Let (X, d) be a metric space. Show that if there is a proper function $f : X \to \mathbb{R}$, then*

$$\tilde{d}(x, y) = d(x, y) + |f(x) - f(y)|$$

is a metric which generates the same topology and for which $(X, \tilde{d})$ is proper.

B.6. Separation

The **distance** between a point $x \in X$ and a subset $Y \subseteq X$ is

$$\operatorname{dist}(x, Y) := \inf_{y \in Y} d(x, y). \tag{B.22}$$

Note that x is a limit point of Y if and only if $\operatorname{dist}(x, Y) = 0$ (Problem B.22).

Lemma B.27. *Let X be a metric space and $Y \subseteq X$ nonempty. Then*

$$|\operatorname{dist}(x, Y) - \operatorname{dist}(z, Y)| \le d(x, z). \tag{B.23}$$

In particular, $x \mapsto \operatorname{dist}(x, Y)$ is continuous.

Proof. Taking the infimum in the triangle inequality $d(x,y) \le d(x,z) + d(z,y)$ shows $\operatorname{dist}(x,Y) \le d(x,z) + \operatorname{dist}(z,Y)$. Hence $\operatorname{dist}(x,Y) - \operatorname{dist}(z,Y) \le d(x,z)$. Interchanging x and z shows $\operatorname{dist}(z,Y) - \operatorname{dist}(x,Y) \le d(x,z)$. □

A topological space is called **normal** if for any two disjoint closed sets C_1 and C_2, there are disjoint open sets O_1 and O_2 such that $C_j \subseteq O_j$, $j = 1, 2$.

Lemma B.28 (Urysohn)**.** *Let X be a topological space. Then X is normal if and only if for every pair of disjoint closed sets C_1 and C_2, there exists a continuous function $f : X \to [0,1]$ which is one on C_1 and zero on C_2.*

If in addition X is locally compact and C_1 is compact, then f can be chosen to have compact support.

Proof. To construct f we choose an open neighborhood O_0 of C_1 (e.g. $O_0 := X \setminus C_2$). Now we could set f equal to one on C_1, equal to zero outside O_0 and equal to $\frac{1}{2}$ on the layer $O_0 \setminus C_1$ in between. Clearly this function is not continuous, but we can successively improve the situation by introducing additional layers in between.

To this end observe that X is normal if and only if for every closed set C and every open neighborhood U of C, there exists an open set O_1 and a closed set C_1 such that $C \subset O_1 \subset C_1 \subset O$ (just use the identification $C_1 \leftrightarrow C$, $C_2 \leftrightarrow X \setminus O$ and $O_1 \leftrightarrow O_1$, $O_2 \leftrightarrow X \setminus C$).

Using our observation we can find an open set $O_{1/2}$ and a closed set $C_{1/2}$ such that $C_1 \subseteq O_{1/2} \subseteq C_{1/2} \subseteq O_0$. Repeating this argument we get two more open and two more closed sets such that

$$C_1 \subseteq O_{3/4} \subseteq C_{3/4} \subseteq O_{1/2} \subseteq C_{1/2} \subseteq O_{1/4} \subseteq C_{1/4} \subseteq O_0.$$

Iterating this construction we get open sets O_q and closed sets C_q for every dyadic rational $q = k/2^n \in [0,1]$ such that $O_q \subseteq C_q$ and $C_p \subseteq O_q$ for $p < q$. Now set $f_n^s(x) := \max\{q = k/2^n | x \in U_q,\ 0 \le k < 2^n\}$ for $x \in O_0$ and $f_n^s(x) := 0$ else as well as $f_n^i(x) := \min\{q = k/2^n | x \notin C_q,\ 0 \le k < 2^n\}$ for $x \notin C_1$ and $f_n^i(x) := 1$ else. Then $f_n^s(x) \nearrow f^s(x) := \sup\{q | x \in U_q\}$ and $f_n^i \searrow f^i(s) := \inf\{q | x \notin C_q\}$. Moreover, if $f_n^s(x) = q$ we have $x \in O_q \setminus O_{q+2^{-n}}$ and depending on $x \in C_{q+2^{-n}}$ or $x \notin C_{q+2^{-n}}$ we have $f_n^i(x) = q + 2^{-n+1}$ or $f_n^i(x) = q + 2^{-n}$, respectively. In particular, $f^s(x) = f^i(x)$. Finally, since $(f^s)^{-1}((r,1]) = \bigcup_{q>r} O_r$ and $(f^i)^{-1}([0,r)) = \bigcap_{q<r} X \setminus C_r$ are open we see that $f := f^s = f^i$ is continuous.

Conversely, given f choose $O_1 := f^{-1}([0,1/2))$ and $O_2 := f^{-1}((1/2,1])$.

For the second claim, observe that there is an open set O_0 such that $\overline{O_0}$ is compact and $C_1 \subset O_0 \subset \overline{O_0} \subset X \setminus C_2$. In fact, for every $x \in C_1$, there is a ball $B_\varepsilon(x)$ such that $\overline{B_\varepsilon(x)}$ is compact and $\overline{B_\varepsilon(x)} \subset X \setminus C_2$. Since C_1

is compact, finitely many of them cover C_1 and we can choose the union of those balls to be O_0. □

Example B.20. In a metric space we can choose $f(x) := \frac{\operatorname{dist}(x,C_2)}{\operatorname{dist}(x,C_1)+\operatorname{dist}(x,C_2)}$ and hence every metric space is normal. ◇

Another important result is the **Tietze extension theorem**:

Theorem B.29 (Tietze)**.** *Suppose C is a closed subset of a normal topological space X. For every continuous function $f : C \to [-1,1]$ there is a continuous extension $\bar{f} : X \to [-1,1]$.*

Proof. The idea is to construct a rough approximation using Urysohn's lemma and then iteratively improve this approximation. To this end we set $C_1 := f^{-1}([\frac{1}{3},1])$ and $C_2 := f^{-1}([-1,-\frac{1}{3}])$ and let g be the function from Urysohn's lemma. Then $f_1 := \frac{2g-1}{3}$ satisfies $|f(x) - f_1(x)| \le \frac{2}{3}$ for $x \in C$ as well as $|f_1(x)| \le \frac{1}{3}$ for all $x \in X$. Applying this same procedure to $f - f_1$ we obtain a function f_2 such that $|f(x) - f_1(x) - f_2(x)| \le \left(\frac{2}{3}\right)^2$ for $x \in C$ and $|f_2(x)| \le \frac{1}{3}\left(\frac{2}{3}\right)$. Continuing this process we arrive at a sequence of functions f_n such that $|f(x) - \sum_{j=1}^n f_j(x)| \le \left(\frac{2}{3}\right)^n$ for $x \in C$ and $|f_n(x)| \le \frac{1}{3}\left(\frac{2}{3}\right)^{n-1}$. By construction the corresponding series converges uniformly to the desired extension $\bar{f} := \sum_{j=1}^\infty f_j$. □

Note that by extending each component we can also handle functions with values in $\mathbb{R}^n$.

A **partition of unity** is a collection of functions $h_j : X \to [0,1]$ such that $\sum_j h_j(x) = 1$. We will only consider the case where there are countably many functions. A partition of unity is **locally finite** if every x has a neighborhood where all but a finite number of the functions h_j vanish. Moreover, given a cover $\{O_j\}$ of X it is called **subordinate** to this cover if every h_j has support contained in some set O_k from this cover. Of course the partition will be locally finite if the cover is locally finite which we can always assume without loss of generality for an open cover if X is a metric space by Lemma B.14.

Lemma B.30. *Let X be a metric space and $\{O_j\}$ a countable open cover. Then there is a continuous **partition of unity** subordinate to this cover. We can even choose the cover such that h_j has support contained in O_j.*

Proof. For notational simplicity we assume $j \in \mathbb{N}$. Now introduce $f_n(x) := \min(1, \sup_{j\le n} d(x, X\setminus O_j))$ and $g_n = f_n - f_{n-1}$ (with the convention $f_0(x) := 0$). Since f_n is increasing we have $0 \le g_n \le 1$. Moreover, $g_n(x) > 0$ implies $d(x, X \setminus O_n)) > 0$ and thus $\operatorname{supp}(g_n) \subset O_n$. Next, by monotonicity $f_\infty := \lim_{n\to\infty} f_n = \sum_n g_n$ exists and $f_\infty(x) = 0$ implies $d(x, X \setminus O_j) = 0$

for all j, that is, $x \notin O_j$ for all j. Hence f_∞ is everywhere positive since $\{O_j\}$ is a cover. Finally, by

$$\begin{aligned}|f_n(x) - f_n(y)| &\le |\sup_{j\le n} d(x, X\setminus O_j) - \sup_{j\le n} d(y, X\setminus O_j)| \\ &\le \sup_{j\le n} |d(x, X\setminus O_j) - d(y, X\setminus O_j)| \le d(x,y)\end{aligned}$$

we see that all f_n (and hence all g_n) are continuous. Moreover, the very same argument shows that f_∞ is continuous and thus we have found the required partition of unity $h_j = g_j/f_\infty$. □

Finally, we also mention that in the case of subsets of $\mathbb{R}^n$ there is a smooth partition of unity. To this end recall that for every point $x \in \mathbb{R}^n$ there is a smooth bump function with values in $[0,1]$ which is positive at x and supported in a given neighborhood of x.

Example B.21. The standard bump function is $\phi(x) := \exp(\frac{1}{|x|^2-1})$ for $|x| < 1$ and $\phi(x) = 0$ otherwise. To show that this function is indeed smooth it suffices to show that all left derivatives of $f(r) = \exp(\frac{1}{r-1})$ at $r = 1$ vanish, which can be done using l'Hôpital's rule. By scaling and translation $\phi(\frac{x-x_0}{r})$ we get a bump function which is supported in $B_r(x_0)$ and satisfies $\phi(\frac{x-x_0}{r})\big|_{x=x_0} = \phi(0) = \mathrm{e}^{-1}$. ⋄

Lemma B.31. *Let $X \subseteq \mathbb{R}^n$ be open and $\{O_j\}$ a countable open cover. Then there is a locally finite partition of unity of functions from $C_c^\infty(X)$ subordinate to this cover. If the cover is finite so will be the partition.*

Proof. Let U_j be as in Lemma B.25 (iv). For $\overline{U}_j$ choose finitely many bump functions $\tilde{h}_{j,k}$ such that $\tilde{h}_{j,1}(x) + \cdots + \tilde{h}_{j,k_j}(x) > 0$ for every $x \in \overline{U}_j \setminus U_{j-1}$ and such that $\operatorname{supp}(\tilde{h}_{j,k})$ is contained in one of the O_k and in $U_{j+1}\setminus U_{j-1}$. Then $\{\tilde{h}_{j,k}\}_{j,k}$ is locally finite and hence $h := \sum_{j,k} \tilde{h}_{j,k}$ is a smooth function which is everywhere positive. Finally, $\{\tilde{h}_{j,k}/h\}_{j,k}$ is a partition of unity of the required type. □

Problem B.22. *Show* $\operatorname{dist}(x,Y) = \operatorname{dist}(x,\overline{Y})$. *Moreover, show* $x \in \overline{Y}$ *if and only if* $\operatorname{dist}(x,Y) = 0$.

Problem B.23. *Let $Y, Z \subseteq X$ and define*

$$\operatorname{dist}(Y,Z) := \inf_{y\in Y, z\in Z} d(y,z).$$

Show $\operatorname{dist}(Y,Z) = \operatorname{dist}(\overline{Y},\overline{Z})$. *Moreover, show that if K is compact, then* $\operatorname{dist}(K,Y) > 0$ *if and only if* $K \cap \overline{Y} = \emptyset$.

Problem B.24. *Let $K \subseteq U$ with K compact and U open. Show that there is some $\varepsilon > 0$ such that $K_\varepsilon := \{x \in X \,|\, \operatorname{dist}(x,K) < \varepsilon\} \subseteq U$.*

Problem B.25. *Let (X, d) be a locally compact metric space. Then X is σ-compact if and only if there exists a proper function $f : X \to [0, \infty)$. (Hint: Let U_n be as in item (iv) of Lemma B.25 and use Uryson's lemma to find functions $f_n : X \to [0, 1]$ such that $f(x) = 0$ for $x \in \overline{U_n}$ and $f(x) = 1$ for $x \in X \setminus U_{n+1}$. Now consider $f = \sum_{n=1}^{\infty} f_n$.)*

B.7. Connectedness

Roughly speaking a topological space X is disconnected if it can be split into two (nonempty) separated sets. This of course raises the question what should be meant by separated. Evidently it should be more than just disjoint since otherwise we could split any space containing more than one point. Hence we will consider two sets separated if each is disjoint form the closure of the other. Note that if we can split X into two separated sets $X = U \cup V$ then $\overline{U} \cap V = \emptyset$ implies $\overline{U} = U$ (and similarly $\overline{V} = V$). Hence both sets must be closed and thus also open (being complements of each other). This brings us to the following definition:

A topological space X is called **disconnected** if one of the following equivalent conditions holds

- X is the union of two nonempty separated sets.
- X is the union of two nonempty disjoint open sets.
- X is the union of two nonempty disjoint closed sets.

In this case the sets from the splitting are both open and closed. A topological space X is called **connected** if it cannot be split as above. That is, in a connected space X the only sets which are both open and closed are $\emptyset$ and X. This last observation is frequently used in proofs: If the set where a property holds is both open and closed it must either hold nowhere or everywhere. In particular, any mapping from a connected to a discrete space must be constant since the inverse image of a point is both open and closed.

A subset of X is called (dis-)connected if it is (dis-)connected with respect to the relative topology. In other words, a subset $A \subseteq X$ is disconnected if there are disjoint nonempty open sets U and V which split A according to $A = (U \cap A) \cup (V \cap A)$.

Example B.22. In $\mathbb{R}$ the nonempty connected sets are precisely the intervals (Problem B.26). Consequently $A = [0, 1] \cup [2, 3]$ is disconnected with $[0, 1]$ and $[2, 3]$ being its components (to be defined precisely below). While you might be reluctant to consider the closed interval $[0, 1]$ as open, it is important to observe that it is the relative topology which is relevant here. $\diamond$

The maximal connected subsets (ordered by inclusion) of a nonempty topological space X are called the **connected components** of X.

Example B.23. Consider $\mathbb{Q} \subseteq \mathbb{R}$. Then every rational point is its own component (if a set of rational points contains more than one point there would be an irrational point in between which can be used to split the set). ⋄

In many applications one also needs the following stronger concept. A space X is called **path-connected** if any two points $x, y \in X$ can be joined by a **path**, that is a continuous map $\gamma : [0,1] \to X$ with $\gamma(0) = x$ and $\gamma(1) = y$. A space is called **locally (path-)connected** if for every given point and every open set containing that point there is a smaller open set which is (path-)connected.

Example B.24. Every normed vector space is (locally) path-connected since every ball is path-connected (consider straight lines). Every open subset of a locally (path-)connected space is locally (path-)connected. ⋄

Every path-connected space is connected. In fact, if $X = U \cup V$ were disconnected but path-connected we could choose $x \in U$ and $y \in V$ plus a path γ joining them. But this would give a splitting $[0,1] = \gamma^{-1}(U) \cup \gamma^{-1}(V)$ contradicting our assumption. The converse however is not true in general as a space might be impassable (an example will follow).

Example B.25. The spaces $\mathbb{R}$ and $\mathbb{R}^n$, $n > 1$, are not homeomorphic. In fact, removing any point form $\mathbb{R}$ gives a disconnected space while removing a point form $\mathbb{R}^n$ still leaves it (path-)connected. ⋄

We collect a few simple but useful properties below.

Lemma B.32. *Suppose X and Y are topological spaces.*

(i) *Suppose $f : X \to Y$ is continuous. Then if X is (path-)connected so is the image $f(X)$.*

(ii) *Suppose $A_\alpha \subseteq X$ are (path-)connected and $\bigcap_\alpha A_\alpha \neq \emptyset$. Then $\bigcup_\alpha A_\alpha$ is (path-)connected*

(iii) *$A \subseteq X$ is (path-)connected if and only if any two points $x, y \in A$ are contained in a (path-)connected set $B \subseteq A$*

(iv) *Suppose $X_1, \ldots, X_n$ are (path-)connected then so is $\bigtimes_{j=1}^n X_j$.*

(v) *Suppose $A \subseteq X$ is connected, then $\overline{A}$ is connected.*

(vi) *A locally path-connected space is path-connected if and only if it is connected.*

Proof. (i). Suppose we have a splitting $f(X) = U \cup V$ into nonempty disjoint sets which are open in the relative topology. Hence, there are open

sets U' and V' such that $U = U' \cap f(X)$ and $V = V' \cap f(X)$ implying that the sets $f^{-1}(U) = f^{-1}(U')$ and $f^{-1}(V) = f^{-1}(V')$ are open. Thus we get a corresponding splitting $X = f^{-1}(U) \cup f^{-1}(V)$ into nonempty disjoint open sets contradicting connectedness of X.

If X is path connected, let $y_1 = f(x_1)$ and $y_2 = f(x_2)$ be given. If γ is a path connecting x_1 and x_2, then $f \circ \gamma$ is a path connecting y_1 and y_2.

(ii). Let $A = \bigcup_\alpha A_\alpha$ and suppose there is a splitting $A = (U \cap A) \cup (V \cap A)$. Since there is some $x \in \bigcap_\alpha A_\alpha$ we can assume $x \in U$ w.l.o.g. Hence there is a splitting $A_\alpha = (U \cap A_\alpha) \cup (V \cap A_\alpha)$ and since A_α is connected and $U \cap A_\alpha$ is nonempty we must have $V \cap A_\alpha = \emptyset$. Hence $V \cap A = \emptyset$ and A is connected.

If the $x \in A_\alpha$ and $y \in A_\beta$ then choose a point $z \in A_\alpha \cap A_\beta$ and paths γ_α from x to z and γ_β from z to y, then $\gamma_\alpha \odot \gamma_\beta$ is a path from x to y, where $\gamma_\alpha \odot \gamma_\beta(t) = \gamma_\alpha(2t)$ for $0 \le t \le \frac{1}{2}$ and $\gamma_\alpha \odot \gamma_\beta(t) = \gamma_\beta(2t-1)$ for $\frac{1}{2} \le t \le 1$ (cf. Problem B.17).

(iii). If X is connected we can choose $B = A$. Conversely, fix some $x \in A$ and let B_y be the corresponding set for the pair x, y. Then $A = \bigcup_{y \in A} B_y$ is (path-)connected by the previous item.

(iv). We first consider two spaces $X = X_1 \times X_2$. Let $x, y \in X$. Then $\{x_1\} \times X_2$ is homeomorphic to X_2 and hence (path-)connected. Similarly $X_1 \times \{y_2\}$ is (path-)connected as well as $\{x_1\} \times X_2 \cup X_1 \times \{y_2\}$ by (ii) since both sets contain $(x_1, y_2) \in X$. But this last set contains both x, y and hence the claim follows from (iii). The general case follows by iterating this result.

(v). Let $x \in \overline{A}$. Then $\{x\}$ and A cannot be separated and hence $\{x\} \cup A$ is connected. The rest follows from (ii).

(vi). Consider the set $U(x)$ of all points connected to a fixed point x (via paths). If $y \in U(x)$ then so is any path-connected neighborhood of y by gluing paths (as in item (ii)). Hence $U(x)$ is open. Similarly, if $y \in \overline{U(x)}$ then any path-connected neighborhood of y will intersect $U(y)$ and hence $y \in U(x)$. Thus $U(x)$ is also closed and hence must be all of X by connectedness. The converse is trivial. □

A few simple consequences are also worth while noting: If two different components contain a common point, their union is again connected contradicting maximality. Hence two different components are always disjoint. Moreover, every point is contained in a component, namely the union of all connected sets containing this point. In other words, the components of any topological space X form a partition of X (i.e, they are disjoint, nonempty, and their union is X). Moreover, every component is a closed subset of the

original space X. In the case where their number is finite we can take complements and each component is also an open subset (the rational numbers from our first example show that components are not open in general). In a locally (path-)connected space, components are open and (path-)connected by (vi) of the last lemma. Note also that in a second countable space an open set can have at most countably many components (take those sets from a countable base which are contained in some component, then we have a surjective map from these sets to the components).

Example B.26. Consider the graph of the function $f : (0,1] \to \mathbb{R}$, $x \mapsto \sin(\frac{1}{x})$. Then $\Gamma(f) \subseteq \mathbb{R}^2$ is path-connected and its closure $\overline{\Gamma(f)} = \Gamma(f) \cup \{0\} \times [-1,1]$ is connected. However, $\overline{\Gamma(f)}$ is not path-connected as there is no path from $(1,0)$ to $(0,0)$. Indeed, suppose γ were such a path. Then, since γ_1 covers $[0,1]$ by the intermediate value theorem (see below), there is a sequence $t_n \to 1$ such that $\gamma_1(t_n) = \frac{2}{(2n+1)\pi}$. But then $\gamma_2(t_n) = (-1)^n \not\to 0$ contradicting continuity. ⋄

Theorem B.33 (Intermediate Value Theorem)**.** *Let X be a connected topological space and $f : X \to \mathbb{R}$ be continuous. For any $x, y \in X$ the function f attains every value between $f(x)$ and $f(y)$.*

Proof. The image $f(X)$ is connected and hence an interval. □

Problem B.26. *A nonempty subset of $\mathbb{R}$ is connected if and only if it is an interval.*

B.8. Continuous functions on metric spaces

Let X, Y be topological spaces and let $C(X,Y)$ be the set of all continuous functions $f : X \to Y$. Set $C(X) := C(X, \mathbb{C})$. Moreover, if Y is a metric space then $C_b(X,Y)$ will denote the set of all bounded continuous functions, that is, those continuous functions for which $\sup_{x\in X} d_Y(f(x), y)$ is finite for some (and hence for all) $y \in Y$. Note that by the extreme value theorem $C_b(X,Y) = C(X,Y)$ if X is compact. For these functions we can introduce a metric via

$$d(f,g) := \sup_{x\in X} d_Y(f(x), g(x)). \tag{B.24}$$

In fact, the requirements for a metric are readily checked. Of course convergence with respect to this metric implies pointwise convergence but not the other way round.

Example B.27. Consider $X := [0,1]$, then $f_n(x) := \max(1 - |nx - 1|, 0)$ converges pointwise to 0 (in fact, $f_n(0) = 0$ and $f_n(x) = 0$ on $[\frac{2}{n}, 1]$) but not with respect to the above metric since $f_n(\frac{1}{n}) = 1$. ⋄

This kind of convergence is known as **uniform convergence** since for every positive ε there is some index N (independent of x) such that $d_Y(f_n(x), f(x)) < \varepsilon$ for $n \geq N$. In contradistinction, in the case of pointwise convergence, N is allowed to depend on x. One advantage is that continuity of the limit function comes for free.

Theorem B.34. *Let X be a topological space and Y a metric space. Suppose $f_n \in C(X, Y)$ converges uniformly to some function $f : X \to Y$. Then f is continuous.*

Proof. Let $x \in X$ be given and write $y := f(x)$. We need to show that $f^{-1}(B_\varepsilon(y))$ is a neighborhood of x for every $\varepsilon > 0$. So fix ε. Then we can find an N such that $d(f_n, f) < \frac{\varepsilon}{2}$ for $n \geq N$ implying $f_N^{-1}(B_{\varepsilon/2}(y)) \subseteq f^{-1}(B_\varepsilon(y))$ since $d(f_n(z), y) < \frac{\varepsilon}{2}$ implies $d(f(z), y) \leq d(f(z), f_n(z)) + d(f_n(z), y) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ for $n \geq N$. □

Corollary B.35. *Let X be a topological space and Y a complete metric space. The space $C_b(X, Y)$ together with the metric d is complete.*

Proof. Suppose f_n is a Cauchy sequence with respect to d, then $f_n(x)$ is a Cauchy sequence for fixed x and has a limit since Y is complete. Call this limit $f(x)$. Then $d_Y(f(x), f_n(x)) = \lim_{m\to\infty} d_Y(f_m(x), f_n(x)) \leq \sup_{m\geq n} d(f_m, f_n)$ and since this last expression goes to 0 as $n \to \infty$, we see that f_n converges uniformly to f. Moreover, $f \in C(X, Y)$ by the previous theorem so we are done. □

Let Y be a vector space. By $C_c(X, Y) \subseteq C_b(X, Y)$ we will denote the set of continuous functions with compact support. Its closure will be denoted by $C_0(X, Y) := \overline{C_c(X, Y)} \subseteq C_b(X, Y)$. Of course if X is compact all these spaces agree $C_c(X, Y) = C_0(X, Y) = C_b(X, Y) = C(X, Y)$. In the general case one at least assumes X to be locally compact since if we take a closed neighborhood V of $f(x) \neq 0$ which does not contain 0, then $f^{-1}(U)$ will be a compact neighborhood of x. Hence without this assumption f must vanish on every point which does not have a compact neighborhood and $C_c(X, Y)$ will not be sufficiently rich.

Example B.28. Let X be a separable and locally compact metric space and $Y = \mathbb{C}^n$. Then

$$C_0(X, \mathbb{C}^n) = \{f \in C_b(X, \mathbb{C}^n) |\ \forall \varepsilon > 0, \exists K \subseteq X \text{ compact} : |f(x)| < \varepsilon,\ x \in X \setminus K\}. \tag{B.25}$$

To see this denote the set on the right-hand side by C. Let K_m be an increasing sequence of compact sets with $K_m \nearrow X$ (Lemma B.25) and let φ_m be a corresponding sequence as in Urysohn's lemma (Lemma B.28). Then for $f \in C$ the sequence $f_m = \varphi_m f \in C_c(X, \mathbb{C}^n)$ will converge to f.

Conversely, if $f_n \in C_c(X, \mathbb{C}^n)$ converges to $f \in C_b(X, \mathbb{C}^n)$, then given $\varepsilon > 0$ choose $K = \operatorname{supp}(f_m)$ for some m with $d(f_m, f) < \varepsilon$.

In the case where X is an open subset of $\mathbb{R}^n$ this says that $C_0(X, Y)$ are those which vanish at the boundary (including the case as $|x| \to \infty$ if X is unbounded). $\diamond$

Lemma B.36. *If X is a separable and locally compact space then $C_0(X, \mathbb{C}^n)$ is separable.*

Proof. Choose a countable base $\mathcal{B}$ for X and let $\mathcal{I}$ the collection of all balls in $\mathbb{C}^n$ with rational radius and center. Given $O_1, \dots, O_m \in \mathcal{B}$ and $I_1, \dots, I_m \in \mathcal{I}$ we say that $f \in C_c(X, \mathbb{C}^n)$ is adapted to these sets if $\operatorname{supp}(f) \subseteq \bigcup_{j=1}^m O_j$ and $f(O_j) \subseteq I_j$. The set of all tuples $(O_j, I_j)_{1 \le j \le m}$ is countable and for each tuple we choose a corresponding adapted function (if there exists one at all). Then the set of these functions $\mathcal{F}$ is dense. It suffices to show that the closure of $\mathcal{F}$ contains $C_c(X, \mathbb{C}^n)$. So let $f \in C_c(X, \mathbb{C}^n)$ and let $\varepsilon > 0$ be given. Then for every $x \in X$ there is some neighborhood $O(x) \in \mathcal{B}$ such that $|f(x) - f(y)| < \varepsilon$ for $y \in O(x)$. Since $\operatorname{supp}(f)$ is compact, it can be covered by $O(x_1), \dots, O(x_m)$. In particular $f(O(x_j)) \subseteq B_\varepsilon(f(x_j))$ and we can find a ball I_j of radius at most 2ε with $f(O(x_j)) \subseteq I_j$. Now let g be the function from $\mathcal{F}$ which is adapted to $(O(x_j), I_j)_{1 \le j \le m}$ and observe that $|f(x) - g(x)| < 4\varepsilon$ since $x \in O(x_j)$ implies $f(x), g(x) \in I_j$. $\square$

Let X, Y be metric spaces. A function $f \in C(X, Y)$ is called **uniformly continuous** if for every $\varepsilon > 0$ there is a $\delta > 0$ such that

$$d_Y(f(x), f(y)) \le \varepsilon \quad \text{whenever} \quad d_X(x, y) < \delta. \tag{B.26}$$

Note that with the usual definition of continuity on fixes x and then chooses δ depending on x. Here δ has to be independent of x. If the domain is compact, this extra condition comes for free.

Theorem B.37. *Let X be a compact metric space and Y a metric space. Then every $f \in C(X, Y)$ is uniformly continuous.*

Proof. Suppose the claim were wrong. Fix $\varepsilon > 0$. Then for every $\delta_n = \frac{1}{n}$ we can find x_n, y_n with $d_X(x_n, y_n) < \delta_n$ but $d_Y(f(x_n), f(y_n)) \ge \varepsilon$. Since X is compact we can assume that x_n converges to some $x \in X$ (after passing to a subsequence if necessary). Then we also have $y_n \to x$ implying $d_Y(f(x_n), f(y_n)) \to 0$, a contradiction. $\square$

Note that a uniformly continuous function maps Cauchy sequences to Cauchy sequences. This fact can be used to extend a uniformly continuous function to boundary points.

Theorem B.38. *Let X be a metric space and Y a complete metric space. A uniformly continuous function $f : A \subseteq X \to Y$ has a unique continuous extension $\bar{f} : \overline{A} \to Y$. This extension is again uniformly continuous.*

Proof. If there is an extension it must be $\bar{f}(x) := \lim_{n\to\infty} f(x_n)$, where $x_n \in A$ is some sequence converging to $x \in \overline{A}$. Indeed, since x_n converges, $f(x_n)$ is Cauchy and hence has a limit since Y is assumed complete. Moreover, uniqueness of limits shows that $\bar{f}(x)$ is independent of the sequence chosen. Also $\bar{f}(x) = f(x)$ for $x \in A$ by continuity. To see that $\bar{f}$ is uniformly continuous, let $\varepsilon > 0$ be given and choose a δ which works for f. Then for given x, y with $d_X(x, y) < \frac{\delta}{3}$ we can find $\tilde{x}, \tilde{y} \in A$ with $d_X(\tilde{x}, x) < \frac{\delta}{3}$ and $d_Y(f(\tilde{x}), \bar{f}(x)) \leq \varepsilon$ as well as $d_X(\tilde{y}, y) < \frac{\delta}{3}$ and $d_Y(f(\tilde{y}), \bar{f}(y)) \leq \varepsilon$. Hence $d_Y(\bar{f}(x), \bar{f}(y)) \leq d_Y(\bar{f}(x), f(\tilde{x})) + d_Y(f(\tilde{x}), f(\tilde{y})) + d_Y(f(x), \bar{f}(y)) \leq 3\varepsilon$. □

Next we want to identify relatively compact subsets in $C(X, Y)$. A family of functions $F \subset C(X, Y)$ is called (pointwise) **equicontinuous** if for every $\varepsilon > 0$ and every $x \in X$ there is a neighborhood $U(x)$ of x such that

$$d_Y(f(x), f(y)) \leq \varepsilon \quad \text{whenever} \quad y \in U(x), \quad \forall f \in F. \tag{B.27}$$

Theorem B.39 (Arzelà–Ascoli)**.** *Let X be a compact space and Y a proper metric space. Let $F \subset C(X, Y)$ be a family of continuous functions. Then every sequence from F has a uniformly convergent subsequence if and only if F is equicontinuous and the set $\{f(x) | f \in F\}$ is bounded for every $x \in X$. In this case F is even bounded.*

Proof. Suppose F is equicontinuous and pointwise bounded. Fix $\varepsilon > 0$. By compactness of X there are finitely many points $x_1, \ldots, x_n \in X$ such that the neighborhoods $U(x_j)$ (from the definition of equicontinuity) cover X. Now first of all note that, F is bounded since $d_Y(f(x), y) \leq \max_j \sup_{f\in F} d_Y(f(x_j), y) + \varepsilon$ for every $x \in X$ and every $f \in F$.

Next consider $P : C(X, Y) \to Y^n$, $P(f) = (f(x_1), \ldots, f(x_n))$. Then $P(F)$ is bounded and $d(f, g) \leq 3\varepsilon$ whenever $d_Y(P(f), P(g)) < \varepsilon$. Indeed, just note that for every x there is some j such that $x \in U(x_j)$ and thus $d_Y(f(x), g(x)) \leq d_Y(f(x), f(x_j)) + d_Y(f(x_j), g(x_j)) + d_Y(g(x_j), g(x)) \leq 3\varepsilon$. Hence F is relatively compact by Lemma 1.12.

Conversely, suppose F is relatively compact. Then F is totally bounded and hence bounded. To see equicontinuity fix $x \in X$, $\varepsilon > 0$ and choose a corresponding ε-cover $\{B_\varepsilon(f_j)\}_{j=1}^n$ for F. Pick a neighbrohood $U(x)$ such that $y \in U(x)$ implies $d_Y(f_j(y), f_j(x)) < \varepsilon$ for all $1 \leq j \leq n$. Then $f \in B_\varepsilon(f_j)$ for some j and hence $d_Y(f(y), f(x)) \leq d_Y(f(y), f_j(y)) + d_Y(f_j(y), f_j(x)) + d_Y(f_j(x), f(x)) \leq 3\varepsilon$, proving equicontinuity. □

In many situations a certain property can be seen for a class of *nice* functions and then extended to a more general class of functions by approximation. In this respect it is important to identify classes of functions which allow to approximate *all* functions. That is, in our present situation we are looking for functions which are dense in $C(X, Y)$. For example, the classical Weierstraß approximation theorem (Theorem 1.3) says that the polynomials are dense in $C([a, b])$ for any compact interval. Here we will present a generalization of this result. For its formulation observe that $C(X)$ is not only a vector space but also comes with a natural product, given by pointwise multiplication of functions, which turns it into an algebra over $\mathbb{C}$. By a subalgebra we will mean a subspace which is closed under multiplication and by a $*$-subalgebra we will mean a subalgebra which is also closed under complex conjugation. The ($*$-)subalgebra generated by a set is of course the smallest ($*$-)subalgebra containing this set.

The proof will use the fact that the absolute value can be approximated by polynomials on $[-1, 1]$. This of course follows from the Weierstraß approximation theorem but can also be seen directly by defining the sequence of polynomials p_n via

$$p_1(t) := 0, \qquad p_{n+1}(t) := p_n(t) + \frac{t^2 - p_n(t)^2}{2}. \tag{B.28}$$

Then this sequence of polynomials satisfies $p_n(t) \le p_{n+1}(t) \le |t|$ and converges pointwise to $|t|$ for $t \in [-1, 1]$. Hence by Dini's theorem (Problem B.29) it converges uniformly. By scaling we get the corresponding result for arbitrary compact subsets of the real line.

Theorem B.40 (Stone–Weierstraß, real version)**.** *Suppose K is a compact topological space and consider $C(K, \mathbb{R})$. If $F \subset C(K, \mathbb{R})$ contains the identity 1 and separates points (i.e., for every $x_1 \neq x_2$ there is some function $f \in F$ such that $f(x_1) \neq f(x_2)$), then the subalgebra generated by F is dense.*

Proof. Denote by A the subalgebra generated by F. Note that if $f \in \overline{A}$, we have $|f| \in \overline{A}$: Choose a polynomial $p_n(t)$ such that $\big||t| - p_n(t)\big| < \frac{1}{n}$ for $t \in f(K)$ and hence $p_n(f) \to |f|$.

In particular, if $f, g \in \overline{A}$, we also have

$$\max\{f, g\} = \frac{(f + g) + |f - g|}{2} \in \overline{A}, \quad \min\{f, g\} = \frac{(f + g) - |f - g|}{2} \in \overline{A}.$$

Now fix $f \in C(K, \mathbb{R})$. We need to find some $f^\varepsilon \in \overline{A}$ with $\|f - f^\varepsilon\|_\infty < \varepsilon$.

First of all, since A separates points, observe that for given $y, z \in K$ there is a function $f_{y,z} \in A$ such that $f_{y,z}(y) = f(y)$ and $f_{y,z}(z) = f(z)$ (show this). Next, for every $y \in K$ there is a neighborhood $U(y)$ such that

$$f_{y,z}(x) > f(x) - \varepsilon, \quad x \in U(y),$$

and since K is compact, finitely many, say $U(y_1), \dots, U(y_j)$, cover K. Then

$$f_z = \max\{f_{y_1,z}, \dots, f_{y_j,z}\} \in \overline{A}$$

and satisfies $f_z > f - \varepsilon$ by construction. Since $f_z(z) = f(z)$ for every $z \in K$, there is a neighborhood $V(z)$ such that

$$f_z(x) < f(x) + \varepsilon, \quad x \in V(z),$$

and a corresponding finite cover $V(z_1), \dots, V(z_k)$. Now

$$f^\varepsilon = \min\{f_{z_1}, \dots, f_{z_k}\} \in \overline{A}$$

satisfies $f^\varepsilon < f + \varepsilon$. Since $f - \varepsilon < f_{z_l}$ for all $1 \le l \le k$ we have $f - \varepsilon < f^\varepsilon$ and we have found a required function. □

Theorem B.41 (Stone–Weierstraß)**.** *Suppose K is a compact topological space and consider $C(K)$. If $F \subset C(K)$ contains the identity 1 and separates points, then the $*$-subalgebra generated by F is dense.*

Proof. Just observe that $\tilde{F} = \{\mathrm{Re}(f), \mathrm{Im}(f) | f \in F\}$ satisfies the assumption of the real version. Hence every real-valued continuous function can be approximated by elements from the subalgebra generated by $\tilde{F}$; in particular, this holds for the real and imaginary parts for every given complex-valued function. Finally, note that the subalgebra spanned by $\tilde{F}$ is contained in the $*$-subalgebra spanned by F. □

Note that the additional requirement of being closed under complex conjugation is crucial: The functions holomorphic on the unit disc and continuous on the boundary separate points, but they are not dense (since the uniform limit of holomorphic functions is again holomorphic).

Corollary B.42. *Suppose K is a compact topological space and consider $C(K)$. If $F \subset C(K)$ separates points, then the closure of the $*$-subalgebra generated by F is either $C(K)$ or $\{f \in C(K) | f(t_0) = 0\}$ for some $t_0 \in K$.*

Proof. There are two possibilities: either all $f \in F$ vanish at one point $t_0 \in K$ (there can be at most one such point since F separates points) or there is no such point.

If there is no such point, then the identity can be approximated by elements in $\overline{A}$: First of all note that $|f| \in \overline{A}$ if $f \in \overline{A}$, since the polynomials $p_n(t)$ used to prove this fact can be replaced by $p_n(t) - p_n(0)$ which contain no constant term. Hence for every point y we can find a nonnegative function in $\overline{A}$ which is positive at y and by compactness we can find a finite sum of such functions which is positive everywhere, say $m \le f(t) \le M$. Now approximate $\min(m^{-1}t, t^{-1})$ by polynomials $q_n(t)$ (again a constant term is not needed) to conclude that $q_n(f) \to f^{-1} \in \overline{A}$. Hence $1 = f \cdot f^{-1} \in \overline{A}$ as claimed and so $\overline{A} = C(K)$ by the Stone–Weierstraß theorem.

If there is such a t_0 we have $\overline{A} \subseteq \{f \in C(K) | f(t_0) = 0\}$ and the identity is clearly missing from $\overline{A}$. However, adding the identity to $\overline{A}$ we get $\overline{A} + \mathbb{C} = C(K)$ by the Stone–Weierstraß theorem. Moreover, if $f \in C(K)$ with $f(t_0) = 0$ we get $f = \tilde{f} + \alpha$ with $\tilde{f} \in \overline{A}$ and $\alpha \in \mathbb{C}$. But $0 = f(t_0) = \tilde{f}(t_0) + \alpha = \alpha$ implies $f = \tilde{f} \in \overline{A}$, that is, $\overline{A} = \{f \in C(K) | f(t_0) = 0\}$. □

Problem B.27. *Suppose X is compact and connected and let $F \subset C(X, Y)$ be a family of equicontinuous functions. Then $\{f(x_0) | f \in F\}$ bounded for one x_0 implies F bounded.*

Problem B.28. *Let X, Y be metric spaces. A family of functions $F \subset C(X, Y)$ is called* ***uniformly equicontinuous*** *if for every $\varepsilon > 0$ there is a $\delta > 0$ such that*

$$d_Y(f(x), f(y)) \leq \varepsilon \quad \text{whenever} \quad d_X(x, y) < \delta, \quad \forall f \in F. \tag{B.29}$$

Show that if X is compact, then a family F is pointwise equicontinuous if and only if it is uniformly equicontinuous.

Problem* B.29 (Dini's theorem)**.** *Suppose X is compact and let $f_n \in C(X)$ be a sequence of decreasing (or increasing) functions converging pointwise $f_n(x) \searrow f(x)$ to some function $f \in C(X)$. Then $f_n \to f$ uniformly. (Hint: Reduce it to the case $f_n \searrow 0$ and apply the finite intersection property to $f_n^{-1}([\varepsilon, \infty)$.)*

Problem B.30. *Let $k \in \mathbb{N}$ and $I \subseteq \mathbb{R}$. Show that the $*$-subalgebra generated by $f_{z_0}(t) = \frac{1}{(t-z_0)^k}$ for one $z_0 \in \mathbb{C}$ is dense in the set $C_0(I)$ of continuous functions vanishing at infinity:*

- *for $I = \mathbb{R}$ if $z_0 \in \mathbb{C} \backslash \mathbb{R}$ and $k = 1$ or $k = 2$,*
- *for $I = [a, \infty)$ if $z_0 \in (-\infty, a)$ and k arbitrary,*
- *for $I = (-\infty, a] \cup [b, \infty)$ if $z_0 \in (a, b)$ and k odd.*

(Hint: Add ∞ to $\mathbb{R}$ to make it compact.)

Problem B.31. *Let $U \subseteq \mathbb{C} \backslash \mathbb{R}$ be a set which has a limit point and is symmetric under complex conjugation. Show that the span of $\{(t - z)^{-1} | z \in U\}$ is dense in the set $C_0(\mathbb{R})$ of continuous functions vanishing at infinity. (Hint: The product of two such functions is in the span provided they are different.)*

Problem B.32. *Let $K \subseteq \mathbb{C}$ be a compact set. Show that the set of all functions $f(z) = p(x, y)$, where $p : \mathbb{R}^2 \to \mathbb{C}$ is polynomial and $z = x + \mathrm{i}y$, is dense in $C(K)$.*

Bibliography

[1] H. W. Alt, *Lineare Funktionalanalysis*, 4th ed., Springer, Berlin, 2002.

[2] H. Bauer, *Measure and Integration Theory*, de Gruyter, Berlin, 2001.

[3] M. Berger and M. Berger, *Perspectives in Nonlinearity*, Benjamin, New York, 1968.

[4] A. Bowers and N. Kalton, *An Introductory Course in Functional Analysis*, Springer, New York, 2014.

[5] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, 2011.

[6] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer, New York, 1982.

[7] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer, New York, 1994.

[8] K. Deimling, *Nichtlineare Gleichungen und Abbildungsgrade*, Springer, Berlin, 1974.

[9] K. Deimling, *Nonlinear Functional Analysis*, Springer, Berlin, 1985.

[10] E. DiBenedetto, *Real Analysis*, Birkhäuser, Boston, 2002.

[11] L. C. Evans, *Weak Convergence Methods for nonlinear Partial Differential Equations*, CBMS 74, American Mathematical Society, Providence, 1990.

[12] L. C. Evans, *Partial Differential Equations*, 2nd ed., American Mathematical Society, Providence, 2010.

[13] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd. ed., Wiley, Hoboken NJ, 1999.

[14] J. Franklin, *Methods of Mathematical Economics*, Springer, New York, 1980.

[15] I. Gohberg, S. Goldberg, and M.A. Kaashoek, *Basic Classes of Linear Opeartors*, Springer, Basel, 2003.

[16] J. Goldstein, *Semigroups of Linear Operators and Appications*, Oxford University Press, New York, 1985.

[17] L. Grafakos, *Classical Fourier Analysis*, 2nd ed., Springer, New York, 2008.

[18] L. Grafakos, *Modern Fourier Analysis*, 2nd ed., Springer, New York, 2009.

[19] G. Grubb, *Distributions and Operators*, Springer, New York, 2009.

[20] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer, Berlin, 1965.

[21] K. Jänich, *Toplogy*, Springer, New York, 1995.

[22] I. Kaplansky, *Set Theory and Metric Spaces*, AMS Chelsea, Providence, 1972.

[23] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1966.

[24] J. L. Kelley, *General Topology*, Springer, New York, 1955.

[25] O. A. Ladyzhenskaya, *The Boundary Values Problems of Mathematical Physics*, Springer, New York, 1985.

[26] P. D. Lax, *Functional Analysis*, Wiley, New York, 2002.

[27] E. Lieb and M. Loss, *Analysis*, 2nd ed., Amer. Math. Soc., Providence, 2000.

[28] G. Leoni, *A First Course in Sobolev Spaces*, Amer. Math. Soc., Providence, 2009.

[29] N. Lloyd, *Degree Theory*, Cambridge University Press, London, 1978.

[30] R. Meise and D. Vogt, *Introduction to Functional Analysis*, Oxford University Press, Oxford, 2007.

[31] F. W. J. Olver et al., *NIST Handbook of Mathematical Functions*, Cambridge University Press, Cambridge, 2010.

[32] I. K. Rana, *An Introduction to Measure and Integration*, 2nd ed., Amer. Math. Soc., Providence, 2002.

[33] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I. Functional Analysis*, rev. and enl. edition, Academic Press, San Diego, 1980.

[34] J. R. Retherford, *Hilbert Space: Compact Operators and the Trace Theorem*, Cambridge University Press, Cambridge, 1993.

[35] J.J. Rotman, *Introduction to Algebraic Topology*, Springer, New York, 1988.

[36] H. Royden, *Real Analysis*, Prencite Hall, New Jersey, 1988.

[37] W. Rudin, *Real and Complex Analysis*, 3rd edition, McGraw-Hill, New York, 1987.

[38] M. Růžička, *Nichtlineare Funktionalanalysis*, Springer, Berlin, 2004.

[39] H. Schröder, *Funktionalanalysis*, 2nd ed., Harri Deutsch Verlag, Frankfurt am Main 2000.

[40] B. Simon, *A Comprehensive Course in Analysis*, Amer. Math. Soc., Providence, 2015.

[41] L. A. Steen and J. A. Seebach, Jr., *Counterexamples in Topology*, Springer, New York, 1978.

[42] M. E. Taylor, *Measure Theory and Integration*, Amer. Math. Soc., Providence, 2006.

[43] G. Teschl, *Mathematical Methods in Quantum Mechanics; With Applications to Schrödinger Operators*, Amer. Math. Soc., Providence, 2009.

[44] J. Weidmann, *Lineare Operatoren in Hilberträumen I: Grundlagen*, B.G.Teubner, Stuttgart, 2000.

[45] D. Werner, *Funktionalanalysis*, 7th edition, Springer, Berlin, 2011.

[46] M. Willem, *Functional Analysis*, Birkhäuser, Basel, 2013.

[47] E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, Springer, New York 1995.

[48] E. Zeidler, *Applied Functional Analysis: Main Principles and Their Applications*, Springer, New York 1995.

Glossary of notation

$AC[a,b]$	... absolutely continuous functions, 348
$\arg(z)$	... argument of $z \in \mathbb{C}$; $\arg(z) \in (-\pi, \pi]$, $\arg(0) = 0$
$B_r(x)$	... open ball of radius r around x, 520
$B(X)$	... Banach space of bounded measurable functions
$BV[a,b]$	... functions of bounded variation, 345
$\mathfrak{B}$	$= \mathfrak{B}^1$
$\mathfrak{B}^n$	... Borel σ-algebra of $\mathbb{R}^n$, 219
$\mathbb{C}$	... the set of complex numbers
$C(U)$	... set of continuous functions from U to $\mathbb{C}$
$C_0(U)$	... set of continuous functions vanishing on the boundary ∂U, 547
$C_c(U)$	... set of compactly supported continuous functions
$C_{per}[a,b]$	... set of periodic continuous functions (i.e. $f(a) = f(b)$)
$C^k(U)$	... set of k times continuously differentiable functions
$C^\infty_{pg}(U)$	... set of smooth functions with at most polynomial growth, 416
$C^\infty_c(U)$	... set of compactly supported smooth functions
$C(U,Y)$	... set of continuous functions from U to Y, 435
$C^r(U,Y)$	... set of r times continuously differentiable functions, 440
$C^r_b(U,Y)$	... functions in C^r with derivatives bounded, 37, 446
$C^r_c(U,Y)$	... functions in C^r with compact support
$c_0(\mathbb{N})$	... set of sequences converging to zero, 10
$\mathscr{C}(X,Y)$	... set of compact linear operators from X to Y, 63
$\mathcal{C}(U,Y)$	... set of compact maps from U to Y, 492
$\mathrm{CP}(f)$	... critical points of f, 471
$\mathrm{CS}(K)$	... nonempty convex subsets of K, 483

$\mathrm{CV}(f)$	... critical values of f, 471
$\chi_\Omega(.)$	... characteristic function of the set Ω
$\mathfrak{D}(.)$	... domain of an operator
$\delta_{n,m}$	... Kronecker delta, 12
$\deg(D,f,y)$	... mapping degree, 471, 478
det	... determinant
dim	... dimension of a linear space
div	... divergence of a vector filed, 272
$\mathrm{diam}(U)$	$= \sup_{(x,y)\in U^2} d(x,y)$ diameter of a set
$\mathrm{dist}(U,V)$	$= \inf_{(x,y)\in U\times V} d(x,y)$ distance of two sets
$D^r_y(\overline{U},Y)$	... functions in $C^r(\overline{U},Y)$ which do not attain y on the boundary, 471
$\bar{\mathcal{C}}_y(U,Y)$	... functions in $\mathcal{C}(\overline{U},Y)$ which do not attain y on the boundary, 494
e	... Napier's constant, $\mathrm{e}^z = \exp(z)$
dF	... derivative of F, 435
$\mathcal{F}(X,Y)$	... set of compact finite dimensional functions, 492
$\Phi(X,Y)$	... set of all linear Fredholm operators from X to Y, 187
$\Phi_0(X,Y)$	... set of all linear Fredholm operators of index 0, 187
$\mathrm{GL}(n)$	... general linear group in n dimensions
$\Gamma(z)$	... gamma function, 268
$\Gamma(f_1,\ldots,f_n)$	... Gram determinant, 47
$\mathfrak{H}$	... a Hilbert space
conv(.)	... convex hull
$\mathcal{H}(U)$	... set of holomorphic functions on a domain $U\subseteq\mathbb{C}$, 469
$H^k(U)$	... Sobolev space, 368, 403
$H^k_0(U)$	... Sobolev space, 370
i	... complex unity, $\mathrm{i}^2=-1$
Im(.)	... imaginary part of a complex number
inf	... infimum
$J_f(x)$	$= \det df(x)$ Jacobi determinant of f at x, 471
$\mathrm{Ker}(A)$	... kernel of an operator A, 27
λ^n	... Lebesgue measure in $\mathbb{R}^n$, 262
$\mathscr{L}(X,Y)$	... set of all bounded linear operators from X to Y, 29
$\mathscr{L}(X)$	$= \mathscr{L}(X,X)$
$L^p(X,d\mu)$	... Lebesgue space of p integrable functions, 284
$L^\infty(X,d\mu)$	... Lebesgue space of bounded functions, 284
$L^p_{loc}(X,d\mu)$	... locally p integrable functions, 285
$\mathcal{L}^1(X)$	... space of integrable functions, 252
$\mathcal{L}^2_{cont}(I)$	... space of continuous square integrable functions, 20
$\ell^p(\mathbb{N})$	... Banach space of p summable sequences, 9
$\ell^2(\mathbb{N})$	... Hilbert space of square summable sequences, 17
$\ell^\infty(\mathbb{N})$	... Banach space of bounded sequences, 10
max	... maximum
$\mathcal{M}(X)$	... finite complex measures on X, 323
$\mathcal{M}_{reg}(X)$	... finite regular complex measures on X, 326
$\mathcal{M}(f)$	... Hardy–Littlewood maximal function, 429

$\mathbb{N}$	... the set of positive integers
$\mathbb{N}_0$	$= \mathbb{N} \cup \{0\}$
$n(\gamma, z_0)$	... winding number
$O(.)$	... Landau symbol, $f = O(g)$ iff $\limsup_{x \to x_0} \lvert f(x)/g(x) \rvert < \infty$
$o(.)$	... Landau symbol, $f = o(g)$ iff $\lim_{x \to x_0} \lvert f(x)/g(x) \rvert = 0$
$\mathbb{Q}$	... the set of rational numbers
$\mathbb{R}$	... the set of real numbers
$\rho(A)$	... resolvent set of an operator A, 160
$\mathrm{RV}(f)$	... regular values of f, 471
$\mathrm{Ran}(A)$	... range of an operator A, 27
$\mathrm{Re}(.)$	... real part of a complex number
$R(I, X)$	... set of regulated functions, 194
$\sigma(A)$	... spectrum of an operator A, 160 ,170
σ^{n-1}	... surface measure on S^{n-1}, 267
S^{n-1}	$= \{x \in \mathbb{R}^n \mid \lvert x \rvert = 1\}$ unit sphere in $\mathbb{R}^n$
S_n	$= n\pi^{n/2}/\Gamma(\frac{n}{2}+1)$, surface area of the unit sphere in $\mathbb{R}^n$, 266
$\mathrm{sign}(z)$	$= z/\lvert z \rvert$ for $z \neq 0$ and 1 for $z = 0$; complex sign function
$S(I, X)$	... simple functions $f : I \to X$, 443
$\mathcal{S}^n$	... semialgebra of rectangles in $\mathbb{R}^n$, 215
$\bar{\mathcal{S}}^n$	... algebra of finite unions of rectangles in $\mathbb{R}^n$, 215
$\mathcal{S}(\mathbb{R}^n)$	... Schwartz space, 391
sup	... supremum
$\mathrm{supp}(f)$	... support of a function f, 530
$\mathrm{supp}(\mu)$	... support of a measure μ, 231
$\mathrm{span}(M)$	... set of finite linear combinations from M, 11
$W^{k,p}(U)$	... Sobolev space, 368
$W_0^{k,p}(U)$	... Sobolev space, 370
V_n	$= \pi^{n/2}/\Gamma(\frac{n}{2}+1)$, volume of the unit ball in $\mathbb{R}^n$, 267
$\mathbb{Z}$	... the set of integers
$\mathbb{I}$	... identity operator
$\sqrt{z}$	... square root of z with branch cut along $(-\infty, 0)$
z^*	... complex conjugation
A^*	... adjoint of A, 52
$\overline{A}$	... closure of A, 101
$\hat{f}$	$= \mathcal{F}f$, Fourier transform/coefficients of f, 58, 389
$\check{f}$	$= \mathcal{F}^{-1}f$, inverse Fourier transform of f, 392
$\lvert x \rvert$	$= \sqrt{\sum_{j=1}^n \lvert x_j \rvert^2}$ Euclidean norm in $\mathbb{R}^n$ or $\mathbb{C}^n$
$\lvert \Omega \rvert$	... Lebesgue measure of a Borel set Ω
$\lVert . \rVert$	... norm, 17
$\lVert . \rVert_p$	... norm in the Banach space ℓ^p and L^p, 9, 283
$\langle ., .. \rangle$	... scalar product in $\mathfrak{H}$, 17

$\oplus$ ... direct/orthogonal sum of vector spaces or operators, 33, 55
$\hat{\oplus}$ ... direct sum of operators with the same image space, 33
$\otimes$ ... tensor product, 57
$\biguplus$ ... union of disjoint sets, 220
$\lfloor x \rfloor$ $= \max\{n \in \mathbb{Z} | n \leq x\}$, floor function
$\lceil x \rceil$ $= \min\{n \in \mathbb{Z} | n \geq x\}$, ceiling function
∂ $= (\partial_1 f, \ldots, \partial_m f)$ gradient in $\mathbb{R}^m$
∂_α ... partial derivative in multi-index notation, 37
$\partial_x F(x, y)$ ... partial derivative with respect to x, 439
∂U $= \overline{U} \setminus U^\circ$ boundary of the set U, 520
$\overline{U}$ ... closure of the set U, 524
U° ... interior of the set U, 524
$M^\perp$ ... orthogonal complement, 48
(λ_1, λ_2) $= \{\lambda \in \mathbb{R} \,|\, \lambda_1 < \lambda < \lambda_2\}$, open interval
$[\lambda_1, \lambda_2]$ $= \{\lambda \in \mathbb{R} \,|\, \lambda_1 \leq \lambda \leq \lambda_2\}$, closed interval
$x_n \to x$ ... norm convergence, 8
$x_n \rightharpoonup x$ ... weak convergence, 120
$x_n \overset{*}{\rightharpoonup} x$ ... weak-$*$ convergence, 127
$A_n \to A$ ... norm convergence
$A_n \overset{s}{\to} A$ ... strong convergence, 125
$A_n \rightharpoonup A$ ... weak convergence, 124

Index