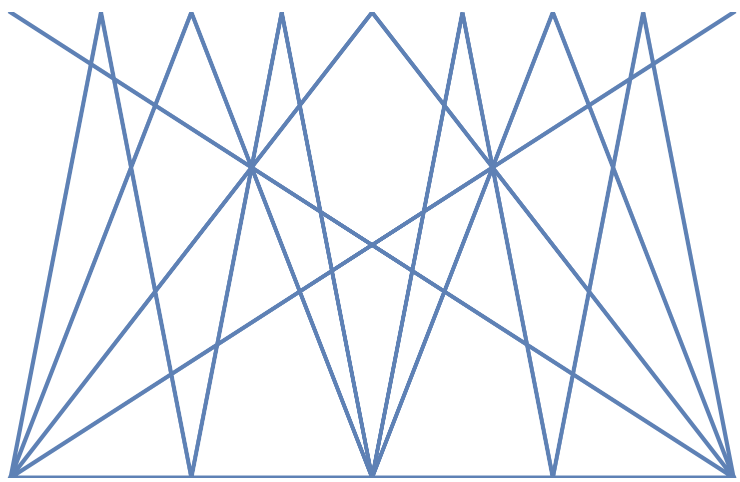
Topics in Linear and Nonlinear Functional Analysis

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, 46E30, 47H10, 47H11, 58Exx, 76D05

Abstract. This manuscript provides a brief introduction to linear and nonlinear Functional Analysis. It covers basic Hilbert and Banach space theory as well as some advanced topics like unbounded operators on Banach spaces, mapping degrees and fixed point theorems.

Keywords and phrases. Functional Analysis, Banach space, Hilbert space, spectral theory, mapping degree, fixed point theorem, differential equation, Navier–Stokes equation.

Typeset by $\mathcal{AMS}$-LaTeX, Ti*k*Z, and Makeindex.
Version: June 27, 2025

Contents

Preface

The present manuscript was written for my course *Functional Analysis* given at the University of Vienna in 2004, 2009, 2022, and 2023 (the last one also included basic *Topology*). The second part are the notes for my course *Nonlinear Functional Analysis* held at the University of Vienna in 1998, 2001, and 2018. The two 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). However, there is an accompanying part on Real Analysis [**40**], where these topics are covered.

The manuscript 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 the reader 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 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 these 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 always be readily available. For convenience, the Appendix contains full proofs. 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. I am aware that this is a decision one might dispute, however, it has some evident advantages. In particular, the examples frequently discuss discrete versions of classical topics thereby not only avoiding Lebesgue integration as a prerequisite, but also avoiding technical difficulties which hide the main ideas. Readers familiar with measure theory should then have no problems to handle the continuous case. At a few places I also assume some basic results from complex analysis, but it will be sufficient to just believe them.

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

Content

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

Chapter 1. I have decided to start with Fourier's treatment of the heat equation, which led 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 like completeness, compactness, or bounded operators. It should be covered in detail except for Section 1.8, which introduces some interesting examples for later use.

Chapter 2. Since my overall strategy is to first understand the special case and then work my way up to the general case, it is natural to start with basic Hilbert space theory. These results 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.

Chapter 3. Once we understand the geometry of Hilbert spaces, it seems natural to develop basic spectral theory. In line with what was said before, we start with the simpelst case of compact self-adjoint operators. The first core result is the spectral theorem for compact 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 are again optional and will not be required later on.

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.

Chapter 5 develops spectral theory for bounded self-adjoint operators via the framework of C^* algebras. The last section contains some optional results establishing the connection with the measure theoretic formulation of the spectral theorem.

The next chapters contain selected advanced topics.

Chapter 6 centers around convexity. Except for the geometric Hahn–Banach theorem, which is a prerequisite for the other sections, the remaining sections are independent of each other to simplify the selection of topics.

Chapter 7 presents some advanced topics from spectral theory: The Gelfand representation theorem, spectral theory for compact operators in Banach spaces and Fredholm theory. Again these sections are independent of each other except for the fact that Section 7.1, which contains the spectral theorem for compact operators, is important for the Fredholm alternative for compact perturbations of the identity.

Chapter 8 touches upon unbounded operators starting with the basic results about closed operators. Since unbounded operators in Banach spaces play an increasing role in applications, I felt it is appropriate to discuss at least some basics.

Finally, there is a part on nonlinear functional analysis.

Chapter 9 discusses analysis in Banach spaces (with a view towards applications in the calculus of variations and infinite dimensional dynamical systems).

Chapter 10 and 11 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 12 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 refer to [**21, 34, 35**] as good starting points.

To the teacher

There are a couple of courses to be taught from this book. First of all, there is 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 (parts of) Chapters 1, 2, and 3 from [**40**]. 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 [**38**] (where one can start with Chapter 2).

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

Problems relevant for the main text are marked with a "*".

Acknowledgments

I wish to thank my readers, Olta Ahmeti, Kerstin Ammann, Phillip Bachler, Batuhan Bayır, Alexander Beigl, Elian Biltri, Mikhail Botchkarev, Ho Boon Suan, Peng Du, Christian Ekstrand, Mischa Elkner, David Fajman, Damir Ferizović, Michael Fischer, Simon Garger, Raffaello Giulietti, Melanie Graf, Josef Greilhuber, Julian Grüber, Matthias Hammerl, Jona Marie Hassenbach, Alexander Jerschow, Nobuya Kakehashi, Jerzy Knopik, Nikolas Knotz, Florian Kogelbauer, Stefan Krieger, Helge Krüger, Reinhold Küstner, Oliver Leingang, Juho Leppäkangas, Annemarie Luger, Joris Mestdagh, Alice Mikikits-Leitner, Claudiu Mîndrilă, Jakob Möller, Caroline Moosmüller, Frederick Moscatelli, Nikola Negovanovic, Matthias Ostermann, Martina Pflegpeter, Mateusz Piorkowski, Piotr Owczarek, Fabio Plaga, Tobias Preinerstorfer, Maximilian H. Ruep, Tidhar Sariel, Chiara Schindler, Christian Schmid, Stephan Schneider, Laura Shou, Bastam Tajik, Bertram Tschiderer, Liam Urban, Vincent Valmorin, David Wallauch, Richard Welke, David Wimmesberger, Gunter Wirthumer, Song Xiaojun, Markus Youssef, Valentin Zlabinger, Rudolf Zeidler, and colleagues Pierre-Antoine Absil, 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, Giacomo Sodini, Marko Stautz, Maxim Zinchenko who have pointed out several typos and made useful suggestions for improvements. Moreover, I am most grateful to Iryna Karpenko who read several parts of the manuscript, provided long lists of typos, and also contributed some of the problems. 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
October, 2024

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[1] was led 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 $u : [0,\infty) \times [0,1] \to \mathbb{R}$ is the temperature distribution in a thin rod at time $t \geq 0$ at the point $x \in [0,1]$. It is usually assumed, that the temperature

[1] Joseph Fourier (1768–1830), French mathematician and physicist

at $x = 0$ and $x = 1$ is fixed, say $u(t, 0) = \alpha$ and $u(t, 1) = \beta$. By considering $u(t, x) \to u(t, x) - \alpha - (\beta - \alpha)x$ it is clearly no restriction to assume $\alpha = \beta = 0$. 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 unfeasible, 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 led 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

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

and the second one (assuming $\lambda > 0$)

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

Here a, b, c are arbitrary real constants and since we are only interested in the product $w\,y$, we can choose $a = 1$ without loss of generality. Moreover, $y(x)$ must also satisfy the boundary conditions $y(0) = y(1) = 0$. The first one $y(0) = 0$ is satisfied if $b = 0$ and the second one yields ($c = 0$ only leads to the trivial solution)

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

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. Similarly,

$\lambda = 0$ also only leads to the trivial solution. In summary, we obtain the family of 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 $c_n = \hat{u}_{0,n}$ (cf. Problem 1.2).

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:[2]

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

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**.[3] 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

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

[2]Erwin Schrödinger (1887–1961), Austrian physicist

[3]Jacques Charles François Sturm (1803–1855), French mathematician

[3]Joseph Liouville (1809–1882), French mathematician and engineer

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 led to Hilbert spaces which are the proper setting for our eigenvalue problem.
- Finally, the spectral theorem for compact self-adjoint operators will provide 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*[4] *M-test. When can you interchange the order of summation and differentiation?)*

Problem 1.2. *Show that for* $n, m \in \mathbb{N}$ *we have*

$$2\int_0^1 \sin(n\pi x)\sin(m\pi x)dx = \begin{cases} 1, & n = m, \\ 0, & n \neq m. \end{cases}$$

Conclude that the Fourier sine coefficients are given by

$$\hat{u}_{0,n} = 2\int_0^1 \sin(n\pi x) u_0(x) dx$$

provided the sum in (1.13) *converges uniformly. Conclude that in this case the solution can be expressed as*

$$u(t,x) = \int_0^1 K(t,x,y) u_0(y) dy, \qquad t > 0,$$

where

$$\begin{aligned} K(t,x,y) &:= 2\sum_{n=1}^{\infty} \mathrm{e}^{-(\pi n)^2 t} \sin(n\pi x)\sin(n\pi y) \\ &= \frac{1}{2}\big(\vartheta(\frac{x-y}{2}, \mathrm{i}\pi t) - \vartheta(\frac{x+y}{2}, \mathrm{i}\pi t)\big). \end{aligned}$$

[4] Karl Weierstrass (1815–1897), German mathematician

Here

$$\vartheta(z,\tau) := \sum_{n\in\mathbb{Z}} \mathrm{e}^{\mathrm{i}\pi n^2\tau + 2\pi\mathrm{i}nz} = 1 + 2\sum_{n\in\mathbb{N}} \mathrm{e}^{\mathrm{i}\pi n^2\tau}\cos(2\pi nz), \quad \mathrm{Im}(\tau) > 0,$$

is the ***Jacobi theta function***.[5]

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 $f \in C(I)$:

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

Note that the fact that $|f(x)|$, as a composition of continuous functions, attains a maximum on every compact set, is the content of the extreme value theorem (Theorem B.21).

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

If positive definiteness is dropped from the requirements, one calls $\|.\|$ a **seminorm**. You will get a normed space once you identify vectors which are indistinguishable for the seminorm (Problem 1.7).

From the triangle inequality we also get the **inverse triangle inequality** (Problem 1.4)

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

which shows that the norm is continuous.

Also note that norms are closely related to convexity. To this end recall that a subset $C \subseteq X$ is called **convex** if for every $f, g \in C$ we also have $\lambda f + (1-\lambda) g \in C$ whenever $\lambda \in (0,1)$.

[5] Carl Gustav Jacob Jacobi (1804–1851), German mathematician

Example 1.1. Of course every subspace is convex. Furthermore, the triangle inequality implies that every **ball**

$$B_r(f) := \{g \in X| \, \|f - g\| < r\} \subseteq X$$

is convex. ◇

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

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

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** $d(f,g) := \|f - g\|$ (in particular, every normed space is a special case of a **metric space**) and hence we know when a sequence of vectors f_n **converges** to a vector f (namely if $\|f_n - f\| \to 0$, that is, for every $\varepsilon > 0$ there is some N such that $\|f_n - f\| \le \varepsilon$ for all $n \ge N$). 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 1.5).

Two normed spaces X and Y are called **isomorphic** if there exists a linear bijection $T : X \to Y$ such that T and its inverse T^{-1} are continuous. We will write $X \cong Y$ in this case. They are called **isometrically isomorphic** if in addition, T is an **isometry**, $\|T(f)\| = \|f\|$ for every $f \in X$.

In addition to the concept of convergence, we also have the concept of a **Cauchy sequence**:[6] A sequence f_n is Cauchy if for every $\varepsilon > 0$ there is some N such that $\|f_n - f_m\| \le \varepsilon$ for all $n, m \ge N$. Of course every convergent sequence is Cauchy but the converse might not be true in general. Hence a normed space is called **complete** if every Cauchy sequence has a limit. A complete normed space is called a **Banach space**.[7]

Example 1.2. The real numbers $\mathbb{R}$ as well as the complex numbers $\mathbb{C}$ with the absolute value as norm are Banach spaces. ◇

Example 1.3. The space $\ell^1(\mathbb{N})$ of all complex-valued sequences $a := (a_j)_{j=1}^\infty$ for which the norm

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

[6] Augustin-Louis Cauchy (1789–1857), French mathematician

[7] Stefan Banach (1892–1945), Polish mathematician

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 positive 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 some N such that $\|a^m - a^n\|_1 \le \varepsilon$ for $m, n \ge N$. 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 $m \to \infty$:

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

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 $\|a - a^n\|_1 \le \varepsilon$ for $n \ge N$, our candidate a is indeed the limit of a^n. ⋄

Example 1.4. 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 1.21.)

To prove the triangle inequality we need **Hölder's inequality**[8]

$$\|ab\|_1 \le \|a\|_p \|b\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad 1 < p, q < \infty, \tag{1.24}$$

[8]Otto Hölder (1859–1937), German mathematician

for $a \in \ell^p(\mathbb{N})$, $b \in \ell^q(\mathbb{N})$. Here the product ab of sequences is of course formed elementwise and the statement $ab \in \ell^1(\mathbb{N})$ is implicit in (1.24). In the case $p, q \in \{1, \infty\}$ this is immediate and for the remaining cases we first observe that 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 Young's inequality[9] (Problem 1.11)

$$\alpha^{1/p}\beta^{1/q} \le \frac{1}{p}\alpha + \frac{1}{q}\beta, \qquad \frac{1}{p} + \frac{1}{q} = 1, \quad \alpha, \beta \ge 0,\ 1 < p, q < \infty, \tag{1.25}$$

and summing over all j.

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 1.12).

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). This is related to the question of equality in the triangle inequality and will be discussed in Problems 1.18 and 1.19. ⋄

Example 1.5. 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 1.13). Note that with this definition, Hölder's inequality (1.24) remains true for the cases $p = 1$, $q = \infty$ and $p = \infty$, $q = 1$. The reason for the notation is explained in Problem 1.20. ⋄

By a **subspace** U of a normed space X, we mean a subset which is closed under the vector operations. If it is also closed in a topological sense, we call it a closed subspace. In this context recall that a subset $U \subseteq X$ is called **open** if for every point $f \in U$ there is also a ball $B_r(f) \subseteq U$ contained within U. The **closed** sets are then defined as the complements of open sets and one has that a set $V \subseteq X$ is closed if and only if for every convergent sequence $f_n \in V$ the limit is also in the set, $\lim_n f_n \in V$. A set $V \subseteq X$ is called **bounded** if it is contained within a (finite) ball, that is, there is some $r > 0$ such that $\|x\| \le r$ for all $x \in V$. Warning: Some authors require

[9] William Henry Young (1863–1942), English mathematician

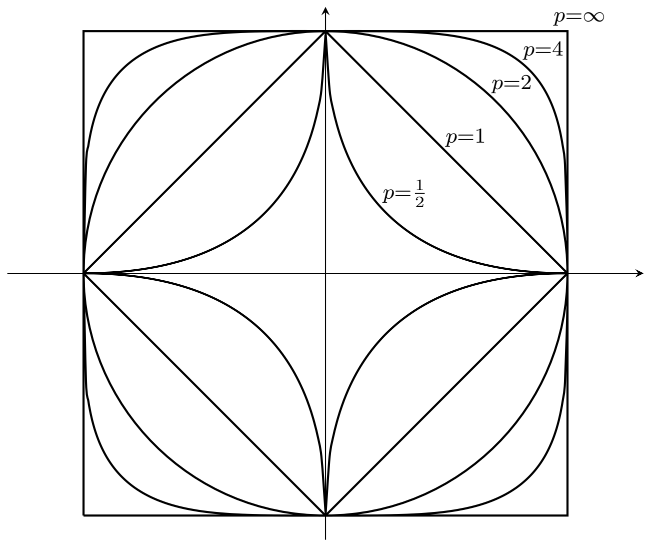


Figure 1.1. Unit balls for $\|.\|_p$ in $\mathbb{R}^2$

subspaces to be closed.

Example 1.6. 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 $a + b \notin c_0(\mathbb{N})$ for every $b \in \ell^\infty(\mathbb{N})$ with $\|b\|_\infty < \varepsilon$. Hence the complement of $c_0(\mathbb{N})$ is open. $\diamond$

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 $f(x) := \lim_{n\to\infty} f_n(x)$. 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 f 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 form 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**;[10] Problem 1.10). 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 1.9 and 4.5). 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. That is every (finite) linear combination yielding the zero vector, $\sum_{j=1}^m \alpha_j u_{n_j} = 0$, must be trivial, $\alpha_j = 0$ for $1 \le j \le m$. 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.

Let $N \in \mathbb{N} \cup \{\infty\}$ with $N \in \mathbb{N}$ in case X is finite dimensional (and in which case N equals the dimension of X) or $N = \infty$ in case X is infinite dimensional. We will call a countable sequence of vectors $(u_n)_{n=1}^N$ from X a **Schauder basis**[11] if every element $f \in X$ can be uniquely written as a countable linear combination of the basis elements:

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

[10] Georg Hamel (1877–1954)), German mathematician
[11] Juliusz Schauder (1899–1943), Polish mathematician

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 4.11). 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**.[12]

Example 1.7. In a finite dimensional space every basis is also a Schauder basis. Note that in this case continuity of the coordinate functionals is immediate since linear maps on finite dimensional spaces are always continuous (see Lemma 1.15 below). ⋄

Example 1.8. The sequence of vectors $\delta^n = (\delta^n_m := \delta_{n,m})_{m\in\mathbb{N}}$ is a Schauder basis for the Banach space $\ell^p(\mathbb{N})$, $1 \le p < \infty$.

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 $a^m_j = a_j$ for $1 \le j \le m$ and $a^m_j = 0$ for $j > m$. Hence

$$a = \lim_{m\to\infty} \sum_{n=1}^m a_n \delta^n =: \sum_{n=1}^\infty a_n \delta^n$$

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 138.

Example 1.9. Every Schauder basis is total and thus every Banach space with a Schauder basis is separable (the converse puzzled mathematicians

[12] Leopold Kronecker (1823–1891), German mathematician

for quite some time and was eventually shown to be false by Enflo[13]). In particular, the Banach space $\ell^p(\mathbb{N})$ is separable for $1 \le p < \infty$.

However, $\ell^\infty(\mathbb{N})$ is not separable (Problem 1.16)! ⋄

While we will not give a Schauder basis for $C(I)$ (Problem 1.25), 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.

Approximating a function f is frequently done by considering the **convolution**

$$(u_n * f)(x) = \int_{\mathbb{R}} u_n(y) f(x-y) dy = \int_{\mathbb{R}} u_n(x-y) f(y) dy \tag{1.33}$$

with some nice functions u_n. In our situation u_n will always have compact support (and both functions will be at least piecewise continuous) such that we do not have to worry about existence of this integral. A typical example would be using $u_n(x) := 2n\chi_{[-1/n,1/n]}$, which amounts to replacing $f(x)$ by its average over a neighborhood $[x-1/n, x+1/n]$. As the neighborhood shrinks, we have $f_n(x) \to f(x)$ for continuous f. Moreover, f_n inherits many properties from u_n and hence this can be used to construct approximations with given properties (e.g. f_n will be as smooth as u_n, which can be seen by interchanging differentiation and integration). A family of functions u_n for which this approximation works is known as an **approximate identity**:

Lemma 1.2 (Approximate identity)**.** *Let u_n be a sequence of nonnegative integrable functions such that*

$$\int_{\mathbb{R}} u_n(x) dx = 1 \quad \textit{and} \quad \int_{\delta \le |x|} u_n(x) dx \to 0, \quad \delta > 0. \tag{1.34}$$

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

Then for every bounded uniformly continuous function f we have that $u_n * f$ converges uniformly to f.

Proof. Since f is uniformly continuous, for given ε we can find a δ (independent of x) such that $|f(x) - f(y)| \le \varepsilon/2$ whenever $|x - y| \le \delta$. Moreover, setting $M := \sup |f(x)|$ we can choose N such that $\int_{\delta \le |y|} u_n(y) dy \le \varepsilon/(4M)$

[13]Per Enflo (*1944), Swedish mathematician

for $n \geq N$. Then,

$$\begin{aligned}
|f(x) - \int_{\mathbb{R}} u_n(y) f(x-y) dy| &= \left| \int_{\mathbb{R}} u_n(y)(f(x-y) - f(x)) dy \right| \\
&\leq \int_{|y|\leq\delta} u_n(y)|f(x-y) - f(x)| dy + \int_{\delta\leq|y|} u_n(y)|f(x-y) - f(x)| dy \\
&\leq \frac{\varepsilon}{2} \int_{|y|\leq\delta} u_n(y) dy + 2M \int_{\delta\leq|y|} u_n(y) dy \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

for $n \geq N$, which shows that $f_n \to f$ uniformly. □

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**[14]

$$u_n(x) := \frac{1}{I_n}(1-x^2)^n, \quad |x| < 1, \qquad u_n(x) := 0, \quad |x| \geq 1,$$

where

$$\begin{aligned}
I_n &:= \int_{-1}^{1} (1-x^2)^n dx = 2\int_0^1 (1-x)^n (1+x)^n dx \\
&\geq 2\int_0^1 (1-x)^n dx = \frac{2}{n+1}.
\end{aligned}$$

Indeed, the first part of (1.34) holds by construction, and the second part follows from $\int_{\delta\leq|x|\leq 1} u_n(x) dx \leq 2u_n(\delta) \leq (n+1)(1-\delta^2)^n \to 0$. Hence u_n is an approximate identity. Now extend f to $\mathbb{R}$ by setting it equal to 0 outside $[-\frac{1}{2}, \frac{1}{2}]$ and notice that $(u_n * f)(x)$ is a polynomial for $x \in [-\frac{1}{2}, \frac{1}{2}]$. Thus the result follows from Lemma 1.2 □

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.

[14]Edmund Landau (1877–1938), German mathematician

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 $C(I)$. If they were, the coefficients would satisfy $|\alpha_n| \|x\|_\infty^n = \|f_n - f_{n-1}\|_\infty \to 0$ and hence the limit must be analytic on the interior of I. 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 1.3. *Let X be a normed space. Show that a Cauchy sequence is bounded and that the limit of a convergent sequence is unique. (Here a subset $U \subset X$ is called bounded if there is a constant $M \geq 0$ such that $\|f\| \leq M$ for all $f \in U$.)*

Problem* 1.4. *Let X be a normed space and $f, g \in X$. Show that $|\|f\| - \|g\|| \leq \|f - g\|$.*

Problem* 1.5. *Let X be a normed 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 1.6. *Let X be a normed space and $f, g \in X$. Show that $\|f\| \leq \max(\|f - g\|, \|f + g\|)$.*

Problem 1.7. *Let X be a seminormed space. Show that the set $N := \{f \in X | \, \|f\| = 0\}$ is a subspace. Show that $\|.\|$ is a well-defined norm on the quotient space X/N.*

Problem 1.8. *Let X be a Banach space and $(f_j)_{j \in \mathbb{N}} \subset X$. 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 1.9. *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[15] *determinant. See Problem 4.5 for a generalization.)*

Problem 1.10. *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, Theorem A.2, to show existence.)*

Problem* 1.11. *Prove Young's inequality* (1.25)*. Show that equality occurs precisely if $\alpha = \beta$. (Hint: Take logarithms on both sides.)*

Problem* 1.12. *Show that $\ell^p(\mathbb{N})$, $1 \le p < \infty$, is complete.*

Problem* 1.13. *Show that $\ell^\infty(\mathbb{N})$ is a Banach space.*

Problem 1.14. *Is $\ell^1(\mathbb{N})$ a closed subspace of $\ell^\infty(\mathbb{N})$ (with respect to the $\|.\|_\infty$ norm)? If not, what is its closure?*

Problem 1.15. *Consider $\ell^1(\mathbb{N})$. Show that $\|a\| := \sup_{n\in\mathbb{N}} |\sum_{j=1}^n a_j|$ is a norm. Is $\ell^1(\mathbb{N})$ complete with this norm?*

Problem* 1.16. *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 1.17. *Show that the set of convergent sequences $c(\mathbb{N})$ is a Banach space isomorphic to the set of convergent sequence $c_0(\mathbb{N})$. (Hint: Hilbert's hotel.)*

Problem* 1.18. *Show that there is equality in the Hölder inequality* (1.24) *for $1 < p < \infty$ if and only if either $a = 0$ or $|b_j|^q = \alpha |a_j|^p$ for all $j \in \mathbb{N}$. Show that we have equality in the triangle inequality for $\ell^1(\mathbb{N})$ if and only if $a_j b_j^* \ge 0$ for all $j \in \mathbb{N}$ (here the '$*$' denotes complex conjugation). Show that we have equality in the triangle inequality for $\ell^p(\mathbb{N})$ with $1 < p < \infty$ if and only if $a = 0$ or $b = \alpha a$ with $\alpha \ge 0$.*

Problem* 1.19. *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 $\alpha \ge 0$ or $x = 0$.*
(ii) *If $\|x\| = \|y\| = 1$ and $x \ne y$ then $\|\lambda x + (1-\lambda) y\| < 1$ for all $0 < \lambda < 1$.*
(iii) *If $\|x\| = \|y\| = 1$ and $x \ne y$ then $\frac{1}{2}\|x + y\| < 1$.*

[15] Alexandre-Théophile Vandermonde (1735–1796), French mathematician, musician and chemist

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

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

Problem 1.21. *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*

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

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

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

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 $\alpha + \beta \le (\alpha^p + \beta^p)^{1/p} \le 2^{1/p-1}(\alpha + \beta)$ for $0 < p < 1$ and $\alpha, \beta \ge 0$.)

Problem 1.22. *Consider $X = C([-1,1])$. Which of the following subsets are subspaces of X? Which of them are closed?*

(i) *monotone functions*
(ii) *even functions*
(iii) *polynomials*
(iv) *polynomials of degree at most k for some fixed $k \in \mathbb{N}_0$*
(v) *continuous piecewise linear functions*
(vi) $C^1([-1,1])$
(vii) $\{f \in C([-1,1]) | f(c) = f_0\}$ *for some fixed $c \in [-1,1]$ and $f_0 \in \mathbb{R}$*

Problem 1.23. *Let I be a compact interval. Show that the set $Y := \{f \in C(I,\mathbb{R}) | f(x) > 0\}$ is open in $X := C(I,\mathbb{R})$. Compute its closure.*

Problem 1.24. *Compute the closure of the following subsets of $\ell^2(\mathbb{N})$: (i) $B_1 := \{a \in \ell^2(\mathbb{N}) | \sum_{j\in\mathbb{N}} |a_j| \le 1\}$. (ii) $B_\infty := \{a \in \ell^2(\mathbb{N}) | \sum_{j\in\mathbb{N}} |a_j| < \infty\}$.*

Problem* 1.25. *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 $1 \le k \le 2^n$, 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 $1 \le k \le 2^n$.*

1.3. The geometry of Hilbert spaces

So far 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,

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

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

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

sesquilinear form is also called a **Hermitian form**[16] 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 1.32). 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 shorthand 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**.[17]

Example 1.10. 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. ◇

---

[16] Charles Hermite (1822–1901), French mathematician

[17] David Hilbert (1862–1943), German mathematician

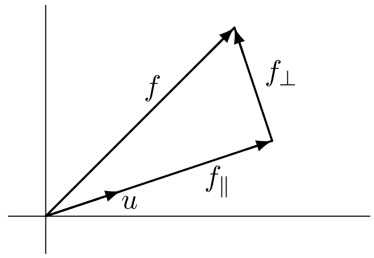


Figure 1.2. Orthogonal projection

Example 1.11. 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

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

That this sum is (absolutely) convergent (and thus well-defined) for $a, b \in \ell^2(\mathbb{N})$ follows from Hölder's inequality (1.24) 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**:[18]

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

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 $\langle u, f_\perp\rangle = \langle u, f - \langle u, f\rangle u\rangle = \langle u, f\rangle - \langle u, f\rangle\langle u, u\rangle = 0$. This is illustrated in Figure 1.2. Taking any other vector parallel to u, we

[18] Pythagoras (c. 570–c. 495 BC), ancient Ionian Greek philosopher

obtain from (1.40)

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

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

As a first consequence we obtain the **Cauchy–Bunyakovsky–Schwarz**[19] 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 1.29).

Theorem 1.6 (Jordan–von Neumann[20])**.** *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***

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

---

[19] Viktor Bunyakovsky (1804–1889), Russian mathematician

[19] Hermann Schwarz (1843 –1921), German mathematician

[20] Pascual Jordan (1902–1980), German theoretical and mathematical physicist

[20] John von Neumann (1903–1957), Hungarian-American mathematician, physicist, computer scientist, and engineer

Proof. If an inner product space is given, verification of the parallelogram law and the polarization identity is straightforward (Problem 1.32).

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 easy 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 $\alpha \geq 0$ 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 1.32).

But is there a way to 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

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

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

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

It is straightforward to check the following.

Lemma 1.7. *Let $\|.\|_1$ and $\|.\|_2$ be two norms on the same vector space. If $\|.\|_2$ is stronger than $\|.\|_1$, then every $\|.\|_2$ Cauchy sequence is also a $\|.\|_1$ Cauchy sequence and every sequence which converges with respect to $\|.\|_2$ also converges with respect to $\|.\|_1$ (to the same limit).*

Hence if a function $F : X \to Y$ is continuous in $(X, \|.\|_1)$, it is also continuous in $(X, \|.\|_2)$. Indeed, $f_n \xrightarrow{\|.\|_2} f$ implies $f_n \xrightarrow{\|.\|_1} f$ and hence $F(f_n) \to F(f)$ by continuity with respect to $\|.\|_1$. Warning: For a function

$F : Y \to X$, the order is reversed, that is, if $F : Y \to X$ is continuous in $(X, \|.\|_2)$, it is also continuous in $(X, \|.\|_1)$.

Similarly, 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 1.12. 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 1.35)! $\diamond$

Example 1.13. The previous example indicates that we should consider (1.48) on a larger class of functions, for example on the class of Riemann[21] 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 1.34). 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 $\mathcal{N}(I)$:

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

Since $\|f\|_2 = \|g\|_2$ for $f - g \in \mathcal{N}(I)$ we have a norm on $\mathcal{L}^2_{Ri}(I)$ (compare also Problem 1.7). 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[22] integral. Hence we will not pursue this further at this point. $\diamond$

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*

[21] Bernhard Riemann (1826–1866), German mathematician

[22] Henri Lebesgue (1875–1941), French mathematician

constants m_1 and m_2 such that

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

In particular, every finite dimensional normed space is isomorphic to $\mathbb{C}^n$ ($\mathbb{R}^n$) and hence a Banach space.

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: $\|f\|_2 := \|\sum_j \alpha_j u_j\|_2 = (\sum_j |\alpha_j|^2)^{1/2}$. In particular, $(X, \|.\|_2)$ is isometrically isomorphic to $\mathbb{C}^n$. 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 $m_2 = \sqrt{\sum_j \|u_j\|_1^2}$.

Now, 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.19). Now choose $m_1 = 1/m$. This shows that $(X, \|.\|_1)$ is isomorphic to $(X, \|.\|_2)$. □

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 1.26. *Which of the following bilinear forms are scalar products on $\mathbb{R}^n$?*

(i) $s(x, y) := \sum_{j=1}^n (x_j + y_j)$.
(ii) $s(x, y) := \sum_{j=1}^n \alpha_j x_j y_j$, $\alpha \in \mathbb{R}^n$.

Problem 1.27. *Show that the norm in a Hilbert space satisfies $\|f + g\| = \|f\| + \|g\|$ if and only if $f = \alpha g$, $\alpha \ge 0$, or $g = 0$. Hence Hilbert spaces are strictly convex (cf. Problem 1.19).*

Problem 1.28 (Generalized parallelogram law). *Show that, in a Hilbert space,*

$$\sum_{1\le j<k\le n} \|f_j - f_k\|^2 + \|\sum_{1\le j\le n} f_j\|^2 = n \sum_{1\le j\le n} \|f_j\|^2$$

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

Problem 1.29. *Show that the maximum norm on $C[0,1]$ does not satisfy the parallelogram law.*

Problem 1.30. *Show that $\ell^p(\mathbb{N})$, $1 \le p \le \infty$, is a Hilbert space if and only if $p=2$.*

Problem 1.31. *Let X and Y be real normed spaces. Show that an additive (i.e. $T(x+y) = T(x) + T(y)$) continuous map $T : X \to Y$ is linear. What about complex spaces?*

Problem* 1.32. *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 1.33. *A sesquilinear form s 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 $q(f) := s(f,f)$ is ***bounded*** *if*

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

is finite. Show

$$\|q\| \le \|s\| \le 2\|q\|$$

with $\|q\| = \|s\|$ if s is symmetric. (Hint: Use the polarization identity from the previous problem. For the symmetric case look at the real part.)

Problem* 1.34. *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* 1.35. *Prove the claims made about f_n in Example 1.12.*

1.4. Completeness

Since $\mathcal{L}^2_{cont}(I)$ 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 $\mathcal{X}$, which inherits the vector space structure from X. Call two Cauchy sequences equivalent if their difference converges to zero and denote by $\bar{X}$ the set of all equivalence classes. To this end observe that the set of null sequences $\mathcal{N}$ (i.e., the sequences converging to zero) is a subspaceof $\mathcal{X}$ and hence the quotient space $\bar{X} = \mathcal{X}/\mathcal{N}$ is a vector space X. Moreover, by continuity of the norm (cf. (1.20)), the norm $\|x_n\|$ of a Cauchy sequence x_n in X is again Cauchy. 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 (compare also Problem 1.7). It contains X as a subspace by virtue of the embedding $x \mapsto [(x)_{n=1}^\infty]$ which identifies x with the constant sequence $x_n := x$.

Theorem 1.9. *Let X be a normed space. $\bar{X}$ 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. To see that constant sequences are dense, note that we can approximate $[(x_n)_{n=1}^\infty]$ by the constant sequence $[(x_{n_0})_{n=1}^\infty]$ as $n_0 \to \infty$.

It remains to show that $\bar{X}$ is complete. Let $\xi_n := [(x_{n,j})_{j=1}^\infty]$ be a Cauchy sequence in $\bar{X}$. Since $(x_{n,j})_{j\in\mathbb{N}}$ is a Cauchy sequence for fixed n, there is some $N_n \in \mathbb{N}$ such that $\|x_{n,j} - x_{n,k}\| \le \frac{1}{n}$ for $j,k \ge N_n$. After dropping some terms we can assume that $N_n \le n$ without loss of generality.

Now we claim that $\xi := [(x_{j,j})_{j=1}^\infty]$ is the limit of ξ_n. Indeed, let $\varepsilon > 0$ be given and choose an N such that $\|\xi_n - \xi_m\| \le \varepsilon/3$ for $n,m \ge N$. By definition this means that

$$\lim_{j\to\infty} \|x_{n,j} - x_{m,j}\| \le \frac{\varepsilon}{3}.$$

Moreover, by increasing N if necessary, we can also assume $\frac{1}{N} \le \frac{\varepsilon}{3}$. Then

$$\begin{aligned}\|x_{n,k} - x_{m,m}\| &\le \|x_{n,k} - x_{n,j}\| + \|x_{n,j} - x_{m,j}\| + \|x_{m,j} - x_{m,m}\| \\ &\le \frac{2}{N} + \|x_{n,j} - x_{m,j}\|, \qquad n, m, k, j \ge N,\end{aligned}$$

and taking the limit $j \to \infty$ we conclude

$$\|x_{n,k} - x_{m,m}\| \le \varepsilon, \qquad n, m, k \ge N.$$

Now choosing $k = n$ we see that $x_{n,n}$ is a Cauchy sequence and hence $\xi \in \bar{X}$. Moreover, choosing $k = m$ shows $\|\xi_n - \xi\| \le \varepsilon$ for $n \ge N$ and hence $\xi_n \to \xi$. □

Notice that the completion $\bar{X}$ is unique. More precisely, every other Banach space which contains X as a dense subset and whose norm, restricted to X, is the given norm on X, 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 1.16 below).

Another way of obtaining the completion is via dual spaces, which will be discussed in Example 4.20.

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 1.14. Let $I = [a, b]$ be a compact interval. 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 sometimes inconvenient to work with equivalence classes of Cauchy sequences. A more convenient characterization can be given with the help of the Lebesgue integral (see Chapter 3 from [**40**] if you are familiar with basic Lebesgue integration; Theorem 3.18 from [**40**] will establish equivalence of both approaches). However, the present approach is still sufficient in many situations. In fact, for many objects/statements in a Banach space it suffices to define/establish them on a dense set and then extend it to the entire space by continuity. In the present context this implies that one establishes a claim first for continuous functions and then concludes that it holds on all of L^2 by an approximation argument.

For example, we can define the set of integrable functions $L^1(I)$ as the completion of $C(I)$ with respect to the norm

$$\|f\|_1 := \int_a^b |f(x)| dx.$$

The straightforward verification that this is indeed a norm is left as an exercise. Now for an arbitrary element $f \in L^1(I)$ we define its integral by choosing a Cauchy sequence of continuous functions f_n representing f and setting

$$\int_c^d f(x)dx := \lim_{n\to\infty} \int_c^d f_n(x)dx.$$

Using the inequality

$$\left|\int_c^d f(x)dx\right| \le \|f\|_1$$

for $f \in C(I)$ one easily checks that $\int_c^d f_n(x)dx$ is a Cauchy sequence in $\mathbb{C}$ and hence has a limit. Moreover, the same inequality shows that the integral of a null sequence is zero and hence (by the triangle inequality) our definition is independent of the particular representative chosen. By construction, the integral is linear and, since the above inequality extends to arbitrary $f \in L^1(I)$, we see that the integral is even (Lipschitz) continuous. Establishing some further properties of this integral is left as an exercise (Problem 1.36). Note however, that the notation $\int_c^d f(x)dx$ is somewhat imprecise, since f is an equivalence class of Cauchy sequences of continuous functions, which will not converge pointwise in general (Problem 1.37). Hence the value $f(x)$ has no meaning for $f \in L^1(I)$ in general.

We remark that the present argument is a special case of a more general approach: The integral is a continuous linear operation defined on a dense set which consequently has a unique extension to the entire Banach space. This will be the content of Theorem 1.16 below. ⋄

Example 1.15. 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. To this end we first establish (Problem 1.38) 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 $f, g \in C(I)$. From this we conclude (Problem 1.38) that $\|.\|_p$ satisfies the triangle inequality (still for continuous functions).

Now consider $f \in L^p(I)$, $g \in L^q(I)$ and let $f = [f_n]$, $g = [g_n]$ be the corresponding Cauchy sequences. Then Hölder's inequality implies that $f_n g_n$ is a Cauchy sequence with respect to $\|.\|_1$ and we can define $fg := [f_n g_n]$. Using Hölder's inequality and the triangle inequality it is easy to check that this definition is independent of the representatives for f and g. In particular, with this definition Hölder's inequality remains true for arbitrary $f \in L^p(I)$, $g \in L^q(I)$ (with $f \in C(I)$ if $p = \infty$ or $g \in C(I)$ if $q = \infty$).

Finally note that, Hölder's inequality implies

$$\|f\|_1 \le \|1\|_q \|f\|_p = (b-a)^{1/q}\|f\|_p,$$

which shows that $L^p(I) \subseteq L^1(I)$ (for this the fact that I is bounded is crucial). In particular, any $f \in L^p(I)$ can be integrated as explained in the previous example. $\diamond$

Problem 1.36. *Show that the integral defined in Example 1.14 satisfies*

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

How should $|f|$ be defined here?

Problem 1.37. *Show that the value $f(x)$ of a* function $f \in L^1([-1,1])$ *is not well-defined. To this end find a null sequence (with respect to $\|.\|_1$) which converges pointwise to any given value at* 0. *Also find a null sequence which diverges at* 0.

Problem* 1.38. *Show Hölder's inequality for continuous functions and conclude that $\|.\|_p$ fulfills the requirements of a norm on $C(I)$.*

Problem 1.39. *Show that $L^p(I)$, $1 \le p < \infty$, is a Hilbert space if and only if $p = 2$.*

Problem 1.40. *Is it possible to define* $\sin(f)$ *for $f \in L^1(I)$? If yes, how should this be done? What about* $\exp(f)$*?*

Problem 1.41. *Show that every $f \in L^1(I)$ has a uniformly continuous primitive*

$$F(x) := \int_c^x f(t)dt,$$

where $c \in I$ is some fixed point. The set of functions arising in this way is known as the set of ***absolutely continuous functions****.*

1.5. Compactness

In analysis, compactness is one of the most ubiquitous tools for showing existence of solutions for various problems.

Before we look into this, please recall that a subset K of a metric space X (or more generally, a topological space) is called **compact** if from any collection of open subsets covering K, we can select a finite number of sets which still cover K. There is also an alternative definition, namely, that K is compact if every sequence from K has a convergent subsequence whose limit is in K. In general topological spaces these two concepts are different, but in a metric space they agree (Lemma B.18). A set U is called **relatively compact** if its closure is compact.

For a subset U of a Banach space (or more generally a complete metric space) the following are equivalent (see Lemma B.18):

- U is relatively compact
- every sequence from U has a convergent subsequence
- U is totally bounded (i.e. for every $\varepsilon > 0$ it can be covered by a finite number of balls of radius ε centered at points in U)

Note that a totally bounded set is in particular bounded (i.e. it is contained in some ball of finite radius — Problem 1.42) and in $\mathbb{C}^n$ (or $\mathbb{R}^n$) the converse is also true, which is the content of the Heine–Borel theorem (Theorem B.19). Moreover, by Theorem 1.8 every finite dimensional Banach space is isomorphic to $\mathbb{C}^n$ (or $\mathbb{R}^n$) and hence the Heine–Borel theorem also applies in this situation:

Theorem 1.10 (Heine–Borel[23])**.** *In a finite dimensional Banach space a set is compact if and only if it is bounded and closed.*

In the infinite dimensional case this breaks down.

Example 1.16. Consider the bounded sequence $(\delta^n)_{n=1}^\infty$ in $\ell^p(\mathbb{N})$. Since $\|\delta^n - \delta^m\|_p = 2^{1/p}$ 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 turns out that it fails in any infinite dimensional space as we will see in Theorem 4.31 below. 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 1.11. *Let X be a metric space and K some subset. Assume that for every $\varepsilon > 0$ there is a metric space Y_ε, a surjective map $P_\varepsilon : X \to Y_\varepsilon$, and some $\delta > 0$ such that $P_\varepsilon(K)$ is totally bounded and $d(x,y) < \varepsilon$ whenever $x, y \in K$ with $d(P_\varepsilon(x), P_\varepsilon(y)) < \delta$. Then K is totally bounded.*

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

[23]Eduard Heine (1821–1881), German mathematician

[23]Émil Borel (1871–1956), French mathematician and politician

Proof. Fix $\varepsilon > 0$. Then there is a corresponding δ and by total boundedness of $P_\varepsilon(K)$ we can find a cover $\{B_\delta(y_j)\}_{j=1}^m$ for $P_\varepsilon(K)$ with $y_j = P_\varepsilon(x_j)$ for some $x_j \in K$. Now note that $\{B_\varepsilon(x_j)\}_{j=1}^n$ is an ε-cover for K since $P_\varepsilon^{-1}(B_\delta(y_j)) \cap K \subseteq B_\varepsilon(x_j)$.

For the last claim consider $P_{\varepsilon/3}$ and note that for $\delta := \varepsilon/3$ we have $\|x-y\| \le \|(1-P_{\varepsilon/3})x\| + \|P_{\varepsilon/3}(x-y)\| + \|(1-P_{\varepsilon/3})y\| < \varepsilon$ for $x, y \in K$. □

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

Theorem 1.12 (Fréchet[24])**.** *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 1.11.

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

This theorem tells us that there are essentially two ways a sequence can fail to have a convergent subsequence. It can be (pointwise) unbounded or its *mass* can run off to infinity. The first case is prevented by condition (i) and the second by condititon (ii). The latter case is what happened in Example 1.16.

Example 1.17. Fix $b \in \ell^p(\mathbb{N})$ if $1 \le p < \infty$ or $b \in c_0(\mathbb{N})$ if $p = \infty$. Then $\mathcal{K} := \{a \,|\, |a_j| \le |b_j|\} \subset \ell^p(\mathbb{N})$ is compact. Indeed, by $\|a\|_\infty \le \|b\|_\infty \le \|b\|_p$ for $a \in \mathcal{K}$ we see that (i) holds. Concerning (ii) note that choosing n such that $\sum_{j>n} |b_j|^p \le \varepsilon^p$ we conclude that $\|(1-P_n)a\|_p \le \|(1-P_n)b\|_p \le \varepsilon$ for all $a \in \mathcal{K}$. ⋄

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

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

[24] Maurice René Fréchet (1878–1973), French mathematician

That is, in this case δ is required to be not only independent of $x, y \in I$ but also of the function $f \in F$.

Warning: Like continuity, equicontinuity can be defined uniformly, that is with δ independent of x, or pointwise, that is with δ dependent of x. In case of a compact interval, both versions agree (cf. Problem B.65).

Theorem 1.13 (Arzelà–Ascoli[25]). *Let $F \subset C(I)$ be a family of continuous functions on a compact interval I. Then F is relatively compact if and only if F is equicontinuous and bounded.*

Proof. Suppose F is equicontinuous and bounded. Fix $\varepsilon > 0$. According to the definition of equicontinuity, there is a corresponding δ and we can choose points $x_1, \dots x_n$ such that the intervals $(x_j - \delta, x_j + \delta)$ cover I. Now consider $P : C(I) \to \mathbb{C}^n$, $P(f) = (f(x_1), \dots, f(x_n))$. Then $P(F)$ is bounded and for $f, g \in F$ we have $\|f - g\|_\infty \le 3\varepsilon$ whenever $|P(f) - P(g)| < \varepsilon$. Indeed, just note that for every x there is some j such that $|x - x_j| < \delta$ 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 1.11.

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. □

Again we see that there are essentially two ways a sequence can fail to have a convergent subsequence. Either it is unbounded or it locally changes faster and faster. An example for the first situation is $f_n(x) = n$ (which is perfectly equicontinuous) while an example for the second situation is the sequence from Example 1.12.

Example 1.18. Consider the solution $f_n(x)$ of the initial value problem

$$f' = \sin(nf), \qquad f(0) = 1.$$

(Assuming this solution exists — it can in principle be found using separation of variables.) Then $|f_n'(x)| \le 1$ and hence the mean value theorem shows that the family $\{f_n\} \subseteq C([0,1])$ is equicontinuous. Hence there is a uniformly convergent subsequence. (Remark: By looking at the fixed points of this differential equation it is not hard to see that $f_n \to 1$.) ⋄

Problem 1.42. *Show that in a metric space X a totally bounded set U is bounded.*

[25]Cesare Arzelá (1847–1912), Italian mathematician
[25]Giulio Ascoli (1843–1896), Italian mathematician

Problem 1.43. *Find a compact subset of $\ell^\infty(\mathbb{N})$ which does not satisfy (ii) from Theorem 1.12.*

Problem 1.44. *Show that a subset $\mathcal{K} \subset c_0(\mathbb{N})$ 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 $b \in \mathcal{K}$.*

Problem 1.45. *Is the family $F := \{f_n(x) = \sin(nx)\}_{n\in\mathbb{N}} \subset C[0,1]$ relatively compact?*

Problem 1.46. *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**)

$$\operatorname{Ker}(A) := \{x \in \mathfrak{D}(A) | Ax = 0\} \subseteq X \tag{1.58}$$

and **range**

$$\operatorname{Ran}(A) := \{Ax | x \in \mathfrak{D}(A)\} = A\mathfrak{D}(A) \subseteq Y \tag{1.59}$$

are again linear subspaces. Note that a linear map A will be continuous if and only if 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

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

is finite. This says that A is bounded if the image of the closed unit ball $\bar{B}_1^X(0) \cap \mathfrak{D}(A)$ is contained in some closed ball $\bar{B}_r^Y(0)$ 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

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

and hence is Lipschitz[26] continuous, that is, $\|Ax - Ay\|_Y \le \|A\|\|x - y\|_X$ for $x, y \in \mathfrak{D}(A)$. Note that $\|A\|$ could also be defined as the optimal constant in the inequality (1.61). In particular, it is continuous. The converse is also true:

Theorem 1.14. *A linear operator A between normed spaces 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 $x_n \in \mathfrak{D}(A)$ such that $\|Ax_n\|_Y \ge n$. Then $y_n := \frac{1}{n}x_n$ converges to 0 but $\|Ay_n\|_Y \ge 1$ does not converge to 0. □

Of course it suffices to check continuity at one point in X, say at 0, since continuity at all other points will then follow by a simple translation.

If X is finite dimensional, then every linear operator is bounded:

Lemma 1.15. *Let X, Y be normed spaces with X finite dimensional. Then every linear operator $A : \mathfrak{D}(A) \subseteq X \to Y$ is bounded.*

Proof. Choose a basis $\{x_j\}_{j=1}^n$ for $\mathfrak{D}(A)$ such that every $x \in \mathfrak{D}(A)$ can be written as $x = \sum_{j=1}^n \alpha_j x_j$. By Theorem 1.8 there is a constant $m > 0$ such that $(\sum_{j=1}^n |\alpha_j|^2)^{1/2} \le m\|x\|_X$. Then

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

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

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 1.19. 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). Note also that that the sup in (1.60) is only attained if a attains its supremum.

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). ⋄

[26]Rudolf Lipschitz (1832–1903), German mathematician

Example 1.20. Consider the vector space of differentiable functions $X := C^1[0,1]$ and equip it with the norm (cf. Problem 1.55)

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

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 1.16 (extension principle)**.** *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 $\mathfrak{D}(A)$ is dense, we can find a convergent sequence $x_n \to x$ from $\mathfrak{D}(A)$ for every $x \in X$. Moreover, since A is bounded, Ax_n is also Cauchy and has a limit since Y is assumed complete. Consequently, this extension can only be given by

$$\overline{A}x := \lim_{n\to\infty} Ax_n, \qquad x_n \in \mathfrak{D}(A), \quad x \in X.$$

To show that this definition is independent of the sequence $x_n \to x$, let $y_n \to x$ be a second sequence and observe

$$\|Ax_n - Ay_n\| = \|A(x_n - y_n)\| \le \|A\|\|x_n - y_n\| \to 0.$$

Since for $x \in \mathfrak{D}(A)$ we can choose $x_n := x$, we see that $\overline{A}x = Ax$ 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 $\mathscr{L}(X,Y)$. Here an operator $A : X \to Y$ (in contradistinction to $A : \mathfrak{D}(A) \subseteq X \to Y$) is implicitly understood to be defined on all of X, that is, $A \in \mathscr{L}(X,Y)$ implies $\mathfrak{D}(A) = X$. 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 (in case of a real Banach space we set of course $X^* := \mathscr{L}(X, \mathbb{R})$). The dual space takes the role of coordinate functions in a Banach space.

Example 1.21. Let X be a finite dimensional Banach space and let $\{u_j\}_{j=1}^n$ be a basis. Then every $x \in X$ can be uniquely written as $x = \sum_{j=1}^n \alpha_j u_j$ and we can consider the dual functionals defined via $u_j'(x) := \alpha_j$ for $1 \le j \le n$. The biorthogonal system $\{u_j'\}_{j=1}^n$ (whose elements are continuous by Lemma 1.15) forms a **dual basis** for X^* since any other linear functional $\ell \in X^*$ can be written as $\ell = \sum_{j=1}^n \ell(u_j) u_j'$. In particular, X and X^* have the same dimension.

Another easy consequence is that finite dimensional subspaces Y of arbitrary normed spaces are closed: Let $\{u_j\}_{j=1}^n$ be a basis for Y and $\{u_j'\}_{j=1}^n$ the corresponding dual basis for Y^*. Suppose $x_m \in Y$ converges to x. Then the sequences $u_j^*(x_m)$ are Cauchy and converge to a limit α_j. Hence $x_m = \sum_{j=1}^n u_j^*(x_m) u_j \to \sum_{j=1}^n \alpha_j u_j \in Y$. ⋄

Example 1.22. 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 $\|\ell_j\| = 1$ (since equality is attained for $a = \delta^j$). More general, let $b \in \ell^q(\mathbb{N})$ where $\frac{1}{p} + \frac{1}{q} = 1$. Then

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

is a bounded linear functional satisfying $\|\ell_b\| \le \|b\|_q$ by Hölder's inequality. In fact, since equality in the Hölder inequality is attained we even have $\|\ell_b\| = \|b\|_q$ (Problem 1.52). Note that the first example is a special case of the second one upon choosing $b = \delta^j$. ⋄

Example 1.23. 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 $C(I) \subset L^2(I)$! 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 $\ell_{x_0}(f_n) = f_n(x_0) = \sqrt{\frac{3n}{2}} \to \infty$. This implies that ℓ_{x_0} cannot be extended to all of $L^2(I)$ 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 1.24. 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 1.17. *Let X, Y be normed spaces. 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 $A_n x$ converges for every x. Define a new operator A via $Ax := \lim_{n\to\infty} A_n x$. By continuity of the vector operations, A is linear and by continuity of the norm $\|Ax\| = \lim_{n\to\infty} \|A_n x\| \le (\lim_{n\to\infty} \|A_n\|)\|x\|$, it is bounded (recall that $\|A_n\|$ is Cauchy by the inverse triangle inequality). Furthermore, given $\varepsilon > 0$, there is some N such that $\|A_n - A_m\| \le \varepsilon$ for $n, m \ge N$ and thus $\|A_n x - A_m x\| \le \varepsilon \|x\|$. Taking the limit $m \to \infty$, we see $\|A_n x - Ax\| \le \varepsilon \|x\|$; that is, $\|A_n - A\| \le \varepsilon$ and hence $A_n \to A$. □

Example 1.25. In the special case $X := \mathbb{C}$ we have $\mathscr{L}(\mathbb{C}, Y) \cong Y$ via the isometry $A \mapsto A1$. In particular, this shows that the theorem fails in general if Y is not complete. ◇

Example 1.26. In the special case $Y := \mathbb{C}$ we see that the dual space $X^* := \mathscr{L}(X, \mathbb{C})$ is always a Banach space, even if X is not complete. Moreover, by Theorem 1.16 the completion $\bar{X}$ satisfies $\bar{X}^* = X^*$. ◇

Warning: Observe that $\mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ is no Hilbert space in general even if $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are Hilbert spaces. To see this observe that by Example 1.19 $\mathscr{L}(\ell^2(\mathbb{N}))$ contains $\ell^\infty(\mathbb{N})$ as a subspace, where the latter is no Hilbert space by Problem 1.30. Based on this observation it is straightforward to construct a counterexample to the parallelogram law provided both $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are at least two dimensional. In the case $\mathfrak{H}_2 = \mathbb{C}$ the dual space turns out to be isomorphic to $\mathfrak{H}_1$ (see Theorem 2.10 below).

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 $\|\mathbb{I}\| = 1$.

Example 1.27. Another example of a Banach algebra is $C(I)$ since we clearly have

$$\|fg\|_\infty \le \|f\|_\infty \|g\|_\infty.$$

◇

Problem 1.47. *Show that two norms on X are equivalent if and only if they give rise to the same convergent sequences.*

Problem 1.48. *Show that a finite dimensional subspace $M \subseteq X$ of a normed space is closed.*

Problem 1.49. *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 1.50. *Let $X := C[0,1]$. Investigate if the following operators $A : X \to X$ are linear and, if yes, compute the norm.*

(i) $f(x) \mapsto (1-x)x\,f(x^2)$.
(ii) $f(x) \mapsto (1-x)x\,f(x)^2$.
(iii) $f(x) \mapsto \int_0^1 (1-x)y\,f(y)dy$.

Problem 1.51. *Let $X := C[0,1]$. Investigate the operator $A : X \to X$, $f(x) \mapsto x\,f(x)$. Show that this is a bounded linear operator and compute its norm. What is the closure of* $\mathrm{Ran}(A)$*?*

Problem* 1.52. *Show that* $\|l_b\| = \|b\|_q$*, where* $l_b \in \ell^p(\mathbb{N})^*$ *as defined in Example 1.22. (Hint: Choose* $a \in \ell^p$ *such that* $a_j b_j = |b_j|^q$*. See also Problem 1.18)*

Problem 1.53. *Let* $X := C[0,1]$*. Show that* $\ell(f) := \int_0^1 f(x)dx$ *is a linear functional. Compute its norm. Is the norm attained? What if we replace* X *by* $X_0 := \{f \in C[0,1] | f(0) = 0\}$ *(in particular, check that this is a closed subspace)?*

Problem 1.54. *Show that the* ***Fredholm 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* $X := L^2(0,1)$*. Show that the norm in the* $X = C[0,1]$ *case is given by*

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

Show that the same is true for the ***Volterra integral operator***[27]

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

with norm in the $X = C[0,1]$ *case is given by*

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

Problem* 1.55. *Let* I *be a compact interval. Show that the set of differentiable functions* $C^1(I)$ *becomes a Banach space if we set* $\|f\|_{\infty,1} := \max_{x\in I}|f(x)| + \max_{x\in I}|f'(x)|$*. In fact, it is even a Banach algebra.*

Problem* 1.56. *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* $A_n B_n \to AB$*.*

Problem 1.57. *Let* $A \in \mathscr{L}(X)$ *be a bijection. Show*

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

Problem* 1.58. *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* $\|B\| < \|A^{-1}\|^{-1}$*.*

[27] Vito Volterra (1860–1940), Italian mathematician

Problem* 1.59 (analytic functional calculus). *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 $\sum_{j=0}^{\infty} |f_j| \|A^j\| < \infty$ (note that by $\|A^j\| \le \|A\|^j$ this holds for example if $\|A\| < R$). 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 1.8.)

Problem* 1.60. *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 $\ell(y) = 1$.)*

Problem 1.61. *Show that a normed space is finite dimensional if and only if all subspaces are closed. (Hint: Problem 1.60.)*

Problem 1.62. *Show that the norm of a nontrivial linear functional $\ell \in X^*$ equals the reciprocal of the distance of the hyperplane $\ell(x) = 1$ to the origin:*

$$\|\ell\| = \frac{1}{\inf\{\|x\| | \ell(x) = 1\}}.$$

1.7. Sums and quotients of Banach spaces

Given two normed 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 normed space and a sequence in X converges if and only if the components converge in X_1 and X_2, respectively. Hence $X_1 \oplus X_2$ will be complete iff both X_1 and X_2 are complete.

Moreover, since all norms on $\mathbb{C}^2$ are equivalent (Theorem 1.8), we could equivalently take the norms $\|(x_1, x_2)\|_p := (\|x_1\|^p + \|x_2\|^p)^{1/p}$ or $\|(x_1, x_2)\|_\infty := \max(\|x_1\|, \|x_2\|)$. 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 associated with the scalar product $\langle (x_1, x_2), (y_1, y_2) \rangle := \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle$.

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 1.64).

Lemma 1.18. *Let X_j, $j = 1, \dots, n$, be Banach spaces and define $X := \bigoplus_{p,j=1}^n X_j$ to be the Cartesian product $X_1 \times \cdots \times X_n$ together with the norm*

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

Then X is a Banach space. Moreover, all norms are equivalent and the sum is associative $(X_1 \oplus_p X_2) \oplus_p X_3 \cong X_1 \oplus_p (X_2 \oplus_p X_3) \cong \bigoplus_{p,j=1}^3 X_j$.

If $A_j : \mathfrak{D}(A_j) \subseteq X_j \to Y_j$, $j = 1, 2$, are linear operators, then we can define a linear operator via

$$\begin{aligned} A_1 \oplus A_2 : \mathfrak{D}(A_1) \times \mathfrak{D}(A_2) \subseteq X_1 \oplus X_2 &\to Y_1 \oplus Y_2 \\ (x_1, x_2) &\mapsto (A_1 x_1, A_2 x_2). \end{aligned} \tag{1.65}$$

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 given by

$$\begin{aligned} A_1 \widehat{\oplus} A_2 : \mathfrak{D}(A_1) \times \mathfrak{D}(A_2) \subseteq X_1 \oplus X_2 &\to Y \\ (x_1, x_2) &\mapsto A_1 x_1 + A_2 x_2. \end{aligned} \tag{1.66}$$

Again $A_1 \widehat{\oplus} A_2$ will be bounded if and only if both A_1 and A_2 are bounded and $\|A_1 \widehat{\oplus} A_2\| = \max(\|A_1\|, \|A_2\|)$. If an index $p \ne 1$ is used to define the direct sum $X_1 \oplus_p X_2$, then $\|A_1 \widehat{\oplus} A_2\| = (\|A_1\|^q + \|A_2\|^q)^{1/q}$ with $\frac{1}{p} + \frac{1}{q} = 1$.

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 $(\ell_1, \ell_2) \in X_1^* \oplus_q X_2^* \mapsto \ell_1 \widehat{\oplus} \ell_2 \in (X_1 \oplus_p X_2)^*$. It is not hard to see that this identification is bijective and preserves the norm (Problem 1.65).

Lemma 1.19. *Let X_j, $j = 1, \dots, n$, be Banach spaces. Then $(\bigoplus_{p,j=1}^n X_j)^* \cong \bigoplus_{q,j=1}^n X_j^*$, where $\frac{1}{p} + \frac{1}{q} = 1$.*

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. Warning: It is important to observe, that $M \dot{+} N$ is in general not isomorphic to $M \oplus N$ 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 1.21 below).

Example 1.28. Consider $X := \ell^p(\mathbb{N})$ with $1 \le p < \infty$. Let $M := \{a \in X | a_{2n} = 0\}$ and $N := \{b \in X | b_{2n-1} = n^3 b_{2n}\}$. Then both subspaces are closed and $M \cap N = \{0\}$. Moreover, if $c = a + b$ with $a \in M$ and $b \in N$, then $b_{2n} = c_{2n}$ (and $b_{2n-1} = n^3 c_{2n}$) as well as $a_{2n-1} = c_{2n-1} - n^3 c_{2n}$ (and $a_{2n} = 0$). Hence there is such a splitting for c if and only if $n^3 c_{2n} \in \ell^p(\mathbb{N})$. In particular, this works for all sequences with compact support and thus $M \dotplus N$ is dense. However, it is not all of X since $c_n = \frac{1}{n^2} \notin M \dotplus N$. Indeed, by the above analysis we had $b_{2n} = \frac{1}{4n^2}$ and hence $b_{2n-1} = \frac{n}{4}$, contradicting $b \in N \subseteq X$. What about the case $p = \infty$? ⋄

A closed subspace M is called **complemented** if we can find another closed subspace N with $M \cap N = \{0\}$ and $M \dotplus N = X$. 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 $M = \mathrm{Ker}(Q) = \mathrm{Ran}(P)$ and $N = \mathrm{Ker}(P) = \mathrm{Ran}(Q)$. Of course one would like P (and hence also Q) to be continuous. If we consider the linear bijection $\phi : M \oplus N \to X$, $(x_1, x_2) \mapsto x_1 + x_2$, then this is equivalent to the question if ϕ^{-1} is continuous. By the triangle inequality ϕ is continuous with $\|\phi\| \le 1$ and the inverse mapping theorem (Theorem 4.8) will answer this question affirmative. In summary, we have $M \oplus N \cong X$.

It is important to emphasize, that N is not unique (consider a two dimensional space and choose a one-dimensional subspace, then any other one-dimensional subspace will complement the given one as long as it is different). Moreover, it is precisely the requirement that N is closed which makes P continuous (conversely observe that $N = \mathrm{Ker}(P)$ 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). Furhtermore, 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 1.72 and 4.26).

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). The dimension of X/M is known as the **codimension** of M.

Note that if M is complemented by N, then $X/M \cong N$ since the map $X/M \to N$, $[x_1 + x_2] \mapsto x_2$ is a bijection (cf. Problem 1.71). In particular, $\dim(X/M) = \dim(N)$ in this case.

IFor a linear operator $A : X \to Y$ the linear space $\operatorname{Coker}(A) := Y/\operatorname{Ran}(A)$ is known as the **cokernel** of A.

Lemma 1.20. *Let M be a closed subspace of a normed space X. Then X/M together with the norm*

$$\|[x]\| := \operatorname{dist}(x, M) = \inf_{y \in M} \|x - y\| \tag{1.67}$$

is a normed space. It is complete if X is.

Proof. First of all we need to show that (1.67) 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

$$\begin{aligned} \|[x] + [y]\| = \|[x + y]\| &= \inf_{z \in M} \|x + y + z\| = \inf_{z_1, z_2 \in M} \|x + z_1 + y + z_2\| \\ &\le \inf_{z_1 \in M} \|x + z_1\| + \inf_{z_2 \in M} \|y + z_2\| = \|[x]\| + \|[y]\|. \end{aligned}$$

Thus (1.67) is a norm and it remains to show that X/M is complete if X is. 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 (1.67) we can choose the representatives x_n such that $\|x_{n+1} - x_n\| < 2^{-n}$ (start with x_1 and then choose 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 $\|[x_n] - [x]\| = \|[x_n - x]\| \le \|x_n - x\|$ we see that $[x]$ is the limit of $[x_n]$. □

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

Example 1.29. If $X := C[0, 1]$ and $M := \{f \in X | f(0) = 0\}$ then $X/M \cong \mathbb{C}$. In fact, note that every $f \in X$ can be written as $f(x) = g(x) + \alpha$ with $g(x) := f(x) - f(0) \in M$ and $\alpha := f(0) \in \mathbb{C}$. ⋄

Example 1.30. 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 $\alpha := \lim_{n\to\infty} x_n \in \mathbb{C}$ its limit. ⋄

The **quotient map** $\pi : X \to X/M$, $x \mapsto [x]$ is a linear surjective map with $\mathrm{Ker}(\pi) = M$. By $\|[x]\| \le \|x\|$ 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 1.21. *Let X be a normed 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 1.63. *Let X be a normed space and suppose $P \in \mathscr{L}(X)$ is a projection (i.e., $P^2 = P$). Show that $Q := \mathbb{I} - P$ is also a projection satisfying $PQ = QP = 0$.*

Problem* 1.64. *Prove Lemma 1.18.*

Problem* 1.65. *Prove Lemma 1.19. (Hint: Hölder's inequality in $\mathbb{C}^n$ and note that equality is attained.)*

Problem 1.66. *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*

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

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 1.67. *Let $X := \ell^p(\mathbb{N})$ and $M := \{a \in X | a_{2n} = 0\}$, $N := \{a \in X | n\, a_{2n} = a_{2n-1}\}$. Is $M \dot{+} N$ closed?*

Problem 1.68. *Let ℓ be a nontrivial linear functional. Then its kernel has codimension one. Find the projection onto its kernel.*

Problem 1.69. *Consider $X := \ell^\infty(\mathbb{N})$ and $M := c_0(\mathbb{N})$. Show $\mathrm{dist}(a, M) = \limsup_j |a_j|$ for $a \in X$.*

Problem 1.70 (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* 1.71. *Let X, Y be normed spaces and 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 $\mathrm{Ker}(\tilde{A}) = \mathrm{Ker}(A)/M$, $\mathrm{Ran}(\tilde{A}) = \mathrm{Ran}(A)$. In particular, $\tilde{A}$ is injective for $M = \mathrm{Ker}(A)$.*

Problem* 1.72. *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 $\ell_k([x_j]) = \delta_{j,k}$ — see Example 1.21.)*

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. Let I be some compact interval, then we can make $C^1(I)$ into a Banach space (Problem 1.55) by introducing the norm $\|f\|_{1,\infty} := \|f\|_\infty + \|f'\|_\infty$. By a straightforward extension we can even get (cf. Problem 1.76)

Theorem 1.22. *Let $I \subseteq \mathbb{R}$ be some interval. The space $C_b^k(I)$ of all functions whose partial derivatives up to order k are bounded and continuous form a Banach space with norm*

$$\|f\|_{k,\infty} := \sum_{j=0}^{k} \sup_{x \in I} |f^{(j)}(x)|. \tag{1.68}$$

Since als norms on finite dimensional spaces are equivalent (Theorem 1.8), the sum in (1.68) could be replaced by (e.g.) a weighted sum or a maximum. Different choices might be more convenient in different situations.

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

$$[f]_\gamma := \sup_{x \neq y \in I} \frac{|f(x) - f(y)|}{|x - y|^\gamma} \tag{1.69}$$

is finite. Clearly, any Hölder continuous function is uniformly continuous (explicitly we can choose $\delta = (\varepsilon/[f]_\gamma)^{1/\gamma}$) 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 1.73).

Example 1.31. By the mean value theorem every function $f \in C_b^1(I)$ is Lipschitz continuous with $[f]_1 \leq \|f'\|_\infty$. ⋄

Example 1.32. 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 $0 \leq x < y$ and set $t = \frac{x}{y} \in [0, 1)$. Then we have

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

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 (1.69) is a seminorm and, once we add the maximum norm, the corresponding space is complete.

Theorem 1.23. *Let $I \subseteq \mathbb{R}$ be an interval. The space $C_b^{k,\gamma}(I)$ 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*

$$\|f\|_{k,\gamma,\infty} := \|f\|_{k,\infty} + [f^{(k)}]_\gamma. \tag{1.70}$$

As already noted before, in the case $\gamma = 0$ we get a norm which is equivalent to $\|f\|_{\infty,k}$ and we will set $C_b^{k,0}(I) := C_b^k(I)$ 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 $\overline{I}$ (cf. Theorem B.39). In this sense, the spaces $C_b^{0,\gamma}(I)$ and $C_b^{0,\gamma}(\overline{I})$ are naturally isomorphic. Finally, since Hölder continuous functions on a bounded domain are automatically bounded, we can drop the subscript b in this situation.

We also note, that if we increase γ, then the Hölder conditions gets stronger. To formulate the corresponding result we call an embedding J :

$Y \hookrightarrow X$ (i.e. J is a linear injective map) **continuous** if J is bounded and **compact** if J maps bounded sets to compact sets. In other words, the embedding is continuous if the Y-norm is stronger than the X-norm, in the sense that $\|J(y)\|_X \le C\|y\|_Y$, and the embedding is compact if every sequence y_n from Y, which is bounded with respect to the Y-norm, has a subsequence which converges with respect to the X-norm.

Theorem 1.24. *Suppose $I \subset \mathbb{R}$ is a compact interval. Then $C^{0,\gamma_2}(I) \subseteq C^{0,\gamma_1}(I) \subseteq C(I)$ for $0 < \gamma_1 < \gamma_2 \le 1$ with the embeddings being compact.*

Proof. That we have continuous embeddings follows since $|x-y|^{-\gamma_1} = |x-y|^{-\gamma_2+(\gamma_2-\gamma_1)} \le (2r)^{\gamma_2-\gamma_1}|x-y|^{-\gamma_2}$ if r denotes the length of I. Hence $[f]_{\gamma_1} \le (2r)^{\gamma_2-\gamma_1}[f]_{\gamma_2}$.

Moreover, that the embedding $C^{0,\gamma_1}(I) \subseteq C(I)$ is compact follows from the Arzelà–Ascoli theorem (Theorem 1.13).

To see the remaining claim let f_m be a bounded sequence in $C^{0,\gamma_2}(I)$, explicitly $\|f_m\|_\infty \le C$ and $[f_m]_{\gamma_2} \le C$. Hence by the Arzelà–Ascoli theorem we can assume that f_m converges uniformly to some $f \in C(I)$. Moreover, taking the limit in $|f_m(x) - f_m(y)| \le C|x-y|^{\gamma_2}$ we see that we even have $f \in C^{0,\gamma_2}(I) \subset C^{0,\gamma_1}(I)$. To see that f is the limit of f_m in $C^{0,\gamma_1}(I)$ we need to show $[g_m]_{\gamma_1} \to 0$, where $g_m := f_m - f$. Now observe that

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

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

As pointed out in Example 1.31, the embedding $C_b^1(I) \subseteq C_b^{0,1}(I)$ is continuous and combining this with the previous result immediately gives

Corollary 1.25. *Suppose $I \subset \mathbb{R}$ is a compact interval, $k_1, k_2 \in \mathbb{N}_0$, and $0 \le \gamma_1, \gamma_2 \le 1$. Then $C^{k_2,\gamma_2}(I) \subseteq C^{k_1,\gamma_1}(I)$ for $k_1 + \gamma_1 \le k_2 + \gamma_2$ with the embeddings being compact if the inequality is strict.*

For now continuous functions on intervals 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.9.

Problem 1.73. *Let I be an interval. Suppose $f : I \to \mathbb{C}$ is Hölder continuous with exponent $\gamma > 1$. Show that f is constant.*

Problem 1.74. *Show that the primitive F of $f \in L^p(I)$, $p > 1$ (cf. Problem 1.41) is Hölder continuous:*

$$|F(x) - F(y)| \le \|f\|_p |x - y|^{1-\frac{1}{p}}.$$

Problem 1.75. *Let $I := [a, b]$ be a compact interval and consider $C^1(I)$. Which of the following is a norm? In case of a norm, is it equivalent to $\|.\|_{1,\infty}$?*

(i) $\|f\|_\infty$
(ii) $\|f'\|_\infty$
(iii) $|f(a)| + \|f'\|_\infty$
(iv) $|f(a) - f(b)| + \|f'\|_\infty$
(v) $\int_a^b |f(x)|dx + \|f'\|_\infty$

Problem* 1.76. *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 1.77. *Show that the product of two bounded Hölder continuous functions is again Hölder continuous with*

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

Conclude that $C_b^{0,\gamma}(I)$ is a Banach algebra.

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 $1 \le j \le n$. In particular,

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

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

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

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**[1]

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

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 $\sum_j \langle u_j, f\rangle u_j$ is a Cauchy sequence since $\sum_j |\langle u_j, f\rangle|^2$ 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

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

[1] Friedrich Bessel (1784–1846), German astronomer, mathematician, physicist, and geodesist

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 (show this).

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

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

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**.[2] 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 $L^2(-1,1)$, 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**[3]

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

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

Example 2.3. The set of functions

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

forms an orthonormal basis for $\mathfrak{H} := L^2(-\pi,\pi)$. 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.*

---

[2]Jørgen Pedersen Gram (1850–1916), Danish actuary and mathematician

[2]Erhard Schmidt (1876–1959), Baltic German mathematician

[3]Adrien-Marie Legendre (1752–1833), French mathematician

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

$$f = \sum_{j \in J} \langle u_j, f \rangle u_j. \tag{2.15}$$

(iii) *For every vector* $f \in \mathfrak{H}$ *we have* ***Parseval's relation***[4]

$$\|f\|^2 = \sum_{j \in J} |\langle u_j, f \rangle|^2. \tag{2.16}$$

(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

$$\begin{aligned}\left| \sum_{j\in J} |\langle u_j, f\rangle|^2 - \sum_{j\in J} |\langle u_j, g\rangle|^2 \right| &\le \sqrt{\sum_{j\in J} |\langle u_j, f-g\rangle|^2} \sqrt{\sum_{j\in J} |\langle u_j, f+g\rangle|^2} \\ &\le \|f-g\| \|f+g\| \end{aligned} \tag{2.17}$$

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 $\sum_{k=1}^n U_{k,j}^* v_k = u_j$ showing that $v_1, \dots, v_n$ span $\mathfrak{H}$ and hence K cannot have more than n elements.

[4] Marc-Antoine Parseval (1755–1836), French mathematician

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 $|\tilde{K}| \le |J \times \mathbb{N}| = |J|$ (Theorem A.9). But $k \in K \setminus \tilde{K}$ implies $v_k = 0$ and hence $\tilde{K} = K$. So $|K| \le |J|$ and reversing the roles of J and K shows $|K| = |J|$ (Theorem A.3). □

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, $\ell^2(\mathbb{N})$ is the only 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:

$$\langle Ug, Uf\rangle_2 = \langle g, f\rangle_1, \qquad g, f \in \mathfrak{H}_1. \tag{2.18}$$

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 that 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 (Theorem A.2): 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. This maximal element is an ONB by Theorem 2.4 (i). □

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 all **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. $\mathrm{e}^{\mathrm{i}t}+\mathrm{e}^{\mathrm{i}\sqrt{2}t}$ is not periodic).

It is not difficult to show that

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\mathrm{e}^{\mathrm{i}\theta t}dt=\begin{cases}1, & \theta=0,\\ 0, & \theta\neq 0,\end{cases}$$

and hence one can conclude that every almost periodic function has a mean value

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

Note that $|M(f)|\le\|f\|_\infty$.

Next one can show that

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

defines a nonnegative sesquilinear form on $AP(\mathbb{R})$ which is positive definite on the set of trigonometric polynomials since

$$\|f\|^2=M(|f|^2)=\sum_{k=1}^n|\alpha_k|^2$$

in this case (with f as above). It is also positive definite on $AP(\mathbb{R})$, but this is more difficult to see ([**22**, Section 2.3]).

Abbreviating $e_\theta(t):=\mathrm{e}^{\mathrm{i}\theta t}$ we see that $\{e_\theta\}_{\theta\in\mathbb{R}}$ is an uncountable orthonormal set and the **Bohr transfom**[5]

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

maps $AP(\mathbb{R})$ isometrically (with respect to $\|.\|$) into $\ell^2(\mathbb{R})$. This map is however not surjective (take e.g. the Fourier series of a square integrable function

[5]Harald Bohr (1887–1951), Danish mathematician

which is discontinuous — see later) and hence $AP(\mathbb{R})$ is not complete with respect to $\|.\|$. ⋄

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: First establish $\Gamma(f_1, \dots, f_j + \alpha f_k, \dots, f_n) = \Gamma(f_1, \dots, f_n)$ for $j \ne k$ and use it to investigate how Γ changes 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.)*

Problem 2.4. *Show that the set of vectors $\{c^n := (1, n^{-1}, n^{-2}, \dots)\}_{n=2}^{\infty}$ is total in $\ell^2(\mathbb{N})$. (Hint: Use that for any $a \in \ell^2(\mathbb{N})$ the functions $f(z) := \sum_{j\in\mathbb{N}} a_j z^{j-1}$ is holomorphic in the unit disc.)*

Problem 2.5 (Orthogonal polynomials)**.** *Let $I = (a, b)$ be some interval and consider the scalar product*

$$\langle f, g\rangle := \int_a^b f(x)^* g(x) w(x) dx$$

associated with some positive weight function $w(x)$. Let $P_j(x) = x^j + \beta_j x^{j-1} + \gamma_j x^{j-2} + \dots$ be the corresponding monic orthogonal polynomials obtained by applying the Gram–Schmidt procedure (without normalization) to the monomials:

$$\int_a^b P_i(x) P_j(x) w(x) dx = \begin{cases} \alpha_j^2, & i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\bar{P}_j(x) := \alpha_j^{-1} P_j(x)$ be the corresponding orthonormal polynomials and show that they satisfy the three term recurrence relation

$$a_j \bar{P}_{j+1}(x) + b_j \bar{P}_j(x) + a_{j-1} \bar{P}_{j-1}(x) = x \bar{P}_j(x),$$

or equivalently

$$P_{j+1}(x) = (x - b_j)P_j(x) - a_{j-1}^2 P_{j-1}(x),$$

where

$$a_j := \int_a^b x\bar{P}_{j+1}(x)\bar{P}_j(x)w(x)dx, \qquad b_j := \int_a^b x\bar{P}_j(x)^2 w(x)dx.$$

Here we set $P_{-1}(x) = \bar{P}_{-1}(x) \equiv 0$ *for notational convenience (in particular we also have* $\beta_0 = \gamma_0 = \gamma_1 = 0$*). Moreover, show*

$$a_j = \frac{\alpha_{j+1}}{\alpha_j} = \sqrt{\gamma_{j+1} - \gamma_{j+2} + (\beta_{j+2} - \beta_{j+1})\beta_{j+1}}, \quad b_j = \beta_j - \beta_{j+1}.$$

Note also

$$\alpha_j = \alpha_0 \prod_{i=0}^{j-1} a_i, \quad \alpha_0 = \sqrt{\int_a^b w(x)dx}, \qquad \beta_j = -\sum_{i=0}^{j-1} b_i.$$

(Note that $w(x)dx$ *could be replaced by a measure* $d\mu(x)$*.)*

2.2. The projection theorem and the Riesz representation theorem

Let $M \subseteq \mathfrak{H}$ be a subset. Then

$$M^\perp := \{f | \langle g, f\rangle = 0, \forall g \in M\} \tag{2.19}$$

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 $(\overline{\mathrm{span}(M)})^\perp = M^\perp$. 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*

$$M \oplus M^\perp = \mathfrak{H} \tag{2.20}$$

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 $(\{u_j\}_{j\in J})^\perp$. $\square$

Given a closed subspace M the operator $P_M f := f_\parallel$ is called the **orthogonal projection** corresponding to M. If $\{u_j\}_{j\in J}$ is an orthonormal basis for M, then it is explixitely given by

$$P_M f = \sum_{j\in J} \langle u_j, f\rangle u_j. \tag{2.21}$$

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 2.14).

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, $M^{\perp\perp} = \operatorname{Ran}(P_{M^{\perp\perp}}) = \operatorname{Ran}(P_M) = M$. 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.

Example 2.5. Consider $\mathfrak{H} = \ell^2(\mathbb{N})$ and $J \subseteq \mathbb{N}$. Then $Z_J := \{a \in \mathfrak{H} | a_j = 0\, \forall j \in J\}$ is a closed subspace and

$$(P_{Z_J} a)_j = \begin{cases} 0, & j \in J, \\ a_j, & j \in \mathbb{N} \setminus J. \end{cases}$$

Of course a basis for Z_J is $\{\delta^j\}_{j\in\mathbb{N}\setminus J}$ and we can also write

$$P_{Z_J} a = \sum_{j\in\mathbb{N}\setminus J} \langle \delta^j, a\rangle \delta^j = \sum_{j\in\mathbb{N}\setminus J} a_j \delta^j.$$

Moreover, $Z_J^\perp = Z_{\mathbb{N}\setminus J}$. $\diamond$

Example 2.6. Let $\mathfrak{H}_{1,2}$ be Hilbert spaces and $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$ with closed range. Then the restriction $A_0 : \operatorname{Ker}(A)^\perp \to \operatorname{Ran}(A)$ is bijective and it will follow from the inverse mapping Theorem (Theorem 4.8) that the **pseudo inverse** $A^+ := A_0^{-1} P_{\operatorname{Ran}(A)} \in \mathscr{L}(\mathfrak{H}_2, \mathfrak{H}_1)$. Of course A^+ reduces to the usual inverse if A is bijective. Form the construction it is not hard to see that $A^+ A = P_{\operatorname{Ker}(A)^\perp}$ and $AA^+ = P_{\operatorname{Ran}(A)}$.

In case A is not bijective, the equation $Af = g$ will not have a solution in general and minimizing the error $\|Af - g\|$ is the best one can do. If we write $g = g_\parallel + g_\perp$ with respect to $\operatorname{Ran}(A)$ we see that $\|Af - g\|^2 = \|Af - g_\parallel\|^2 + \|g_\perp\|^2$ and hence any solution of $Af = g_\parallel$ (which exists since thanks to our assumption that the range of G is closed) will minimize the error. Of course this solution is only unique up to an element from the kernel of A and writing $f = f_\parallel + f_\perp$ with respect to $\mathrm{Ker}(A)$ we see that $f_\perp$ is the smallest minimizer. In particular, this smallest minimizer is given by $f_\perp = A^+ g$. $\diamond$

Next 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[6] representation theorem)**.** *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 $g_1 - g_2 \in \mathfrak{H}^\perp = \{0\}$. $\square$

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

$$Cf := \sum_{j\in J} \langle u_j, f\rangle^* u_j = \sum_{j\in J} \langle f, u_j\rangle u_j. \tag{2.24}$$

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 $\mathfrak{H}^*$.

Example 2.7. Consider $\mathfrak{H} = \ell^2(\mathbb{N})$. With respect to the standard basis $\{\delta^j\}_{j\in\mathbb{N}}$, the conjugation is given by $(Ca)_j = a_j^*$. With respect to the basis $\{\mathrm{i}\delta^j\}_{j\in\mathbb{N}}$, the conjugation is given by $(Ca)_j = -a_j^*$. $\diamond$

[6] Frigyes Riesz (1880–1956), Hungarian mathematician

Finally, we remark that projections cannot 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 $P_K(f) \in K$ such that $\|P_K(f) - f\| = \inf_{g\in K} \|f - g\|$. If we extend $P_K : \mathfrak{H} \to K$ by setting $P_K(g) = g$ 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

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

which shows that f_n is Cauchy and hence converges to some point in K 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 $\|P(f) - \tilde{P}(f)\|^2 \le 0$ 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

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

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

If K is a closed subspace then this projection will of course coincide with the orthogonal projection defined before. By inspection of the proof, note that $P_K(f)$ is alternatively characterized by $\mathrm{Re}(\langle f - P_K(f), g - P_K(f)\rangle) \le 0$ for all $g \in K$.

Example 2.8. In case of the closed unit ball $K := \bar{B}_1(0)$ we have $P_K(f) = \frac{f}{\|f\|}$ for $\|f\| > 1$ and $P_K(f) = f$ for $\|f\| \le 1$. ⋄

Problem 2.6. *Show that $M \subseteq N$ implies $N^\perp \subseteq M^\perp$.*

Problem 2.7. *Consider $\mathfrak{H} := \ell^2(\mathbb{Z})$. A sequence $a \in \mathfrak{H}$ is called even if $a_j = a_{-j}$ for all $j \in \mathbb{Z}$. Show that the set of even sequences M forms a closed subspace. Compute P_M and $M^\perp$.*

Problem 2.8. *Consider* $M := \{f \in L^2(0,1) | \int_0^1 f(x)dx = 0\} \subset L^2(0,1)$. *Compute* $M^\perp$.

Problem 2.9. *Consider the subspace* $M_e := \{f \in C(-1,1) | f(x) = f(-x)\}$ *of all even continuous functions in* $L^2(-1,1)$. *Is* M_e *closed? What is* $M_e^\perp$*? Compute* $\operatorname{dist}(\exp(x), M_e)$.

Problem 2.10. *Let* M_1, M_2 *be two subspaces of a Hilbert space* $\mathfrak{H}$. *Show that* $(M_1 + M_2)^\perp = M_1^\perp \cap M_2^\perp$. *If in addition* M_1 *and* M_2 *are closed, show that* $(M_1 \cap M_2)^\perp = \overline{M_1^\perp + M_2^\perp}$.

Problem 2.11. *Let* $M \subseteq \mathfrak{H}$ *be a closed subspace. Show that the quotient space* $\mathfrak{H}/M \cong M^\perp$ *(isometrically).*

Problem 2.12. *Show that* $\ell(a) = \sum_{j=1}^\infty \frac{a_j + a_{j+2}}{2^j}$ *defines a bounded linear functional on* $X := \ell^2(\mathbb{N})$. *Compute its norm.*

Problem 2.13. *Suppose* $U : \mathfrak{H} \to \mathfrak{H}$ *is unitary and* $M \subseteq \mathfrak{H}$. *Show that* $UM^\perp = (UM)^\perp$.

Problem* 2.14. *Suppose* $P \in \mathscr{L}(\mathfrak{H})$ *is a projection, that is,* $P^2 = P$. *Note that in this case every* $f \in \mathfrak{H}$ *can be written as* $f = Pf + (\mathbb{I} - P)f$ *with* $Pf \in \operatorname{Ran}(P)$ *and* $(\mathbb{I} - P)f \in \operatorname{Ker}(P)$, *where both* $\operatorname{Ker}(P)$ *and* $\operatorname{Ran}(P)$ *are closed (cf. Section 1.7).*

Show that the following are equivalent:

(i) $P = P_M$ *is the orthogonal projection onto* $M := \operatorname{Ran}(P)$.

(ii) $\|P\| = 1$.

(iii) $\operatorname{Ran}(P) = \operatorname{Ker}(P)^\perp$.

(iv) $\langle Pf, g\rangle = \langle f, Pg\rangle$ *for all* $f, g \in \mathfrak{H}$.

(Hint: To see (ii) $\Rightarrow$ *(iii) note that if you can find two normalized vectors* $f \in \operatorname{Ran}(P)$ *and* $g \in \operatorname{Ker}(P)$ *with* $|\langle f, g\rangle| = \varepsilon > 0$, *then* $\|P\| \geq (1 - \varepsilon^2)^{-1/2}$.*)*

Problem 2.15. *Let* $\mathfrak{H}$ *be a Hilbert space and* K *a nonempty closed convex subset. Prove that* K *has a unique element of minimal norm.*

Problem 2.16. *Consider* $\mathfrak{H} := \ell^2(\mathbb{N})$ *and set* $K := \{a \in \mathfrak{H} | |a_1| \leq 1, a_j = 0, j > 1\}$. *Compute* P_K.

Problem 2.17. *Compute* P_K *for a closed ball* $K := \bar{B}_r(g)$.

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

$$|s(g, f)| \le C\|g\|_{\mathfrak{H}_2} \|f\|_{\mathfrak{H}_1}. \tag{2.25}$$

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

$$s(g, f) = \langle g, Af\rangle_{\mathfrak{H}_2}. \tag{2.26}$$

Moreover, the norm of A is given by

$$\|A\| = \sup_{\|g\|_{\mathfrak{H}_2}=\|f\|_{\mathfrak{H}_1}=1} |\langle g, Af\rangle_{\mathfrak{H}_2}| \le C. \tag{2.27}$$

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

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

we infer $\|Af\|_{\mathfrak{H}_2} \le C\|f\|_{\mathfrak{H}_1}$, which shows that A is bounded with $\|A\| \le C$. Equation (2.27) is left as an exercise (Problem 2.18). □

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 2.9. 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

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

Since $(A^*_{jk})_{k=1}^{\infty}$ are the expansion coefficients of $A^*\delta^j$ (see (2.28) below), we have $\sum_{k=1}^{\infty} |A_{jk}|^2 = \|A^*\delta^j\|^2$ and by the Cauchy–Schwarz inequality the above sum is even absolutely convergent. ⋄

Moreover, in a complex Hilbert space the polarization identity (Problem 1.32) implies that $A \in \mathscr{L}(\mathfrak{H})$ 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. *Let $\mathfrak{H}_1$, $\mathfrak{H}_2$ be Hilbert spaces. 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*

$$\langle f, A^* g\rangle_{\mathfrak{H}_1} = \langle Af, g\rangle_{\mathfrak{H}_2}. \tag{2.28}$$

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

Lemma 2.14. *Let $\mathfrak{H}$ be a complex Hilbert space. A bounded operator $A \in \mathscr{L}(\mathfrak{H})$ is self-adjoint if and only if its quadratic form is real-valued. Moreover, in this case*

$$\|A\| = \sup_{\|f\|=1} |\langle f, Af\rangle|. \tag{2.29}$$

Proof. The formula for the norm follows by using the polarisation identity to estimate the real part of $\langle g, Af\rangle$ (see Problem 1.33). □

Warning: This result fails in a real Hilbert space.

Example 2.10. If $\mathfrak{H} := \mathbb{C}^n$ and $A := (a_{jk})_{1\le j,k\le n}$, then $A^* = (a_{kj}^*)_{1\le j,k\le n}$. Clearly A is self-adjoint if and only if $a_{jk} = a_{kj}^*$. ⋄

Example 2.11. If $\mathbb{I} \in \mathscr{L}(\mathfrak{H})$ is the identity, then $\mathbb{I}^* = \mathbb{I}$. ⋄

Example 2.12. 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 2.13. 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^*. In particular, A is self-adjoint if and only if a is real-valued. ⋄

Example 2.14. 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:

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

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 2.15. 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 2.15. *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* $\|A\|^2 = \|A^*A\| = \|AA^*\|$.

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 (2.27) from

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

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). In particular, by virtue of (1.64), we have $\|A\| \le \|A^*\|$. Finally replacing A by A^* and using (ii) finishes the claim. □

Note that $\|A\| = \|A^*\|$ implies that taking adjoints is a continuous operation. For later use also note that (Problem 2.23)

$$\operatorname{Ker}(A^*) = \operatorname{Ran}(A)^{\perp}. \tag{2.30}$$

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 $s(f, f) \ge 0$

and it is called **coercive** if

$$\mathrm{Re}(s(f,f)) \geq \varepsilon \|f\|^2, \qquad \varepsilon > 0. \tag{2.31}$$

We will call $A \in \mathscr{L}(\mathfrak{H})$ nonnegative, coercive if its associated sesquilinear form is, respectively. We will write $A \geq 0$ if A is nonnegative and $A \geq B$ 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; in case of a real space a nonnegative form is required to be symmetric).

Example 2.16. For any operator A the operators A^*A and AA^* 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 2.16. *Let $\mathfrak{H}$ be a Hilbert space. Suppose $A \in \mathscr{L}(\mathfrak{H})$ satisfies $\|Af\| \geq \varepsilon\|f\|$ for some $\varepsilon > 0$. Then $\mathrm{Ran}(A)$ is closed and $A : \mathfrak{H} \to \mathrm{Ran}(A)$ is a bijection with bounded inverse, $\|A^{-1}\| \leq \frac{1}{\varepsilon}$. If we have the stronger condition $|\langle f, Af\rangle| \geq \varepsilon\|f\|^2$, then $\mathrm{Ran}(A) = \mathfrak{H}$.*

Proof. Since $Af = 0$ implies $f = 0$ our operator is injective and thus for every $g \in \mathrm{Ran}(A)$ there is a unique $f = A^{-1}g$. Moreover, by $\|A^{-1}g\| = \|f\| \leq \varepsilon^{-1}\|Af\| = \varepsilon^{-1}\|g\|$ the operator A^{-1} is bounded. So if $g_n \in \mathrm{Ran}(A)$ 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.

By $\varepsilon\|f\|^2 \leq |\langle f, Af\rangle| \leq \|f\|\|Af\|$ the second condition implies the first. To show that $\mathrm{Ran}(A) = \mathfrak{H}$ we pick $h \in \mathrm{Ran}(A)^\perp$. Then $0 = \langle h, Ah\rangle \geq \varepsilon\|h\|^2$ shows $h = 0$ and thus $\mathrm{Ran}(A)^\perp = \{0\}$. □

As a consequence we obtain the famous **Lax–Milgram theorem**[7] which plays an important role in theory of elliptic partial differential equations.

Theorem 2.17 (Lax–Milgram)**.** *Let $s : \mathfrak{H} \times \mathfrak{H} \to \mathbb{C}$ be a sesquilinear form on a Hilbert space $\mathfrak{H}$ which*

- *is bounded, $|s(f,g)| \leq C\|f\|\,\|g\|$, and*
- *satisfies $|s(f,f)| \geq \varepsilon\|f\|^2$ for some $\varepsilon > 0$.*

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

$$s(h,f) = \langle h, g\rangle, \qquad \forall h \in \mathfrak{H}. \tag{2.32}$$

Moreover, $\|f\| \leq \frac{1}{\varepsilon}\|g\|$.

[7] Peter Lax (*1926), American mathematician of Hungarian origin
[7] Arthur Milgram (1912–1961), American mathematician

Proof. Let A be the operator associated with s by Lemma 2.12. Then A is a bijection by Lemma 2.16 and $f = A^{-1}g$ has the required properties. □

Instead of the second condition one frequently requires that s is coercive, which is clearly stronger. Moreover, note that the first condition could be replaced by $|s(f,f)| \le C\|f\|^2$ (by Problem 1.33).

Note that (2.32) can also be phrased as a minimizing problem if s is nonnegative — Problem 2.28.

Example 2.17. 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 $b \in \ell^2(\mathbb{N})$. Moreover, since $(A+Q)^{-1}$ is bounded, this solution depends continuously on b. ⋄

Problem* 2.18. *Prove* (2.27). *(Hint: Use* $\|f\| = \sup_{\|g\|=1} |\langle g, f\rangle|$ *— compare Theorem 1.5.)*

Problem* 2.19. *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 2.20. *Let $\mathfrak{H}$ be a Hilbert space and $\{f_j\}_{j=1}^n \subset \mathfrak{H}$ some vectors. Show that $A : \mathbb{C}^n \to \mathfrak{H}$, $\alpha \mapsto \sum_{j=1}^n \alpha_j f_j$ is bounded and compute A^*.*

Problem 2.21. *Let $A \in \mathscr{L}(\ell^2(\mathbb{N}))$. Show that the matrix elements of the adjoint A^* are given by*

$$(A^*)_{jk} = (A_{kj})^*.$$

Problem 2.22. *Let $\mathfrak{H}_1$, $\mathfrak{H}_2$ be Hilbert spaces and 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* 2.23. *Show* (2.30).

Problem 2.24. *Let $\mathfrak{H}$ be a Hilbert space and $A \in \mathscr{L}(A)$, $g \in \mathfrak{H}$. Show that f minimizes $\|Af - g\|$ if and only if $A^*Af = A^*g$.*

Problem 2.25. *Show that under the assumptions of Problem 1.59 one has $f(A)^* = f^\#(A^*)$ where $f^\#(z) = f(z^*)^*$.*

Problem 2.26. *Show that every positive operator A has a positive square root $\sqrt{A}$. (Hint: Use that the Taylor series for $\sqrt{1+z}$ converges absolutely for $|z| \le 1$ and apply Problem 1.59.)*

Problem* 2.27. *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 1.59, 2.25.)*

Problem 2.28 (Abstract Dirichlet problem)**.** *Show that the solution of* (2.32) *is also the unique minimizer of*

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

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

Problem 2.29. *Consider $A \in \mathscr{L}(\mathfrak{H})$ and denote the set of fixed points by* $\mathrm{Fix}(A) := \mathrm{Ker}(A - \mathbb{I}) = \{f \in \mathfrak{H} | Af = f\}$. *Show* $\mathrm{Fix}(A) = \mathrm{Fix}(A^*)$ *provided* $\|A\| \le 1$.

*Use this to establish the **mean ergodic theorem** for $A \in \mathscr{L}(\mathfrak{H})$ with $\|A\| \le 1$:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n A^n f = P_{\mathrm{Fix}} f,$$

where P_{Fix} denotes the orthogonal projection onto $\mathrm{Fix}(A)$. (Hint: To compute the limit consider the cases $f \in \mathrm{Fix}(A)$ and $f \in \mathrm{Ran}(A - \mathbb{I})$. Also (2.30) *might be useful.)*

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 $\mathfrak{H}_1 \times \mathfrak{H}_2$ together with the usual vector operations and the scalar product

$$\langle (g_1, g_2), (f_1, f_2) \rangle := \langle g_1, f_1 \rangle_{\mathfrak{H}_1} + \langle g_2, f_2 \rangle_{\mathfrak{H}_2}. \tag{2.33}$$

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

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

$$\bigoplus_{j=1}^\infty \mathfrak{H}_j := \{\bigoplus_{j=1}^\infty f_j | \, f_j \in \mathfrak{H}_j, \, \sum_{j=1}^\infty \|f_j\|_{\mathfrak{H}_j}^2 < \infty\}, \tag{2.34}$$

which becomes a Hilbert space with the scalar product

$$\langle \bigoplus_{j=1}^\infty g_j, \bigoplus_{j=1}^\infty f_j \rangle := \sum_{j=1}^\infty \langle g_j, f_j \rangle_{\mathfrak{H}_j}. \tag{2.35}$$

The proof is a straightforward adaption of the completeness proof of $\ell^2(\mathbb{N})$ from Example 1.4.

Example 2.19. $\bigoplus_{j=1}^\infty \mathbb{C} = \ell^2(\mathbb{N})$. $\diamond$

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 $\mathfrak{H} \times \tilde{\mathfrak{H}}$ (here $\mathfrak{H} \times \tilde{\mathfrak{H}}$ is considered as a set without any vector operations)

$$\mathcal{F}(\mathfrak{H}, \tilde{\mathfrak{H}}) := \{\sum_{j=1}^n \alpha_j (f_j, \tilde{f}_j) | (f_j, \tilde{f}_j) \in \mathfrak{H} \times \tilde{\mathfrak{H}}, \, \alpha_j \in \mathbb{C}\}. \tag{2.36}$$

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

$$\mathcal{N}(\mathfrak{H}, \tilde{\mathfrak{H}}) := \operatorname{span}\Big\{ \sum_{j,k=1}^n \alpha_j \beta_k (f_j, \tilde{f}_k) - (\sum_{j=1}^n \alpha_j f_j, \sum_{k=1}^n \beta_k \tilde{f}_k) \Big\} \tag{2.37}$$

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

$$\langle f \otimes \tilde{f}, g \otimes \tilde{g} \rangle = \langle f, g \rangle_{\mathfrak{H}} \langle \tilde{f}, \tilde{g} \rangle_{\tilde{\mathfrak{H}}} \tag{2.38}$$

holds. To this end we set

$$s\Big(\sum_{j=1}^{n} \alpha_j (f_j, \tilde{f}_j), \sum_{k=1}^{n} \beta_k (g_k, \tilde{g}_k)\Big) = \sum_{j,k=1}^{n} \alpha_j^* \beta_k \langle f_j, g_k \rangle_{\mathfrak{H}} \langle \tilde{f}_j, \tilde{g}_k \rangle_{\tilde{\mathfrak{H}}}, \tag{2.39}$$

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

$$\Big\langle \sum_{j=1}^{n} \alpha_j \, f_j \otimes \tilde{f}_j, \sum_{k=1}^{n} \beta_k \, g_k \otimes \tilde{g}_k \Big\rangle = \sum_{j,k=1}^{n} \alpha_j^* \beta_k \langle f_j, g_k \rangle_{\mathfrak{H}} \langle \tilde{f}_j, \tilde{g}_k \rangle_{\tilde{\mathfrak{H}}} \tag{2.40}$$

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

$$f = \sum_{j,k} \alpha_{jk} u_j \otimes \tilde{u}_k, \quad \alpha_{jk} = \sum_i \alpha_i \langle u_j, f_i \rangle_{\mathfrak{H}} \langle \tilde{u}_k, \tilde{f}_i \rangle_{\tilde{\mathfrak{H}}} \tag{2.41}$$

and we compute

$$\langle f, f \rangle = \sum_{j,k} |\alpha_{jk}|^2 > 0. \tag{2.42}$$

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 2.18. *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 (2.38). 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 2.20. We have $\mathfrak{H} \otimes \mathbb{C}^n = \mathfrak{H}^n$. ⋄

Example 2.21. A quantum mechanical particle which can only attain two possible states is called a **qubit**. Its state space is accordingly $\mathbb{C}^2$ and the two states, usually written as $|0\rangle$ and $|1\rangle$, are an orthonormal basis for $\mathbb{C}^2$. The state space for two qubits is given by the tensor product $\mathbb{C}^2 \otimes \mathbb{C}^2 \cong \mathbb{C}^4$. An orthonormal basis is given by $|00\rangle := |0\rangle \otimes |0\rangle$, $|01\rangle := |0\rangle \otimes |1\rangle$, $|10\rangle :=$

$|1\rangle \otimes |0\rangle$, and $|11\rangle := |1\rangle \otimes |1\rangle$. The state space of n qubits is the n fold tensor product of $\mathbb{C}^2$ (isomorphic to $\mathbb{C}^{2^n}$). ◇

Example 2.22. 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 and we have $(\mathfrak{H}_1 \otimes \mathfrak{H}_2) \otimes \mathfrak{H}_3 \cong \mathfrak{H}_1 \otimes (\mathfrak{H}_2 \otimes \mathfrak{H}_3)$ where one identifies $(f_1 \otimes f_2) \otimes f_3$ with $f_1 \otimes (f_2 \otimes f_3)$.

We also note

$$(\bigoplus_{j=1}^\infty \mathfrak{H}_j) \otimes \mathfrak{H} = \bigoplus_{j=1}^\infty (\mathfrak{H}_j \otimes \mathfrak{H}), \tag{2.43}$$

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

$$(\sum_{j=1}^\infty f_j) \otimes f = \sum_{j=1}^\infty f_j \otimes f. \tag{2.44}$$

Problem 2.30. *Show that $f \otimes \tilde{f} = 0$ if and only if $f = 0$ or $\tilde{f} = 0$.*

Problem 2.31. *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* 2.32. *Show* (2.43).

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 $f : [-\pi,\pi] \to \mathbb{C}$ we can define its **Fourier series**

$$S(f)(x) := \frac{a_0}{2} + \sum_{k\in\mathbb{N}} \big(a_k \cos(kx) + b_k \sin(kx)\big), \tag{2.45}$$

where the corresponding Fourier coefficients are given by

$$a_k := \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(kx) f(x) dx, \qquad b_k := \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kx) f(x) dx. \tag{2.46}$$

At this point (2.45) is just a formal expression and the question in what sense the above series converges lead to the development of harmonic analysis. 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.

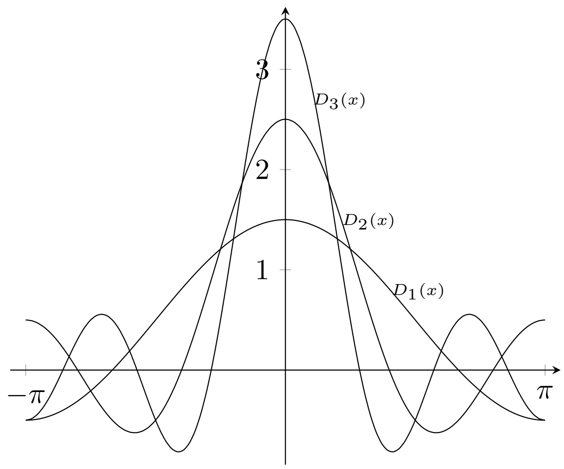


Figure 2.1. The Dirichlet kernels D_1, D_2, and D_3

For our purpose the complex form

$$S(f)(x) = \sum_{k\in\mathbb{Z}} \hat{f}_k \mathrm{e}^{\mathrm{i}kx}, \qquad \hat{f}_k := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}ky} f(y)dy \tag{2.47}$$

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

$$S_n(f)(x) := \sum_{k=-n}^{n} \hat{f}_k \mathrm{e}^{\mathrm{i}kx} = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_n(x-y) f(y)dy, \tag{2.48}$$

where

$$D_n(x) = \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}kx} = 1 + 2\sum_{k=1}^{n} \cos(kx) = \frac{\sin((n+1/2)x)}{\sin(x/2)} \tag{2.49}$$

is known as the **Dirichlet kernel**[8] (to obtain the third form observe that the first form is a geometric series) depicted in Figure 2.1. 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 $\int_{-\pi}^{\pi} D_n(x)dx = 2\pi$.

Since

$$\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}kx} \mathrm{e}^{\mathrm{i}lx} dx = 2\pi\delta_{k,l} \tag{2.50}$$

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

[8] Peter Gustav Lejeune Dirichlet (1805 –1859), German mathematician

Theorem 2.19. *For every square integrable function $f \in L^2(-\pi,\pi)$, the Fourier coefficients $\hat{f}_k$ are square summable*

$$\sum_{k\in\mathbb{Z}} |\hat{f}_k|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2 dx \tag{2.51}$$

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 2.21 below (see the discussion after this theorem). It will also follow as a special case of Theorem 3.12 below (see the examples after this theorem) as well as from the Stone–Weierstraß theorem — Problem 2.38. □

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 2.20 (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. Since a continuous function is square integrable, its Fourier coefficients are square summable by Bessel's inequality (2.4). Hence the map $f \to \hat{f}$ maps $C[-\pi,\pi] \subset L^1(-\pi,\pi)$ to $c_0(\mathbb{Z})$ (the sequences vanishing as $|k| \to \infty$). Moreover, by $|\hat{f}_k| \le (2\pi)^{-1}\|f\|_1$ this map is bounded and hence the unique extension to all of $L^1(-\pi,\pi)$ (by Theorem 1.16) still maps into $c_0(\mathbb{Z})$. □

It turns out that this result is best possible in general and we cannot say more about the decay without additional assumptions on f. For example, if f is periodic of period 2π and continuously differentiable, then integration by parts shows

$$\hat{f}_k = \frac{1}{2\pi\mathrm{i}k}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}kx} f'(x) dx. \tag{2.52}$$

Then, since both k^{-1} and the Fourier coefficients of f' are square summable, we conclude that $\hat{f}$ is absolutely summable and hence the Fourier series converges uniformly. So we have a simple sufficient criterion for summability of the Fourier coefficients, but can we do better? Of course continuity of f is a necessary condition for absolute summability but this alone will not even be enough for pointwise convergence as we will see in Example 4.7. Moreover, continuity does not provide any uniform decay estimates (see Example 4.8).

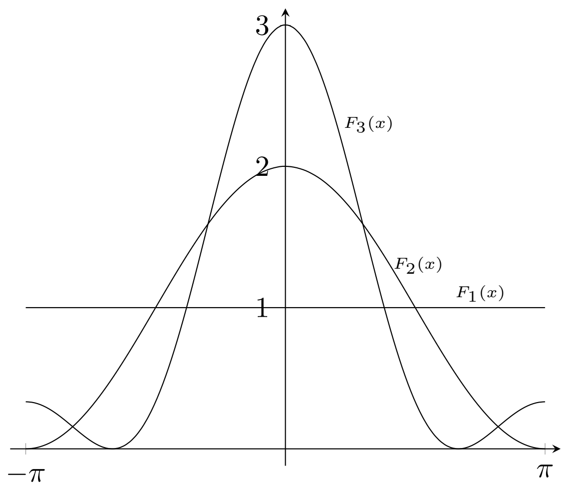


Figure 2.2. The Fejér kernels F_1, F_2, and F_3

A few improvements are easy: (2.52) holds for any class of functions for which integration by parts holds, e.g., piecewise continuously differentiable functions or, slightly more general, absolutely continuous functions (cf. Lemma 4.30 from [**40**]) provided one assumes that the derivative is square integrable. So this gives a convenient sufficient criterion. There are further criteria for absolute summability of the Fourier coefficients, but no simple necessary and sufficient one. We refer to [**40**] for further results or to classical references like [**46**] for a comprehensive treatment.

Note however, that the situation looks much brighter if one looks at mean values

$$\bar{S}_n(f)(x) := \frac{1}{n}\sum_{k=0}^{n-1} S_k(f)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} F_n(x-y)f(y)dy, \tag{2.53}$$

where

$$F_n(x) = \frac{1}{n}\sum_{k=0}^{n-1} D_k(x) = \frac{1}{n}\left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^2 \tag{2.54}$$

is the **Fejér kernel**[9] depicted in Figure 2.2. To see the second form we use the closed form for the Dirichlet kernel to obtain

$$\begin{aligned} nF_n(x) &= \sum_{k=0}^{n-1} \frac{\sin((k+1/2)x)}{\sin(x/2)} = \frac{1}{\sin(x/2)}\mathrm{Im}\sum_{k=0}^{n-1} \mathrm{e}^{\mathrm{i}(k+1/2)x} \\ &= \frac{1}{\sin(x/2)}\mathrm{Im}\left(\mathrm{e}^{\mathrm{i}x/2}\frac{\mathrm{e}^{\mathrm{i}nx}-1}{\mathrm{e}^{\mathrm{i}x}-1}\right) = \frac{1-\cos(nx)}{2\sin(x/2)^2} = \left(\frac{\sin(nx/2)}{\sin(x/2)}\right)^2. \end{aligned}$$

[9]Lipót Fejér (1880–1959), Hungarian mathematician

The main difference to the Dirichlet kernel is positivity: $F_n(x) \geq 0$. Of course the property $\int_{-\pi}^{\pi} F_n(x)dx = 2\pi$ is inherited from the Dirichlet kernel.

Theorem 2.21 (Fejér)**.** *Suppose f is continuous and periodic with period 2π. Then the mean values of the partial Fourier sums $\bar{S}_n(f)$ converge uniformly to f.*

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 F_n form an approximate identity and Lemma 1.2 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 2.37).

Note that for a given continuous function f this result shows that if $S_n(f)(x)$ converges, then it must converge to $\lim_{n\to\infty} \bar{S}_n(f)(x) = f(x)$. We also remark that one can extend this result (see Lemma 3.21 from [**40**]) to show that for $f \in L^p(-\pi,\pi)$, $1 \leq p < \infty$, 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 2.19).

Finally, we look at pointwise convergence.

Theorem 2.22. *Let $x_0 \in [-\pi,\pi]$ be fixed. Suppose*

$$\frac{f(x) - f(x_0)}{x - x_0} \tag{2.55}$$

is in $L^1(-\pi,\pi)$ as a function of x (e.g. f is Hölder continuous), then the partial Fourier sums converge pointwise at x_0,

$$\lim_{m,n\to\infty} \sum_{k=-m}^{n} \hat{f}_k \mathrm{e}^{\mathrm{i}kx_0} = f(x_0). \tag{2.56}$$

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 2.23 (Dirichlet–Dini[10] criterion)**.** *Suppose there is some α such that*

$$\frac{f(x_0+x)+f(x_0-x)-2\alpha}{x} \tag{2.57}$$

is in $L^1(-\pi,\pi)$ as a function of x. 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 2.33. *Compute the Fourier series of $f(x):=\operatorname{sign}(x)$. For which $x\in[-\pi,\pi]$ does $S_n(f)(x)$ converge to $f(x)$?*

Problem 2.34. *Compute the Fourier series of the Dirichlet kernel D_n and the Fejér kernel F_n.*

Problem 2.35. *Compute the Fourier series of $f(x):=x$ on $[-\pi,\pi]$ and use this to show (**Basel problem**)*

$$\sum_{n=1}^{\infty}\frac{1}{n^2}=\frac{\pi^2}{6}.$$

(Hint: Parseval's relation.)

Problem 2.36. *Compute the Fourier series of $f(x):=x^2$ on $[-\pi,\pi]$ and use this to solve again the **Basel problem** (see the previous problem). (Hint: Evaluate the series at $x=\pi$.)*

Problem 2.37. *Show that $C_{per}[-\pi,\pi]$ is dense in $L^p(-\pi,\pi)$ for $1\le p<\infty$.*

Problem 2.38. *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.)*

[10] Ulisse Dini (1845–1918), Italian mathematician and politician

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. Lemma B.18), A is compact if it maps bounded sets to relatively compact ones. The set of all compact operators is denoted by $\mathscr{K}(X,Y)$. If $X = Y$ we will just write $\mathscr{K}(X) := \mathscr{K}(X,X)$ 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. $\diamond$

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, $\mathscr{K}(X,Y) \subseteq \mathscr{L}(X,Y)$. Linear combinations of compact operators are compact, that is, $\mathscr{K}(X,Y)$ is a subspace of $\mathscr{L}(X,Y)$. Moreover, the product of a bounded and a compact operator is again compact, that is, $A \in \mathscr{L}(X,Y)$, $B \in \mathscr{K}(Y,Z)$ or $A \in \mathscr{K}(X,Y)$, $B \in \mathscr{L}(Y,Z)$ implies $BA \in \mathscr{K}(X,Z)$.*

In particular, the set of compact operators $\mathscr{K}(X)$ is an ideal within the set of bounded operators. Moreover, if Y is a Banach space this ideal is even closed:

Theorem 3.2. *Suppose X is a normed and Y a Banach space. Let $A_n \in \mathscr{K}(X,Y)$ 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 1.55) 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 1.13). Of course the same conclusion holds if we take Hölder continuous functions $X := C^{0,\gamma}[0,1]$ (cf. Theorem 1.23). ◇

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 1.16. Moreover, if $A \in \mathscr{K}(X,Y)$, then $A \in \mathscr{K}(X,\overline{Y})$ is immediate. That we also have $\overline{A} \in \mathscr{K}(\overline{X},\overline{Y})$ 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 $A \in \mathscr{K}(X,Y)$. Let $\overline{X}$, $\overline{Y}$ be the completion of X, Y, respectively. Then $\overline{A} \in \mathscr{K}(\overline{X},\overline{Y})$, where $\overline{A}$ is the unique extension of A (cf. Theorem 1.16).*

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 $g_n \in X$ such that $\|g_n - f_n\| \le \frac{1}{n}$ and by compactness of A we can assume that $A g_n \to g$. But then $\|\overline{A} f_n - g\| \le \|A\| \|f_n - g_n\| + \|A g_n - g\|$ 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 1.13).

Lemma 3.4. *Let $X := C([a,b])$ or $X := L^2(a,b)$. The **Fredholm integral operator**[1] $K : X \to X$ 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 $|K(y,t) - K(x,t)| \le \varepsilon$ for any $t \in [a,b]$ whenever $|y - x| \le \delta$. Moreover, $\|K\|_\infty = \sup_{x,y \in [a,b]} |K(x,y)| < \infty$.

We begin with the case $X : L^2(a,b)$. Let $g := Kf$. Then

$$|g(x)| \le \int_a^b |K(x,t)| \, |f(t)| dt \le \|K\|_\infty \int_a^b |f(t)| dt \le \|K\|_\infty \|1\| \, \|f\|,$$

where we have used Cauchy–Schwarz in the last step (note that $\|1\| = \sqrt{b-a}$). Similarly,

$$\begin{aligned} |g(x) - g(y)| &\le \int_a^b |K(y,t) - K(x,t)| \, |f(t)| dt \\ &\le \varepsilon \int_a^b |f(t)| dt \le \varepsilon \|1\| \, \|f\|, \end{aligned}$$

whenever $|y - x| \le \delta$. Hence, if $f_n(x)$ is a bounded sequence in $L^2(a,b)$, then $g_n := K f_n$ is bounded and equicontinuous and hence has a uniformly convergent subsequence by the Arzelà–Ascoli theorem (Theorem 1.13). Since the embedding $C([a,b]) \hookrightarrow L^2(a,b)$ is continuous, it follows that K is compact.

The case $X := C([a,b])$ follows by the same argument upon observing $\int_a^b |f(t)| dt \le (b-a) \|f\|_\infty$. □

Compact operators share many similarities with (finite) matrices as we will see in the next section.

Problem* 3.1. *Show Theorem 3.1.*

[1] Erik Ivar Fredholm (1866–1927), Swedish mathematician

Problem 3.2. *Is the left shift* $(a_1, a_2, a_3, \dots) \mapsto (a_2, a_3, \dots)$ *compact in* $\ell^2(\mathbb{N})$*?*

Problem 3.3. *Let* $\mathfrak{H}$ *be a Hilbert space and* $A, B \in \mathscr{L}(\mathfrak{H})$*. Show that if* B *is compact and* $0 \le A \le B$*, then also* A *is compact. (Hint: Show that* $\sqrt{A}$ *— see Problem 2.26 — is compact.)*

Problem 3.4 (Lions' lemma).[2] *Let* X*,* Y*, and* Z *be Banach spaces. Assume* X *is compactly embedded into* Y *and* Y *is continuously embedded into* Z*. Show that for every* $\varepsilon > 0$ *there exists some* $C(\varepsilon)$ *such that*

$$\|x\|_Y \le \varepsilon \|x\|_X + C(\varepsilon)\|x\|_Z.$$

Problem 3.5. *Is the operator* $\frac{d}{dx} : C^k[0,1] \to C[0,1]$ *compact for* $k = 1, 2$*? (Hint: Problem 1.46 and Example 3.3.)*

Problem 3.6. *Is the multiplication operator* $M_t : C^k[0,1] \to C[0,1]$ *with* $M_t f(t) = t f(t)$ *compact for* $k = 0, 1$*? (Hint: Problem 1.46 and Example 3.3.)*

Problem 3.7. *Let* $X := C([a,b])$ *or* $X := L^2(a,b)$*. Show that the* ***Volterra integral operator*** $K : X \to X$ *defined by*

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

where $K(x,y) \in C([a,b] \times [a,b])$*, is compact.*

Problem* 3.8. *Show that the adjoint of the Fredholm integral operator* K *on* $L^2(a,b)$ *from Lemma 3.4 is the integral operator with kernel* $K(y,x)^*$*:*

$$(K^* f)(x) = \int_a^b K(y,x)^* f(y) dy.$$

Similarly, the adjoint of the Volterra integral operator from Problem 3.7 is the integral operator:

$$(K^* f)(x) = \int_x^b K(y,x)^* f(y) dy.$$

(Hint: Fubini.)

3.2. The spectral theorem for compact self-adjoint 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

[2] Jacques-Louis Lions (1928–2001), French mathematician

subtle but important difference between symmetry and self-adjointness (see also Example 3.7 below).

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 $A - z\mathbb{I}$. The dimension of $\operatorname{Ker}(A - z)$ is also known as the (geometric) **multiplicity** of z and z is called (geometrically) **simple** if its multiplicity is one, that is, 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. $\diamond$

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 linearly 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. $\diamond$

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 a symmetric operator in an inner product space. 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 a 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**[3] 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. ◇

Example 3.7. Let $\mathfrak{H} := L^2(-\pi, \pi)$ and consider the operator

$$A := \frac{1}{\mathrm{i}} \frac{d}{dx}, \qquad \mathfrak{D}(A) := C^1[-\pi, \pi].$$

To compute its eigenvalues we need to solve the differential equation

$$-\mathrm{i}u'(x) = zu(x)$$

whose solution is given by

$$u(x) := \mathrm{e}^{\mathrm{i}zx}$$

[3] Pafnuty Chebyshev (1821–1894), Russian mathematician

and hence every $z \in \mathbb{C}$ is an eigenvalue. To investigate symmetry we use integration by parts which shows

$$\begin{aligned}\langle g, Af\rangle &= \frac{1}{\mathrm{i}}\int_{-\pi}^{\pi} g(x)^* f'(x) dx\\ &= \frac{1}{\mathrm{i}}\big(g(\pi)^* f(\pi) - g(-\pi)^* f(-\pi)\big) - \frac{1}{\mathrm{i}}\int_{-\pi}^{\pi} g'(x)^* f(x) dx\\ &= \frac{1}{\mathrm{i}}\big(g(\pi)^* f(\pi) - g(-\pi)^* f(-\pi)\big) + \langle Ag, f\rangle\end{aligned}$$

for $g, f \in \mathfrak{D}(A)$. Since the boundary terms will not vanish in general, we conclude that our operator is not symmetric. This could also be verified by observing that the eigenfunctions for (e.g.) $z = 0$ and $z = 1/2$ are not orthogonal. However, the above formula also shows that we can obtain a symmetric operator by further restricting the domain. For example, we can impose **Dirichlet boundary conditions**[4] and consider

$$A_0 := \frac{1}{\mathrm{i}}\frac{d}{dx}, \qquad \mathfrak{D}(A_0) := \{f \in C^1[-\pi,\pi] | f(-\pi) = f(\pi) = 0\}.$$

Then the above computation shows that A_0 is symmetric since the boundary terms vanish for $g, f \in \mathfrak{D}(A_0)$. Moreover, note that this domain is still dense (see Lemma 3.11 below). However, note that since the exponential function has no zeros, we lose all eigenvalues!

The reason for this unfortunate behavior is that A and A_0 are adjoint to each other in the sense that $\langle g, A_0 f\rangle = \langle Ag, f\rangle$ for $f \in \mathfrak{D}(A_0)$ and $g \in \mathfrak{D}(A)$. Hence, at least formally, the adjoint of A_0 is (at least) A and hence A_0 is symmetric but not *self-adjoint*. This gives a first hint at the fact, that symmetry is not the same as self-adjointness for unbounded operators.

Returning to our original problem, another choice are periodic boundary conditions

$$A_p := \frac{1}{\mathrm{i}}\frac{d}{dx}, \qquad \mathfrak{D}(A_p) := \{f \in C^1[-\pi,\pi] | f(-\pi) = f(\pi)\}.$$

Now we have increased the domain (in comparison to A_0) such that we are still symmetric, but such that A is no longer adjoint to A_p. Moreover, we lose some of the eigenfunctions, but not all:

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

In fact, the eigenfunctions are just the orthonormal basis from the Fourier series.

[4] Peter Gustav Lejeune Dirichlet (1805 –1859), German mathematician

Note that A_p is not the only self-adjoint extension of A_0. Another one would be antiperiodic boundary conditions $f(-\pi) = -f(\pi)$ or, even more general, $f(-\pi) = \alpha f(\pi)$ with $|\alpha| = 1$. ⋄

The previous examples show 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 compact self-adjoint operator $A \in \mathscr{K}(\mathfrak{H})$ 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 $\lim_{n\to\infty} \frac{1}{\alpha^2} A^2 u_n =: u$. Now

$$\begin{aligned}\|(A^2 - \alpha^2)u_n\|^2 &= \|A^2 u_n\|^2 - 2\alpha^2\langle u_n, A^2 u_n\rangle + \alpha^4\\ &\le 2\alpha^2(\alpha^2 - \langle u_n, A^2 u_n\rangle)\end{aligned}$$

(where we have used $\|A^2 u_n\| \le \|A\|\|Au_n\| \le \|A\|^2\|u_n\| = \alpha^2$) 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. □

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 3.9).

Now consider a compact self-adjoint 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 compact self-adjoint 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 $|\alpha_{j+1}| = \|A_j\| \le \|A_{j-1}\| = |\alpha_j|$. This procedure will not stop unless $\mathfrak{H}$ is finite dimensional. However, note that $\alpha_j = 0$ for $j > n$ might happen if $A_n = 0$.

Theorem 3.7 (Hilbert–Schmidt; Spectral theorem for compact self-adjoint operators)**.** *Suppose $\mathfrak{H}$ is a Hilbert space and $A \in \mathscr{K}(\mathfrak{H})$ is a compact symmetric operator. Then there exists a (possibly finite) sequence of real nonzero eigenvalues $(\alpha_j)_{j=1}^N$ (listed with multiplicity) which converges to 0 if $N = \infty$. The corresponding normalized eigenvectors u_j form an orthonormal set and every $f \in \mathfrak{H}$ can be written as*

$$f = \sum_{j=1}^{N} \langle u_j, f\rangle u_j + h, \tag{3.7}$$

where h is in the kernel of A, that is, $Ah = 0$. Moreover, $u_j \in \operatorname{Ker}(A)^{\perp}$.

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. If the sequence is finite, we are done.

Otherwise, since the sequence $|\alpha_j|$ is decreasing, it has a limit $\varepsilon \ge 0$ and we have $|\alpha_j| \ge \varepsilon$. If this limit is nonzero, then $v_j := \alpha_j^{-1} u_j$ is a bounded sequence ($\|v_j\| \le \frac{1}{\varepsilon}$) 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

$$\|A(f - f_n)\| \le |\alpha_{n+1}| \|f - f_n\| \le |\alpha_{n+1}| \|f\|$$

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 self-adjoint operators.

Corollary 3.8. *Every compact self-adjoint operator A can be written as*

$$Af = \sum_{j=1}^{N} \alpha_j \langle u_j, f\rangle u_j, \tag{3.8}$$

where $(\alpha_j)_{j=1}^N$ are the nonzero eigenvalues with corresponding eigenvectors u_j from the previous theorem. In particular, the range of A is separable.

Remark: Our procedure constructs all nonzero eigenvalues plus the corresponding eigenvectors. In principal we could continue once $\alpha_n = 0$ but this will still not produce an orthonormal basis of eigenfunctions in general. In fact, if there is an infinite number of nonzero eigenvalues we never reach $\alpha_n = 0$ and the entire kernel is missed. But even if 0 is reached, we might still miss some of the eigenvectors corresponding to 0 (if the kernel is not separable or if we do not choose the vectors u_j properly). 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 self-adjoint 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 3.8. Let $a, b \in c_0(\mathbb{N})$ be real-valued sequences and consider the operator

$$(Jc)_j := a_j c_{j+1} + b_j c_j + a_{j-1} c_{j-1}$$

in $\ell^2(\mathbb{N})$. If A, B denote the multiplication operators by the sequences a, b, respectively, then we already know that A and B are compact (Example 3.2). 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. Note that in this case it is not possible to get a closed expression for either the eigenvalues or the eigenvectors. $\diamond$

In particular, in the new picture it is easy to define functions of our operator (thus extending the functional calculus from Problem 1.59). To this end set $\Sigma := \{\alpha_j\}_{j\in J} \subset \mathbb{C}$ and denote by $B(\Sigma)$ the Banach algebra of bounded functions $F: \Sigma \to \mathbb{C}$ together with the sup norm. We regard $\mathbb{C} \subseteq B(\Sigma)$ by identifying a complex numbers $z \in \mathbb{C}$ with the constant function $z: \Sigma \to \mathbb{C}$, $\alpha \mapsto z$ and we write $\mathbb{I}: \Sigma \to \mathbb{C}$, $\alpha \mapsto \alpha$ for the identity function.

Corollary 3.10 (Functional calculus)**.** *Let A be a compact self-adjoint 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 Banach 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 some subset of $\mathbb{R}$ containing Σ) and, since only the values $F(\alpha_j)$ are used, two functions which agree on all eigenvalues will give the same result.

Finally, we will discuss these results within the context of 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. In finite dimension the rank-nullity theorem implies that $A-z$ is surjective if and only if it is injective and hence in finite dimensions the spectrum is precisely the set of eigenvalues.

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 1.58 one can show that if $(A - z_0)^{-1} \in \mathscr{L}(X)$, then $(A - z)^{-1} \in \mathscr{L}(X)$ for all z with $|z - z_0| < \|(A - z_0)^{-1}\|^{-1}$. Consequently the complement of the spectrum is open, and hence the spectrum is closed. Since we will discuss this in detail in Chapter 5, we will not pursue this here but only look at our special case of compact self-adjoint 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

$z \notin \overline{\Sigma}$, then $\mathrm{dist}(z,\Sigma) > 0$ and F is bounded on Σ. Hence $F \in B(\Sigma)$ and

$$R_A(z)f := \sum_{j\in J} \frac{1}{\alpha_j - z}\langle u_j, f\rangle u_j \tag{3.11}$$

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. Morever, 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 converge if and only 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.

Before we apply these results to Sturm–Liouville operators, we look at a toy problem which illustrates the main ideas.

Example 3.9. We continue Example 3.7 and would like to apply the spectral theorem to our operator A_p. However, since A_p is unbounded (its eigenvalues are not bounded), it cannot be compact and hence we cannot apply Theorem 3.7 directly to A_p. The trick is to apply it to the resolvent. Computing the resolvent boils down to solving the inhomogeneous differential equation

$$-\mathrm{i}f'(x) - z\, f(x) = g(x),$$

whose solution is (variation of constants; [**39**, Theorem 3.13])

$$f(x) = f(-\pi)\mathrm{e}^{\mathrm{i}z(x+\pi)} + \mathrm{i}\int_{-\pi}^{x} \mathrm{e}^{\mathrm{i}z(x-y)} g(y)dy$$

at least for $g \in C[0,1]$. The requirement $f \in \mathfrak{D}(A_p)$ gives

$$f(-\pi) = f(\pi) = f(-\pi)\mathrm{e}^{2\pi\mathrm{i}z} + \mathrm{i}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}z(\pi-y)} g(y)dy$$

implying

$$(A_p - z)^{-1}g(x) = \frac{\mathrm{i}\mathrm{e}^{2\pi\mathrm{i}z}}{1-\mathrm{e}^{2\pi\mathrm{i}z}}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}z(x-y)} g(y)dy + \mathrm{i}\int_{-\pi}^{x} \mathrm{e}^{\mathrm{i}z(x-y)} g(y)dy$$

for $z \notin \mathbb{Z}$ and $g \in C[0,1]$. Indeed, for every $g \in C[0,1]$ we have constructed $f \in \mathfrak{D}(A_p)$ such that $(A_p - z)f = g$ provided $z \notin \mathbb{Z}$. In particular, $(A_p - z) : \mathfrak{D}(A_p) \to C[0,1]$ is surjective in this case. Moreover, since $z \notin \mathbb{Z}$ is not an eigenvalue, $A_p - z$ is also injective and hence bijective in this case. Thus the inverse exists and is given by $(A_p - z)^{-1} g := f$ as claimed. (Alternatively we could have also checked $(A_p - z)^{-1}(A_p - z)f$ for $f \in \mathfrak{D}(A_p)$ — please remember that a right inverse might not be a left inverse in the infinite dimensional case; cf. Problem 2.14).

Note that while $(A_p - z)^{-1}$ is not defined on all of $L^2(0,1)$ it has a unique extension (Theorem 1.16) which is is compact by Problem 3.7. Moreover, that it is symmetric for $z \in \mathbb{R} \setminus \mathbb{Z}$ could be checked using Fubini, but this is not necessary since it comes for free from symmetry of A_p (cf. Problem 3.14).

Finally, observe that the eigenvalues of the resolvent are $\frac{1}{n-z}$ while the eigenvectors are the same as those of A_p. This shows that $u_n(x) := (2\pi)^{-1/2} \mathrm{e}^{inx}$ is an orthonormal basis for $L^2(-\pi,\pi)$.

It is interesting to observe, that, while the algebraic formulas remain the same, most of our claims break down if we try to replace the Hilbert space $L^2(-\pi,\pi)$ by the Banach space $C[-\pi,\pi]$: Of course we have the same eigenvalues and eigenfunctions and also the formula for the resolvent remains valid. The resolvent is still compact, but of course symmetry is not defined in this context. However, we already know that the Fourier series of a continuous function does not converge in general. This emphasizes the special role of Hilbert spaces and symmetry. $\diamond$

Problem 3.9. *Show that if $A \in \mathscr{L}(\mathfrak{H})$, then every eigenvalue α satisfies $|\alpha| \le \|A\|$.*

Problem 3.10. *Suppose $A \in \mathscr{L}(\mathfrak{H})$ is involutive, that is, $A^2 = \mathbb{I}$. Show that the only possible eigenvalues are ± 1. Show that there are projections $P_\pm$ onto the corresponding eigenspaces with $P_+ + P_- = \mathbb{I}$.*

Problem 3.11. *Find the eigenvalues and eigenfunctions of the integral operator $K \in \mathscr{L}(L^2(0,1))$ 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 3.12. *Find the eigenvalues and eigenfunctions of the integral operator $K \in \mathscr{L}(L^2(0,1))$ given by*

$$(Kf)(x) := 2\int_0^1 (2xy - x - y + 1)f(y)dy.$$

Problem 3.13. *Let $\mathfrak{H} := L^2(0,1)$. Show that the Volterra integral operator $K : \mathfrak{H} \to \mathfrak{H}$ from Problem 3.7 has no eigenvalues except for 0. Show that 0 is not an eigenvalue if $K(x,y)$ is C^1 and satisfies $K(x,x) > 0$. Why does this not contradict Theorem 3.6? (Hint: Gronwall's inequality.)*

Problem* 3.14. *Show that the inverse $(A-z)^{-1}$ (provided it exists and is densely defined) of a symmetric operator A is again symmetric for $z \in \mathbb{R}$.*

Show that a densely defined bounded operator is symmetric if and only if its extension, according to Theorem 1.16, is.

(Hint: $g \in \mathfrak{D}((A-z)^{-1})$ if and only if $g = (A-z)f$ for some $f \in \mathfrak{D}(A)$.)

Problem 3.15. *Show that a compact self-adjoint operator in an infinite-dimensional Hilbert space cannot be surjective.*

Problem 3.16. *Is every self-adjoint operator $A \in \mathscr{L}(\ell^2(\mathbb{N}))$ unitarily equivalent to a multiplication operator Q as in Example 3.5?*

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

$$\mathfrak{D}(L) := \{f \in C^2[0,1] | f(0) = f(1) = 0\} \subset L^2(0,1), \tag{3.14}$$

has an orthonormal basis of eigenfunctions. Here we have chosen the interval to be $[0,1]$ for notational simplicity. Of course it can be replaced by any compact interval. This is known as a regular Sturm–Liouville problem to distinguish it from the singular case, where either the interval is unbounded or q has a (non-integrable) singularity near one of the endpoints.

We will not need this, but we mention that the domain of L is dense:

Lemma 3.11. *The set of twice differentiable functions with compact support $C_c^2((0,1))$ is dense in $L^2(0,1)$.*

Proof. Let $P(x) = 30\int_0^x y^2(y-1)^2 dy = x^3(6x^2 - 15x + 10)$. Note that by construction $P(x)$ is monotone increasing from 0 to 1 (in particular $0 \le P(x) \le 1$ for $0 \le x \le 1$) and both $P'(x)$ as well as $P''(x)$ vanish at $x = 0, 1$. We set $P(x) = 0$ for $x \le 0$ and $P(x) = 1$ for $x \ge 1$ such that $P(x) \in C^2(\mathbb{R})$.

Next pick $f \in C[0,1]$. Since f is uniformly continuous we can find a $\delta > 0$ for every $\varepsilon > 0$ such that $|f(x) - f(y)| \le \varepsilon/2$ whenever $|x - y| \le \delta$. By decreasing δ we can assume $\delta = \frac{1}{n}$ for some integer n and $\delta \le \varepsilon$. Now set $x_j = j\delta$, $0 \le j \le n$, and define

$$\begin{aligned} f_\varepsilon(x) =& f(x_1)P(\frac{x-\delta/2}{\delta/2}) + \sum_{j=1}^{n-2}(f(x_{j+1}) - f(x_j))P(\frac{x-x_j}{\delta}) \\ & - f(x_{n-1})P(\frac{x-1+\delta}{\delta/2}). \end{aligned}$$

Then $f_\varepsilon \in C_c^2((0,1))$ and $\max_{x\in[x_1,x_{n-1}]} |f(x) - f_\varepsilon(x)| \le \varepsilon$. Hence

$$\|f - f_\varepsilon\|^2 \le 8M^2\delta + \varepsilon^2 \le \varepsilon(8M^2 + \varepsilon),$$

where $M = \|f\|_\infty$ and the claim follows. □

It is not hard to show that the same is true for $C_c^\infty((0,1))$ (Problem 3.17).

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 equation 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.

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 1.20)!

So here comes the trick (cf. Example 3.9): 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 the resolvent $R_L(z) := (L-z)^{-1}$, $z \in \mathbb{C}$.

In order to compute the resolvent, we need to solve the inhomogeneous equation $(L-z)f = g$, where it suffices to consider the case $g \in C[0,1]$. 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 ([**39**, Theorem 3.13]). 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**,[5] which is independent of x (compute its derivative!).

Of course this formula implicitly assumes $W(z) \neq 0$. This condition is not surprising since the zeros of the Wronskian are precisely the eigenvalues: In fact, $W(z)$ evaluated at $x = 0$ gives $W(z) = u_+(z,0)$ and hence shows that the Wronskian vanishes if and only if $u_+(z,x)$ satisfies both boundary conditions and is thus an eigenfunction.

Returning to (3.17) we clearly have $f(0) = 0$ since $u_-(z,0) = 0$ and similarly $f(1) = 0$ since $u_+(z,1) = 0$. Furthermore, f is differentiable and a straightforward computation verifies

$$\begin{aligned} f'(x) =& \frac{u_+'(z,x)}{W(z)} \Big(\int_0^x u_-(z,t) g(t) dt \Big) \\ &+ \frac{u_-'(z,x)}{W(z)} \Big(\int_x^1 u_+(z,t) g(t) dt \Big). \end{aligned} \tag{3.19}$$

[5] Józef Maria Hoene-Wroński (1776–1853), Polish philosopher and mathematician

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 $(L-z)f = g$. In particular, $(L-z) : \mathfrak{D}(L) \to C[0,1]$ is bijective for $W(z) \neq 0$.

Introducing the **Green function**[6]

$$
G(z,x,t) := \frac{1}{W(u_+(z),u_-(z))}\begin{cases} u_+(z,x)u_-(z,t), & x \geq t,\\ u_+(z,t)u_-(z,x), & x \leq t,\end{cases}
\tag{3.21}
$$

we see that $(L-z)^{-1}$ is given by

$$
(L-z)^{-1}g(x) = \int_0^1 G(z,x,t)g(t)dt.
\tag{3.22}
$$

Symmetry of $L-z$ for $z \in \mathbb{R}$ also implies symmetry of $R_L(z)$ for $z \in \mathbb{R}$ (Problem 3.14) but this can also be verified directly using $G(z,x,t) = G(z,t,x)$ (Problem 3.18). From Lemma 3.4 it follows that it $(L-z)^{-1}$ 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 3.12 (Steklov[7]). *The regular 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 $L^2(0,1)$, that is, every $f \in L^2(0,1)$ 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

$$
E_j = \langle u_j, Lu_j\rangle = \int_0^1 \big(|u_j'(x)|^2 + q(x)|u_j(x)|^2\big)dx \geq \min_{x\in[0,1]} q(x),
\tag{3.25}
$$

[6]George Green (1793–1841), British mathematical physicist

[7]Vladimir Andreevich Steklov (1864–1926), Soviet mathematician, mechanician and physicist.

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 (2.16), $\sum_{j=0}^{\infty} |\langle u_j, g\rangle|^2 = \|g\|^2$ 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 (2.16). $\square$

Moreover, it is even possible to weaken our assumptions for uniform convergence. To this end we consider the sequilinear form associated with L:

$$s_L(f, g) := \langle f, Lg\rangle = \int_0^1 \left(f'(x)^* g'(x) + q(x)f(x)^* g(x) \right) dx \qquad (3.26)$$

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 associated quadratic form q_L is (defined and) finite and such the partial integration needed to obtain (3.26) can be justified would be good enough (e.g. the set of absolutely continuous functions with square integrable derivatives discussed in Section 4.4 from [**40**]).

Lemma 3.13. *For a regular Sturm–Liouville problem* (3.24) *converges absolutely uniformly provided $f \in \mathfrak{Q}(L)$.*

Proof. By replacing $L \to L - E_0 + 1$ (this will shift the eigenvalues $E_n \to E_n - E_0 + 1$ and leave the eigenvectors unchanged) we can assume $q_L(f) := s_L(f,f) > 0$ and $E_j > 0$ without loss of generality.

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 from the proof of Theorem 3.12 (with $\lambda = 0$ and $\alpha_j = E_j^{-1}$) we have $u_j(x) =$

$E_j\langle u_j, G(0,x,.)\rangle$ and hence

$$\begin{aligned}\sum_{j=m}^{n} |\langle u_j, f\rangle u_j(x)| &= \sum_{j=m}^{n} E_j |\langle u_j, f\rangle\langle u_j, G(0,x,.)\rangle| \\ &\le \left(\sum_{j=m}^{n} E_j|\langle u_j, f\rangle|^2 \sum_{j=m}^{n} E_j|\langle u_j, G(0,x,.)\rangle|^2\right)^{1/2} \\ &\le \left(\sum_{j=m}^{n} E_j|\langle u_j, f\rangle|^2\right)^{1/2} q_L(G(0,x,.))^{1/2},\end{aligned}$$

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 3.14. *For z not an eigenvalue 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.66) 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 3.10. 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 $f \in L^2(0,1)$ 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 $f \in C_p^1[0,1]$ with $f(0) = f(1) = 0$ the series will converge uniformly. For an application of the trace formula see Problem 3.20. ◇

Example 3.11. 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, (Problem 3.21) every function $f \in L^2(0,1)$ 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 3.12. 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 2.19. ◇

Problem 3.17. *Show that the set of infinitely differentiable functions with compact support $C_c^\infty((0,1))$ is dense in $L^2(0,1)$. (Hint: Replace $P(x)$ in the proof of Lemma 3.11 by $\int_0^x \exp((y(y-1))^{-1})dy / \int_0^1 \exp((y(y-1))^{-1})dy$.)*

Problem* 3.18. *Show that for our Sturm–Liouville operator $u_\pm(z,x)^* = u_\pm(z^*,x)$. Conclude $R_L(z)^* = R_L(z^*)$. (Hint: Which differential equation does $u_\pm(z,x)^*$ solve? For the second part use Problem 3.8.)*

Problem 3.19. *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 3.20. *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[8] 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**[9] defined via $\frac{z}{e^z-1} = \sum_{k=0}^\infty \frac{B_k}{k!} z^k$. (Hint: Use the trace formula (3.29).)*

Problem 3.21. *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 3.12 continues to hold except for the lower bound on the eigenvalues.

Problem 3.22. *Find the norm of the integral operator*

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

in $L^2(0,1)$. (Hint: Use that K^*K is the inverse of a Sturm-Liouville operator to compute its norm and recall Lemma 2.15.)

[8]Leonhard Euler (1707–1783), Swiss mathematician, physicist, astronomer, geographer, logician and engineer

[9]Jacob Bernoulli (1655–1705), Swiss mathematician

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 in some Hilbert space 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**.[10]

Example 3.13. 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.

[10]John William Strutt, 3rd Baron Rayleigh (1842–1919), English physicist
[10]Walther Ritz (1878–1909), Swiss theoretical physicist

First of all we can improve the above bound upon observing $0 \le q(x) \le 1$ which implies

$$\langle f, L_0 f\rangle \le \langle f, Lf\rangle \le \langle f, (L_0+1)f\rangle, \qquad f \in \mathfrak{D}(L) = \mathfrak{D}(L_0),$$

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

$$\pi^2 \le E_1 \le \pi^2 + 1$$

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

$$E_1 \le \langle f_1, L f_1\rangle = \pi^2 + \frac{1}{2} \approx 10.3696.$$

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

$$E_1 \le \frac{5}{2}\pi^2 + \frac{1}{2} - \frac{\sqrt{1024+729\pi^8}}{18\pi^2} \approx 10.3685$$

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 3.15 (Courant[11] Max-min principle)**.** *Let A be a symmetric operator in some Hilbert space and let $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_N$ be eigenvalues of A with corresponding orthonormal eigenvectors $u_1, u_2, \dots, u_N$. Suppose*

$$A = \sum_{j=1}^{N} \alpha_j \langle u_j, .\rangle u_j + \tilde{A} \tag{3.33}$$

[11]Richard Courant (1888–1972), German American mathematician

with $\langle f, \tilde{A}f\rangle \geq \alpha_N \|f\|^2$ *for all* $f \in \mathfrak{D}(A) \cap \mathrm{span}\{u_1, \dots, u_N\}^\perp$ *and* $u_1, \dots, u_N \in \mathrm{Ker}(\tilde{A})$. *Then*

$$\alpha_j = \sup_{f_1,\dots,f_{j-1}} \inf_{f \in U(f_1,\dots,f_{j-1})} \langle f, Af\rangle, \qquad 1 \leq j \leq N, \tag{3.34}$$

where

$$U(f_1, \dots, f_j) := \{f \in \mathfrak{D}(A) |\, \|f\| = 1,\ f \in \mathrm{span}\{f_1, \dots, f_j\}^\perp\}. \tag{3.35}$$

Proof. We have

$$\inf_{f \in U(f_1,\dots,f_{j-1})} \langle f, Af\rangle \leq \alpha_j.$$

In fact, set $f := \sum_{k=1}^{j} \gamma_k u_k$ and choose γ_k such that $f \in U(f_1, \dots, f_{j-1})$. To this end note that the map $\mathrm{span}\{u_k\}_{k=1}^j \cong \mathbb{C}^j \to \mathbb{C}^{j-1}$, $f \mapsto (\langle u_k, f\rangle)_{k=1}^{j-1}$ must have a nontrivial kernel. Then

$$\langle f, Af\rangle = \sum_{k=1}^{j} |\gamma_k|^2 \alpha_k \leq \alpha_j$$

and the claim follows.

Conversely, let $\gamma_k = \langle u_k, f\rangle$ and write $f = \sum_{k=1}^N \gamma_k u_k + \tilde{f}$. Then

$$\inf_{f \in U(u_1,\dots,u_{j-1})} \langle f, Af\rangle = \inf_{f \in U(u_1,\dots,u_{j-1})} \left(\sum_{k=j}^{N} |\gamma_k|^2 \alpha_k + \langle \tilde{f}, \tilde{A}\tilde{f}\rangle \right) = \alpha_j. \quad \square$$

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

$$A \leq B \qquad \Leftrightarrow \qquad \langle f, Af\rangle \leq \langle f, Bf\rangle, \quad f \in \mathfrak{D}(A) \cap \mathfrak{D}(B). \tag{3.36}$$

Corollary 3.16. *Suppose A and B are symmetric operators with corresponding eigenvalues α_j and β_j as in the previous theorem. If $A \leq B$ and $\mathfrak{D}(B) \subseteq \mathfrak{D}(A)$ then $\alpha_j \leq \beta_j$.*

Proof. By assumption we have $\langle f, Af\rangle \leq \langle f, Bf\rangle$ for $f \in \mathfrak{D}(B)$ implying

$$\inf_{f \in U_A(f_1,\dots,f_{j-1})} \langle f, Af\rangle \leq \inf_{f \in U_B(f_1,\dots,f_{j-1})} \langle f, Af\rangle \leq \inf_{f \in U_B(f_1,\dots,f_{j-1})} \langle f, Bf\rangle,$$

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 3.14. Let L be again our Sturm–Liouville operator and L_0 the corresponding operator with $q(x) = 0$. Set $q_- = \min_{0 \leq x \leq 1} q(x)$ and $q_+ = \max_{0 \leq x \leq 1} q(x)$. Then $L_0 + q_- \leq L \leq L_0 + q_+$ implies

$$\pi^2 n^2 + q_- \leq E_n \leq \pi^2 n^2 + q_+.$$

In particular, we have proven the famous **Weyl asymptotic**[12]

$$E_n = \pi^2 n^2 + O(1)$$

for the eigenvalues. ◇

There is also an alternative version which can be proven similar (Problem 3.23):

Theorem 3.17 (Courant Min-max principle)**.** *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* 3.23. *Prove Theorem 3.17.*

Problem 3.24. *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 $K \in \mathscr{K}(\mathfrak{H}_1, \mathfrak{H}_2)$ is compact, then $K^*K \in \mathscr{K}(\mathfrak{H}_1)$ 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 3.18 (Schmidt; Singular value decomposition of compact operators)**.** *Let $\mathfrak{H}_1$ and $\mathfrak{H}_2$ be Hilbert spaces. Let $K \in \mathscr{K}(\mathfrak{H}_1, \mathfrak{H}_2)$ 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*

$$K = \sum_j s_j \langle u_j, .\rangle v_j, \tag{3.40}$$

[12] Hermann Weyl (1885–1955), German mathematician, theoretical physicist and philosopher

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 $f_\perp \in \mathrm{Ker}(K^*K) = \mathrm{Ker}(K)$ (Problem 3.25). 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,

$$\langle v_j, v_k\rangle = (s_j s_k)^{-1} \langle K u_j, K u_k\rangle = (s_j s_k)^{-1} \langle K^* K u_j, u_k\rangle = s_j s_k^{-1} \langle u_j, u_k\rangle$$

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 $K \in \mathscr{K}(\mathfrak{H})$ 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 $\overline{\mathrm{Ran}(K^*)} = \overline{\mathrm{span}\{u_j\}}$ onto $\overline{\mathrm{Ran}(K)} = \overline{\mathrm{span}\{v_j\}}$.

From the min-max theorem (Theorem 3.17) we obtain:

Lemma 3.19. *Let $K \in \mathscr{K}(\mathfrak{H}_1, \mathfrak{H}_2)$ be compact operator between Hilbert spaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$. Then*

$$s_j(K) = \min_{f_1,\dots,f_{j-1}} \max_{f \in U(f_1,\dots,f_{j-1})} \|Kf\|, \tag{3.45}$$

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.19).

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 3.20 (Schmidt)**.** *Let $K \in \mathscr{K}(\mathfrak{H}_1, \mathfrak{H}_2)$ be compact operator between Hilbert spaces $\mathfrak{H}_1$ and $\mathfrak{H}_2$ 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 3.21. *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 3.18. □

Corollary 3.22. *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 2.19). The rest is straightforward. □

From this last lemma one easily gets a number of useful inequalities for the singular values:

Corollary 3.23 (Weyl). *Let K_1 and K_2 be compact and let $s_j(K_1)$ and $s_j(K_2)$ be ordered. Then*

(i) $s_{j+k-1}(K_1 + K_2) \le s_j(K_1) + s_k(K_2)$,
(ii) $s_{j+k-1}(K_1 K_2) \le s_j(K_1) s_k(K_2)$,
(iii) $|s_j(K_1) - s_j(K_2)| \le \|K_1 - K_2\|$.

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 $s_{j+k-1}(K_1 + K_2) \le \|(K_1 + K_2) - (F_1 + F_2)\| = \|K_1 - F_1\| + \|K_2 - F_2\| = s_j(K_1) + s_k(K_2)$ 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 $s_{j+k-1}(K_1 K_2) \le \|K_1 K_2 - F\| = \|(K_1 - F_1)(K_2 - F_2)\| \le \|K_1 - F_1\| \|K_2 - F_2\| = s_j(K_1) s_k(K_2)$.

Next, choosing $k = 1$ and replacing $K_2 \to K_2 - K_1$ in (i) gives $s_j(K_2) \le s_j(K_1) + \|K_2 - K_1\|$. Reversing the roles gives $s_j(K_1) \le s_j(K_2) + \|K_1 - K_2\|$ and proves (iii). □

Example 3.15. On might hope that item (i) from the previous corollary can be improved to $s_j(K_1 + K_2) \le s_j(K_1) + s_j(K_2)$. 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* 3.25. *Show that* $\mathrm{Ker}(A^*A) = \mathrm{Ker}(A)$ *for any* $A \in \mathscr{L}(\mathfrak{H}_1, \mathfrak{H}_2)$.

Problem 3.26. *Compute the singular value decomposition of the rank-one operator* $K := \langle f, .\rangle g$.

Problem 3.27. *Find the spectrum of a finite rank operator in a Hilbert space.*

Problem 3.28. *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 3.29. *Let* K *be multiplication by a sequence* $k \in c_0(\mathbb{N})$ *in the Hilbert space* $\ell^2(\mathbb{N})$ *and consider* $L := KS^-$, *where* $S^\pm$ *are the shift operators. What are the singular values of* L*? Does* L *have any eigenvalues?*

Problem 3.30. *Let* $K \in \mathscr{K}(\mathfrak{H}_1, \mathfrak{H}_2)$ *be compact and let its singular values be ordered. Let* $M \subseteq \mathfrak{H}_1$, $N \subseteq \mathfrak{H}_1$ *be subspaces with 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

$$M = \mathrm{span}\{u_k\}_{k=1}^{j-1}, \qquad N = \mathrm{span}\{v_k\}_{k=1}^{j-1}.$$

3.6. Hilbert–Schmidt and trace class operators

We can further subdivide the class of compact operators $\mathscr{K}(\mathfrak{H})$ 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

$$\mathcal{J}_p(\mathfrak{H}) = \{K \in \mathscr{K}(\mathfrak{H}) | \|K\|_p < \infty\}, \tag{3.52}$$

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)

$$\|K\| \leq \|K\|_p \tag{3.53}$$

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 $p = 2$: $\mathcal{J}_2(\mathfrak{H})$ is the space of **Hilbert–Schmidt operators** and $\mathcal{J}_1(\mathfrak{H})$ is the space of **trace class** operators.

Example 3.16. 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 3.17. By virtue of the Weyl asymptotics (see Example 3.14) the resolvent of a regular Sturm–Liouville operator is trace class. ⋄

Example 3.18. 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 3.24. *A bounded operator $K \in \mathscr{L}(\mathfrak{H})$ on a Hilbert space $\mathfrak{H}$ 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 $\overline{\mathrm{span}\{v_k\}} = \mathrm{Ker}(K^*)^\perp = \overline{\mathrm{Ran}(K)}$ in the first step. □

Corollary 3.25. *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 3.26. *The set of Hilbert–Schmidt operators forms an ideal in $\mathscr{L}(\mathfrak{H})$ and*

$$\|KA\|_2 \le \|A\|\|K\|_2, \quad \textit{respectively,} \quad \|AK\|_2 \le \|A\|\|K\|_2. \tag{3.57}$$

Proof. If K_1 and K_2 are Hilbert–Schmidt operators, then so is their sum since

$$\begin{aligned}\|K_1+K_2\|_2 &= \Big(\sum_{j\in J} \|(K_1+K_2)w_j\|^2\Big)^{1/2} \le \Big(\sum_{j\in J} (\|K_1 w_j\| + \|K_2 w_j\|)^2\Big)^{1/2} \\ &\le \|K_1\|_2 + \|K_2\|_2,\end{aligned}$$

where we have used the triangle inequality for $\ell^2(J)$.

Let K be Hilbert–Schmidt and A bounded. Then AK is compact and

$$\|AK\|_2^2 = \sum_j \|AKw_j\|^2 \le \|A\|^2 \sum_j \|Kw_j\|^2 = \|A\|^2\|K\|_2^2.$$

For KA just consider adjoints. □

Example 3.19. 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**.[13]

Of course the same calculation shows that a bounded operator is Hilbert–Schmidt if and only if its matrix elements $\langle w_j, K w_k\rangle$ 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. $\diamond$

Example 3.20. 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 $\|K\|_2 \le \sqrt{b-a}M$, 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 $|(Kf)(x)| \le M\|f\|$. By the Riesz lemma there is some $K_x \in L^2(I)$ with $\|K_x\| \le M$ such that

$$(Kf)(x) = \langle K_x, f\rangle$$

and hence for any orthonormal basis $\{w_j\}_{j\in\mathbb{N}}$ we have

$$\sum_{j\in\mathbb{N}} |(Kw_j)(x)|^2 = \sum_{j\in\mathbb{N}} |\langle K_x, w_j\rangle|^2 = \|K_x\|^2 \le M^2.$$

But then

$$\begin{aligned}\sum_{j\in\mathbb{N}} \|Kw_j\|^2 &= \sum_{j\in\mathbb{N}} \int_a^b |(Kw_j)(x)|^2 dx = \int_a^b \Big(\sum_{j\in\mathbb{N}} |(Kw_j)(x)|^2\Big) dx \\ &\le (b-a)M^2\end{aligned}$$

as claimed. $\diamond$

Since Hilbert–Schmidt operators turn out easy to identify (cf. also Section 3.5 from [**40**]), it is important to relate $\mathcal{J}_1(\mathfrak{H})$ with $\mathcal{J}_2(\mathfrak{H})$:

Lemma 3.27. *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*

$$\|K\|_1 \le \|K_1\|_2 \|K_2\|_2. \tag{3.58}$$

In fact, K_1, K_2 can be chosen such that $\|K\|_1 = \|K_1\|_2\|K_2\|_2$.

[13] Ferdinand Georg Frobenius (1849 –1917), German mathematician

Proof. Using (3.40) (where we can extend u_n and v_n to orthonormal bases if necessary) and Cauchy–Schwarz we have

$$\|K\|_1 = \sum_n \langle v_n, K u_n\rangle = \sum_n |\langle K_1^* v_n, K_2 u_n\rangle| \\ \le \Big(\sum_n \|K_1^* v_n\|^2 \sum_n \|K_2 u_n\|^2\Big)^{1/2} = \|K_1\|_2 \|K_2\|_2$$

and hence $K = K_1 K_2$ 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 3.28. *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

$$\sum_n \langle w_n, K_1 K_2 w_n\rangle = \sum_n \langle K_1^* w_n, K_2 w_n\rangle = \sum_{n,m} \langle K_1^* w_n, \tilde{w}_m\rangle \langle \tilde{w}_m, K_2 w_n\rangle \\ = \sum_{m,n} \langle K_2^* \tilde{w}_m, w_n\rangle \langle w_n, K_1 \tilde{w}_m\rangle = \sum_m \langle K_2^* \tilde{w}_m, K_1 \tilde{w}_m\rangle \\ = \sum_m \langle \tilde{w}_m, K_2 K_1 \tilde{w}_m\rangle.$$

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 Lidskii[14] trace theorem (see **[28]** for an easy to read introduction).

[14] Victor Lidskii (1924–2008), Soviet mathematician

Example 3.21. 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 3.22. For our integral operator K from Example 3.18 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 3.29. *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 2.27 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 3.30. *An operator $K \in \mathscr{L}(\mathfrak{H})$ in a Hilbert space $\mathfrak{H}$ 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

$$\begin{aligned}\sum_{k=1}^N s_k &= \sum_{k=1}^N \langle v_k, K u_k \rangle = \sum_{k=1}^N \sum_j \langle v_k, g_j \rangle \langle f_j, u_k \rangle = \sum_j \sum_{k=1}^N \langle v_k, g_j \rangle \langle f_j, u_k \rangle \\ &\le \sum_j \left(\sum_{k=1}^N |\langle v_k, g_j \rangle|^2 \right)^{1/2} \left(\sum_{k=1}^N |\langle f_j, u_k \rangle|^2 \right)^{1/2} \le \sum_j \|f_j\| \|g_j\|.\end{aligned}$$

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

$$\begin{aligned}\operatorname{tr}(K) &= \sum_k \langle w_k, K w_k \rangle = \sum_k \sum_j \langle w_k, \langle f_j, w_k \rangle g_j \rangle = \sum_j \sum_k \langle \langle w_k, f_j \rangle w_k, g_j \rangle \\ &= \sum_j \langle f_j, g_j \rangle.\end{aligned}$$

□

An immediate consequence of (3.64) is:

Corollary 3.31. *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 3.31. *Show that the set of rank-one operators $\mathfrak{H}_1 \to \mathfrak{H}_2$ together with the Hilbert–Schmidt norm is a Hilbert space isomorphic to $\mathfrak{H}_1 \otimes \mathfrak{H}_2$.*

Problem 3.32. *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 3.33. *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 $\|K\|_2 \le \|c\|_1$ if $|k_j| \le c_j$, 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

Let X be a topological space. 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.

Example 4.1. The easiest example of a nowhere dense set in $\mathbb{R}$ is a discrete set. Similarly, $\{\frac{1}{n}|n \in \mathbb{N}\}$ is nowhere dense in $\mathbb{R}$. A somewhat more complex example is the Cantor set obtained by starting with $[0,1]$ and successively removing the open middle third of every remaining interval. If we denote the set in the n'th step by C_n, then the resulting set $C := \bigcap_{n\in\mathbb{N}} C_n$ is closed. Moreover, since C_n consists of 2^n intervals of length 3^{-n}, the Cantor set contains no interval and hence has empty interior. In other words, it is nowhere dense.

Moreover, any closed set of Lebesgue measure zero is nowhere dense. The converse is however not true. To see this change the above construction by removing open subintervals of total length $1/4^{n+1}$ from each of the 2^n intervals of C_n. Then C_n consists of 2^n intervals of length $(2^n+1)/(2\cdot 4^n)$ and the Lebesgue measure of C is $\frac{1}{2}$ but it still contains no intervals. This last set is known as fat Cantor set.

In a Banach space, every subspace is either dense or nowhere dense. In fact, every closed subspace containing a ball, also contains a ball centered at zero (by translation) and hence equals the entire space by scaling.

In an infinite dimensional Banach space a compact subset is nowhere dense. This follows since (nontrivial) closed balls are compact if and only if the space is finite dimensional (see Theorem 4.31 below). ⋄

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. This was first proven for $\mathbb{R}$ by Osgood[1] and independently for $\mathbb{R}^n$ by Baire[2] and in its general form by Hausdorff[3].

Theorem 4.1 (Bair 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; in particular, $X \setminus X_n$ 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 axiom 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 $x \in \overline{B_{r_n}(x_n)} \subseteq X \setminus X_n$ implies $x \notin X_n$ 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 a complete metric space X, at least one X_n contains a ball of radius $\varepsilon > 0$.

Example 4.2. 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** (also meagre). All

[1]William Fogg Osgood (1864–1943), American mathematician
[2]René-Louis Baire (1874–1932), French mathematician
[3]Felix Hausdorff (1868–1942), German mathematician

other sets are **second category** or **fat**. Hence explaining the name category theorem. Be aware that this distinction is typically only useful for a complete metric space: In $\mathbb{Q}$ every set is meager.

Moreover, it might be worth while emphasizing that the complement of a fat set is not meager in general (e.g. in $\mathbb{R}$ every nonempty interval is fat). The complement of a meager set is called **comeager** or **residual**.

Next observe that the following topological concepts are dual in the sense that, if a set has one property, the complement has the other (cf. Problem B.12):

closed	open
dense	empty interior
dense interior	nowhere dense

In particular, a residual set is the countable intersection of sets whose interior is dense.

Using this correspondence and the definition we get the following (Problem 4.2):

- Any subset of a meager set is meager and the countable union of meager sets is meager (in other words, the meager sets form a **σ-ideal**).
- Any superset of a residual set is residual and the countable intersection of residual sets is residual.

In this respect another reformulation of the category theorem is also worthwhile noting:

Corollary 4.2. *Let X be a complete metric space.*

(i) *The countable intersection of open dense sets is dense.*

(ii) *A residual set is dense*

(iii) *The countable union of closed nowhere dense sets has empty interior.*

(iv) *A meager set has empty interior.*

Proof. (i). 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 $X \setminus \bigcap_n O_n = \bigcup_n X_n$, 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 $\overline{B_\varepsilon}$ (show this). But $\overline{B_\varepsilon}$ is a complete metric space, a contradiction.

(ii). If R is residual it can be written as $R = \bigcap_{j\in\mathbb{N}} R_j$ where R_j has dense interior. Hence $R \supseteq \bigcap_j R_j^\circ$ and the claim follows from (i).

(iii) and (iv) follow from (i) and (ii) by taking complements, respectively. □

Remark: Since (ii) is a special case of (i), the proof shows that the above items are equivalent in a nonempty topological space. Moreover, it is also evident that (iv) implies the Bair category theorem. Consequently, a topological space satisfying one of these conditions is called a **Bair space**.

Countable intersections of open sets are in some sense the next general sets after open sets 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). A dense G_δ is residual and properties which hold on a residual set are considered *generic* in this context. Note that, as with residual sets, the countable intersection of dense G_δ sets is a dense G_δ set.

Example 4.3. The irrational numbers are a dense G_δ set in $\mathbb{R}$. To see this, let $\{x_n\}_{n\in\mathbb{N}}$ 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. ◇

Example 4.4. Note that a dense G_δ can still be small in some other sense. For example, it can have (Lebesgue) measure zero.

To this end recall that a subset $N \subset \mathbb{R}$ is called a null set (or a set of Lebesgue measure zero) if for every $\varepsilon > 0$ it can be covered by a countable number of intervals whose total length is at most ε.

Now let $\{x_n\}_{n\in\mathbb{N}}$ be an enumeration of the rationals and let $I_{n,m} := (x_n - 2^{-n-m-1}, x_n + 2^{-n-m-1})$. Then $O_m := \bigcup_{n\in\mathbb{N}} I_{n,m}$ is open and dense (since $\mathbb{Q} \subset O_m$). Hence $G := \bigcap_{m\in\mathbb{N}} O_m$ is a dense G_δ. Moreover, since the total length of all intervals in O_m is 2^{-m}, it is also a null set.

In particular, since the complement of a dense G_δ is meager, $\mathbb{R}$ can be written as the disjoint union of a null set and a meager set. ◇

Now we are ready for the first important consequence:

Theorem 4.3 (Banach–Steinhaus[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. 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 set.*

[4]Hugo Steinhaus (1887–1972), Polish mathematician

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. □

Warning: There is also a variant, sometimes also called Banach–Steinhaus theorem, for pointwise limits of bounded operators which will be discussed in Lemma 4.34 (iii).

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. If $\sup_\alpha \|A_\alpha x\| \le C_x$ is uniformly bounded with respect to α for every fixed $x \in X$, then $\{A_\alpha\}$ is uniformly bounded, $\|A_\alpha\| \le C$.*

Example 4.5. A subset $U \subset X$ of a normed space is called strongly bounded if $\sup_{x\in U} \|x\| < \infty$ and weakly bounded if $\sup_{x\in U} |\ell(x)| < \infty$ for all $\ell \in X^*$. Clearly a strongly bounded subset is weakly bounded. To see the converse we consider the linear map $J : X \to X^{**}$ defined by $J(x)(\ell) := \ell(x)$ (i.e., $J(x)$ is evaluation at x). Then by assumption we have $\sup_{x\in U} |J(x)(\ell)| = \sup_{x\in U} |\ell(x)| < \infty$ for all $\ell \in X^*$ and hence the uniform boundedness principle implies that the family of functionals $J(U) \subset X^{**}$ is uniformly bounded, $\sup_{x\in U} \|J(x)\| < \infty$. In Theorem 4.18 we will show $\|J(x)\| = \|x\|$ and consequently U is strongly bounded, $\sup_{x\in U} \|x\| < \infty$. ⋄

Conversely, if there is no uniform bound, the pointwise bound must fail on a dense G_δ. This is illustrated in the following examples.

Example 4.6. The set of functions in $C[0,1]$ which fail to be differentiable at a given fixed point $x_0 \in [0,1]$ is a dense G_δ. In fact, consider the linear functionals $\ell_n(f) := n(f(x_0 + 1/n) - f(x_0))$ (for $x_0 < 1$; for $x_0 = 1$ replace $+1/n$ by $-1/n$). Since we have $\|\ell_n\| = 2n$ the Banach–Steinhaus theorem implies that the set $\{f \in C[0,1] | \sup_n |\ell_n(f)| = \infty\}$ is a dense G_δ. But functions in this set are not differentiable at x_0, since $\ell_n(f) \to f'(x_0)$ in case f is differentiable at x_0.

Moreover, since a countable intersection of dense G_δ sets is still a dense G_δ, the set of functions which fail to be differentiable at all rational points is also a dense G_δ. See Problem 4.8 for an even stronger statement. ⋄

Example 4.7. Consider the Fourier series (2.47) 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 1.24 (adapted to our present periodic setting) we have

$$\|\ell_n\| = \frac{1}{2\pi} \|D_n\|_1.$$

Now we estimate

$$\begin{aligned}\|D_n\|_1 &= 2\int_0^\pi |D_n(x)| dx \ge 2\int_0^\pi \frac{|\sin((n+1/2)x)|}{x/2} dx\\ &= 4\int_0^{(n+1/2)\pi} |\sin(y)| \frac{dy}{y} \ge 4\sum_{k=1}^n \int_{(k-1)\pi}^{k\pi} |\sin(y)| \frac{dy}{k\pi} = \frac{8}{\pi}\sum_{k=1}^n \frac{1}{k}\end{aligned}$$

and note that the harmonic series diverges.

As in the previous example, there is a dense G_δ of functions whose Fourier series diverges on a given countable set of points. ⋄

Example 4.8. Recall (cf. (2.52)) that the Fourier coefficients of an absolutely continuous function satisfy the estimate

$$|\hat{f}_k| \le \frac{1}{2\pi} \begin{cases} \|f\|_\infty, & k = 0,\\ \frac{\|f'\|_\infty}{|k|}, & k \ne 0.\end{cases}$$

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

$$|\hat{f}_k| \le C_f c_k,$$

where C_f is some constant depending on f. Now consider the linear functionals

$$\ell_k(f) := \frac{\hat{f}_k}{c_k}, \qquad k \in \mathbb{Z},$$

on $C_{per}[-\pi,\pi]$ (continuous periodic functions with sup norm). If the above estimate holds for all $f \in C_{per}[-\pi,\pi]$, then this family satisfies the assumptions of the uniform boundedness principle implying $\|\ell_k\| \le C$. In other words, we must have an estimate of the type

$$|\hat{f}_k| \le C\, c_k \|f\|_\infty,$$

which implies $\sqrt{2\pi} \le C\, c_k$ 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. $\diamond$

The next application is

Theorem 4.5 (Open mapping)**.** *Let $A \in \mathscr{L}(X,Y)$ be a continuous linear operator between Banach spaces X and Y. 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 $B_r^Y := B_r^Y(0)$. Let $O \subseteq X$ be open. To prove that $A(O) \subseteq Y$ is open it suffices to show that for every $x \in O$ and every $\varepsilon > 0$ there is a δ such that $B_\delta^Y(Ax) \subseteq A(B_\varepsilon^X(x))$. After a translation (using linearity of A), we can assume $x = 0$ without loss of generality. (By scaling we could also assume $\varepsilon = 1$ without loss of generality.)

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, $\overline{nA(B_\varepsilon^X)}$ 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 $z \in B^Y_{\delta_1} \subset \overline{A(B^X_{\varepsilon_1})}$ we have $x_1 \in B^X_{\varepsilon_1}$ such that Ax_1 is arbitrarily close to z, say $z - Ax_1 \in B^Y_{\delta_2} \subset \overline{A(B^X_{\varepsilon_2})}$. Hence we can find $x_2 \in B^X_{\varepsilon_2}$ such that $(z - Ax_1) - Ax_2 \in B^Y_{\delta_3} \subset \overline{A(B^X_{\varepsilon_3})}$ and proceeding like this a sequence $x_n \in B^X_{\varepsilon_n}$ such that

$$z - \sum_{k=1}^n Ax_k \in B^Y_{\delta_{n+1}}.$$

By construction the limit $x := \sum_{k=1}^\infty x_k$ exists and satisfies $x \in B^X_\varepsilon$ as well as $z = Ax \in A(B^X_\varepsilon)$. That is, $B^Y_{\delta_1} \subseteq A(B^X_\varepsilon)$ as desired.

Conversely, if A is open, then the image of the unit ball contains again some ball $B^Y_\varepsilon \subseteq A(B^X_1)$. Hence by scaling $B^Y_{r\varepsilon} \subseteq A(B^X_r)$ and letting $r \to \infty$ we see that A is onto: $Y = A(X)$. □

Example 4.9. Let X be a Banach space and M a closed subspace. Then the quotient map $\pi : X \to X/M$ is open. ⋄

Example 4.10. 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 $x_{2n-1} = 1$). ⋄

As a by-product of the proof we record two consequences:

Corollary 4.6. *For a continuous linear operator $A \in \mathscr{L}(X,Y)$ between Banach spaces the following are equivalent:*

(i) *A is open.*
(ii) *$B^Y_\delta(0) \subseteq AB^X_1(0)$ for some $\delta > 0$.*
(iii) *$B^Y_\delta(0) \subseteq \overline{AB^X_1(0)}$ for some $\delta > 0$.*

Corollary 4.7. *Let $A \in \mathscr{L}(X,Y)$ be a continuous linear operator between Banach spaces. Then either* $\operatorname{Ran}(A)$ *is meager or* $\operatorname{Ran}(A) = Y$.

Proof. As shown in the proof, if A is not onto, none of the sets $\overline{AB^X_n}$ will contain a ball and hence the sets AB^X_n are nowhere dense. Consequently, $\operatorname{Ran}(A) = \bigcup_n AB^X_n$ is meager in this case. □

Example 4.11. For example, $\ell^{p_0}(\mathbb{N})$ is meager as a subset of $\ell^p(\mathbb{N})$ for $p_0 < p$ (which follows from applying the above corollary to the natural embedding operator — Problem 1.20). ⋄

As another immediate consequence (cf. Lemma B.10) we get the inverse mapping theorem:

Theorem 4.8 (Inverse mapping)**.** *Let $A \in \mathscr{L}(X,Y)$ be a continuous linear bijection between Banach spaces. Then A^{-1} is continuous.*

Example 4.12. Consider the operator $(Aa)_{j=1}^\infty := (\frac{1}{j} a_j)_{j=1}^\infty$ in $\ell^2(\mathbb{N})$. Then its inverse $(A^{-1}a)_{j=1}^\infty = (j\, a_j)_{j=1}^\infty$ 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 8.2 below.

Note that if one considers A on the subspace of sequences with finitely many nonzero terms, then A would be surjective. This shows that the open mapping theorem and the inverse mapping theorem fail if the spaces are not complete. ⋄

Example 4.13. Consider the Fourier series (2.47) of an integrable function. Using the inverse mapping 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 1.16. Moreover, it can be shown that F is injective (see the discussion after Theorem 2.21). 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 4.7. ⋄

Another important consequence is the closed graph theorem. The **graph** of an operator $A : \mathfrak{D}(A) \subseteq X \to Y$ between Banach spaces is

$$\Gamma(A) := \{(x, Ax) | x \in \mathfrak{D}(A)\}. \tag{4.1}$$

If A is linear, the graph is a subspace of the Banach space $X \oplus Y$, 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 $X \oplus Y$. Explicitly this says that if $(x_n, Ax_n) \in \Gamma(A)$ converges to $(x, y) \in X \oplus Y$, we must have $(x, y) \in \Gamma(A)$, that is, $x \in \mathfrak{D}(A)$ and $y = Ax$. In the case of a bounded operator one gets convergence of Ax_n for free from convergence of x_n and hence the graph will be closed if and only if the domain $\mathfrak{D}(A)$ is closed. In the case of an unbounded operator closedness of the graph will not guarantee convergence of Ax_n but it will ensure that if this sequence converges, then it will converge to the right object, namely Ax.

Example 4.14. Let $X := C[0,1]$ and consider the unbounded operator (cf. Example 1.20)

$$\mathfrak{D}(A) := C^1[0,1], \qquad Af := f'.$$

Then $\Gamma(A)$ is closed since $f_n \to f$ and $f_n' \to g$ implies that f is differentiable and $f' = g$. ⋄

Theorem 4.9 (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 4.10 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. These operators play an important role and we will have a closer look at them in Section 8.1. For now we only point out that the closed graph theorem tells us that a closed linear operator can be defined on all of X if and only if it is 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 4.10 (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. *The finite union of nowhere dense sets is nowhere dense and the finite intersections of sets with dense interior has dense interior.*

Problem* 4.2. *Let X be a topological space. Prove that the meager sets form a σ-ideal. Moreover, prove that every superset of a fat set is fat.*

Problem 4.3. *Consider $X := C[-1,1]$. Show that $M := \{x \in X | x(-t) = x(t)\}$ is meager.*

Problem 4.4. *Let X be a complete metric space without isolated points. Show that a residual set cannot be countable. (Hint: A single point is nowhere dense.)*

Problem 4.5. *An infinite dimensional Banach space cannot have a countable Hamel basis (see Problem 1.10). (Hint: Apply Baire's theorem to $X_n := \operatorname{span}\{u_j\}_{j=1}^n$.)*

Problem 4.6. *Consider $f = \chi_{\mathbb{Q}}$ on $[0,1]$. Show that there cannot be a sequence of continuous functions $f_n \in C[0,1]$ which converges to f pointwise. (Hint: Apply Baire's theorem with $F_n := \{t \in [0,1] | \, |f_n(t) - f_m(t)| \le \frac{1}{3} \forall m \ge n\}$.)*

Problem 4.7. *Let $f \in C(0,\infty)$. Show that $\lim_{n\to\infty} f(n\,t) = 0$ for all $t > 0$ implies $\lim_{t\to\infty} f(t) = 0$. (Hint: By Baire's theorem the sets $F_n := \{t > 0 | \, |f(n\,t)| \le \varepsilon \, \forall m \ge n\}$ eventually must contain an interval.)*

Problem 4.8. *Let $X := C[0,1]$. Show that the set of functions which are nowhere differentiable contains a dense G_δ. (Hint: Consider $F_k := \{f \in X | \, \exists x \in [0,1] : \; |f(x) - f(y)| \le k|x-y|, \, \forall y \in [0,1]\}$. 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 P_m of piecewise linear functions whose slopes are bounded below by m in absolute value are dense. Now observe that $F_k \cap P_m = \emptyset$ for $m > k$.)*

Problem 4.9. *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 4.10. *Let X be a Banach space and Y, Z normed spaces. 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 4.11. *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 $\alpha = (\alpha_n)_{n=1}^N$ by $\mathcal{A}$ and equip it with the norm $\|\alpha\| := \sup_n \|\sum_{k=1}^n \alpha_k u_k\|$. 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 4.12. *Show that a relatively compact set in an infinite dimensional Banach space is meager. Conclude that a compact operator $A \in \mathscr{K}(X,Y)$ between Banach spaces cannot be surjective if Y is infinite dimensional. (Hint: Use Theorem 4.31 below.)*

Problem 4.13. *Show that the operator*

$$\mathfrak{D}(A) := \{a \in \ell^p(\mathbb{N}) | j\, a_j \in \ell^p(\mathbb{N})\}, \qquad (Aa)_j := j\, a_j,$$

is a closed operator in $X := \ell^p(\mathbb{N})$.

Problem 4.14. *Suppose $A : X \to Y$ and $B : Y \to Z$ are linear operators between Banach spaces X, Y, Z. Show that A is bounded if BA and B are bounded with B injective.*

4.2. The Hahn–Banach theorem and its consequences

Let X be a normed 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 1.17).

Example 4.15. Consider the Banach space $\ell^p(\mathbb{N})$, $1 \le p < \infty$. Taking the Kronecker deltas δ^j as a Schauder basis the j'th term a_j of a sequence $a \in \ell^p(\mathbb{N})$ can also be considered as the j'th coordinate of a with respect to this basis. Moreover, the map $l_j(a) := a_j$ is a bounded linear functional, that is, $l_j \in \ell^p(\mathbb{N})^*$, since $|l_j(a)| = |a_j| \le \|a\|_p$. It is a special case of the following more general example below (in fact, we have $l_j = l_{\delta^j}$). Since the coordinates of a vector carry all the information this explains why understanding linear functionals is of key importance. ◇

Example 4.16. Consider the Banach space $\ell^p(\mathbb{N})$, $1 \le p \le \infty$ and let q be the corresponding dual index satisfying $\frac{1}{p} + \frac{1}{q} = 1$. We have already seen in Example 1.22 that by Hölder's inequality (1.24) every $b \in \ell^q(\mathbb{N})$ gives rise to a bounded linear functional

$$l_b(a) := \sum_{j\in\mathbb{N}} b_j a_j \tag{4.3}$$

whose norm is $\|l_b\| = \|b\|_q$. 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

$$b_j := l(\delta^j).$$

Then

$$|b_j| = |l(\delta^j)| \le \|l\| \, \|\delta^j\|_1 = \|l\|$$

shows $\|b\|_\infty \le \|l\|$, that is, $b \in \ell^\infty(\mathbb{N})$. By construction $l(a) = l_b(a)$ for every $a \in \operatorname{span}\{\delta^j\}_{j\in\mathbb{N}}$. By continuity of l it even holds for $a \in \overline{\operatorname{span}\{\delta^j\}_{j\in\mathbb{N}}} = \ell^1(\mathbb{N})$. Hence the map $b \mapsto l_b$ 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 4.11 (Hahn–Banach[5], 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

$$\tilde{\ell}(y + \alpha x) := \ell(y) + \alpha\tilde{\ell}(x), \qquad y \in Y.$$

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

$$\sup_{\alpha>0, y\in Y} \frac{\varphi(y - \alpha x) - \ell(y)}{-\alpha} \le \tilde{\ell}(x) \le \inf_{\alpha>0, y\in Y} \frac{\varphi(y + \alpha x) - \ell(y)}{\alpha}$$

[5] Hans Hahn (1879–1934), Austrian mathematician

and is hence only possible if and only 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 (Theorem A.2): Let E be the collection of all extensions $\tilde{\ell}$ satisfying $\tilde{\ell}(x)\le\varphi(x)$. Then E can be partially ordered by inclusion (with respect to the domain, i.e., $\tilde{\ell}_1\subseteq\tilde{\ell}_2$ if $\mathfrak{D}(\tilde{\ell}_1)\subseteq\mathfrak{D}(\tilde{\ell}_2)$ and $\tilde{\ell}_2|_{\mathfrak{D}(\tilde{\ell}_1)}=\tilde{\ell}_1$) 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 $-\varphi(-x)\le\overline{\ell}(x)$, $x\in X$, for free. In particular, if $\varphi(x)=\varphi(-x)$ then $|\overline{\ell}(x)|\le\varphi(x)$.

Theorem 4.12 (Hahn–Banach, complex version)**.** *Let X be a complex vector space and $\varphi: X\to\mathbb{R}$ a convex function satisfying $\varphi(\alpha x)\le\varphi(x)$ if $|\alpha|=1$.*

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. 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 $\overline{\ell}_r(x)\le\varphi(x)$. 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 $|\overline{\ell}(x)|\le\varphi(x)$ we abbreviate $\alpha:=\frac{\overline{\ell}(x)^*}{|\overline{\ell}(x)|}$ and use

$$|\overline{\ell}(x)|=\alpha\overline{\ell}(x)=\overline{\ell}(\alpha x)=\overline{\ell}_r(\alpha x)\le\varphi(\alpha x)\le\varphi(x),$$

which finishes the proof. □

Note that $\varphi(\alpha x) \le \varphi(x)$, $|\alpha| = 1$ 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 4.13. *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 4.17. Note that in a Hilbert space this result is trivial: First of all there is a unique extension to $\overline{Y}$ by Theorem 1.16. 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 4.18. In a Banach space this extension will in general not be unique: Consider $X := \ell^1(\mathbb{N})$ and $\ell(a) := a_1$ on $Y := \mathrm{span}\{\delta^1\}$. Then by Example 4.16 any extension is of the form $\overline{\ell} = l_b$ with $b \in \ell^\infty(\mathbb{N})$ and $b_1 = 1$, $\|b\|_\infty \le 1$. (Sometimes it still might be unique: Problems 4.15 and 4.17). ⋄

Moreover, we can now easily prove our anticipated result

Corollary 4.14. *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 $\mathrm{span}\{x\}$ by setting $\ell(\alpha x) = \alpha$ and extending it to X^* using the previous corollary. But this contradicts our assumption. □

Example 4.19. 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

$$l(a) := \lim_{j\to\infty} a_j, \qquad a \in c(\mathbb{N}), \tag{4.4}$$

then l is bounded since

$$|l(a)| = \lim_{j\to\infty} |a_j| \le \|a\|_\infty. \tag{4.5}$$

Hence we can extend it to $\ell^\infty(\mathbb{N})$ by Corollary 4.13. Then l cannot be written as $l = l_b$ for some $b \in \ell^1(\mathbb{N})$ (as in (4.3)) since $b_j = l(\delta^j) = 0$ shows $b = 0$ and hence $l_b = 0$. The problem is that $\overline{\mathrm{span}\{\delta^j\}_{j\in\mathbb{N}}} = c_0(\mathbb{N}) \neq \ell^\infty(\mathbb{N})$, 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 4.19 (iii) below. ⋄

Another useful consequence is

Corollary 4.15 (Mazur[6]). *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 $\operatorname{dist}(x_0, Y) \le \|x_0 - \frac{-y}{\alpha}\|$ for every $y \in Y$ we have

$$|\ell(y + \alpha x_0)| = |\alpha| \operatorname{dist}(x_0, Y) \le \|y + \alpha x_0\|, \quad y \in Y.$$

Hence $\|\ell\| \le 1$ and there is an extension to X by Corollary 4.13. To see that the norm is in fact equal to one, take a sequence $y_n \in Y$ such that $\operatorname{dist}(x_0, Y) \ge (1 - \frac{1}{n})\|x_0 + y_n\|$. Then

$$|\ell(y_n + x_0)| = \operatorname{dist}(x_0, Y) \ge (1 - \frac{1}{n})\|y_n + x_0\|$$

establishing (iii). □

Two more straightforward consequences of the last corollary are also worthwhile noting:

Corollary 4.16. *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 4.17. *Let $Y \subset X$ be a closed subspace and let $\{x_j\}_{j=1}^n$ be a linearly independent subset of a normed vector space 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 $j \ne k$, $\ell_j(x_j) = 1$ and $\ell(y) = 0$ for $y \in Y$.*

Proof. Fix j_0. Since $Y_{j_0} = Y \dot{+} \operatorname{span}\{x_j\}_{1 \le j \le n; j \ne j_0}$ is closed (Corollary 1.21), $x_{j_0} \notin Y_{j_0}$ implies $\operatorname{dist}(x_{j_0}, Y_{j_0}) > 0$. Hence we can apply Corollary 4.15 and rescale the obtained functional such that $\ell_{j_0}(x_j) = 1$. □

Problem 4.15. *Let $X := \mathbb{C}^3$ equipped with the norm $|(x, y, z)|_1 := |x| + |y| + |z|$ and $Y := \{(x, y, z) \in X | x + y = 0,\ z = 0\}$. Find at least two extensions of $\ell(x, y, z) := x$ from Y to X which preserve the norm. What if we take $Y := \{(x, y, z) \in X | x + y = 0\}$?*

Problem 4.16. *Consider $X := C[0, 1]$ and let $f_0(x) := 1 - 2x$. Find at least two linear functionals with norm one such that $\ell(f_0) = 1$.*

---

[6]Stanisław Mazur (1905–1981), Polish mathematician

Problem 4.17. *Show that the extension from Corollary 4.13 is unique if X^* is strictly convex. (Hint: Problem 1.19.)*

Problem* 4.18. *Let X be some normed space. Show that*

$$\|x\| = \sup_{\ell \in V,\, \|\ell\|=1} |\ell(x)|,$$

where $V \subset X^*$ is some dense subspace. Show that equality is attained if $V = X^*$.

Problem 4.19. *Let X, Y be some normed spaces and $A : \mathfrak{D}(A) \subseteq X \to Y$. Show*

$$\|A\| = \sup_{x \in \mathfrak{D}(A),\, \|x\|=1;\, \ell \in V,\, \|\ell\|=1} |\ell(Ax)|,$$

where $V \subset Y^*$ is a dense subspace.

Problem* 4.20. *Show that every $l \in \ell^p(\mathbb{N})^*$, $1 \le p < \infty$, can be written as*

$$l(a) = l_b(a) := \sum_{j\in\mathbb{N}} b_j a_j$$

with some unique $b \in \ell^q(\mathbb{N})$. (Hint: To see $b \in \ell^q(\mathbb{N})$ consider a^N defined such that $a^n_j := |b_j|^q / b_j$ for $j \le n$ with $b_j \ne 0$ and $b^n_j := 0$ else. Now look at $|l(a^n)| \le \|l\| \|a^n\|_p$.)

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*

$$l(a) = \sum_{j\in\mathbb{N}} b_j a_j$$

with some unique $b \in \ell^1(\mathbb{N})$ which satisfies $\|b\|_1 = \|\ell\|$.

(iii) *Show that every $l \in c(\mathbb{N})^*$ can be written as*

$$l(a) = \sum_{j\in\mathbb{N}} b_j a_j + b_0 \lim_{j\to\infty} a_j$$

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 (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 4.24. *Let X be a Banach space. Show that a subset $U \subseteq X$ is bounded if and only if $\ell(U) \subseteq \mathbb{C}$ is bounded for every $\ell \in X^*$. (Hint: Uniform boundedness principle.)*

Problem 4.25. *Show that there exists a nontrivial bounded linear functional in $\ell^\infty(\mathbb{N})$ which vanishes on all periodic sequences.*

Problem* 4.26. *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^*.)*

Problem* 4.27. *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 $x - \sum_{j=1}^n \ell_j(x) x_j$.)*

4.3. Reflexivity

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 4.18. *Let X be a normed space. Then $J : X \to X^{**}$ is isometric (norm preserving).*

Proof. Fix $x_0 \in X$. By $|J(x_0)(\ell)| = |\ell(x_0)| \le \|\ell\|_* \|x_0\|$ we have at least $\|J(x_0)\|_{**} \le \|x_0\|$. 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 4.20. 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 1.17). ⋄

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. Of course a reflexive space must necessarily be complete.

Warning: The condition $J(X) = X^{**}$ is stronger than the condition $X \cong X^{**}$. James[7] gave an example (known as **James' space**) of a non-reflexive Banach space which is isometric to its bidual space.

Example 4.21. Every finite dimensional space is reflexive since $\dim(X) = \dim(X^*) = \dim(X^{**})$ (cf. Example 1.21). ⋄

Example 4.22. The Banach spaces $\ell^p(\mathbb{N})$ with $1 < p < \infty$ are reflexive: Identify $\ell^q(\mathbb{N})$ with $\ell^p(\mathbb{N})^*$ by virtue of $b \sim l_b$ as in Problem 4.20 and choose $c \in \ell^p(\mathbb{N})^{**}$. Then there is some $a \in \ell^p(\mathbb{N})$ such that

$$c(l_b) = \sum_{j\in\mathbb{N}} b_j a_j, \qquad l_b \in \ell^p(\mathbb{N})^* \cong \ell^q(\mathbb{N}).$$

But this implies $c(l_b) = l_b(a)$, that is, $c = J(a)$, and thus J is surjective. (Warning: It does not suffice to just argue $\ell^p(\mathbb{N})^{**} \cong \ell^q(\mathbb{N})^* \cong \ell^p(\mathbb{N})$ as already pointed out above.)

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 in Example 4.19. 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 $J(c_0(\mathbb{N})) = c_0(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$. ⋄

Example 4.23. 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 $f \mapsto \langle f, .\rangle$. Taking $h \in \mathfrak{H}^{**}$ we have, again by the Riesz lemma, that $h(g) = \langle\langle f, .\rangle, \langle g, .\rangle\rangle_{\mathfrak{H}^*} = \langle f, g\rangle^* = \langle g, f\rangle = J(f)(g)$. ⋄

Example 4.24. The sum of reflexive spaces is reflexive. Indeed, recall $(X \oplus Y)^* \cong X^* \oplus Y^*$ with $(x', y')(x, y) := x'(x) + y'(y)$ and hence also $(X \oplus Y)^{**} \cong X^{**} \oplus Y^{**}$ with $(x'', y'')(x', y') := x''(x') + y''(y')$. Hence for $(x'', y'') \in (X \oplus Y)^{**}$ there is $x \in X$ and $y \in Y$ such that $(x'', y'') = (J_X(x), J_Y(y)) = J(x, y)$ and hence J is surjective. This even extends to countable sums — Problem 4.30. ⋄

[7] Robert C. James (1918–2004), American mathematician

Lemma 4.19. *Let X be a normed 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 $(j_{**}(y''))(\ell) = y''(\ell\restriction_Y)$ for $y'' \in Y^{**}$ and $\ell \in X^*$. Note that j_{**} is isometric. Indeed, $\|j_{**}(y'')\| = \sup_{\ell\in X^*,\|\ell\|=1} |y''(\ell\restriction_Y)| = \sup_{\tilde{\ell}\in Y^*,\|\tilde{\ell}\|=1} |y''(\tilde{\ell})| = \|y''\|$ since for every normalized $\tilde{\ell} \in Y^*$ there is some normalized $\ell \in X^*$ with $\ell\restriction_Y = \tilde{\ell}$ by Corollary 4.13. Then

$$\begin{array}{ccc} X & \stackrel{J_X}{\longrightarrow} & X^{**} \\ j\uparrow & & \uparrow j_{**} \\ Y & \underset{J_Y}{\longrightarrow} & Y^{**} \end{array}$$

commutes. In fact, we have $j_{**}(J_Y(y))(\ell) = J_Y(y)(\ell\restriction_Y) = \ell(y) = J_X(y)(\ell)$. 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 $j_{**}(y'')(\ell) = y''(\ell\restriction_Y)$ 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 4.16. That is, $j_{**}(Y^{**}) = J_X(Y)$ and $J_Y = j_{**}^{-1} \circ J_X \circ j$ is surjective.

(ii) Suppose X is reflexive. Then the two maps

$$\begin{array}{cccc} (J_X)_*: & X^* & \to & X^{***} \\ & x' & \mapsto & x' \circ J_X^{-1} \end{array} \qquad \begin{array}{cccc} (J_X)^*: & X^{***} & \to & X^* \\ & x''' & \mapsto & x''' \circ J_X \end{array}$$

are inverse of each other. Moreover, fix $x'' \in X^{**}$ and let $x = J_X^{-1}(x'')$. Then $J_{X^*}(x')(x'') = x''(x') = J_X(x)(x') = x'(x) = x'(J_X^{-1}(x''))$, that is $J_{X^*} = (J_X)_*$ respectively $(J_{X^*})^{-1} = (J_X)^*$, which shows that X^* is reflexive if X is 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 4.15. 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 4.25. 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 (see Theorem 6.5 from [**40**]). $\diamond$

Finally we discuss the analog of the orthogonal complement of a set. Given subsets $M \subseteq X$ and $N \subseteq X^*$ we define their **annihilator** as

$$\begin{aligned} M^\perp &:= \{\ell \in X^* | \ell(x) = 0 \,\forall x \in M\} = \{\ell \in X^* | M \subseteq \mathrm{Ker}(\ell)\} \\ &= \bigcap_{x\in M} \{\ell \in X^* | \ell(x) = 0\} = \bigcap_{x \in M} \{x\}^\perp, \\ N_\perp &:= \{x \in X | \ell(x) = 0 \,\forall \ell \in N\} = \bigcap_{\ell \in N} \mathrm{Ker}(\ell) = \bigcap_{\ell\in N} \{\ell\}_\perp. \end{aligned} \tag{4.6}$$

In particular, $\{\ell\}_\perp = \mathrm{Ker}(\ell)$ while $\{x\}^\perp = \mathrm{Ker}(J(x))$ (with $J : X \hookrightarrow X^{**}$ the canonical embedding). Note $\{0\}^\perp = X^*$ and $\{0\}_\perp = X$.

Example 4.26. In a Hilbert space the annihilator is simply the orthogonal complement. $\diamond$

The following properties are immediate from the definition (by linearity and continuity):

- $M^\perp$ is a closed subspace of X^* and $M^\perp = (\overline{\mathrm{span}(M)})^\perp$.
- $N_\perp$ is a closed subspace of X and $N_\perp = (\overline{\mathrm{span}(N)})_\perp$.

Note that we can also consider $N^\perp = \{x'' \in X^{**} | N \subseteq \mathrm{Ker}(x'')\} \subseteq X^{**}$ and that we have $J(N_\perp) \subseteq N^\perp$ with equality if X is reflexive. Similarly we have $J(M)_\perp = M^\perp$.

Note also that

$$\begin{aligned} \overline{\mathrm{span}(M)} = X \quad &\Leftrightarrow \quad M^\perp = \{0\}, \\ \overline{\mathrm{span}(N)} = X^* \quad &\Rightarrow \quad N_\perp = \{0\} \end{aligned} \tag{4.7}$$

by Corollary 4.15 and Corollary 4.14, respectively. The converse of the last statement is wrong in general (unless X is reflexive, see the following lemma).

Example 4.27. Consider $X := \ell^1(\mathbb{N})$ and $N := \{\delta^n\}_{n\in\mathbb{N}} \subset \ell^\infty(\mathbb{N}) \cong X^*$. Then $\overline{\mathrm{span}(N)} = c_0(\mathbb{N})$ but $N_\perp = \{0\}$. $\diamond$

Lemma 4.20. *We have $(M^\perp)_\perp = \overline{\mathrm{span}(M)}$ and $(N_\perp)^\perp \supseteq \overline{\mathrm{span}(N)}$ with equality if X is reflexive.*

Proof. By the preceding remarks we can assume M, N to be closed subspaces. The first part

$$(M^\perp)_\perp = \{x \in X | \ell(x) = 0 \,\forall \ell \in X^* \text{ with } M \subseteq \mathrm{Ker}(\ell)\} = M$$

is Corollary 4.16 and for the second part one just has to spell out the definition:

$$(N_\perp)^\perp = \{\ell \in X^* | \bigcap_{\tilde{\ell}\in N} \operatorname{Ker}(\tilde{\ell}) \subseteq \operatorname{Ker}(\ell)\} \supseteq N.$$

If X is reflexive we can use the first part to conclude

$$(N_\perp)^\perp = (J(N_\perp))_\perp = (N^\perp)_\perp = N. \qquad \square$$

Note that we also have equality in the preceding lemma if N is finite dimensional (Problem 4.27). For non-reflexive spaces the inclusion can be strict as the previous example shows. Moreover, with a little more machinery one can identify $(N_\perp)^\perp$ as the weak-$*$ closure of $\overline{\operatorname{span}(N)}$ (Problem 6.20).

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.

With the help of annihilators we can also describe the dual spaces of subspaces.

Theorem 4.21. *Let M be a closed subspace of a normed space X. Then there are canonical isometries*

$$(X/M)^* \cong M^\perp, \qquad M^* \cong X^*/M^\perp. \tag{4.8}$$

Proof. In the first case the isometry is given by $\ell \mapsto \ell \circ \pi$, where $\pi : X \to X/M$ is the quotient map, and in the second case by $x' + M^\perp \mapsto x' \restriction_M$. The details are easy to check. $\square$

Corollary 4.22. *Suppose X is a reflexive Banach space and M a closed subspace. Then X/M is reflexive.*

Proof. By Lemma 4.19 (i) $M^\perp \cong (X/M)^*$ is reflexive and hence so is X/M by Lemma 4.19 (ii). $\square$

Problem 4.28. *Show that X is finite dimensional if and only X^* is. In this case the dimensions agree and X is reflexive.*

Problem 4.29. *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 6.39. (Hint: For the first part apply Problem 4.18 to X^*. For the second part consider Problem 4.21)

Problem* 4.30. *Let $X := \bigoplus_{p,j\in\mathbb{N}} X_j$ be defined as in Problem 1.66 and let $\frac{1}{p}+\frac{1}{q}=1$. Show that for $1 \le p < \infty$ we have $X^* \cong \bigoplus_{q,j\in\mathbb{N}} X_j^*$, where the identification is given by*

$$y(x) = \sum_{j\in\mathbb{N}} y_j(x_j), \quad x = (x_j)_{j\in\mathbb{N}} \in \bigoplus\nolimits_{p,j\in\mathbb{N}} X_j, \quad y = (y_j)_{j\in\mathbb{N}} \in \bigoplus\nolimits_{q,j\in\mathbb{N}} X_j^*.$$

Moreover, if all X_j are reflexive, so is X for $1 < p < \infty$.

Problem 4.31. *Suppose X is separable. Show that there exists a countable set $N \subset X^*$ with $N_\perp = \{0\}$.*

Problem 4.32. *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

$$(M_1 \cap M_2)^\perp \supseteq \overline{(M_1^\perp + M_2^\perp)}, \qquad (M_1^\perp \cap M_2^\perp)_\perp = \overline{(M_1 + M_2)}.$$

4.4. 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** (also **dual**) $A' : Y^* \to X^*$ via $A'y' = y' \circ A$, $y' \in Y^*$.

Example 4.28. If $X := \mathbb{C}^n$ and $\{\delta_j\}_{j=1}^n$ is the standard basis with corresponding dual basis $\{\delta_j'\}_{j=1}^n$, then $(A'\ell_k')(\delta_j) = \ell_k'(A\delta_j) = A_{kj}$, where $A_{jk} := \delta_j'(A\delta_k)$ is the matrix corresponding to A with respect to the standard basis. In particular, A' corresponds to the transposed matrix. ⋄

Warning: In the case of complex Hilbert spaces there is subtile difference between the adjoint operator A^* as defined in Section 2.3 and the adjoint operator A' as defined above:

Example 4.29. 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 $A' = C_1 A^* C_2^{-1}$.

In the finite dimensional case this implies that A' corresponds to the transposed matrix while A^* corresponds to the conjugate transposed matrix. ⋄

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 4.18 to obtain the fourth equality. In summary,

Theorem 4.23. *Suppose X, Y are normed spaces. 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

$$(\alpha A + \beta B)' = \alpha A' + \beta B' \tag{4.9}$$

and for $A \in \mathscr{L}(X,Y)$ and $B \in \mathscr{L}(Y,Z)$ we have

$$(BA)' = A'B' \tag{4.10}$$

which is immediate from the definition. Note also that $\mathbb{I}_X' = \mathbb{I}_{X^*}$.

Example 4.30. Let $X := Y := \ell^p(\mathbb{N})$, $1 \le p < \infty$, such that $X^* \cong \ell^q(\mathbb{N})$, $\frac{1}{p} + \frac{1}{q} = 1$. Consider the right shift $R \in \mathscr{L}(\ell^p(\mathbb{N}))$ given by

$$Ra := (0, a_1, a_2, \dots).$$

Then for $b \in \ell^q(\mathbb{N})$

$$l_b(Ra) = \sum_{j=1}^{\infty} b_j (Ra)_j = \sum_{j=2}^{\infty} b_j a_{j-1} = \sum_{j=1}^{\infty} b_{j+1} a_j$$

which shows $(R'b)_k = b'_{k+1}$ upon choosing $a = \delta^k$. Hence $R' = L$ is the left shift: $Lb := (b_2, b_3, \dots)$. Similarly, $L' = R$. ◇

Example 4.31. Let $c \in \ell^\infty(\mathbb{N})$ and consider the multiplication operator A in $\ell^p(\mathbb{N})$ with $1 \le p < \infty$ defined by $(Aa)_j := c_j a_j$. As in the previous example $X^* \cong \ell^q(\mathbb{N})$ with $\frac{1}{q} + \frac{1}{p} = 1$ and for $b \in \ell^q(\mathbb{N})$ we have

$$l_b(Aa) = \sum_{j=1}^{\infty} b_j (c_j a_j) = \sum_{j=1}^{\infty} (c_j b_j) a_j,$$

which shows $(A'b)_j = c_j b_j$ and hence A' is multiplication with c but now in $\ell^q(\mathbb{N})$. Also note that in the case $p = 2$ the Hilbert space adjoint A^* would be multiplication by the complex conjugate sequence c^*. ◇

Example 4.32. 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 4.26), 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

$$A''(J_X(x))(y') = J_X(x)(A'y') = (A'y')(x) = y'(Ax) = J_Y(Ax)(y') \tag{4.11}$$

shows that the following diagram commutes

$$\begin{array}{ccc} X & \xrightarrow{A} & Y \\ J_X \downarrow & & \downarrow J_Y \\ X^{**} & \xrightarrow[A'']{} & Y^{**} \end{array}$$

Consequently

$$A'' \restriction_{\mathrm{Ran}(J_X)} = J_Y A J_X^{-1}, \qquad A = J_Y^{-1} A'' J_X. \tag{4.12}$$

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 4.34).

Theorem 4.24 (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 $f_n := y_n'\restriction_K \in C(K)$. 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.40) 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 $A = J_Y^{-1} A'' J_X$. □

Theorem 4.25. *Suppose X, Y are Banach spaces. If $A \in \mathscr{L}(X, Y)$, then A^{-1} exists and is in $\mathscr{L}(Y, X)$ if and only if $(A')^{-1}$ exists and is in $\mathscr{L}(X^*, Y^*)$. Moreover, in this case we have*

$$(A')^{-1} = (A^{-1})'. \tag{4.13}$$

Proof. If A has a bounded inverse, then $A'(A^{-1})' = (A^{-1}A)' = \mathbb{I}_X' = \mathbb{I}_{X^*}$ and $(A^{-1})'A' = (A^{-1}A)' = \mathbb{I}_Y' = \mathbb{I}_{Y^*}$ shows that A' is invertible with $(A')^{-1} = (A^{-1})'$.

Conversely, let $(A')^{-1} \in \mathscr{L}(X^*, Y^*)$. Then by the first part $(A'')^{-1}$ exists and is in $\mathscr{L}(X^{**}, Y^{**})$. Moreover, $A^{-1} = J_X^{-1}(A'')^{-1} J_Y \in \mathscr{L}(Y, X)$. □

Finally we discuss the relation between solvability of $Ax = y$ and the corresponding adjoint equation $A'y' = x'$. We begin with the following analog of (2.30) (Problem 4.37):

Lemma 4.26. *Let X, Y be Banach spaces and $A \in \mathscr{L}(X, Y)$, then* $\operatorname{Ran}(A)^\perp = \operatorname{Ker}(A')$ *and* $\operatorname{Ran}(A')_\perp = \operatorname{Ker}(A)$.

Taking annihilators in these formulas Lemma 4.20 implies

$$\operatorname{Ker}(A')_\perp = (\operatorname{Ran}(A)^\perp)_\perp = \overline{\operatorname{Ran}(A)} \tag{4.14}$$

and

$$\operatorname{Ker}(A)^\perp = (\operatorname{Ran}(A')_\perp)^\perp \supseteq \overline{\operatorname{Ran}(A')}, \tag{4.15}$$

with equality if X is reflexive.

Note that the first identity tells us that, for an operator A with closed range, a necessary and sufficient solvability criterion for the equation $Ax = y$ is $y \in \operatorname{Ker}(A')_\perp$ (that is, $\ell(y) = 0$ for all $\ell \in \operatorname{Ker}(A')$). Equality in the second identity would imply an analogous criterion for the adjoint equation.

Example 4.33. Suppose X, Y are Banach spaces and $A \in \mathscr{L}(X, Y)$. When trying to solve the abstract problem $Ax = y$ an estimate of the type (for some $C > 0$)

$$\|x\| \le C\|y\|$$

for a solution x (provided there is a solution at all) in terms of the inhomogeneous datum y is known as an **a priori estimate**.

Of course an a priori estimate shows that $x = 0$ whenever $y = 0$ and hence immediately gives injectivity of A. Moreover, if we choose a convergent sequence $y_n = Ax_n \to y$ from $\operatorname{Ran}(A)$, then the a priori estimate implies convergence of $x_n \to x$ and thus $y = Ax \in \operatorname{Ran}(A)$. Hence we see that $\operatorname{Ran}(A)$ is closed. Moreover, chosing $x = A^{-1}y$ in the a priori estimate shows that the inverse of A is bounded and the optimal constant C is precisely $\|A^{-1}\|$.

As the name insinuates ("a priori" means "from before" in Latin), the estimate is derived before a solution is known to exist, and hence should better be written as

$$\|x\| \le C\|Ax\|, \qquad x \in X,$$

it cannot be used to establish surjectivity of A. For this one could show injectivity of the adjoint problem, that is, $\operatorname{Ker}(A') = \{0\}$, and then appeal to (4.14).

These considerations should be compared with Lemma 2.16 and the Lax–Milgram theorem (Theorem 2.17). In particular, note that coercivity combines the a priori estimate and the fact that the kernel of the adjoint operator is trivial into one convenient condition. ◇

Unfortunately, if X is not reflexive, then equality in (4.15) might fail (Problem 4.40). However, it turns out that this can only happen if the range of A is not closed. To show this we first establish a criterion for the range of A to be closed. Of course we would like to use the a priori estimate from the previous example. Since this will only work if A is injective, we will factor out the kernel (hence it will no longer hold for all x but only for a suitable chosen representative from the equivalence class). Moreover, it will be convenient to also allow Ax to deviate from y, as long as the difference remains under control.

Lemma 4.27. *Let X, Y be Banach spaces and suppose $A \in \mathscr{L}(X,Y)$. Then* $\mathrm{Ran}(A)$ *is closed if and only if there exists some $0 \le \delta < 1$ and some $\varepsilon > 0$ such that for every $y \in \overline{\mathrm{Ran}(A)}$ we can find a corresponding $x \in X$ satisfying*

$$\varepsilon\|x\| + \|y - Ax\| \le \delta\|y\|. \tag{4.16}$$

Proof. If $\mathrm{Ran}(A)$ is closed, then we can factor out its kernel and restrict Y to obtain a bijective operator $\tilde{A}$ as in Problem 1.71. By the inverse mapping theorem (Theorem 4.8) $\tilde{A}$ has a bounded inverse. Fix $\delta < 1$ and choose $\varepsilon < \|\tilde{A}^{-1}\|^{-1}\delta$. Then for every $y \in \mathrm{Ran}(A)$ there is some $x \in X$ with $y = Ax$ and $\|Ax\| \ge \frac{\varepsilon}{\delta}\|x\|$ after maybe adding an element from the kernel to x. This x satisfies $\varepsilon\|x\| + \|y - Ax\| = \varepsilon\|x\| \le \delta\|y\|$ as required.

Conversely, fix $y \in \overline{\mathrm{Ran}(A)}$ and recursively choose a sequence x_n (starting with $x_0 = 0$) such that

$$\varepsilon\|x_n\| + \|(y - A\tilde{x}_{n-1}) - Ax_n\| \le \delta\|y - A\tilde{x}_{n-1}\|, \quad \tilde{x}_n := \sum_{m \le n} x_m.$$

In particular, $\|y - A\tilde{x}_n\| \le \delta^n\|y\|$ as well as $\varepsilon\|x_n\| \le \delta^n\|y\|$, which shows $\tilde{x}_n \to x$ and $A\tilde{x}_n \to Ax = y \in \mathrm{Ran}(A)$. □

Theorem 4.28 (Banach; Closed range)**.** *Suppose X, Y are Banach spaces and $A \in \mathscr{L}(X,Y)$. Then the following items are equivalent:*

(i) $\mathrm{Ran}(A)$ *is closed.*
(ii) $\mathrm{Ker}(A)^\perp = \mathrm{Ran}(A')$.
(iii) $\mathrm{Ran}(A')$ *is closed.*
(iv) $\mathrm{Ker}(A')_\perp = \mathrm{Ran}(A)$.

Proof. Consider $\tilde{X} = X/\mathrm{Ker}(A)$ and $\tilde{Y} = \overline{\mathrm{Ran}(A)}$ and the corresponding operator $\tilde{A}$ as in Problem 1.71. Then $\tilde{A} \in \mathscr{L}(\tilde{X},\tilde{Y})$ is a bounded injective operator whose range is dense. In particular, $\mathrm{Ran}(A)$ is closed if and only if $\tilde{A}^{-1} \in \mathscr{L}(\tilde{Y},\tilde{X})$. Moreover, $\mathrm{Ran}(\tilde{A}') \subseteq \tilde{X}^* \cong \mathrm{Ker}(A)^\perp$ (cf. Theorem 4.21) and the canonical isometry (i.e. composition with the quotient map) will map $\mathrm{Ran}(A') \subseteq \mathrm{Ker}(A)^\perp$ onto $\mathrm{Ran}(\tilde{A}')$ — to see this observe $(\tilde{A}'\tilde{y}')(\tilde{x}) =$

$y'(Ax) = (A'y')(x)$, where $\tilde{x} = x + \mathrm{Ker}(A)$ and $\tilde{y}' = y' + \mathrm{Ran}(A)^\perp$ (using $\tilde{Y}^* \cong Y^*/\tilde{Y}^\perp$ by Theorem 4.21). In particular, $\mathrm{Ran}(A') = \mathrm{Ker}(A)^\perp$ if and only if $(\tilde{A}')^{-1} \in \mathscr{L}(\tilde{X}^*, \tilde{Y}^*)$.

Hence (i) $\Leftrightarrow$ (ii) follows from Theorem 4.25 applied to $\tilde{A}$. (ii) $\Rightarrow$ (iii) is clear since annihilators are closed. (i) $\Leftrightarrow$ (iv) is immediate from (4.14).

(iii) $\Rightarrow$ (i): Without loss of generality we can replace A by $\tilde{A}$ and assume that both A and A' are injective. Then, if $\mathrm{Ran}(A')$ is closed, A' has a bounded inverse (defined on the range) by the inverse mapping theorem.

Suppose $\mathrm{Ran}(A)$ were not closed. Then, given $\varepsilon > 0$ and $0 \le \delta < 1$, by Lemma 4.27 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 Lemma 4.15 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 $\bar{\ell}(0, y) = \mathrm{dist}((0, y), V)$ (note that $(0, y) \notin V$ since $y \ne 0$). Then $\ell(.) = \bar{\ell}(0, .)$ is the functional we are looking for since $\mathrm{dist}((0, y), V) \ge \delta\|y\|$ 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 $\|A'\ell_n\| \to 0$ contradicting the fact that A' has a bounded inverse. □

Problem 4.33. *Consider $\ell^p(\mathbb{N})$, $1 \le p < \infty$, and let $A \in \mathscr{L}(\ell^p(\mathbb{N}))$ be some bounded operator. Let $A_{jk} = (A\delta^k)_j$ be its matrix elements such that*

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

Show that this series converges for every $a \in \ell^p(\mathbb{N})$.

Identify $\ell^p(\mathbb{N})^* \cong \ell^q(\mathbb{N})$, $\frac{1}{p} + \frac{1}{q} = 1$ and show that $A'_{jk} = A_{kj}$.

Problem* 4.34. *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*

$$Aa := (\sum_{j\in\mathbb{N}} a_j, 0, \dots).$$

Show that

$$A'b = (b_1, b_1, \dots).$$

Conclude that A is not the adjoint of an operator from $\mathscr{L}(c_0(\mathbb{N}))$.

Problem 4.35. *Show that that the adjoint of a projection $P \in \mathscr{L}(X)$ is again a projection.*

Problem 4.36. *Show that for $A \in \mathscr{L}(X, Y)$ we have*

$$\mathrm{rank}(A) = \mathrm{rank}(A').$$

Problem 4.37. *Show Lemma 4.26.*

Problem 4.38. *Let $A_j \in \mathscr{L}(X_j, Y_j)$, $j = 1, 2$, and consider $A_1 \oplus A_2$ (cf. Section 1.7). Show that $(A_1 \oplus A_2)' = A_1' \oplus A_2'$ with respect to the identification $(X_1 \oplus X_2)^* \cong X_1^* \oplus X_2^*$ and $(Y_1 \oplus Y_2)^* \cong Y_1^* \oplus Y_2^*$ (cf. Lemma 1.19).*

Problem 4.39. *Let X be a Banach space and let $\ell_n, \ell \in X^*$. 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 $\ell \in (N_\perp)^\perp$.*

Problem* 4.40. *Consider the multiplication operator $A : \ell^1(\mathbb{N}) \to \ell^1(\mathbb{N})$ with $(Aa)_j := \frac{1}{j} a_j$ (cf. Example 4.31). Show that* $\mathrm{Ran}(A)$ *is not closed but dense while* $\mathrm{Ran}(A')$ *is neither closed nor dense. In particular, show* $\mathrm{Ker}(A)^\perp = \{0\}^\perp = \ell^\infty(\mathbb{N}) \supset \overline{\mathrm{Ran}(A')} = c_0(\mathbb{N})$.

4.5. Weak convergence

Let X be a normed space. 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 infinite dimensions this is no longer true in general:

Example 4.34. 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 $\|u_n\| = 1$.) $\diamond$

If $\ell(x_n) \to \ell(x)$ for every $\ell \in X^*$ we say that x_n **converges weakly** to x and write

$$\operatorname*{w-lim}_{n\to\infty} x_n = x \quad \text{or} \quad x_n \rightharpoonup x. \tag{4.17}$$

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 4.29. *Let X be a normed 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 4.18). □

Item (ii) says that the norm is sequentially weakly lower semicontinuous (cf. Problem B.34) while the previous example shows that it is not sequentially weakly continuous. However, bounded linear operators turn out to be sequentially weakly continuous (Problem 4.44). Nonlinear operations are more tricky as the next example shows:

Example 4.35. Consider $L^2(0,1)$ and recall (see Example 3.10) 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

$$\langle u_m, v_n \rangle = \frac{\sqrt{2}(1-(-1)^m)}{m\pi} \frac{4n^2}{(4n^2 - m^2)} \to \frac{\sqrt{2}(1-(-1)^m)}{m\pi} = \langle u_m, 1\rangle$$

and the claim follows from Problem 4.49 since $\|v_n\| = \sqrt{\frac{3}{2}}$. ◇

Example 4.36. Let $X := c_0(\mathbb{N})$ and hence $X^* \cong \ell^1(\mathbb{N})$. Let $a_j^n := 1$ for $1 \le j \le n$ and $a_j^n := 0$ for $j > n$. Then for every $b \in \ell^1(\mathbb{N})$ we have

$$\lim_{n\to\infty} l_b(a^n) = \lim_{n\to\infty} \sum_{j=1}^\infty b_j a_j^n = \lim_{n\to\infty} \sum_{j=1}^n b_j = \sum_{j=1}^\infty b_j$$

and hence a^n is a weak Cauchy sequence which, does not converge. Indeed, $a^n \rightharpoonup a$ would imply $a_j = 1$ for all $j \in \mathbb{N}$ (upon choosing $b = \delta^j$) which is clearly not in X. The limit is however in $X^{**} \cong \ell^\infty(\mathbb{N})$. ◇

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 4.15 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 6.3.

In a Hilbert space there is also a simple criterion for a weakly convergent sequence to converge in norm (see Theorem 6.19 for a generalization).

Lemma 4.30. *Let $\mathfrak{H}$ be a Hilbert space and let $f_n \rightharpoonup f$. Then $f_n \to f$ if and only if* $\limsup \|f_n\| \le \|f\|$.

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.19). In infinite dimensions this breaks down as we have already seen in Section 1.5. We even have

Theorem 4.31 (F. Riesz)**.** *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 $\{X \setminus \mathrm{Ker}(\ell)\}_{\ell \in X^*}$ is an open cover since for every $x \in S$ there is some $\ell \in X^*$ with $\ell(x) \ne 0$ by Corollary 4.13. This cover has a finite subcover, $S \subset \bigcup_{j=1}^n (X \setminus \mathrm{Ker}(\ell_j)) = X \setminus \bigcap_{j=1}^n \mathrm{Ker}(\ell_j)$. Hence $\bigcap_{j=1}^n \mathrm{Ker}(\ell_j) = \{0\}$ and the map $X \to \mathbb{C}^n$, $x \mapsto (\ell_1(x), \cdots, \ell_n(x))$ is injective, that is, $\dim(X) \le n$. □

Of course in the formulation of the above theorem the unit ball could be replaced with any ball (of positive radius). In particular, if X is infinite dimensional, a compact set cannot contain a ball, that is, it must have empty interior. Hence compact sets are always meager in infinite dimensional spaces.

However, if we are willing to trade convergence for weak convergence, the situation looks much brighter!

Theorem 4.32 (Šmulian[8]). *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 $Y := \overline{\operatorname{span}\{x_n\}}$. Then Y is reflexive by Lemma 4.19 (i). Moreover, by construction Y is separable and so is Y^* by the remark after Lemma 4.19.

Let ℓ_k be a dense set in Y^* and use Cantor's diagonal sequence trick to obtain a subsequence on which all functionals converge: To this end we begin by selecting a subsequence x_n^1 from x_n such that $\ell_1(x_n^1)$ converges. Next we select a subsequence x_n^2 from x_n^1 such that $\ell_2(x_n^2)$ converges. Continuing like this we obtain subsequences x_n^m and the diagonal sequence x_n^n has the property that $\ell_k(x_n^n)$ converges for every k. Denote this subsequence again by x_n for notational simplicity. Then,

$$\begin{aligned}|\ell(x_n) - \ell(x_m)| \le& |\ell(x_n) - \ell_k(x_n)| + |\ell_k(x_n) - \ell_k(x_m)| \\ &+ |\ell_k(x_m) - \ell(x_m)| \\ \le& 2C\|\ell - \ell_k\| + |\ell_k(x_n) - \ell_k(x_m)|\end{aligned}$$

shows that $\ell(x_n)$ converges for every $\ell \in \overline{\operatorname{span}\{\ell_k\}} = Y^*$. Thus there is a limit by Lemma 4.29 (iv). □

Note that this theorem breaks down if X is not reflexive.

Example 4.37. Consider the sequence of vectors δ^n in $\ell^p(\mathbb{N})$, $1 \le p < \infty$. Then $\delta^n \rightharpoonup 0$ for $1 < p < \infty$. In fact, since every $l \in \ell^p(\mathbb{N})^*$ is of the form $l = l_b$ for some $b \in \ell^q(\mathbb{N})$ we have $l_b(\delta^n) = b_n \to 0$.

If we consider the same sequence in $\ell^1(\mathbb{N})$ there is no weakly convergent subsequence. In fact, since $l_b(\delta^n) \to 0$ for every sequence $b \in c_0(\mathbb{N})$, the only possible weak limit is zero. On the other hand, choosing the constant sequence $b := (1)_{j=1}^\infty$ we see $l_b(\delta^n) = 1 \not\to 0$, a contradiction. ⋄

Example 4.38. Let $X := L^1(-1,1)$. Every continuous function φ gives rise to a linear functional

$$\ell_\varphi(f) := \int_{-1}^1 f(x)\varphi(x)\,dx$$

[8] Vitold Shmulyan (1914–1944), Soviet mathematician

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

$$\int u_k(x)\varphi(x)\,dx \to \varphi(0)$$

(cf. Lemma 1.2) 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 strictly weaker than strong convergence since otherwise every bounded sequence would have a weakly, and thus by assumption also norm, convergent subsequence contradicting Theorem 4.31. In a non-reflexive space this situation can however occur.

Example 4.39. In $\ell^1(\mathbb{N})$ every weakly convergent sequence is in fact (norm) convergent (such Banach spaces are said to have the **Schur property**[9]). First of all recall that $\ell^1(\mathbb{N})^* \cong \ell^\infty(\mathbb{N})$ and $a^n \rightharpoonup 0$ implies

$$l_b(a^n) = \sum_{k=1}^\infty b_k a_k^n \to 0, \qquad \forall b \in \ell^\infty(\mathbb{N}).$$

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 $a_j^n = l_{\delta^j}(a^n) \to 0$ 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 $\sum_{k_j \leq k < k_{j+1}} |a_k^{n_j}| \geq \frac{2}{3}$. Now define b such that

$$b_k := \operatorname{sign}(a_k^{n_j})^*, \qquad k_j \leq k < k_{j+1}, \quad j \in \mathbb{N},$$

where we assume $k_1 = 1$ without loss of generality (to cover all $k \in \mathbb{N}$). Then

$$|l_b(a^{n_j})| \geq \sum_{k_j \leq k < k_{j+1}} |a_k^{n_j}| - \Bigg| \sum_{1 \leq k < k_j;\, k_{j+1} \leq k} b_k a_k^{n_j} \Bigg| \geq \frac{2}{3} - \frac{1}{3} = \frac{1}{3},$$

[9] Issai Schur (1875–1941), Russian mathematician

contradicting $a^{n_j} \rightharpoonup 0$. $\diamond$

It is also useful to observe that compact operators will turn weakly convergent into (norm) convergent sequences.

Theorem 4.33. *Let $A \in \mathscr{K}(X,Y)$ be compact. Then $x_n \rightharpoonup x$ implies $Ax_n \to Ax$. If X is reflexive the converse is also true.*

Proof. By Lemma 4.29 (iii) every weakly convergent sequence $x_n \rightharpoonup x$ is bounded and by compactness of A we can pass to a subsequence such that $Ax_{n_k} \to y$. Moreover, by Problem 4.44 we even have $y = Ax$ and Lemma B.5 shows $Ax_n \to Ax$.

Conversely, if X is reflexive, then by Theorem 4.32 every bounded sequence x_n has a subsequence $x_{n_k} \rightharpoonup x$ and by assumption $Ax_{n_k} \to Ax$. Hence A is compact. $\square$

Operators which map weakly convergent sequences to convergent sequences are also called **completely continuous**. Warning: 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 4.31.

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,

$$\operatorname*{s-lim}_{n\to\infty} A_n = A \quad :\Leftrightarrow \quad A_n x \to Ax \quad \forall x \in \mathfrak{D}(A) \subseteq \mathfrak{D}(A_n). \tag{4.18}$$

It is said to **converge weakly** to A,

$$\operatorname*{w-lim}_{n\to\infty} A_n = A \quad :\Leftrightarrow \quad A_n x \rightharpoonup Ax \quad \forall x \in \mathfrak{D}(A) \subseteq \mathfrak{D}(A_n). \tag{4.19}$$

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 4.40. Consider the operator $S_n \in \mathscr{L}(\ell^2(\mathbb{N}))$ which shifts a sequence n places to the left, that is,

$$S_n\,(x_1, x_2, \dots) = (x_{n+1}, x_{n+2}, \dots) \tag{4.20}$$

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,

$$S_n^*\,(x_1, x_2, \dots) = (\underbrace{0, \dots, 0}_{n \text{ places}}, x_1, x_2, \dots). \tag{4.21}$$

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 4.34. *Let X, Y be Banach spaces. 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) $\lim_{n\to\infty}(\alpha_n A_n + B_n)x = \lim_{n\to\infty}(\alpha_n A_n x + B_n x) = \alpha A x + Bx$.
(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 4.46). □

Item (iii) of this lemma is sometimes also known as Banach–Steinhaus theorem.

Example 4.41. Consider a **quadrature rule**

$$Q_n f := \sum_{j=1}^{n} w_{n,j} f(x_j)$$

for a function $f \in X := C[0,1]$, where $w_{n,j}$ are given weights and $x_j \in [a,b]$ given nodes. The purpose of a quadrature rule is to approximate the integral, that is, we want to have

$$\lim_{n\to\infty} Q_n f = \int_0^1 f(x)dx$$

for all $f \in X$. A simple case would be the **trapezoidal rule**

$$Q_n f := \frac{1}{n}\Big(\frac{f(0)}{2} + f(\frac{1}{n}) + f(\frac{2}{n}) + \cdots + f(\frac{n-1}{n}) + \frac{f(1)}{2}\Big),$$

which is obtained by integrating a piecewise linear approximation with equidistant nodes. In particular, the result is exact for linear functions.

Considering Q_n as a linear functional we have $\|Q_n\| = \sum_{j=1}^n |w_{n,j}|$ and items (iii) and (iv) from our lemma imply that $Q_n f$ will converge to $\int_0^1 f(x)dx$ for all $f \in X$ if and only if $\sup_n \sum_{j=1}^n |w_{n,j}| < \infty$ and convergence holds for all monomials. ⋄

Example 4.42. Let X be a Banach space of functions $f : [-\pi,\pi] \to \mathbb{C}$ such that 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 (2.47) 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 4.7. 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 (2.54) 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 4.35. *Let X, Y, Z be normed spaces. Suppose $A_n \in \mathscr{L}(Y,Z)$, $B_n \in \mathscr{L}(X,Y)$ are 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 4.43. 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

$$S_n S_n^*(x_1, x_2, \dots) = (x_1, x_2, \dots)$$

converges to the identity! Hence the order in the second claim is important. $\diamond$

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

$$\operatorname*{w^*\text{-}lim}_{n\to\infty} \ell_n = \ell \quad \text{or} \quad \ell_n \stackrel{*}{\rightharpoonup} \ell \tag{4.22}$$

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

$$j(\ell_n) \to j(\ell) \qquad \forall j \in X^{**}, \tag{4.23}$$

whereas for the weak-$*$ limit this is only required for $j \in J(X) \subseteq X^{**}$ (recall $J(x)(\ell) = \ell(x)$).

Example 4.44. 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. $\diamond$

Example 4.45. 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 4.50). $\diamond$

With this notation the proof of Lemma 4.29 (iv) shows that (without assuming X to be refelxive) every weak Cauchy sequence has a limit in X^{**}. Similarly, it is also possible to slightly generalize Theorem 4.32 (Problem 4.51):

Lemma 4.36 (Helly[10])**.** *Suppose X is a separable Banach space. Then every bounded sequence $\ell_n \in X^*$ has a weak-$*$ convergent subsequence.*

[10] Eduard Helly (1884–1943), Austrian mathematician

This last result is also known as **sequential Banach–Alaglou theorem**.

Example 4.46. Let us return to Example 4.38. 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. ⋄

Example 4.47. Consider $X := \ell^\infty(\mathbb{N})$. Then the sequence of projections $l_k \in X^*$ given by $l_k(x) := x_k$ has no weak-$*$ convergent subsequence (if there were such a subsequence k_j, choose $x \in X$ such that x_{k_j} does not converge to get a contradiction). Hence the assumption that X is separable cannot be dropped in Lemma 4.36. ⋄

Problem 4.41. *Show that in a finite dimensional space weak and strong convergence agree.*

Problem 4.42. *Let X be a normed space. Suppose $\ell_n \to \ell$ in X^* and $x_n \rightharpoonup x$ in X. Prove that $\ell_n(x_n) \to \ell(x)$. Similarly, suppose* $\text{s-lim}\, \ell_n = \ell$ *and $x_n \to x$. Prove that $\ell_n(x_n) \to \ell(x)$. Does this still hold if* $\text{s-lim}\, \ell_n = \ell$ *and $x_n \rightharpoonup x$?*

Problem 4.43. *Let $B : X \times Y \to \mathbb{C}$ be a continuous bilinear form (cf. also Problem 4.10). Show that if $x_n \to x$ and $y_n \rightharpoonup y$ we have $B(x_n, y_n) \to B(x, y)$. The same conclusion holds if $x_n \rightharpoonup x$ and $y_n \to y$.*

Problem* 4.44. *Let X, Y be normed spaces. Show that $x_n \rightharpoonup x$ in X implies $Ax_n \rightharpoonup Ax$ for $A \in \mathscr{L}(X, Y)$. Conversely, show that if $x_n \to 0$ in X implies $Ax_n \rightharpoonup 0$ for a linear operator $A : X \to Y$, then $A \in \mathscr{L}(X, Y)$.*

Problem 4.45. *Let X_1, X_2 be normed spaces. Let $X := X_1 \oplus X_2$ and 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 4.46. *Establish Lemma 4.34 in the case of weak convergence. (Hint: Problem 4.19 might be useful.)*

Problem 4.47. *Let X, Y be normed spaces. Suppose $A_n, A \in \mathscr{L}(X, Y)$. Show that* $\text{s-lim}\, A_n = A$ *and $\lim x_n = x$ implies $\lim A_n x_n = Ax$.*

Problem 4.48. *Consider the the sequence of operators $A_n : \ell^1(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ given by $A_n a = (a_1, a_2, \dots, a_n, a_n, a_n, \dots)$ (i.e. $(Aa)_j = a_j$ for $j \le n$ and $(Aa)_j = a_n$ for $j > n$). Does A_n converge in norm/strongly/weakly? If yes, what is its limit?*

Problem* 4.49. *Show that if $\{\ell_j\} \subseteq X^*$ is total in X^* for a Banach space X, then $x_n \rightharpoonup x$ in 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* 4.50. *Let X be a normed space. 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 a bounded sequence and $\ell_n(x_j) \to \ell(x_j)$ for all j.*

Problem* 4.51. *Prove Lemma 4.36.*

Chapter 5

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.

5.1. Banach algebras

A Banach space X together with a multiplication satisfying

$$(x+y)z = xz + yz, \quad x(y+z) = xy + xz, \quad x, y, z \in X, \tag{5.1}$$

and

$$(xy)z = x(yz), \qquad \alpha\,(xy) = (\alpha x)y = x\,(\alpha y), \quad \alpha \in \mathbb{C}, \tag{5.2}$$

and

$$\|xy\| \le \|x\|\|y\| \tag{5.3}$$

is called a **Banach algebra**. In particular, note that (5.3) ensures that multiplication is continuous (Problem 5.1). Conversely one can show that (separate) continuity of multiplication implies existence of an equivalent norm satisfying (5.3) (Problem 5.2).

An element $e \in X$ satisfying

$$ex = xe = x, \qquad \forall x \in X \tag{5.4}$$

is called **identity** (show that e is unique) and we will assume $\|e\| = 1$ in this case (by Problem 5.2 this can be done without loss of generality).

Example 5.1. Let X be a Banach space. The bounded linear operators $\mathscr{L}(X)$ form a Banach algebra with identity $\mathbb{I}$. ⋄

Example 5.2. The continuous functions $C(I)$ over some compact interval form a commutative Banach algebra with identity 1.

Slightly more general, the differentiable functions $C^k(I)$ over some compact interval do not form a commutative Banach algebra since (1.68) does not fulfill (5.3) for $k > 1$. However, the equivalent norm

$$\|f\|_{k,\infty} := \sum_{j=0}^{k} \frac{\|f^{(j)}\|_\infty}{j!}$$

remedies this problem. ⋄

Example 5.3. The space $\ell^p(\mathbb{N})$ together with the component-wise product is a commutative Banach algebra since $\|a\,b\|_p \le \|a\|_\infty \|b\|_p \le \|a\|_p\|b\|_p$ (cf. Problem 1.20). However, except for $p = \infty$ there is no identity.

Note that this does not work for $L^p(I)$, $1 \le p < \infty$, since singularities get worse by squaring and hence the square of a p-integrable function will not be p-integrable in general. ⋄

Example 5.4. 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**.[1] Of course as a Banach space it is isomorphic to $\ell^1(\mathbb{Z})$ by definition. To see that it is a

[1] Norbert Wiener (1894–1964), American mathematician and philosopher

Banach algebra note that

$$\begin{aligned} f(x)g(x) &= \sum_{k\in\mathbb{Z}} \hat{f}_k \mathrm{e}^{\mathrm{i}kx} \sum_{j\in\mathbb{Z}} \hat{g}_j \mathrm{e}^{\mathrm{i}jx} = \sum_{k,j\in\mathbb{Z}} \hat{f}_k \hat{g}_j \mathrm{e}^{\mathrm{i}(k+j)x} \\ &= \sum_{k\in\mathbb{Z}} \Big(\sum_{j\in\mathbb{Z}} \hat{f}_j \hat{g}_{k-j} \Big) \mathrm{e}^{\mathrm{i}kx}. \end{aligned}$$

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 5.5. The space $L^1(\mathbb{R}^n)$ together with the convolution

$$(g * f)(x) := \int_{\mathbb{R}^n} g(x-y) f(y) dy = \int_{\mathbb{R}^n} g(y) f(x-y) dy \tag{5.5}$$

is a commutative Banach algebra (Problem 5.18) 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 5.3).

An element $x \in X$ is called **invertible** if there is some $y \in X$ such that

$$xy = yx = e. \tag{5.6}$$

In this case y is called the inverse of x and is denoted by x^{-1}. It is straightforward to show (Problem 5.4) that the inverse is unique (if one exists at all) and that

$$(xy)^{-1} = y^{-1} x^{-1}, \qquad (x^{-1})^{-1} = x. \tag{5.7}$$

In particular, the set of invertible elements $\mathcal{G}(X)$ forms a group under multiplication.

Example 5.6. 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 5.7. Let $X := \mathscr{L}(\ell^p(\mathbb{N}))$ and recall the shift operators $S^\pm$ defined via $(S^\pm a)_j = a_{j\pm 1}$ with the convention that $a_0 = 0$. Then $S^+S^- = \mathbb{I}$ but $S^-S^+ \ne \mathbb{I}$. 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 all elements in a neighborhood. This will be a consequence of the following straightforward generalization of the geometric series to our abstract setting.

Lemma 5.1. *Let X be a Banach algebra with identity e. Suppose $\|x\| < 1$. Then $e - x$ is invertible and*

$$(e-x)^{-1} = \sum_{n=0}^{\infty} x^n. \tag{5.8}$$

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

$$\left(\sum_{n=0}^{\infty} x^n\right)(e-x) = \sum_{n=0}^{\infty} x^n - \sum_{n=1}^{\infty} x^n = e.$$ □

Corollary 5.2. *Let $x, y \in X$. Suppose x is invertible and $\|x^{-1}y\| < 1$ or $\|yx^{-1}\| < 1$. Then $(x-y)$ is invertible as well and*

$$(x-y)^{-1} = \sum_{n=0}^{\infty} (x^{-1}y)^n x^{-1} \quad \textit{or} \quad (x-y)^{-1} = \sum_{n=0}^{\infty} x^{-1}(yx^{-1})^n. \tag{5.9}$$

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:

$$\|(x-y)^{-1} - x^{-1}\| \le \frac{\|x^{-1}\|\|x^{-1}y\|}{1 - \|x^{-1}y\|}. \tag{5.10}$$

Proof. Just observe $x - y = x(e - x^{-1}y) = (e - yx^{-1})x$. □

For the rest of this section it will be important that X is a complex vector space. The **resolvent set** is defined as

$$\rho(x) := \{\alpha \in \mathbb{C} | (x - \alpha) \text{ is invertible in } X\} \subseteq \mathbb{C}, \tag{5.11}$$

where we have used the shorthand notation $x - \alpha := x - \alpha e$. Its complement is called the **spectrum**

$$\sigma(x) := \mathbb{C} \setminus \rho(x). \tag{5.12}$$

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 5.8. 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 (cf. Problem 4.12). ◇

Example 5.9. 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 $x(t_0) = \alpha$. ◇

Example 5.10. 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 7.22. Hence we again have $\sigma(x) = x([-\pi,\pi])$. ◇

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 (5.9) which implies

$$(x-\alpha)^{-1} = \sum_{n=0}^{\infty} (\alpha-\alpha_0)^n (x-\alpha_0)^{-n-1}, \quad |\alpha-\alpha_0| < \|(x-\alpha_0)^{-1}\|^{-1}. \tag{5.13}$$

This shows that $(x-\alpha)^{-1}$ has a convergent power series with coefficients in X around every point $\alpha_0 \in \rho(x)$. In particular, the resolvent set is open. As in the case of coefficients in $\mathbb{C}$, such functions will be called **analytic**.

Furthermore, 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

$$\|(x-\alpha)^{-1}\| \ge \frac{1}{\operatorname{dist}(\alpha, \sigma(x))} \tag{5.14}$$

as α approaches the spectrum.

Example 5.11. If $A \in \mathscr{L}(\mathbb{C}^n)$ is an n by n matrices, then the resolvent is given by

$$(A-\alpha)^{-1} = \frac{1}{\det(A-\alpha)} (A-\alpha)^{\mathrm{adj}},$$

where A^{adj} denotes the **adjugate** (transpose of the cofactor matrix) of A. Since $(A-\alpha)^{\mathrm{adj}}$ is a polynomial in α, the resolvent has a pole at each eigenvalue whose order is at most the algebraic multiplicity of the eigenvalue. In fact the order of the pole equals the algebraic multiplicity which can be seen using (e.g.) the Jordan[2] canonical form. ◇

[2]Camille Jordan (1838–1922), French mathematician

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 5.3. *For every element x of a Banach algebra X, the spectrum $\sigma(x)$ is compact, nonempty and satisfies*

$$\sigma(x) \subseteq \{\alpha | \, |\alpha| \le \|x\|\}. \tag{5.15}$$

Proof. We already noted that (5.13) shows that $\rho(x)$ is open. Hence $\sigma(x)$ is closed. Moreover, $x - \alpha = -\alpha(e - \frac{1}{\alpha}x)$ together with Lemma 5.1 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. □

You can think of this result as a generalization of the **fundamental theorem of algebra** which in finite dimensions is equivalent to the fact that every matrix has at least one eigenvalue. In particular, note that in the case of real Banach algebras, the spectrum can be empty.

As another simple consequence we obtain:

Theorem 5.4 (Gelfand–Mazur[3])**.** *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

$$p(x) := \sum_{j=0}^{n} p_j x^j. \tag{5.16}$$

In fact, we could easily extend this definition to arbitrary convergent power series whose radius of convergence is larger than $\|x\|$ (cf. Problem 1.59).

[3]Israel Gelfand (1913–2009), Soviet mathematician

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 according to 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 5.5 (Spectral mapping)**.** *Let X be a Banach algebra. For every polynomial p and $x \in X$ we have*

$$\sigma(p(x)) = p(\sigma(x)), \tag{5.17}$$

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 5.5 and hence $p(\alpha) \in \sigma(p(x))$.

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**

$$r(x) := \sup_{\alpha \in \sigma(x)} |\alpha| \tag{5.18}$$

of x. Note that by (5.15) we have

$$r(x) \le \|x\|. \tag{5.19}$$

As our next theorem shows, it is related to the radius of convergence of the **Neumann series** for the resolvent

$$(x - \alpha)^{-1} = -\frac{1}{\alpha} \sum_{n=0}^{\infty} \left(\frac{x}{\alpha}\right)^n \tag{5.20}$$

encountered in the proof of Theorem 5.3 (which is just the Laurent[4] expansion around infinity).

Theorem 5.6 (Beurling–Gelfand)**.** *The spectral radius in a Banach algebra satisfies*

$$r(x) = \inf_{n \in \mathbb{N}} \|x^n\|^{1/n} = \lim_{n \to \infty} \|x^n\|^{1/n}. \tag{5.21}$$

[4]Pierre Alphonse Laurent (1813–1854), French mathematician and engineer

[4]Arne Beurling (1905–1986), Swedish mathematician

Proof. By spectral mapping we have $r(x)^n = r(x^n) \le \|x^n\|$ and hence

$$r(x) \le \inf \|x^n\|^{1/n}.$$

Conversely, fix $\ell \in X^*$, and consider

$$\ell((x-\alpha)^{-1}) = -\frac{1}{\alpha}\sum_{n=0}^{\infty}\frac{1}{\alpha^n}\ell(x^n). \tag{5.22}$$

Then $\ell((x-\alpha)^{-1})$ is analytic in $|\alpha| > r(x)$ and hence (5.22) converges absolutely for $|\alpha| > r(x)$ by Cauchy's formula for the coefficients of the Laurent expansion. 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

$$\frac{\|x^n\|}{|\alpha|^n} \le C(\alpha)$$

and thus

$$\limsup_{n\to\infty} \|x^n\|^{1/n} \le \limsup_{n\to\infty} C(\alpha)^{1/n}|\alpha| = |\alpha|.$$

Since this holds for every $|\alpha| > r(x)$ we have

$$r(x) \le \inf \|x^n\|^{1/n} \le \liminf_{n\to\infty} \|x^n\|^{1/n} \le \limsup_{n\to\infty} \|x^n\|^{1/n} \le r(x),$$

which finishes the proof. □

It might be tempting to conjecture that the sequence $\|x^n\|^{1/n}$ is monotone, however this is false in general – see Problem 5.14. By the ratio test, the Neumann series (5.20) converges for $|\alpha| > r(x)$.

Next, let us look at some examples illustrating these ideas.

Example 5.12. In $X := C(I)$ we have $\sigma(x) = x(I)$ and hence $r(x) = \|x\|_\infty$ for all x. ⋄

Example 5.13. For $X := \mathscr{L}(\mathbb{C}^2)$ and $x := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ we have $x^2 = 0$ 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 $r(x) = 0$. Note that in this case the Neumann series terminates after n terms,

$$(x-\alpha)^{-1} = -\frac{1}{\alpha}\sum_{j=0}^{n-1}\left(\frac{x}{\alpha}\right)^j, \qquad \alpha \ne 0,$$

and the resolvent has a pole of order n at 0.

Elements of a Banach algebra with $r(x) = 0$ are called **quasinilpotent**. Since the Neumann series (5.20) converges for $|\alpha| > 0$ in this case, the resolvent has an essential singularity at 0 if x is quasinilpotent (but not nilpotent).

Example 5.14. Consider the **Volterra integral operator**

$$K(x)(t) := \int_0^t k(t,s)x(s)ds, \qquad x \in C[0,1]. \tag{5.23}$$

Then, using induction, it is not hard to verify (Problem 5.17)

$$|K^n(x)(t)| \le \frac{\|k\|_\infty^n t^n}{n!} \|x\|_\infty. \tag{5.24}$$

Consequently

$$\|K^n x\|_\infty \le \frac{\|k\|_\infty^n}{n!} \|x\|_\infty,$$

that is $\|K^n\| \le \frac{\|k\|_\infty^n}{n!}$, which shows

$$r(K) \le \lim_{n\to\infty} \frac{\|k\|_\infty}{(n!)^{1/n}} = 0.$$

Hence $r(K) = 0$ and for every $\lambda \in \mathbb{C}$ and every $y \in C[0,1]$ the equation

$$x - \lambda K\, x = y \tag{5.25}$$

has a unique solution given by

$$x = (\mathbb{I} - \lambda K)^{-1} y = \sum_{n=0}^{\infty} \lambda^n K^n\, y. \tag{5.26}$$

Note that $\sigma(K) = \{0\}$ but 0 is in general not an eigenvalue (consider e.g. $k(t,s) = 1$ or see Problem 3.13). $\diamond$

In the last two examples we have seen a strict inequality in (5.19). 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 $f \in C(\sigma(x))$. This is nothing but the content of the spectral theorem and self-adjointness will be the property which will make all this work.

Of course for analytic functions f there are no problems since the analytic functional calculus from Problem 1.59 carries over verbatim. In fact, using Cauchy's integral formula one can generalize this approach to the case where f is analytic in a neighborhood of the spectrum. This is known as **holomorphic functional calculus** or Riesz–Dunford[5] calculus.

We end this section with the remark that neither the spectrum nor the spectral radius is continuous. All one can say is

Lemma 5.7. *Let $x_n \in X$ be a convergent sequence in a Banach algebra X and $x := \lim_{n\to\infty} x_n$ its limit. Then whenever $\alpha \in \rho(x)$ we have $\alpha \in \rho(x_n)$ eventually and*

$$(x_n - \alpha)^{-1} \to (x - \alpha)^{-1}. \tag{5.27}$$

Moreover,

$$\lim_{n\to\infty} \sigma(x_n) \subseteq \sigma(x), \tag{5.28}$$

where $\lim_{n\to\infty} \sigma(x_n) := \{\alpha \in \mathbb{C} | \exists \alpha_n \in \sigma(x_n) \to \alpha\}$, and

$$r(x) \geq \limsup_{n\to\infty} r(x_n). \tag{5.29}$$

Proof. The first claim is immediate since taking the inverse is continuous by Corollary 5.2. Furthermore, Corollary 5.2 also shows that for $\alpha \in \rho(x)$ and $\|x - x_n\| + |\alpha - \alpha_n| < \|(x-\alpha)^{-1}\|^{-1}$ we have $\alpha_n \in \rho(x_n)$, which implies the second claim.

Concerning the last claim, observe that $r(x_k) \leq \|x_k^n\|^{1/n}$ implies that $\limsup_{k\to\infty} r(x_k) \leq \|x^n\|^{1/n}$. □

Example 5.15. That the spectrum can expand is shown by the following example due to Kakutani[6]. We consider the bounded linear operators on $\ell^2(\mathbb{N})$ and look at shift-type operators of the form

$$(Aa)_j := q_j a_{j+1},$$

where $q \in \ell^\infty(\mathbb{N})$. Then we have $\|A\| = \sup_{j\in\mathbb{N}} |q_j|$ and

$$(A^n a)_j = (q_j q_{j+1} \cdots q_{j+n-1}) a_{j+n}$$

with $\|A^n\| = \sup_{j\in\mathbb{N}} |q_j q_{j+1} \cdots q_{j+n-1}|$.

Now note that every integer can be written as $j = 2^k(2l+1)$ and write $k(j) := k$ in this case. Choose

$$q_j := \mathrm{e}^{-k(j)}.$$

To compute the above products we group integers into blocks $2^{m-1}, \dots, 2^m - 1$ of 2^{m-1} elements and observe that $K_m := \sum_{j=2^{m-1}}^{2^m-1} k(j) = 2^{m-1} - 1$. Indeed, note that since odd numbers do not contribute to this sum, we can drop them and divide the remaining even ones by 2 to get the previous block.

[5]Nelson Dunford (1906–1986), American mathematician

[6]Shizuo Kakutani (1911–2004), Japanese-American mathematician

This shows $K_m = 2^{m-2} + K_{m-1}$ and establishes the claim. Summing over all blocks we have $\sum_{m=1}^{n} K_m = 2^n - n - 1$ implying

$$\|A^{2^n}\|^{1/2^n} = (q_1 q_2 \cdots q_{2^n-1})^{1/2^n} = \exp(-1 + (n+1)2^{-n}).$$

Taking the limit $n \to \infty$ shows $r(A) = \frac{1}{\mathrm{e}}$.

Next define

$$(A_k a)_j := \begin{cases} 0, & k(j) = k, \\ q_j a_{j+1}, & \text{else}, \end{cases}$$

and observe that A_k is nilpotent since $A_k^{2^{k+1}} = 0$. Indeed note that $(A_k a)_j = 0$ for $j = 2^k, 2^k 3, 2^k 5, \dots$ which are a distance 2^{k+1} apart. Hence applying A_k once more the result will vanish at the previous points as well. Moreover,

$$((A - A_k)a)_j = \begin{cases} q_j a_{j+1}, & k(j) = k, \\ 0, & \text{else}, \end{cases}$$

implying $\|A_k - A\| = \mathrm{e}^{-k}$. Hence we have $A_k \to A$ with $r(A_k) = 0 \to 0 < \mathrm{e}^{-1} = r(x)$ and $\sigma(A_k) = \{0\} \to \{0\} \subsetneq \sigma(A)$. ◇

Problem* 5.1. *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* 5.2. *Suppose that X satisfies all requirements for a Banach algebra except that* (5.3) *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 (5.3) *and $\|e\|_0 = 1$.*

Finally, note that for this construction to work it suffices to assume that multiplication is separately continuous by Problem 4.10.

(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* 5.3 (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* 5.4. *Let X be a Banach algebra with unit e. Show that e is unique. Suppose $x \in X$ has both a right inverse $y \in X$ (i.e., $xy = e$) and a left inverse $z \in X$ (i.e., $zx = e$). Show that $y = z = x^{-1}$.*

Problem* 5.5. *Let X be a unital Banach algebra and $x, y \in X$. 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 5.6. *Let X be a Banach algebra with identity e. Compute $\sigma(e)$.*

Problem 5.7. *Let X be a Banach algebra with identity. Show $\sigma(x^{-1}) = \sigma(x)^{-1}$ if $x \in X$ is invertible.*

Problem 5.8. *An element $x \in X$ from a Banach algebra satisfying $x^2 = x$ is called a* ***projection****. Compute the resolvent and the spectrum of a projection.*

Problem 5.9. *Let $\mathfrak{H}$ be a Hilbert space and $A \in \mathscr{L}(\mathfrak{H})$. Show that $\mathrm{Re}(\langle f, Af\rangle) \geq c\|f\|^2$ implies $\{z|\mathrm{Re}(z) < c\} \subseteq \rho(A)$. (Hint: Lax–Milgram)*

Problem 5.10. *If $X := \mathscr{L}(L^p(I))$, then every $x \in C(I)$ gives rise to a multiplication operator $M_x \in X$ defined as $M_x f := x f$. Show $r(M_x) = \|M_x\| = \|x\|_\infty$ and $\sigma(M_x) = \mathrm{Ran}(x)$.*

Problem 5.11. *If $X := \mathscr{L}(\ell^p(\mathbb{N}))$, $1 \leq p \leq \infty$, then every $m \in \ell^\infty(\mathbb{N})$ gives rise to a multiplication operator $M \in X$ defined as $(Ma)_n := m_n a_n$. Show $r(M) = \|M\| = \|m\|_\infty$ and $\sigma(M) = \overline{\mathrm{Ran}(m)}$.*

Problem 5.12. *Let $X := C_b(\mathbb{R})$ and consider the operator*

$$(A_c f)(x) := f(x - c),$$

where $c \in \mathbb{R} \setminus \{0\}$ is some fixed constant. Find the spectrum of A_c. (Hint: Problem 5.7.)

Problem 5.13. *Can every compact set $K \subset \mathbb{C}$ arise as the spectrum of an element of some Banach algebra?*

Problem* 5.14. *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 5.15. *Let $X := \ell^\infty(\mathbb{N})$. Show $\sigma(x) = \overline{\{x_n\}_{n\in\mathbb{N}}}$. Also show that $r(x) = \|x\|_\infty$ for all $x \in X$.*

Problem 5.16. *Consider the rank-one operator*

$$Ax := x_0'(x) x_0, \qquad x_0 \in X,\ x_0' \in X^*,$$

on some Banach space X. Compute the resolvent, $\sigma(A)$, $\|A\|$, and $r(A)$. When do you have $\|A\| = r(A)$?

Problem* 5.17. *Show* (5.24).

Problem 5.18. *Show that $L^1(\mathbb{R}^n)$ with convolution as multiplication is a commutative Banach algebra without identity (Hint: Lemma 3.20 from* **[40]***).*

Problem 5.19. *Let X be a Banach algebra and $x \in X$. Show the* ***first resolvent identity***

$$(x-\alpha)^{-1} - (x-\beta)^{-1} = (\alpha-\beta)(x-\alpha)^{-1}(x-\beta)^{-1} \\ = (\alpha-\beta)(x-\beta)^{-1}(x-\alpha)^{-1},$$

for $\alpha, \beta \in \rho(x)$.

Problem 5.20. *Let X be a Banach algebra and $x, y \in X$. Show $\sigma(xy)\setminus\{0\} = \sigma(yx)\setminus\{0\}$. (Hint: Find a relation between $(xy-\alpha)^{-1}$ and $(yx-\alpha)^{-1}$.)*

Problem 5.21. *Let X be a Banach algebra and $x, y \in X$. Show $r(xy) \le r(x)r(y)$ and $r(x+y) \le r(x)+r(y)$ provided $xy = yx$. (Hint: For the second claim use the binomial formula and estimate $\|x^n\|$ and $\|y^n\|$.)*

Problem 5.22. *Let X be a Banach space and $D \subseteq \mathbb{C}$ a domain. A function $x : D \to X$ is called* ***weakly analytic*** *if $\ell(x) : D \to \mathbb{C}$ is analytic for all $\ell \in X^*$. Show that a weakly analytic function is analytic and satisfies the Cauchy integral formula*

$$x(z) = \frac{1}{2\pi\mathrm{i}} \int_\Gamma \frac{x(\zeta)}{\zeta - z} d\zeta,$$

where $\Gamma \subset D$ is a circle containing z and the integral is defined as the limit of the corresponding Riemann sums. (Hint: First use the uniform boundedness principle to show that $\|x(z)\|$ is bounded on compact sets. Then use the Cauchy integral formula for $\ell(x(z))$ and Lemma 4.29 (v) to show that $x(z)$ is (norm) continuous. Hence the integral can be defined as a limit of Riemann sums and we can interchange ℓ with the integral.)

5.2. 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 complex Banach algebra X together with an **involution** satisfying

$$(x+y)^* = x^* + y^*, \quad (\alpha x)^* = \alpha^* x^*, \quad x^{**} = x, \quad (xy)^* = y^* x^*, \tag{5.30}$$

and

$$\|x\|^2 = \|x^* x\| \tag{5.31}$$

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.

Example 5.16. The continuous functions $C(I)$ together with complex conjugation form a commutative C^* algebra. $\diamond$

Example 5.17. The Banach algebra $\mathscr{L}(\mathfrak{H})$ is a C^* algebra by Lemma 2.15. The compact operators $\mathscr{K}(\mathfrak{H})$ are a $*$-subalgebra. $\diamond$

Example 5.18. 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.

However, the set $\ell^p(\mathbb{N})$, $1 \le p < \infty$ is not a C^* algebra as it fails (5.31). It satisfies $\|a\|_p = \|a^*\|_p$ though. $\diamond$

Note that (5.31) 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

$$\|x\| = \|x^*\|, \qquad \|x\|^2 = \|x^*x\| = \|xx^*\|. \tag{5.32}$$

The condition (5.31) 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 $\mathscr{L}(\mathfrak{H})$. Furthermore, it is important to emphasize that (5.31) is a rather strong condition as it implies that the norm is already uniquely determined by the algebraic structure.

Example 5.19. Recall that the norm of a matrix $A \in \mathscr{L}(\mathbb{C}^n)$ can be computed as the largest eigenvalue of A^*A (cf. (3.41)). $\diamond$

Indeed, Lemma 5.8 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.

If X has an identity e, we clearly have $e^* = e$, $\|e\| = 1$, $(x^{-1})^* = (x^*)^{-1}$ (show this), and

$$\sigma(x^*) = \sigma(x)^*. \tag{5.33}$$

We will always assume that we have an identity and we note that it is always possible to add an identity (Problem 5.23).

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 $y = y^* \in X$ (the connection with our definition for $\mathscr{L}(\mathfrak{H})$ from Section 2.3 will be established in Problem 5.27). 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 (5.31) 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 5.8. *Let X be a C^* algebra. If $x \in X$ is normal, then $\|x^2\| = \|x\|^2$ and $r(x) = \|x\|$.*

Proof. Using (5.31) 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 5.9. *Let X be a C^* algebra. If $x \in X$ 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 5.5). □

Example 5.20. 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

$$C^*(x) = \overline{\{p(x, x^*) | p : \mathbb{C}^2 \to \mathbb{C} \text{ polynomial}\}}, \qquad xx^* = x^* x, \tag{5.34}$$

and, in particular, $C^*(x)$ is commutative (Problem 5.24). In the self-adjoint case this simplifies to

$$C^*(x) := \overline{\{p(x) | p : \mathbb{C} \to \mathbb{C} \text{ polynomial}\}}, \qquad x = x^*. \tag{5.35}$$

Moreover, in this case $C^*(x)$ is isomorphic to $C(\sigma(x))$ (the continuous functions on the spectrum):

Theorem 5.10 (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

$$\sigma(f(x)) = f(\sigma(x)), \tag{5.36}$$

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.28 to extend f to I and then approximate the extension using polynomials (Theorem 1.3) or use the Stone–Weierstraß theorem (Theorem B.42). Thus Φ uniquely extends to a map on all of $C(\sigma(x))$ by Theorem 1.16. 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 1.59).

We remark that the spectral theorem holds for normal operators and we will establish this in Theorem 7.26.

Problem* 5.23 (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* 5.24. *Let X be a C^* algebra and Y a $*$-subalgebra. Show that if Y is commutative, then so is $\overline{Y}$.*

Problem 5.25. *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 5.26. *Let X be a C^* algebra and $x \in X$ 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 5.27. *Show that an operator $A \in \mathscr{L}(\mathfrak{H})$ is positive as defined in Section 2.3 (i.e., $\langle f, Af\rangle \ge 0$) if and only if it is positive in the sense of C^* algebras. (Hint: One direction is easy. For the other direction use Problem 5.26 and Theorem 2.17.)*

Problem 5.28. *Let $A \in \mathscr{L}(\mathfrak{H})$. Show that A is normal if and only if*

$$\|Af\| = \|A^* f\|, \quad \forall f \in \mathfrak{H}.$$

In particular, $\mathrm{Ker}(A) = \mathrm{Ker}(A^*)$. *(Hint: Problem 1.32.)*

Problem 5.29. *Let X be a C^* algebra. Show that the* ***Cayley transform*** *of a self-adjoint element $x \in X$,*

$$y := (x - \mathrm{i})(x + \mathrm{i})^{-1}$$

is unitary. Show that $1 \notin \sigma(y)$ and

$$x = \mathrm{i}(1 + y)(1 - y)^{-1}.$$

Problem 5.30. *Let X be a C^* algebra. Show if $x \in X$ is unitary then $\sigma(x) \subseteq \{\alpha \in \mathbb{C} | |\alpha| = 1\}$.*

Problem 5.31. *Let X be a C^* algebra and suppose that $x \in X$ is self-adjoint. Show that*

$$\|(x - \alpha)^{-1}\| = \frac{1}{\mathrm{dist}(\alpha, \sigma(x))},$$

with the convention that $1/0 = \infty$ and that $\|(x - \alpha)^{-1}\| = \infty$ if $x - \alpha$ is not invertible.

Problem 5.32. *Show that a C^* algebra (with identity) is one-dimensional if and only if the spectrum of every self-adjoint element consists of one point. Conclude that the norm of a C^* algebra (with identity) can only stem from a scalar product if it is one-dimensional. (Hint: Look at the $*$-subalgebra of a suitable self-adjoint element.)*

5.3. 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 5.11. *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*

$$\langle u, f(A)v\rangle = \int_{\sigma(A)} f(t) d\mu_{u,v}(t), \qquad f \in C(\sigma(A)). \tag{5.37}$$

We have

$$\mu_{u,v_1+v_2} = \mu_{u,v_1} + \mu_{u,v_2}, \quad \mu_{u,\alpha v} = \alpha\mu_{u,v}, \quad \mu_{v,u} = \mu_{u,v}^* \tag{5.38}$$

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. For fixed $u, v \in \mathfrak{H}$ the map $f \mapsto \ell_{u,v}(f) := \langle u, f(A)v\rangle$ is a bounded linear functional on $C(I)$ by Theorem 5.10. Hence existence of a corresponding measure $\mu_{u,v}$ with $|\mu_{u,v}|(I) = \|\ell_{u,v}\| \le \|u\|\|v\|$ follows from the Riesz representation theorem (Theorem 6.5 from [**40**]). 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

$$\mu_{u,v}(\Omega) = \frac{1}{4}(\mu_{u+v}(\Omega) - \mu_{u-v}(\Omega) + \mathrm{i}\mu_{u-\mathrm{i}v}(\Omega) - \mathrm{i}\mu_{u+\mathrm{i}v}(\Omega)). \tag{5.39}$$

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 5.12 (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*

$$\langle u, f(A)v\rangle = \int_{\sigma(A)} f(t) d\mu_{u,v}(t), \qquad f \in B(\sigma(A)). \tag{5.40}$$

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

$$\Big| \int_{\sigma(A)} f(t) d\mu_{u,v}(t) \Big| \le \|f\|_\infty |\mu_{u,v}|(\sigma(A)) \le \|f\|_\infty \|u\| \|v\|$$

and (5.38). 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. $\square$

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**

$$P_A(\Omega) = \chi_\Omega(A), \qquad \Omega \in \mathfrak{B}(\mathbb{R}), \tag{5.41}$$

such that

$$\mu_{u,v}(\Omega) = \langle u, P_A(\Omega) v\rangle. \tag{5.42}$$

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

$$\mathfrak{H} = \mathrm{Ker}(P) \oplus \mathrm{Ran}(P), \tag{5.43}$$

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

In addition, the spectral projections satisfy

$$P_A(\mathbb{R}) = \mathbb{I}, \quad P_A(\dot{\bigcup_{n=1}^{\infty}} \Omega_n) u = \sum_{n=1}^{\infty} P_A(\Omega_n) u, \quad \Omega_n \cap \Omega_m = \emptyset,\ m \ne n, \tag{5.44}$$

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 $\chi_{\Omega_1 \dot\cup \Omega_2} = \chi_{\Omega_1} + \chi_{\Omega_2}$ 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 $\sum_{n=1}^N \chi_{\Omega_n} = \chi_{\dot\bigcup_{n=1}^N \Omega_n} \to \chi_{\dot\bigcup_{n=1}^\infty \Omega_n}$ 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

$$P_A(\sigma(A)) = \mathbb{I}. \tag{5.45}$$

I also remark that in this connection the corresponding distribution function

$$P_A(t) := P_A((-\infty, t]) \tag{5.46}$$

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

$$\int_{\mathbb{R}} f(t) dP_A(t) := \sum_{j=1}^n \alpha_j P_A(\Omega_j). \tag{5.47}$$

By (5.42) we conclude that this definition agrees with $f(A)$ from Theorem 5.12:

$$\int_{\mathbb{R}} f(t) dP_A(t) = f(A). \tag{5.48}$$

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

$$A = \int t\, dP_A(t). \tag{5.49}$$

If $P_A(\{\alpha\}) \neq 0$, then α is an eigenvalue and $\operatorname{Ran}(P_A(\{\alpha\}))$ is the corresponding eigenspace (Problem 5.34). The fact that eigenspaces to different eigenvalues are orthogonal now generalizes to

Lemma 5.13. *Suppose* $\Omega_1 \cap \Omega_2 = \emptyset$. *Then*

$$\operatorname{Ran}(P_A(\Omega_1)) \perp \operatorname{Ran}(P_A(\Omega_2)). \tag{5.50}$$

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. $\square$

Example 5.21. Let $A \in \mathscr{L}(\mathbb{C}^n)$ be some self-adjoint 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 $\mathrm{Ker}(A - \alpha_j)$ corresponding to the eigenvalue α_j by Problem 5.34. 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 5.22. Let $A \in \mathscr{L}(\mathfrak{H})$ 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 (5.42) we have $\mu_{u,v}(\{\alpha\}) = \langle u, Pv\rangle = 0$ for $u, v \in \tilde{\mathfrak{H}}$. Hence the integral in (5.40) 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 $\alpha \in \rho(\tilde{A})$. ⋄

Example 5.23. Let us consider a finite probability measure μ on $\mathfrak{B}(\mathbb{R})$ such that all monomials are integrable with respect to this measure, that is, $\int_{\mathbb{R}} |t|^n d\mu(t) < \infty$ for all $n \in \mathbb{N}$. For example μ could be given by a density $d\mu(t) = m(t)dt$. Moreover, to obtain a bounded operator, we will also assume that μ is supported on a compact interval.

Consider the orthonormal polynomials $p_n(t)$ with respect to this measure, which can be obtained using the Gram–Schmidt procedure with respect to the scalar product

$$\langle p, q\rangle = \int_{\mathbb{R}} p(t)^* q(t)d\mu(t).$$

Note that p_n will have real coefficients with the highest coefficient being positive. Then, expanding the polynomial $t\,p_n(t)$ with respect to this orthonormal basis shows

$$t\,p_n(t) = a_n p_{n+1}(t) + b_n p_n(t) + a_{n-1} p_{n-1}(t), \quad n \in \mathbb{N},$$

where

$$a_n := \langle p_{n+1}, t\,p_n \rangle, \qquad b_n := \langle p_n, t\,p_n \rangle.$$

Concerning the third coefficient observe that $\langle p_{n-1}, t\,p_n \rangle = \langle t\,p_{n-1}, p_n \rangle = \langle p_n, t\,p_{n-1} \rangle$. Moreover, all remaining coefficients vanish since $\langle p_j, t\,p_n \rangle = \langle t\,p_j, p_n \rangle = 0$ for $j < n-1$ by orthogonality. Finally, $a_n > 0$ since the highest coefficient of p_n is positive.

Next denote by $\mathfrak{H}$ the closure of the set of polynomials with respect to the above scalar product. Using the Weierstraß theorem (Theorem 1.3) one can show that $\mathfrak{H}$ can be identified with the L^2 space with respect to this measure, but we won't need this here. Next, consider the unitary map

$$U : \ell^2(\mathbb{N}_0) \to \mathfrak{H}, \qquad u \mapsto \sum_{n=0}^{\infty} u_n p_n(t).$$

Then, by construction, multiplication with t in $\mathfrak{H}$ (which is bounded since we assumed the support of μ bounded) corresponds to the **Jacobi operator**

$$(Ju)_n := \begin{cases} a_n u_{n+1} + b_n u_n + a_{n-1} u_{n-1}, & n > 0 \\ a_0 u_1 + b_0 u_0, & n = 0. \end{cases}$$

Conversely, if we start with some bounded real sequences $a_n > 0$ and b_n, then J is a bounded self-adjoint operator and we can define the polynomials p_n recursively using $p_0 = 1$. Moreover, let μ be the spectral measure associated with δ^0:

$$\langle \delta^0, f(J)\delta^0 \rangle = \int_{\mathbb{R}} f(f) d\mu(t).$$

Then one has $p_n(J)\delta^0 = \delta^n$. Indeed, the case $n = 0$ is trivial and the general case follows from induction

$$\begin{aligned} a_n p_{n+1}(J)\delta^0 &= J p_n(J)\delta^0 - b_n p_n(J)\delta^0 - a_{n-1} p_{n-1}(J)\delta^0 \\ &= J\delta^n - b_n \delta^n - a_{n-1}\delta^{n-1} = a_n \delta^{n+1}. \end{aligned}$$

Consequently, the polynomials p_n are orthonormal with respect to μ and we have established the converse.

This solves the corresponding **Hamburger moment problem**: Given the moments $m_n := \int_{\mathbb{R}} t^n d\mu(t)$, find a corresponding measure. Indeed observe that the moments are all that is needed to construct the polynomials p_n via Gram–Schmidt (provided the moments give rise to a positive sesquilinear form). $\diamond$

It should be mentioned, that the present section can be extended to the case of unbounded operators. This is however beyond our present scope and we refer to [**34**] for further details.

Problem 5.33. *Suppose $A \in \mathscr{L}(\mathfrak{H})$ 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* 5.34. *Suppose $A \in \mathscr{L}(\mathfrak{H})$ is self-adjoint. Show*

$$\mathrm{Ran}(P_A(\{\alpha\})) = \mathrm{Ker}(A - \alpha).$$

(Hint: Start by verifying $\mathrm{Ran}(P_A(\{\alpha\})) \subseteq \mathrm{Ker}(A - \alpha)$. To see the converse, let $u \in \mathrm{Ker}(A - \alpha)$ and use the previous example.)

Part 2

Advanced Functional Analysis

Chapter 6

More on convexity

6.1. 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**[1] (or **gauge**)

$$p_U(x) := \inf\{t > 0 | x \in t\,U\}. \tag{6.1}$$

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. Also observe that for $U \subseteq V$ we have $p_V(x) \leq p_U(x)$ for all $x \in X$.

Example 6.1. 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. ⋄

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

[1] Hermann Minkowski (1864 –1909), German mathematician

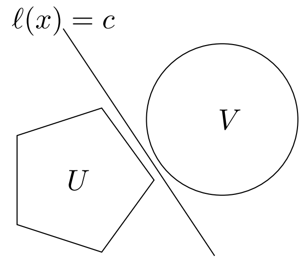


Figure 6.1. Separation of convex sets via a hyperplane

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 6.1. *Let X be a vector space and U a convex subset containing 0. Then*

$$p_U(x+y) \le p_U(x) + p_U(y), \qquad p_U(\lambda\, x) = \lambda p_U(x), \qquad \lambda \ge 0. \tag{6.2}$$

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, suppose $p_U(x)$ and $p_U(y)$ are both finite (otherwise there is nothing to do) and 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 $x = t(t^{-1}x) + (1-t)0 \in U$ by convexity. Similarly, if $x \in U$ then $t^{-1}x = t^{-1}x + (1-t^{-1})0 \in U$ 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 $p(x) \le (1+\varepsilon)^{-1}$. $\square$

Note that (6.2) implies convexity

$$p_U(\lambda x + (1-\lambda)y) \le \lambda p_U(x) + (1-\lambda)p_U(y), \qquad \lambda \in [0,1]. \tag{6.3}$$

Theorem 6.2 (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*

$$\ell(x) < c \le \ell(y), \qquad x \in U,\ y \in V. \tag{6.4}$$

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 (Theorem 4.11) 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, given $\varepsilon > 0$ we consider $W_\varepsilon := (\varepsilon W) \cap (-\varepsilon W)$ which is an open neighborhood of 0. Then, for $x \in W_\varepsilon$ we have $p_W(\pm x) \le \varepsilon$ implying $|\ell(x)| \le \varepsilon$. This shows that ℓ is continuous at 0 and hence continuous everywhere by linearity.

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 $x_1 \in U$ 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 $y \in V$. □

This result is illustrated in Figure 6.1. Of course there is also a complex version.

Theorem 6.3 (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 continuous linear functional $\ell \in X^*$ and some $c \in \mathbb{R}$ such that*

$$\operatorname{Re}(\ell(x)) < c \le \operatorname{Re}(\ell(y)), \qquad x \in U,\ y \in V. \tag{6.5}$$

If in addition V is open, then the second inequality is also strict.

Proof. Consider X as a real topological vector 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 6.2. The assumption that one set is open is crucial. If you let $X := \mathbb{R}^2$ and consider $U_0 := \{x \in \mathbb{R}^2 | x_1 > 0\}$, $V := \{0\}$. Then both sets can be separated as in the theorem using $\ell(x) := -x_1$ (and $c = 0$). However, if we consider $U := U_0 \cup \{(0, x_2) | x_2 > 0\}$ this is no longer possible. ⋄

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 6.4. For now we just remark that every normed vector space is locally convex since balls are convex.

Corollary 6.4. *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*

$$\mathrm{Re}(\ell(x)) \le d < c \le \mathrm{Re}(\ell(y)), \qquad x \in U,\ y \in V. \tag{6.6}$$

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 $\mathrm{Re}(\ell(x)) < c \le \mathrm{Re}(\ell(y))$ for all $x \in \tilde{U}$ and $y \in V$. Moreover, since $\mathrm{Re}(\ell(U))$ is a compact interval $[e, d]$, the claim follows. □

Note that if U is **absolutely convex**, that is, $\alpha U + \beta U \subseteq U$ for $|\alpha| + |\beta| \le 1$, then we can write the previous condition equivalently as

$$|\ell(x)| \le d < c \le |\ell(y)|, \qquad x \in U,\ y \in V, \tag{6.7}$$

since $x \in U$ implies $\theta x \in U$ for $\theta = \mathrm{sign}(\ell(x))$ and thus $|\ell(x)| = \theta\ell(x) = \ell(\theta x) = \mathrm{Re}(\ell(\theta x))$.

From the last corollary we can also obtain versions of Corollaries 4.15 and 4.13 for locally convex vector spaces.

Corollary 6.5. *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 6.4 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 $\mathrm{Re}(\ell(x_0)) < 0$. Finally $\ell(x_0)^{-1}\ell$ is the required functional. □

Corollary 6.6. *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 6.5 gives a functional $\bar{\ell} \in X^*$ 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 $x - \ell(x)x_0 \in \mathrm{Ker}(\ell) \subseteq Y_0$. □

Problem 6.1. *Suppose $\alpha \in \mathbb{C} \setminus \{0\}$. Show that $\alpha U = U$ implies $p_U(\alpha x) = p_U(x)$. Show that if U is symmetric, $U = -U$, and convex, then p_U satisfies the inverse triangle inequality $|p_U(x) - p_U(y)| \le p_U(x - y)$.*

Problem 6.2. *Show that in a Banach space $B_r(0) \subseteq U$ implies $p_U(x) \le \frac{1}{r}\|x\|$.*

Problem* 6.3. *Let X be a topological vector space. Show that $U \pm V$ is convex if both sets are convex and open if one of the sets is open.*

Problem 6.4. *Show that Corollary 6.4 fails even in $\mathbb{R}^2$ unless one set is compact.*

Problem 6.5. *Let X be a topological vector space and $M \subseteq X$, $N \subseteq X^*$. Then the corresponding* ***polar, prepolar sets*** *are*

$$M^\circ := \{\ell \in X^* \,|\, |\ell(x)| \le 1 \,\forall x \in M\}, \quad N_\circ := \{x \in X \,|\, |\ell(x)| \le 1 \,\forall \ell \in N\},$$

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 6.6 (Bipolar theorem)**.** *Let X be a locally convex space and suppose $M \subseteq X$ is absolutely convex. Show $(M^\circ)_\circ = \overline{M}$. (Hint: Use Corollary 6.4 to show that for every $y \notin \overline{M}$ there is some $\ell \in X^*$ with $\mathrm{Re}(\ell(x)) \le 1 < \ell(y)$, $x \in \overline{M}$.)*

6.2. 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 6.7) that the convex hull is given by

$$\mathrm{conv}(U) := \{\sum_{j=1}^n \lambda_j x_j | n \in \mathbb{N},\ x_j \in U,\ \sum_{j=1}^n \lambda_j = 1,\ \lambda_j \geq 0\}. \tag{6.8}$$

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 set 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: Given $x, y \in K$ and $\lambda \in (0,1)$ we have that

$$\lambda x + (1-\lambda) y \in M \quad \Rightarrow x, y \in M. \tag{6.9}$$

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 and M is extremal, then $K \setminus M$ is convex. Conversely, if K and $K \setminus \{x\}$ are convex, then x is an extremal point. 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 6.8).

Example 6.3. 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 1.19 says that the extremal points of the unit ball are precisely its boundary points. ◇

Example 6.4. 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 6.7. *Let X be locally convex vector space. Suppose $K \subseteq X$ and $\ell \in X^*$. If*

$$K_\ell := \{x \in K | \mathrm{Re}(\ell(x)) = \inf_{y \in K} \mathrm{Re}(\ell(y))\}$$

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 $H_\ell := \{x \in X | \mathrm{Re}(\ell(x)) = \inf_{y\in K} \mathrm{Re}(\ell(y))\}$ 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 6.5. 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). ◇

Example 6.6. 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 $\|x\| \leq \lambda\|y\| + (1-\lambda)\|z\| \leq \|x\|$ shows that we have equality in the triangle inequality and hence $x = y = z$ by Problem 1.19 (i). ◇

Example 6.7. In a not strictly convex space the situation is quite different. For example, consider the closed unit ball in $\ell^\infty(\mathbb{N})$. Let $a \in \ell^\infty(\mathbb{N})$ with $\|a\|_\infty = 1$. 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 $\varepsilon \leq 1 - |a_j|$. 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 $1 = |\lambda b_n + (1-\lambda)c_n| \leq \lambda|b_n| + (1-\lambda)|c_n| \leq 1$ and hence $a_n = b_n = c_n$ by strict convexity of the absolute value. 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 6.10) this will show that $c_0(\mathbb{N})$ is not the dual of any Banach space. ◇

Lemma 6.8 (Krein–Milman[2]). *Let X be a locally convex Hausdorff 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 $x \neq y$ then $\{x\}$, $\{y\}$ are two disjoint and compact sets to which we can apply Corollary 6.4 to obtain a linear functional $\ell \in X^*$ with $\mathrm{Re}(\ell(x)) \neq \mathrm{Re}(\ell(y))$. Then by Lemma 6.7 $M_\ell \subset M$ is extremal in M and hence also in K. But by $\mathrm{Re}(\ell(x)) \neq \mathrm{Re}(\ell(y))$ 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 6.11) the closed convex hull is simply the closure of the convex hull.

Theorem 6.9 (Krein–Milman). *Let X be a locally convex Hausdorff 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 $M := \overline{\mathrm{conv}(E)} \subseteq K$ be its closed convex hull. Suppose $x \in K \setminus M$ and use Corollary 6.4 to choose a linear functional $\ell \in X^*$ with

$$\min_{y \in M} \mathrm{Re}(\ell(y)) > \mathrm{Re}(\ell(x)) \geq \min_{y \in K} \mathrm{Re}(\ell(y)).$$

Now consider K_ℓ from Lemma 6.7 which is nonempty and hence contains an extremal point $y \in E$. But $y \notin M$, a contradiction. □

[2]Mark Krein (1907–1989), Soviet mathematician

[2]David Milman (1912–1982), Soviet mathematician

While in the finite dimensional case the closure is not necessary (Problem 6.13), it is important in general as the following example shows.

Example 6.8. 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 $1 = |\lambda b_n + (1-\lambda)c_n| \le \lambda|b_n| + (1-\lambda)|c_n| \le 1$ and hence $a_n = b_n = c_n$ by strict convexity of the absolute value. 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 6.10). To this end note that $\ell^1(\mathbb{N}) \cong c_0(\mathbb{N})^*$. $\diamond$

Also note that in the infinite dimensional case the extremal points can be dense.

Example 6.9. 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 hull 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 1.13) $\bar{K}$ 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 $\|f - f_n\|_\infty \le \frac{1}{n}$. $\diamond$

Problem* 6.7. *Show that the convex hull is given by* (6.8).

Problem* 6.8. *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 6.9. *Show that the closed unit ball in $L^1(0,1)$ has no extremal points.*

Problem 6.10. *Find the extremal points of the closed unit ball in $C([0,1],\mathbb{R})$.*

Problem 6.11. *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.29.)*

Problem 6.12. *Let X be a separable Banach space. 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 $\{y|\mathrm{Re}(\ell(y-x)) \le 0\}$. (Hint: Lemma 4.36.)*

Problem 6.13 (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.)*

6.3. Weak topologies

In Section 4.5 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 a Banach space 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

$$x + \bigcap_{j=1}^{n} |\ell_j|^{-1}\big([0,\varepsilon_j)\big) = \bigcap_{j=1}^{n} \{\tilde{x} \in X | |\ell_j(x-\tilde{x})| < \varepsilon_j,\ 1 \le j \le n\},$$
$$x \in X,\ \ell_j \in X^*,\ \varepsilon_j > 0. \qquad (6.10)$$

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 4.14) the weak topology is Hausdorff. Moreover, the sets in the above base are clearly convex and hence we even have a locally convex space.

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

$$d(x,\tilde{x}) := \sum_{n=1}^{\infty} \frac{1}{2^n} |\ell_n(x-\tilde{x})| \tag{6.11}$$

defines a metric on the unit ball $B_1(0) \subset X$ which can be shown to generate the weak topology (Problem 6.17). However, on all of X the weak topology is not first countable unless X is finite dimensional (Problem 6.18).

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. Like the weak topology it is Hausdorff (since different linear functionals must differ at least at one point) and not first countable unless X is finite dimensional (Problem 6.18). It also turns out that the continuous linear functionals with respect to the weak-$*$ topology are precisely $J(X)$ (Problem 6.19).

A base for the weak-$*$ topology is given by sets of the form

$$\ell + \bigcap_{j=1}^{n} |J(x_j)|^{-1}\big([0,\varepsilon_j)\big) = \bigcap_{j=1}^{n} \{\tilde{\ell} \in X^* | |(\ell-\tilde{\ell})(x_j)| < \varepsilon_j,\ 1 \le j \le n\},$$
$$\ell \in X^*,\ x_j \in X,\ \varepsilon_j > 0. \tag{6.12}$$

Again these sets are convex and we have a locally convex space. 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

$$d(\ell,\tilde{\ell}) := \sum_{n=1}^{\infty} \frac{1}{2^n} |(\ell-\tilde{\ell})(x_n)| \tag{6.13}$$

defines a metric on the unit ball $B_1^*(0) \subset X^*$ which can be shown to generate the weak-$*$ topology (Problem 6.17). Hence Lemma 4.36 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 6.10 (Banach–Alaoglu[3])**.** *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, $|\Phi(\ell)(x)| \le \|\ell\|\|x\|$ implies $\Phi(B^*) \subset \bigtimes_{x\in X} B_x$, where the last product is

[3]Leonidas Alaoglu (1914–1981), Canadian-American mathematician

compact by Tychonoff's theorem (Theorem B.32). Hence it suffices to show that $\Phi(B^*)$ is closed. To this end let $l \in \overline{\Phi(B^*)}$. We need to show that $l \in \Phi(B^*)$. Fix $x_1, x_2 \in X$, $\alpha \in \mathbb{C}$, and consider the open neighborhood

$$U(l) = \left\{ h \in \mathbb{C}^X \middle| \begin{array}{l} |h(x_1 + \alpha x_2) - l(x_1 + \alpha x_2)| < \varepsilon, \\ |h(x_1) - l(x_1)| < \varepsilon,\ |\alpha||h(x_2) - l(x_2)| < \varepsilon \end{array} \right\}$$

of l. Since $U(l) \cap \Phi(B^*)$ 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 $l \in \Phi(B^*)$. □

Please note that Example 4.47 shows that $\bar{B}_1^*(0) \subset X^*$ might not be sequentially compact in the weak-$*$ topology unless X is separable.

If X is a reflexive space and we apply this to X^*, we get that the closed unit ball is compact in the weak topology. In fact, the converse is also true.

Theorem 6.11 (Kakutani). *A Banach space X is reflexive if and only if the closed unit ball $\bar{B}_1(0)$ is weakly compact.*

Proof. Suppose X is not reflexive and choose $x'' \in \bar{B}_1^{**}(0) \setminus J(\bar{B}_1(0))$ with $\|x''\| = 1$. Then, if $\bar{B}_1(0)$ is weakly compact, $J(\bar{B}_1(0))$ is weak-$*$ compact (note that J is a homeomorphism if we equip X with the weak and X^{**} with the weak-* topology). Moreover, if we equip X^{**} with the weak-* topology, then its dual is isomorphic to X^* (Problem 6.19) and by Corollary 6.4 we can find some $\ell \in X^*$ with $\|\ell\| = 1$ and

$$\mathrm{Re}(x''(\ell)) < \inf_{y'' \in J(\bar{B}_1(0))} \mathrm{Re}(y''(\ell)) = \inf_{y \in \bar{B}_1(0)} \mathrm{Re}(\ell(y)) = -1.$$

But this contradicts $|x''(\ell)| \le 1$. □

Note that in this context Theorem 4.32 says that in a reflexive Banach space X the closed unit ball $\bar{B}_1(0)$ is weakly sequentially compact. The converse is also true and known as the Eberlein theorem.

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 6.12 (Mazur). *The weak as well as the norm closure of a convex set K in a Banac space X is the intersection of all half-spaces containing K. In particular, for a convex set $K \subseteq X$ the following are equivalent:*

- *K is weakly closed,*

- *K is weakly sequentially closed,*
- *K is (norm) 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. Furthermore, 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 6.3 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 $x \notin \mathrm{Re}(\ell)^{-1}([c,\infty))$.

For the last claim note that a weakly closed set is weakly sequentially closed (Example B.21). Moreover, a weakly sequentially closed set is also sequentially closed and hence closed. Finally, a closed convex set is weakly closed by the first part. □

Example 6.10. 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 theorem 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 $x_0 \in \bigcap_{j=1}^n \mathrm{Ker}(\ell_j)$ (by Problem 4.27 the ℓ_j would otherwise be a basis) 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 $\bar{B}_1(0) \subseteq \overline{S}^w$. ◇

Example 6.11. Consider the subset $M := \{\sqrt{n}\delta^n\}_{n\in\mathbb{N}} \subset \ell^2(\mathbb{N})$ which is is sequentially weakly closed (since weakly convergent sequences are bounded). It is however not weakly closed. To see this, observe that any neighborhood of 0 intersects M. Indeed, any neighborhood of 0 contains a neighborhood of the form

$$\{a \in \ell^2(\mathbb{Z}) | |\langle b_j, a\rangle| < \varepsilon,\ 1 \le j \le n\}.$$

Now the sequence $b_k := \max_{1\le j\le n} |b^j_k|$ is square summable and hence there is some n such that $\sqrt{n} b_n = \max_{1\le j\le n} |\langle b^j, \sqrt{n}\delta^n\rangle| < \varepsilon$. ◇

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 6.13 (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 $\lambda_{k,j} \ge 0$) such that $y_k \to x$.*

Proof. Let $K := \{\sum_{j=1}^n \lambda_j x_j | n \in \mathbb{N},\ \sum_{j=1}^n \lambda_j = 1,\ \lambda_j \ge 0\}$ be the convex hull of the points $\{x_k\}$. Then by the previous result $x \in \overline{K}$. □

Example 6.12. 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 $\|\psi_j\| = j^{-1/2}$. $\diamond$

For the last result 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 6.14 (Goldstine[4]). *Let X be a Banach space. 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 $\{\tilde{\jmath} \in X^{**} \mid |j(\ell_k) - \tilde{\jmath}(\ell_k)| < \varepsilon, 1 \le k \le n\}$ 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 $1 \le k \le n$ since then $(1+\varepsilon)^{-1}J(x)$ will be in the above neighborhood. Without the requirement $\|x\| \le 1+\varepsilon$ existence of such an x 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 $Y := \bigcap_k \mathrm{Ker}(\ell_k)$. But if $(x+Y) \cap \bar{B}_{1+\varepsilon}(0)$ were empty, we would have $\mathrm{dist}(x, Y) \ge 1+\varepsilon$ and by Corollary 4.15 we could find some normalized $\ell \in X^*$ which vanishes on Y and satisfies $\ell(x) \ge 1+\varepsilon$. By Problem 4.27 we have $\ell \in \mathrm{span}(\ell_1, \dots, \ell_n)$ implying

$$1+\varepsilon \le \ell(x) = j(\ell) \le \|j\|\|\ell\| \le 1$$

a contradiction. $\square$

Note that if $\bar{B}_1(0) \subset X$ is weakly compact, then $J(\bar{B}_1(0))$ is compact (and thus closed) in the weak-$*$ topology on X^{**}. Hence Goldstine's theorem implies $J(\bar{B}_1(0)) = \bar{B}_1^{**}(0)$ and we get an alternative proof of Kakutani's theorem.

Example 6.13. 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 4.50). 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).

[4] Herman Goldstine (1913 –2004), mathematician and computer scientist

Also observe that $c_0(\mathbb{N}) \subseteq \ell^\infty(\mathbb{N})$ is closed but not weak-$*$ closed and hence Mazur's theorem does not hold if we replace the weak by the weak-$*$ topology. ⋄

Problem 6.14. *Show that in an infinite dimensional space, a weakly open neighborhood of* 0 *contains a nontrivial subspace. Show the analogue statement for weak-$*$ open neighborhoods of* 0.

Problem 6.15. *Show that a weakly sequentially compact set is bounded. Similarly, show that a weakly compact set is bounded.*

Problem 6.16 (von Neumann). *Consider $X := \ell^2(\mathbb{N})$ with the canonical basis δ^n. Let $A := \{\delta^n + n\delta^m\}_{m,n\in\mathbb{N}}$. Show that the weak sequential closure of A contains $\{\delta^n\}_{n\in\mathbb{N}}$ but not* 0*. Hence the weak sequential closure of A is not weakly sequentially closed.*

Problem* 6.17. *Show that* (6.11) *generates the weak topology on $B_1(0) \subset X$. Show that* (6.13) *generates the weak-$*$ topology on $B_1^*(0) \subset X^*$.*

Problem 6.18. *Show that neither the weak nor the weak-$*$ topology is first countable if X is infinite dimensional. (Hint: If there is a countable neighborhood base, you can find, using Problem 6.14, an unbounded sequence of vectors which converge weakly to zero.)*

Problem* 6.19. *Show that set of weak-$*$ continuous liner functionals on X^* equals $J(X)$. (Hint: Use that a continuous linear functional is bounded and Problem 4.27.)*

Problem* 6.20. *Show that the annihilator $M^\perp$ of a set $M \subseteq X$ is weak-$*$ closed. Moreover show that $(N_\perp)^\perp = \overline{\operatorname{span}(N)}^{\text{weak-}*}$. (Hint: The first part and hence one inclusion of the second part are straightforward. For the other inclusion use Problem 6.19 and apply Corollary 6.5.)*

6.4. 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 6.4 (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

$$x + \bigcap_{j=1}^{n} q_{\alpha_j}^{-1}([0,\varepsilon_j)) \tag{6.14}$$

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 $\varepsilon := \min_{1\le j\le n} \frac{\varepsilon_j}{2q_{\alpha_j}(x)}$.

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 $\alpha \in A$.

Example 6.14. Of course every Banach space equipped with the norm topology is a locally convex vector space if we choose the single seminorm $q(x) = \|x\|$. $\diamond$

Example 6.15. 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 $x + q_\ell^{-1}([0,\varepsilon)) = \ell^{-1}(B_\varepsilon(\ell(x)))$ in this case. The same is true for X^* equipped with the weak or the weak-$*$ topology. $\diamond$

Example 6.16. 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. $\diamond$

Example 6.17. The space of 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. $\diamond$

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 $x \in X \setminus \{0\}$ 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 6.27).

It turns out crucial to understand when a seminorm is continuous.

Lemma 6.15. *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 corresponding constants $c_j > 0$, $1 \le j \le n$, such that $q(x) \le \sum_{j=1}^n c_j q_{\alpha_j}(x)$.*

Proof. If q is continuous, then $q^{-1}([0,1))$ contains an open neighborhood of 0 of the form $\bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j))$ and choosing $c_j = \max_{1\le j\le n}\varepsilon_j^{-1}$ we obtain that $\sum_{j=1}^n c_j q_{\alpha_j}(x) < 1$ implies $q(x) < 1$ and the claim follows from Problem 6.22. Conversely note that if $q(x) = r$ then $q^{-1}((r-\varepsilon, r+\varepsilon))$ contains the set $U(x) := x + \bigcap_{j=1}^n q_{\alpha_j}^{-1}([0,\varepsilon_j))$ provided $\sum_{j=1}^n c_j\varepsilon_j \le \varepsilon$ since by the inverse triangle inequality $|q(y) - q(x)| \le q(y-x) \le \sum_{j=1}^n c_j q_{\alpha_j}(x-y) < \varepsilon$ for $y \in U(x)$. □

Example 6.18. The weak topology on an infinite dimensional vector 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 $x \ne 0$ implying that $q(x) \le \sum_{j=1}^n c_j q_{\alpha_j}(x) = 0$. 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.33). Finally, since $p_\beta \circ A$ is a seminorm, the previous lemma implies:

Corollary 6.16. *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 $c_j > 0$, $1 \le j \le n$, such that $p_\beta(Ax) \le \sum_{j=1}^n c_j q_{\alpha_j}(x)$.*

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 $\alpha U + \beta U \subseteq U$.

Example 6.19. The absolutely convex sets in $\mathbb{C}$ are precisely the (open and closed) balls. ⋄

Since the sets $q_\alpha^{-1}([0,\varepsilon))$ are absolutely convex, a locally convex space always has a basis of absolutely convex sets. Conversely, a topological space which has a basis of absolutely convex sets is a locally convex space. To see this note that such a neighborhood U of 0 is also absorbing (Problem 6.21) and hence the corresponding Minkowski functional (6.1) is a seminorm (Problem 6.26). 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 6.25), 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 trivial, 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 4.12) and also the geometric Hahn–Banach theorem (Theorem 6.3) 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,

$$(A'y^*)(x) := y^*(Ax). \tag{6.15}$$

A brief calculation

$$q_x(A'y^*) = |(A'y^*)(x)| = |y^*(Ax)| = q_{Ax}(y^*) \tag{6.16}$$

verifies that A' is continuous in the weak-$*$ topology by virtue of Corollary 6.16.

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 6.17. *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*

$$d(x,y) := \sup_{n\in\mathbb{N}} \frac{1}{2^n} \frac{q_n(x-y)}{1+q_n(x-y)}. \tag{6.17}$$

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 (6.17) defines a metric (see also Problem B.3). Moreover, the metric balls $B_r(x) := \bigcap_{n:2^{-n}>r}\{y | q_n(y-x) < \frac{r}{2^{-n}-r}\}$ are clearly open and convex (note that the intersection is finite). Conversely, for every set of the form (6.14) 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.3). □

In general, a locally convex vector space X which has a separated countable family of seminorms is called a **Fréchet space**[5] if it is complete with respect to the metric (6.17). Note that the metric (6.17) is **translation invariant**

$$d(f,g) = d(f-h, g-h). \tag{6.18}$$

Example 6.20. The continuous functions $C(\mathbb{R})$ together with local uniform convergence are a Fréchet space. A countable family of seminorms is for example

$$\|f\|_j := \sup_{|x|\le j} |f(x)|, \qquad j \in \mathbb{N}. \tag{6.19}$$

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 6.21. The space $C^\infty(\mathbb{R}^m)$ together with the seminorms

$$\|f\|_{j,k} := \sum_{|\alpha|\le k} \sup_{|x|\le j} |\partial_\alpha f(x)|, \qquad j \in \mathbb{N},\ k \in \mathbb{N}_0, \tag{6.20}$$

is a Fréchet space.

[5] Maurice René Fréchet (1878–1973), French mathematician

Note that $\partial_\alpha : C^\infty(\mathbb{R}^m) \to C^\infty(\mathbb{R}^m)$ is continuous. Indeed by Corollary 6.16 it suffices to observe that $\|\partial_\alpha f\|_{j,k} \le \|f\|_{j,k+|\alpha|}$. ⋄

Example 6.22. The **Schwartz space**[6]

$$\mathcal{S}(\mathbb{R}^m) := \{f \in C^\infty(\mathbb{R}^m) | \sup_x |x^\alpha(\partial_\beta f)(x)| < \infty, \ \forall \alpha,\beta \in \mathbb{N}_0^m\} \tag{6.21}$$

together with the seminorms

$$q_{\alpha,\beta}(f) := \|x^\alpha(\partial_\beta f)(x)\|_\infty, \qquad \alpha,\beta \in \mathbb{N}_0^m, \tag{6.22}$$

is a Fréchet space. 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 $f \in \mathcal{S}(\mathbb{R}^m)$.

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 6.23. The space of all entire functions f (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 6.24. 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 by Lemma 6.15 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 Lemma 6.15 this will then be true for any continuous seminorm on X.

Theorem 6.18 (Kolmogorov[7]). *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 seminormed 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

[6] Laurent Schwartz (1915 –2002), French mathematician
[7] Andrei Kolmogorov (1903–1987), Soviet mathematician

$q_\alpha(x) \le C_\alpha q(x)$ (Problem 6.22) 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 4.8) and the closed graph theorem (Theorem 4.9) 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 6.21 and 6.28).

Problem* 6.21. *In a topological vector space every neighborhood U of 0 is absorbing.*

Problem* 6.22. *Let p, q be two seminorms. Then $p(x) \le Cq(x)$ if and only if $q(x) < 1$ implies $p(x) < C$.*

Problem 6.23. *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* 6.24. *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* 6.25. *In a topological vector space every convex open neighborhood U of zero contains a balanced open neighborhood. Moreover, 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* 6.26. *Let X be a vector space. Show that the Minkowski functional of a balanced, convex, absorbing set is a seminorm.*

Problem* 6.27. *If a locally convex space is Hausdorff then any corresponding family of seminorms is separated.*

Problem* 6.28. *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

$$d(0,\sum_{j=1}^\infty x_j) \le \sum_{j=1}^\infty d(0,x_j)$$

in this case (compare also Problem 1.8).

Problem 6.29. *Let X be a locally convex Hausdorff space. Then for a nonzero linear functional ℓ the following are equivalent:*

(i) *ℓ is continuous.*
(ii) $\mathrm{Ker}(\ell)$ *is closed.*
(iii) $\mathrm{Ker}(\ell)$ *is not dense in X.*
(iv) *There exists a neighborhood U of zero such that $\ell(U)$ is bounded.*

Problem 6.30. *Instead of* (6.17) *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 6.31. *Suppose X is a metric vector space. Then balls are convex if and only if the metric is quasiconvex:*

$$d(\lambda x+(1-\lambda)y,z) \le \max\{d(x,z),d(y,z)\}, \qquad \lambda\in(0,1).$$

(See also Problem 9.15.)

Problem 6.32. *Consider $\ell^p(\mathbb{N})$ for $p\in(0,1)$ — compare Problem 1.21. 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 6.33. *Show that $C_c^\infty(\mathbb{R}^m)$ is dense in $\mathcal{S}(\mathbb{R}^m)$.*

Problem 6.34. *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.*

6.5. 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 1.19 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 6.25. By Problem 1.19 it follows that $\ell^p(\mathbb{N})$ is strictly convex for $1 < p < \infty$ but not for $p = 1, \infty$. $\diamond$

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

$$\delta(\varepsilon) := \inf \left\{ 1 - \|\tfrac{x+y}{2}\| \,\middle|\, \|x\| = \|y\| = 1,\ \|x-y\| \geq \varepsilon \right\}, \quad 0 \leq \varepsilon \leq 2, \quad (6.23)$$

is called the modulus of convexity. Of course every uniformly convex space is strictly convex. In finite dimensions the converse is also true (Problem 6.38).

Note that δ is nondecreasing and

$$1 - \|\tfrac{x+y}{2}\| = 1 - \|x - \tfrac{x-y}{2}\| \leq \frac{\|x-y\|}{2}$$

for $\|x\| = \|y\| = 1$ 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 6.26. Every Hilbert space is uniformly convex with modulus of convexity $\delta(\varepsilon) = 1 - \sqrt{1 - \frac{\varepsilon^2}{4}}$ (Problem 6.36). $\diamond$

Example 6.27. Consider $C[0,1]$ with the norm

$$\|x\| := \|x\|_\infty + \|x\|_2 = \max_{t\in[0,1]} |x(t)| + \left(\int_0^1 |x(t)|^2 dt \right)^{1/2}.$$

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 1.19. 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 4.25). ◇

Example 6.28. It can be shown that $\ell^p(\mathbb{N})$ is uniformly convex for $1 < p < \infty$ (see Theorem 3.11 from [**40**]). ◇

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 4.30) uses this observation:

Theorem 6.19 (Radon–F. Riesz[8])**.** *Let X be a uniformly convex Banach space and let $x_n \rightharpoonup x$. Then $x_n \to x$ if and only if $\limsup \|x_n\| \le \|x\|$.*

Proof. By Lemma 4.29 (ii) we have in fact $\lim \|x_n\| = \|x\|$. 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

$$\ell\big(\frac{y_n + y}{2}\big) \le \left\| \frac{y_n + y}{2} \right\| \le 1$$

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 the following equivalent condition.

Lemma 6.20. *Let X be a Banach space. Then*

$$\delta(\varepsilon) = \inf\left\{1 - \|\tfrac{x+y}{2}\| \,\Big|\, \|x\| \le 1, \|y\| \le 1,\ \|x - y\| \ge \varepsilon\right\} \tag{6.24}$$

for $0 \le \varepsilon \le 2$.

Proof. It suffices to show that for given x and y which are not both on the unit sphere there is an equivalent pair on the unit sphere within 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 we must have $\|v(t_0)\| > 1$ and by the intermediate value theorem there is some $0 < t_1 < t_0$ with $\|v(t_1)\| = 1$.

[8] Johann Radon (1887–1956), Austrian mathematician

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, consider the convex function $f(t) := \|t\,u + (1-t)x\|$ which satisfies $f(0) = f(1) = 1$. Then for $0 \leq \lambda \leq 1$, $\alpha \geq 1$ we have $f(\lambda) \leq 1 \leq f(\lambda + \alpha)$ and for $\lambda \geq 1$, $\alpha \geq 1$ we have $f(\lambda) \leq f(\lambda + \alpha)$. Hence we always have $f(\lambda) \leq f(\lambda + \alpha)$ or equivalently $\|\frac{1}{2}(x + y)\| \leq \|\frac{1}{2}(u + v)\|$ and u, v is as required. □

Now we can prove:

Theorem 6.21 (Milman–Pettis[9]). *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 6.14) 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 $J(x), J(y) \in U$ 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 6.35. *Find an equivalent norm for $\ell^1(\mathbb{N})$ such that it becomes strictly convex (cf. Problems 1.19 and 1.27).*

Problem* 6.36. *Show that a Hilbert space is uniformly convex. (Hint: Use the parallelogram law.)*

Problem 6.37. *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 $\|x_n - y_n\| \to 0$.*

---

[9]Billy James Pettis (1913–1979), American mathematician

Problem* 6.38. *Show that a finite dimensional space is uniformly convex if and only if it is strictly convex.*

Problem 6.39. *Let X be strictly convex. Show that every nonzero linear functional attains its norm for at most one unit vector (cf. Problem 4.29).*

Chapter 7

Advanced Spectral theory

7.1. Spectral theory for compact 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 4.8), 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**

$$\sigma_p(A) := \{\alpha \in \sigma(A) | \operatorname{Ker}(A-\alpha) \neq \{0\}\} \tag{7.1}$$

as the set of all eigenvalues, the **continuous spectrum**

$$\sigma_c(A) := \{\alpha \in \sigma(A) \setminus \sigma_p(A) | \overline{\operatorname{Ran}(A-\alpha)} = X\} \tag{7.2}$$

and finally the **residual spectrum**

$$\sigma_r(A) := \{\alpha \in \sigma(A) \setminus \sigma_p(A) | \overline{\operatorname{Ran}(A-\alpha)} \neq X\}. \tag{7.3}$$

Clearly we have

$$\sigma(A) = \sigma_p(A) \mathbin{\dot\cup} \sigma_c(A) \mathbin{\dot\cup} \sigma_r(A). \tag{7.4}$$

Here the dot indicates that the union is disjoint. Note that in a Hilbert space $\sigma_x(A^*) = \sigma_x(A')^*$ for $x \in \{p, c, r\}$.

Example 7.1. In the finite dimensional case, $X := \mathbb{C}^n$, we have $\sigma(A) = \sigma_p(A)$ and $\sigma_c(A) = \sigma_r(A) = \emptyset$ since the rank-nullity theorem implies that $A - \alpha$ is bijective if and only if it is injective. $\diamond$

Example 7.2. Suppose $X := \ell^p(\mathbb{N})$ and let $a \in \ell^\infty(\mathbb{N})$ be a bounded sequence. Denote by $S := \bigcup_{n\in\mathbb{N}}\{a_j\}$ the range of a. Consider the operator $(Ax)_n := a_n x_n$ given by multiplication with a in X. Then $\sigma_p(A) = S$ and $\sigma(A) = \overline{S}$. Since for every $\alpha \in \mathbb{C}\setminus\overline{S}$ the resolvent $(A-\alpha)^{-1}$ is given by multiplication with the sequence $(a-\alpha)^{-1}$. Indeed note that this is bounded sequence with $\|(a-\alpha)^{-1}\|_\infty = \operatorname{dist}(\alpha, S)^{-1}$. Moreover, for $\alpha \in \overline{S}\setminus S$ the range of $A-\alpha$ contains at least $\ell_c(\mathbb{N})$ and hence is dense for $1 \le p < \infty$. Here $\ell_c(\mathbb{N})$ denotes the sequences with compact support, i.e., which are eventually zero. Thus $\sigma_c(A) = \overline{S}\setminus S$ and $\sigma_r(A) = \emptyset$. If $p = \infty$ then there must be a subsequence $n_j \to \infty$ such that $(a_{n_j} - \alpha)^{-1}$ is unbounded and hence the closure of the range of $A-\alpha$ cannot contain a sequence with $a_{n_j} = 1$. Hence $\sigma_r(A) = \overline{S}\setminus S$ and $\sigma_c(A) = \emptyset$ in this case.

Note that this example shows that none of σ_p, σ_c, σ_r will be closed in general. $\diamond$

Example 7.3. Suppose $\mathfrak{H}$ is a Hilbert space and $A = A^*$ is bounded and self-adjoint. Then by (2.30), $\sigma_r(A) = \emptyset$. $\diamond$

Example 7.4. Suppose $X := \ell^p(\mathbb{N})$ and $(Lx)_j := x_{j+1}$ is the left shift. Then $\sigma(L) = \bar{B}_1(0)$. Indeed, a simple calculation shows that $\operatorname{Ker}(L-\alpha) = \operatorname{span}\{(\alpha^{j-1})_{j\in\mathbb{N}}\}$ for $|\alpha| < 1$ if $1 \le p < \infty$ and for $|\alpha| \le 1$ if $p = \infty$. Hence $\sigma_p(L) = B_1(0)$ for $1 \le p < \infty$ 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 $\operatorname{Ran}(L-\alpha)$ is dense for $1 \le p < \infty$. Note that for $\alpha \in B_1(0)$ this formula defines an $x \in X$ for every $y \in X$ by the discrete version of Young's inequality for convolutions (see Problem 10.1 from [**40**]) and hence $\operatorname{Ran}(L-\alpha) = X$ for $\alpha \in B_1(0)$. Thus $\sigma_c(L) = \partial B_1(0)$ for $1 \le p < \infty$. Consequently, $\sigma_r(L) = \emptyset$. $\diamond$

Since A is invertible if and only if A' is by Theorem 4.25 we obtain:

Lemma 7.1. *Let X be a Banach space. Suppose $A \in \mathscr{L}(X)$. Then*

$$\sigma(A) = \sigma(A'). \tag{7.5}$$

Moreover,

$$\begin{aligned} \sigma_p(A') &\subseteq \sigma_p(A) \cup\!\!\!\cdot\, \sigma_r(A), & \sigma_p(A) &\subseteq \sigma_p(A') \cup\!\!\!\cdot\, \sigma_r(A'),\\ \sigma_r(A') &\subseteq \sigma_p(A) \cup\!\!\!\cdot\, \sigma_c(A), & \sigma_r(A) &\subseteq \sigma_p(A'),\\ \sigma_c(A') &\subseteq \sigma_c(A), & \sigma_c(A) &\subseteq \sigma_r(A') \cup\!\!\!\cdot\, \sigma_c(A'). \end{aligned} \tag{7.6}$$

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. As already indicated, the first claim follows from Theorem 4.25. The remaining items follow from Lemma 4.26 and (4.7). For example, if $\alpha \in \sigma_p(A')$, then $\mathrm{Ran}(A-\alpha)^\perp = \mathrm{Ker}(A'-\alpha) \ne \{0\}$ and hence $\overline{\mathrm{Ran}(A-\alpha)} \ne X$, so $\alpha \notin \sigma_c(A)$. If $\alpha \in \sigma_r(A')$, then $\mathrm{Ker}(A'-\alpha) = \{0\}$ and hence $\overline{\mathrm{Ran}(A-\alpha)} = X$, so $\alpha \notin \sigma_r(A)$. Etc. In the reflexive case use $A \cong A''$ by (4.12). □

Example 7.5. 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 $\mathrm{Ran}(L'-\alpha)$ for $|\alpha| = 1$: If we have $(L'-\alpha)x = y$ with some $y \in \ell^\infty(\mathbb{N})$, 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 $\mathrm{Ran}(L'-\alpha)$. 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 $\tilde{y} \notin \mathrm{Ran}(L'-\alpha)$, which shows that the range is not dense and hence $\sigma_r(L') = \bar{B}_1(0)$, $\sigma_c(L') = \emptyset$. ⋄

Example 7.6. Consider the bilateral left shift $(Sx)_j := x_{j+1}$ on $X := \ell^p(\mathbb{Z})$. By $\|S\| = 1$ we conclude $\sigma(S) \subseteq \bar{B}_1(0)$. Moreover, since its inverse is the corresponding right shift $(S^{-1}x)_j := x_{j-1}$, we also have $\|S^{-1}\| = 1$ and thus $\sigma(S^{-1}) \subseteq \bar{B}_1(0)$. Since we also have $\sigma(S^{-1}) = \sigma(S)^{-1}$ (cf. Problem 5.7), we arrive at $\sigma(S) \subseteq \partial B_1(0)$ as well as $\sigma(S^{-1}) \subseteq \partial B_1(0)$.

Now for $|\alpha| = 1$ there are two cases: If $p = \infty$, then clearly $x_j := \alpha^j$ is in $\mathrm{Ker}(S-\alpha)$ and hence $\sigma(S) = \sigma_p(S) = \partial B_1(0)$. Similarly, $\sigma(S^{-1}) = \sigma_p(S^{-1}) = \partial B_1(0)$. If $p < \infty$, then one concludes that if $y = (S-\alpha)x$ has compact support, so has x (as outside the support of y the sequence x must be equal to a multiple of α^j and this multiple must be 0 since $x \in \ell^p(\mathbb{Z})$). Moreover, in this case we further infer $\sum_{j\in\mathbb{Z}} \alpha^{-j} y_j = \sum_{j\in\mathbb{Z}} \alpha^{-j}(x_{j+1} - \alpha x_j) = 0$. Hence any sequence y with compact support violating this condition cannot be in the range of $S - \alpha$. Hence $\sigma(S) = \partial B_1(0)$. Moreover, since $S' = S^{-1}$ we see using Lemma 7.1 together with $\sigma_p(S) = \sigma_p(S^{-1}) = \emptyset$ that $\sigma_r(S) = \sigma_r(S^{-1}) = \emptyset$. Hence $\sigma_c(S) = \sigma_c(S^{-1}) = \partial B_1(0)$. ⋄

Moreover, for compact operators the spectrum is particularly simple (cf. also Theorem 3.7). We start with the following observation:

Lemma 7.2. *Let X be a Banach space. Suppose that $K \in \mathscr{K}(X)$ and $\alpha \in \mathbb{C}\setminus\{0\}$. Then $\mathrm{Ker}(K-\alpha)$ is finite dimensional and the range $\mathrm{Ran}(K-\alpha)$ 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 4.31. In particular, it is complemented (Problem 4.26), 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 8.4 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 7.7). In Theorem 3.7 we have seen that for a compact self-adjoint 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 invariant subspaces associated to $A \in \mathscr{L}(X)$ (where we have assumed $\alpha = 0$ without loss of generality):

$$\{0\} \subseteq \operatorname{Ker}(A) \subseteq \operatorname{Ker}(A^2) \subseteq \operatorname{Ker}(A^3) \subseteq \cdots \tag{7.7}$$

and

$$X \supseteq \operatorname{Ran}(A) \supseteq \operatorname{Ran}(A^2) \supseteq \operatorname{Ran}(A^3) \supseteq \cdots \tag{7.8}$$

We will say that the kernel chain **stabilizes** at $n \in \mathbb{N}_0$ 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. Note that if $\operatorname{Ker}(A) = \{0\}$, then the kernel chain stabilizes at 0. Similarly, will say that the range chain **stabilizes** at $m \in \mathbb{N}_0$ if $\operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ and call m the **descent** of A. Again, if $x = A^{m+1}y \in \operatorname{Ran}(A^{m+1})$ we can write $A^m y = A^{m+1}z$ for some z which shows $x = A^{m+2}z \in \operatorname{Ran}(A^{m+2})$ 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 7.7. For the left shift operator L we have $\operatorname{Ran}(L^k) = \ell^p(\mathbb{N})$ for all $k \in \mathbb{N}$ while the kernel chain does not stabilize as $\operatorname{Ker}(L^k) = \{a \in \ell^p(\mathbb{N}) | a_j =$

$0,\ j > k\}$. Similarly, for the right shift operator R we have $\operatorname{Ker}(R^k) = \{0\}$ while the range chain does not stabilize as $\operatorname{Ran}(R^k) = \{a \in \ell^p(\mathbb{N}) | a_j = 0,\ 1 \le j \le k\}$. $\diamond$

Lemma 7.3. *Let X be a vector space and 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 $\operatorname{Ran}(A^n) \cap \operatorname{Ker}(A^n) = \{0\}$ and A restricted to $\operatorname{Ran}(A^n) \to \operatorname{Ran}(A^n)$ is injective.*

(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 $\operatorname{Ker}(A^m) + \operatorname{Ran}(A^m) = X$ and A restricted to $\operatorname{Ran}(A^m) \to \operatorname{Ran}(A^m)$ is surjective.*

(iii) *If both chains stabilize, then $m = n$ and $\operatorname{Ker}(A^m) \dotplus \operatorname{Ran}(A^m) = X$ and A restricted to $\operatorname{Ran}(A^m) \to \operatorname{Ran}(A^m)$ is bijective.*

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, $x = A^n y = 0$. Moreover, if $Ax = 0$ for some $x = A^n y \in \operatorname{Ran}(A^n)$, then $y \in \operatorname{Ker}(A^{n+1}) = \operatorname{Ker}(A^n)$ implying $x = 0$.

(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 $x = A^m y + (x - A^m y)$. Moreover, $A \operatorname{Ran}(A^m) = \operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ shows surjectivity.

(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 $0 \ne A^m x = A^{m+1} y$ there is an $y \in \operatorname{Ker}(A^{m+2}) \setminus \operatorname{Ker}(A^{m+1})$. Iterating this argument would show that the kernel chain does not stabilize, contradiction our assumption. Hence $n \le m$.

Conversely, suppose $\operatorname{Ker}(A^{n+1}) = \operatorname{Ker}(A^n)$ and $\operatorname{Ran}(A^{m+1}) = \operatorname{Ran}(A^m)$ for $m \ge n$. Then for every x there is some y such that

$$A^m x = A^{m+1} y \ \Rightarrow\ x - Ay \in \operatorname{Ker}(A^m) = \operatorname{Ker}(A^n) \ \Rightarrow\ A^n x = A^{n+1} y$$

shows $\operatorname{Ran}(A^{n+1}) = \operatorname{Ran}(A^n)$, that is, $m \le n$. The rest follows by combining (i) and (ii). $\square$

Of course the desired case is (iii), where we have a splitting of X into a direct sum of invariant subspaces such that A restricted to $\operatorname{Ker}(A^m) \to \operatorname{Ker}(A^m)$ is nilpotent and restricted to $\operatorname{Ran}(A^m) \to \operatorname{Ran}(A^m)$ is bijective.

Example 7.8. In a finite dimensional space we are of course always in case (iii). $\diamond$

Example 7.9. Let $A \in \mathscr{L}(\mathfrak{H})$ be a self-adjoint operator in a Hilbert space. Then by (2.30) the kernel chain always stabilizes at $n = 1$ and the range chain stabilizes at $n = 1$ if $\operatorname{Ran}(A)$ is closed. ⋄

As a further preparation we establish the following result which should be thought of a replacement of an orthogonal vector.

Lemma 7.4 (F. Riesz lemma)**.** *Let X be a normed vector space and $Y \subset X$ some subspace which is not dense $\overline{Y} \neq X$. Then for every $\varepsilon \in (0,1)$ there exists an x_ε with $\|x_\varepsilon\| = 1$ and*

$$\inf_{y \in Y} \|x_\varepsilon - y\| \geq 1 - \varepsilon. \tag{7.9}$$

Proof. Pick $x \in X \setminus \overline{Y}$ and abbreviate $d := \operatorname{dist}(x, Y) > 0$. Choose $y_\varepsilon \in Y$ such that $\|x - y_\varepsilon\| \leq \frac{d}{1-\varepsilon}$. Then $x_\varepsilon := \frac{x - y_\varepsilon}{\|x - y_\varepsilon\|}$ is the vector we are looking for since

$$\|x_\varepsilon - y\| = \frac{1}{\|x - y_\varepsilon\|}\|x - (y_\varepsilon + \|x - y_\varepsilon\| y)\| \geq \frac{d}{\|x - y_\varepsilon\|} \geq 1 - \varepsilon$$

as required. □

As already indicated, in a Hilbert space the claim holds with $\varepsilon = 0$ for any normalized x in the orthogonal complement of Y. Slightly more general, if for every $x \in X \setminus Y$ we can find some $y_0 \in \overline{Y}$ with $\|x - y_0\| = \operatorname{dist}(x, Y)$, then the lemma holds with $\varepsilon = 0$. This is the case in reflexive spaces — Example 9.24.

Now we can apply this to our situation.

Lemma 7.5. *Let X be a Banach space. Suppose that $K \in \mathscr{K}(X)$ and $\alpha \in \mathbb{C} \setminus \{0\}$. Then the space $\operatorname{Ker}(K - \alpha)^m$ is finite dimensional and the space $\operatorname{Ran}(K - \alpha)^m$ is closed for every $m \in \mathbb{N}$. Moreover, there is some $n = n(\alpha) \in \mathbb{N}_0$ such that the kernel and range chain of both $K - \alpha$ and $K' - \alpha$ stabilize and hence*

$$\begin{aligned} X &= \operatorname{Ker}(K - \alpha)^n \dotplus \operatorname{Ran}(K - \alpha)^n, \\ X^* &= \operatorname{Ker}(K' - \alpha)^n \dotplus \operatorname{Ran}(K' - \alpha)^n. \end{aligned} \tag{7.10}$$

Proof. Since $\alpha \neq 0$ we can consider $\mathbb{I} - \alpha^{-1} K$ and assume $\alpha = 1$ without loss of generality. Moreover, since $(\mathbb{I} - K)^n - \mathbb{I} \in \mathscr{K}(X)$ 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 Lemma 7.4, 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

$$\|Kx_n - Kx_m\| = \|x_n - (\mathbb{I} - K)x_n - Kx_m\| \geq \operatorname{dist}(x_n, K_n) \geq \frac{1}{2}$$

for $n > m$ and hence the image of the bounded sequence Kx_n has no convergent subsequence, a contradiction.

Consequently the kernel sequence for $\mathbb{I} - K'$ also stabilizes. Moreover, by the closed range theorem (Theorem 4.28) we have

$$\operatorname{Ker}((\mathbb{I} - K')^n)_\perp = \operatorname{Ran}(\mathbb{I} - K)^n, \quad \operatorname{Ker}((\mathbb{I} - K)^n)^\perp = \operatorname{Ran}(\mathbb{I} - K')^n$$

and hence the kernel chain of $\mathbb{I} - K'$ stabilizes at n when the range chain of $\mathbb{I} - K$ stabilizes at n and the kernel chain of $\mathbb{I} - K$ stabilizes at n when the range chain of $\mathbb{I} - K'$ stabilizes at n. The rest follows from the previous lemma. □

As an immediate consequence we get the famous Fredholm[1] alternative:

Theorem 7.6 (Fredholm alternative)**.** *Let X be a Banach space. Suppose that $K \in \mathscr{K}(X)$ and $\alpha \in \mathbb{C} \setminus \{0\}$. Then the following are equivalent:*

- $\operatorname{Ker}(K - \alpha) = \{0\}$.
- $\operatorname{Ran}(K - \alpha) = X$.
- $\operatorname{Ker}(K' - \alpha) = \{0\}$.
- $\operatorname{Ran}(K' - \alpha) = X^*$.

Of course in terms of equations this can equivalently be phrased as, either the inhomogeneous equation

$$(K - \alpha)x = y \tag{7.11}$$

has a unique solution for every $y \in X$ or the corresponding homogeneous equation

$$(K - \alpha)x = 0 \tag{7.12}$$

has a nontrivial solution. Moreover, by the closed range theorem (Theorem 4.28), there will be a solution if and only if $\ell(y) = 0$ for all $\ell \in \operatorname{Ker}(K' - \alpha)$. Of course the analogous statement holds for K'.

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.

Note that this also implies

$$\sigma_p(K) \setminus \{0\} = \sigma_p(K') \setminus \{0\}. \tag{7.13}$$

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 equal to infinity. The **order** of

[1]Erik Ivar Fredholm (1866–1927), Swedish mathematician

a generalized eigenvector u corresponding to an eigenvalue α is the smallest n such that $(A-\alpha)^n u = 0$.

Example 7.10. For the eigenvalues of a self-adjoint operator in a Hilbert space geometric and algebraic multiplicities always agree by Example 7.9. ⋄

Example 7.11. Consider $X := \ell^p(\mathbb{N})$ and $K \in \mathscr{K}(X)$ given by $(Ka)_n := \frac{1}{n} a_{n+1}$. Then $Ka = \alpha\, a$ implies $a_n = \alpha^{n-1}(n-1)! a_1$ and hence $a_1 = 0$ for $\alpha \neq 0$ and $a = a_1 \delta^1$ for $\alpha = 0$. Hence $\sigma_p(K) = \{0\}$ with 0 being an eigenvalue of geometric multiplicity one. Since $K^n a = 0$ implies $a_j = 0$ for $j > n$ we see that its index as well as its algebraic multiplicity is ∞. Moreover, δ^n is a generalized eigenvalue of order n. ⋄

Theorem 7.7 (Spectral theorem for compact operators; F. Riesz)**.** *Let X be a Banach space and suppose that $K \in \mathscr{K}(X)$. Then every $\alpha \in \sigma(K)\setminus\{0\}$ is an eigenvalue of finite algebraic multiplicity and X can be decomposed into invariant closed subspaces according to* (7.10)*, 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 $\operatorname{Ker}(K-\alpha) = \{0\}$, then the kernel chain stabilizes as $n = 0$ and hence $\operatorname{Ran}(K-\alpha) = X$, 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 Lemma 7.4 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, since $(K-\alpha_n)X_n \subset X_{n-1}$, $\|K\tilde{x}_n - K\tilde{x}_m\| = \|\alpha_n \tilde{x}_n + ((K-\alpha_n)\tilde{x}_n - K\tilde{x}_m)\| \geq \frac{|\alpha_n|}{2} \geq \frac{\varepsilon}{2}$ 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 $\tilde{K} : X/\operatorname{Ker}(K) \to \operatorname{Ran}(K)$ (cf. Problem 1.71) which is again compact. Hence $\mathbb{I}_{X/\operatorname{Ker}(K)} = \tilde{K}^{-1}\tilde{K}$ is compact and thus $X/\operatorname{Ker}(K)$ is finite dimensional. □

Note that in contradistinction to the symmetric case, there might be no eigenvalues at al.

Example 7.12. Consider $X := \ell^p(\mathbb{N})$ and $K \in \mathscr{K}(X)$ given by $(Ka)_n := \frac{1}{n} a_{n-1}$ for $n > 1$ and $(Ka)_1 := 0$. Then $Ka = \alpha\, a$ implies $a = 0$ and hence $\sigma_p(K) = \emptyset$. Thus $\sigma(K) = \{0\}$.

Another example with no eigenvalues is the Volterra integral operator from Example 5.14. ⋄

This result says in particular, that for $\alpha \in \sigma_p(K) \setminus \{0\}$ we can split $X = X_1 \oplus X_2$ where both $X_1 := \operatorname{Ker}(K-\alpha)^n$ and $X_2 := \operatorname{Ran}(K-\alpha)^n$ are invariant subspaces for K. Consequently we can split $K = K_1 \oplus K_2$, where K_1 is the restriction of K to the finite dimensional subspace X_1 with $K_1 - \alpha$ a nilpotent matrix and K_2 is the restriction of K to X_2 with $K_2 - \alpha$ bijective and hence $\alpha \in \rho(K_2)$.

Problem 7.1. *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 7.2. *Compute the point, continuous, and residual spectrum of the multiplication operator $A : C[0,1] \to C[0,1]$, $x(t) \mapsto a(t)x(t)$, where $a(t) := |3t-2| - |3t-1|$.*

Problem 7.3. *Compute the point, continuous, and residual spectrum of the operator $A : C[0,1] \to C[0,1]$, $x(t) \mapsto x(0) + t\,x(1)$. Compute the kernel and the range for each point in the spectrum. Is the range closed or dense?*

Problem 7.4. *Compute the point, continuous, and residual spectrum of the operator $A : C[0,1] \to C[0,1]$, $x(t) \mapsto \int_0^t x(s)ds$.*

Problem 7.5. *Consider the bilateral left shift $(Sx)_j := x_{j+1}$ on $X := \ell^p(\mathbb{Z})$ and let $c \in c_0(\mathbb{Z})$ be a nonvanishing sequence converging to 0 (i.e., $\lim_{|j|\to 0} c_j = 0$ and $c_j \neq 0$ for all $j \in \mathbb{Z}$). Compute the point, continuous, and residual spectrum of $K := cS$.*

Problem 7.6. *Suppose $A \in \mathscr{L}(X)$. Is it true that $\dim(\operatorname{Ker}(A-\alpha)) = \dim(\operatorname{Ker}(A'-\alpha))$ if one of these kernels is finite dimensional?*

Problem* 7.7. *Suppose $A \in \mathscr{L}(X)$. Show that generalized eigenvectors corresponding to different eigenvalues or with different order are linearly independent.*

Problem 7.8. *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 5.28.)*

Problem 7.9. *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 7.10. *Let $A : X \to Y$, $B : Y \to Z$. Show* $\dim(\operatorname{Ker}(BA)) \le \dim(\operatorname{Ker}(A)) + \dim(\operatorname{Ker}(B))$ *and hence* $\dim(\operatorname{Ker}(A^n)) \le n \dim(\operatorname{Ker}(A))$ *if $A : X \to X$.*

7.2. Fredholm operators

Let X and Y be Banach spaces. In this section we want to investigate solvability of the equation

$$Ax = y \tag{7.14}$$

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**

$$\dim \operatorname{Ker}(A) + \dim \operatorname{Ran}(A) = \dim X \tag{7.15}$$

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,

$$\operatorname{ind}(A) := \dim \operatorname{Ker}(A) - \dim \operatorname{Coker}(A) = \dim(X) - \dim(Y), \tag{7.16}$$

at least the left-hand side will be finite if we assume both $\operatorname{Ker}(A)$ and $\operatorname{Coker}(A) := Y/\operatorname{Ran}(A)$ 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 7.6). It took a while until Noether[2] 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 7.8. *Let X and Y be Banach spaces. 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 1.71). Moreover, the assumption that the cokernel is finite says that there is a finite subspace $Y_0 \subset Y$ such that $Y = Y_0 \dot{+} \operatorname{Ran}(A)$. Then

$$\widehat{A} : X/\operatorname{Ker}(A) \oplus Y_0 \to Y, \qquad \widehat{A}(x,y) = \tilde{A}x + y$$

is bijective and hence a homeomorphism by Theorem 4.8. Since $\tilde{X} := X/\operatorname{Ker}(A) \oplus \{0\}$ is a closed subspace of $X/\operatorname{Ker}(A) \oplus Y_0$ we see that $\operatorname{Ran}(A) = \widehat{A}(\tilde{X})$ 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

$$\operatorname{ind}(A) := \dim \operatorname{Ker}(A) - \dim \operatorname{Coker}(A). \tag{7.17}$$

[2] Fritz Noether (1884–1941), German mathematician

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 $\Phi_0(X, Y)$.

Example 7.13. We have $\mathbb{I} \in \Phi_0(X)$ but clearly $0 \notin \Phi(X)$ unless X is finite dimensional. Moreover, $A \in \Phi(X, Y)$ implies $\alpha A \in \Phi(X, Y)$ for $\alpha \in \mathbb{C} \setminus \{0\}$ with $\operatorname{ind}(\alpha A) = \operatorname{ind}(A)$ but the sum of two Fredholm operators is in general not Fredholm (e.g. $A - A = 0$). In particular, $\Phi(X, Y)$ is *not* a linear subspace of $\mathscr{L}(X, Y)$ unless both X and Y are finite dimensional. $\diamond$

Lemma 7.9. *Let X and Y be Banach spaces. Suppose $A \in \mathscr{L}(X, Y)$ with* $\operatorname{Ran}(A)$ *closed. Then*

$$\operatorname{Ker}(A') \cong \operatorname{Coker}(A)^*, \qquad \operatorname{Coker}(A') \cong \operatorname{Ker}(A)^*. \tag{7.18}$$

In particular, we have

$$\dim \operatorname{Ker}(A') = \dim \operatorname{Coker}(A), \qquad \dim \operatorname{Ker}(A) = \dim \operatorname{Coker}(A'). \tag{7.19}$$

Proof. Using Lemma 4.26 and Theorems 4.28, 4.21 we have $\operatorname{Ker}(A') = \operatorname{Ran}(A)^\perp \cong (Y/\operatorname{Ran}(A))^* = \operatorname{Coker}(A)^*$ and $\operatorname{Coker}(A') = X^*/\operatorname{Ran}(A') = X^*/\operatorname{Ker}(A)^\perp \cong \operatorname{Ker}(A)^*$. The second claim follows since for a finite dimensional space the dual space has the same dimension. $\square$

An immediate consequence is:

Theorem 7.10 (Riesz)**.** *Let X and Y be Banach spaces. A bounded operator $A \in \mathscr{L}(X, Y)$ is Fredholm if and only if A' is and in this case*

$$\operatorname{ind}(A') = -\operatorname{ind}(A). \tag{7.20}$$

Example 7.14. The left shift operator L in $X = Y := \ell^p(\mathbb{N})$, $1 \le p < \infty$ is Fredholm. In fact, we have $\operatorname{Ker}(L) = \operatorname{span}\{\delta^1\}$ and $\operatorname{Ran}(L) = X$ implying $\operatorname{ind}(L) = 1$. Consequently L', which is just the right shift, is also Fredholm with $\operatorname{ind}(L') = -1$. Of course the last fact can also be checked directly. $\diamond$

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

$$\operatorname{ind}(A) = \dim \operatorname{Ker}(A) - \dim \operatorname{Ran}(A)^\perp \tag{7.21}$$

and $\operatorname{ind}(A^*) = -\operatorname{ind}(A)$ as is immediate from (2.30).

Example 7.15. Suppose $\mathfrak{H}$ is a Hilbert space and $A = A^*$ is a self-adjoint Fredholm operator, then (2.30) 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 5.22, $A - \lambda$ 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 linear operators $A_j : X_j \to X_{j+1}$

$$X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \cdots X_n \xrightarrow{A_n} X_{n+1}$$

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 7.11) 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 7.12).

Lemma 7.11 (Atkinson[3]). *Let X, Y and Z be Banach spaces. Suppose $A \in \mathscr{L}(X,Y)$, $B \in \mathscr{L}(Y,Z)$. If two of the operators A, B, BA are Fredholm, so is the third and we have*

$$\operatorname{ind}(BA) = \operatorname{ind}(A) + \operatorname{ind}(B). \tag{7.22}$$

Proof. It is straightforward to check that the sequence

$$\begin{aligned} 0 \longrightarrow \operatorname{Ker}(A) \longrightarrow \operatorname{Ker}(BA) \xrightarrow{A} \operatorname{Ker}(B) \longrightarrow \operatorname{Coker}(A) \\ \xrightarrow{\tilde{B}} \operatorname{Coker}(BA) \longrightarrow \operatorname{Coker}(B) \longrightarrow 0 \end{aligned}$$

is exact. Here the maps which are not explicitly stated are canonical inclusions/quotient maps. Hence by Problem 7.11, if two operators are Fredholm, so is the third. Moreover, the formula for the index follows from Problem 7.12. □

Next we want to look a bit further into the structure of a Fredholm operator $A : X \to Y$. First of all, since $\operatorname{Ker}(A)$ is finite dimensional, it is complemented (Problem 4.26), 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 1.72) and there exists a finite dimensional 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 $\operatorname{Ker}(A) \dotplus X_0 \to Y_0 \dotplus \operatorname{Ran}(A)$ our Fredholm operator is given by

$$A = \begin{pmatrix} 0 & 0 \\ 0 & A_0 \end{pmatrix}, \tag{7.23}$$

[3]Frederick Valentine Atkinson (1916 –2002), British mathematician

where A_0 is the restriction of A to $X_0 \to \mathrm{Ran}(A)$. By construction A_0 is bijective and hence a homeomorphism (Theorem 4.8).

Example 7.16. In case of an Hilbert space we can choose $X_0 = \mathrm{Ker}(A)^\perp = \mathrm{Ran}(A^*)$ and $Y_0 = \mathrm{Ran}(A)^\perp = \mathrm{Ker}(A^*)$ by (2.30). In particular, $A : \mathrm{Ker}(A)^\perp \to \mathrm{Ker}(A^*)^\perp$ has a bounded inverse. $\diamond$

Defining

$$B := \begin{pmatrix} 0 & 0 \\ 0 & A_0^{-1} \end{pmatrix} \tag{7.24}$$

we get

$$AB = \mathbb{I} - Q, \qquad BA = \mathbb{I} - P \tag{7.25}$$

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 7.12 (Dieudonné)**.** *Let X and Y be Banach spaces. The set of Fredholm operators $\Phi(X,Y)$ is open in $\mathscr{L}(X,Y)$ and the index is locally constant.*

Proof. Let $A \in \Phi(X,Y)$ with corresponding decomposition as in (7.23). Consider $C \in \mathscr{L}(X,Y)$ and write it as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

with respect to this decomposition. Then if $\|C_{22}\| < \|A_0^{-1}\|^{-1}$ we have that $A_0 + C_{22}$ is still invertible (Problem 1.58). 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 (the first part $\mathrm{Ker}(A) \to Y_0$ is finite dimensional and the second part $X_0 \to \mathrm{Ran}(A)$ is bijective as noted above). Moreover, $\mathrm{ind}(A + C) = \mathrm{ind}(C_{11} - C_{12}(A_0 + C_{22})^{-1} C_{21}) = \dim(\mathrm{Ker}(A)) - \dim(Y_0) = \mathrm{ind}(A)$ since the second operator is between finite dimensional spaces and the index can be evaluated using (7.16). $\square$

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 7.13 (Riesz). *Let X be a Banach space. For every $K \in \mathscr{K}(X)$ we have $\mathbb{I} - K \in \Phi_0(X)$.*

Proof. That $\mathbb{I} - K$ is Fredholm follows from Lemma 7.2 since K' is compact as well and $\mathrm{Coker}(\mathbb{I} - K)^* \cong \mathrm{Ker}(\mathbb{I} - K')$ by Lemma 7.9. Furthermore, the index is constant along $[0,1] \to \Phi(X)$, $\alpha \mapsto \mathbb{I} - \alpha K$ and hence $\mathrm{ind}(\mathbb{I} - K) = \mathrm{ind}(\mathbb{I}) = 0$. □

Next we show that an operator is Fredholm if and only if it has a left/right inverse up to compact operators.

Theorem 7.14 (Atkinson). *Let X and Y be Banach spaces. 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 $B_1 A - \mathbb{I} \in \mathscr{K}(X)$ and $AB_2 - \mathbb{I} \in \mathscr{K}(Y)$.*

Proof. If A is Fredholm we have already given an operator B in (7.24) such that $BA - \mathbb{I}$ and $AB - \mathbb{I}$ are finite rank. Conversely, according to Theorem 7.13 $B_1 A$ and AB_2 are Fredholm. Since $\mathrm{Ker}(A) \subseteq \mathrm{Ker}(B_1 A)$ and $\mathrm{Ran}(AB_2) \subseteq \mathrm{Ran}(A)$ 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 7.13:

Corollary 7.15 (Yood). *For every $A \in \Phi(X,Y)$ and $K \in \mathscr{K}(X,Y)$ we have $A + K \in \Phi(X,Y)$ with $\mathrm{ind}(A+K) = \mathrm{ind}(A)$.*

Proof. Using (7.24) we see that $B(A+K) - \mathbb{I} = -P + BK \in \mathscr{K}(X)$ and $(A+K)B - \mathbb{I} = -Q + KB \in \mathscr{K}(Y)$ and hence $A + K$ is Fredholm. In fact, $A + \alpha K$ is Fredholm for $\alpha \in [0,1]$ and hence $\mathrm{ind}(A+K) = \mathrm{ind}(A)$ since the index is locally constant. □

Fredholm operators are also used to split the spectrum. For $A \in \mathscr{L}(X)$ one defines the **essential spectrum**

$$\sigma_{ess}(A) := \{\alpha \in \mathbb{C} | A - \alpha \not\in \Phi_0(X)\} \subseteq \sigma(A) \tag{7.26}$$

and the **Fredholm spectrum**

$$\sigma_{\Phi}(A) := \{\alpha \in \mathbb{C} | A - \alpha \not\in \Phi(X)\} \subseteq \sigma_{ess}(A). \tag{7.27}$$

By Dieudonné's theorem both $\sigma_{ess}(A)$ and $\sigma_{\Phi}(A)$ are closed. Also note that we have $\sigma_c(A) \subseteq \sigma_{\Phi}(A)$.

Warning: Using various refinements on how an operator fails to be Fredholm one can obtain further variants, all of which might be (and sometimes are) used as a definition of the essential spectrum. Specifically, consider the following subsets of the set of bounded operators $\mathscr{L}(X)$: F_1 the set of operators with closed range (normally solvable operators), F_2 the set of operators where either the kernel or the cokernel is finite (semi-Fredholm operators), F_3 Fredholm operators, F_4 Fredholm operators of index zero, and finally F_5 Fredholm operators of index zero for which a punctured neighborhood of 0 belongs to the resolvent set. Then $\sigma^j_{ess}(A) := \{\alpha \in \mathbb{C} | A - \alpha \notin F_j\} \subseteq \sigma(A)$ and clearly $\sigma^1_{ess}(A) \subseteq \sigma^2_{ess}(A) \subseteq \sigma^3_{ess}(A) \subseteq \sigma^4_{ess}(A) \subseteq \sigma^5_{ess}(A)$.

Example 7.17. Let X be infinite dimensional and $K \in \mathscr{K}(X)$. Then $\sigma_{ess}(K) = \sigma_\Phi(K) = \{0\}$. $\diamond$

Example 7.18. If X is a Hilbert space and A is self-adjoint, then $\sigma(A) \subseteq \mathbb{R}$ and for $\alpha \in \mathbb{R} \setminus \sigma_\Phi(A)$ the identity $\operatorname{ind}(A-\alpha) = -\operatorname{ind}((A-\alpha)^*) = -\operatorname{ind}(A-\alpha)$ shows that the index is always zero. Thus $\sigma_{ess}(A) = \sigma_\Phi(A)$ for self-adjoint operators. $\diamond$

Example 7.19. Consider the left shift L from Example 7.4 for $1 \le p < \infty$. Then the considerations there show that $L - \alpha \in \Phi(X)$ with $\operatorname{ind}(L-\alpha) = 1$ for $\alpha \in B_1(0)$ and $L - \alpha \notin \Phi(X)$ for $\alpha \in \partial B_1(0)$. Hence $\sigma_{ess}(L) = \bar{B}_1(0)$ and $\sigma_\Phi(L) = \partial B_1(0)$. $\diamond$

By Corollary 7.15 both the Fredholm spectrum and the essential spectrum are invariant under compact perturbations:

Theorem 7.16 (Weyl)**.** *Let X be a Banach space. Let $A \in \mathscr{L}(X)$, then*

$$\sigma_\Phi(A+K) = \sigma_\Phi(A), \quad \sigma_{ess}(A+K) = \sigma_{ess}(A), \quad K \in \mathscr{K}(X). \tag{7.28}$$

Moreover, if $\alpha \notin \sigma_{ess}(A)$, using the notation from (7.23) for $A - \alpha$, we can find an operator $K_0 : \operatorname{Ker}(A-\alpha) \to Y_0$ such that $K_0 - \alpha$ is invertible 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

$$\sigma_{ess}(A) = \bigcap_{K \in \mathscr{K}(X)} \sigma(A+K). \tag{7.29}$$

It is important to emphasize, that this invariance with respect to a large class of perturbations makes it possible to determine the essential and Fredholm spectra even in nontrivial situations.

The remainder of the spectrum $\sigma(A) \setminus \sigma_{ess}(A)$ consists of eigenvalues with finite geometric multiplicity.

If A is compact, we have seen in Lemma 7.5 that for every nonzero eigenvalue α we can split X into a direct sum of invariant subspaces $X_0 \dot{+} X_1$ such that X_0 is finite dimensional and $A_0 := A|_{X_0}$ has $\sigma(A_1) = \{\alpha\}$ while $A_1 := A|_{X_1}$ has $\alpha \in \rho(A_1)$. Such eigenvalues are called **normal**. Note that a normal eigenvalue must be necessarily isolated and of finite algebraic multiplicity. Indeed the algebraic multiplicity of α as an eigenvalue of A equals the algebraic multiplicity of α as an eigenvalue of A_0 and hence equals

$\dim(X_0)$. Moreover, since $\rho(A) = \rho(A_0) \cap \rho(A_1) = \rho(A_1) \setminus \{\alpha\}$ we see that $\rho(A)$ is a punctured neighborhood of α. The converse also holds as we will see in Section 8.3.

This motivates the definition of the **discrete spectrum**

$$\begin{aligned}\sigma_d(A) &:= \{\alpha \in \sigma_p(A) | \alpha \text{ is normal}\} \\ &= \{\alpha \in \sigma_p(A) | \alpha \text{ is isolated and of finite algebraic multiplicity}\}\end{aligned} \tag{7.30}$$

Warning: Again different definitions of $\sigma_d(A)$ can be found in the literature.

Since for a normal eigenvalue we have $A - \alpha \in \Phi_0(X)$, we have

$$\sigma_d(A) \subseteq \sigma(A) \setminus \sigma_{ess}(A) \tag{7.31}$$

but in general we do not have equality.

Example 7.20. 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 $\sigma(A+K) \subseteq \partial B_1(0)$. Consequently $\sigma_{ess}(A) \subseteq \partial B_1(0)$ and since no point of $\partial B_1(0)$ is an eigenvalue we conclude $\sigma_{ess}(A) = \partial B_1(0)$. In particular, all $\alpha \in B_1(0)$ are eigenvalues of infinite algebraic multiplicity with $A - \alpha \in \Phi_0$. ⋄

Finally, we give a third definition of the discrete spectrum, namely as

$$\sigma_d(A) = \{\alpha \in \sigma(A) \setminus \sigma_{ess}(A) | \alpha \text{ has finite algebraic multiplicity}\}. \tag{7.32}$$

Indeed, if the algebraic multiplicity of an eigenvalue $\alpha \in \sigma(A) \setminus \sigma_{ess}(A)$ is finite, then by definition the kernel chain stabilizes and so does the range chain (Problem 7.14). Hence by Lemma 7.3 we have $X = \operatorname{Ker}((A-\alpha)^n) \dot{+} \operatorname{Ran}((A-\alpha)^n)$, where n is the index of α.

Example 7.21. For self-adjoint operators in a Hilbert space geometric and algebraic multiplicities coincide (cf. Example 7.10) and the discrete spectrum is precisely the set of isolated eigenvalues with finite multiplicity. ⋄

Problem* 7.11. *Suppose $X \xrightarrow{A} Y \xrightarrow{B} Z$ is exact. Show that if X and Z are finite dimensional, so is Y.*

Problem* 7.12. *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 7.13. *Let $A \in \Phi(X,Y)$ with a corresponding parametrix $B_1, B_2 \in \mathscr{L}(Y,X)$. Set $K_1 := \mathbb{I} - B_1 A \in \mathscr{K}(X)$, $K_2 := \mathbb{I} - AB_2 \in \mathscr{K}(Y)$ and show*

$$B_1 - B = B_1 Q - K_1 B \in \mathscr{K}(Y,X), \quad B_2 - B = PB_2 - BK_2 \in \mathscr{K}(Y,X).$$

Hence a parametrix is unique up to compact operators. Moreover, $B_1, B_2 \in \Phi(Y,X)$.

Problem* 7.14. *Suppose $A \in \Phi(X)$. If the kernel chain stabilizes then* $\operatorname{ind}(A) \le 0$. *If the range chain stabilizes then* $\operatorname{ind}(A) \ge 0$. *Moreover, if $A \in \Phi_0(X)$, then the kernel chain stabilizes if and only if the range chain stabilizes.*

7.3. 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 as 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 homomorphism $m : X \to \mathbb{C}$ will be called a **multiplicative linear functional** or **character**:

$$m(xy) = m(x)m(y), \qquad m(e) = 1. \tag{7.33}$$

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 7.18 below).

As we will see, multiplicative linear functionals 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 7.22. 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 $\mathrm{Ker}(m_{x_0}) = \{f \in C[a,b] | f(x_0) = 0\}$ is a maximal ideal (we will prove this in more generality in Lemma 7.18 below). ⋄

Example 7.23. 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 7.17. *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). Let I be a proper ideal. 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 5.1 all elements in the ball $B_1(e)$ are invertible and hence every proper ideal must be contained in the closed set $X \setminus B_1(e)$. 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 1.20) is again a Banach algebra if we define

$$[x][y] = [xy]. \tag{7.34}$$

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,

$$\|[xy]\| = \inf_{a\in I} \|xy + a\| = \inf_{b,c\in I} \|(x+b)(y+c)\| \le \inf_{b\in I} \|x+b\| \inf_{c\in I} \|y+c\|$$
$$= \|[x]\|\|[y]\|. \tag{7.35}$$

In particular, the quotient map $\pi : X \to X/I$ is a Banach algebra homomorphism.

Example 7.24. Consider the Banach algebra $\mathscr{L}(X)$ together with the ideal of compact operators $\mathscr{K}(X)$. Then the Banach algebra $\mathscr{L}(X)/\mathscr{K}(X)$ is known as the **Calkin algebra**.[4] Atkinson's theorem (Theorem 7.14) says that the invertible elements in the Calkin algebra are precisely the images of the Fredholm operators. ⋄

Lemma 7.18. *Let X be a unital Banach algebra and m a character. Then* $\mathrm{Ker}(m)$ *is a maximal ideal and m is continuous with $\|m\| = m(e) = 1$.*

Proof. It is straightforward to check that $\mathrm{Ker}(m)$ is an ideal. Moreover, every x can be written as

$$x = m(x)e + y, \qquad y \in \mathrm{Ker}(m).$$

Let I be an ideal containing $\mathrm{Ker}(m)$. If there is some $x \in I \setminus \mathrm{Ker}(m)$ then $e = m(x)^{-1}(x - y) \in I$ and thus $\mathrm{Ker}(m)$ is maximal. Since maximal ideals are closed by the previous lemma, we conclude that m is continuous by Problem 1.60. Clearly $\|m\| \ge m(e) = 1$. Conversely, suppose we can find some $x \in X$ with $\|x\| < 1$ and $m(x) = 1$. Consequently $\|x^n\| \le \|x\|^n \to 0$ contradicting $m(x^n) = m(x)^n = 1$. □

In a commutative algebra the other direction is also true.

Lemma 7.19. *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 a corresponding maximal ideal $\mathrm{Ker}(m)$. Conversely, let I be a maximal ideal and consider the quotient map $\pi : X \to X/I$. We first claim that every nontrivial element in X/I is invertible. To this end suppose $[x_0] \ne [0]$ 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 5.4), 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 $\mathrm{Ker}(m) = I$. □

[4] John Williams Calkin (1909–1964), American mathematician

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**:

$$\Gamma : X \to C(\mathcal{M}), \quad x \mapsto \hat{x}(m) := m(x). \tag{7.36}$$

Theorem 7.20 (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 $\|\hat{x}\|_\infty \le r(x) \le \|x\|$ 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 6.10).

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

$$x \in \operatorname{Ker}(\Gamma) \quad \Leftrightarrow \quad x \in \bigcap_{m \in \mathcal{M}} \operatorname{Ker}(m) \tag{7.37}$$

and it can only be injective if X is commutative (if $xy \ne yx$, then $xy - yx \in \bigcap_{m \in \mathcal{M}} \operatorname{Ker}(m)$). In this case Lemma 7.19 implies

$$x \in \operatorname{Ker}(\Gamma) \quad \Leftrightarrow \quad x \in \operatorname{Rad}(X) := \bigcap_{I \text{ maximal ideal}} I, \tag{7.38}$$

where $\operatorname{Rad}(X)$ is known as the **Jacobson radical**[5] 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. Three examples where this can be done are given below. The first one is not very surprising.

Example 7.25. 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.14). 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 an open neighborhood $B(x)$ and a function $f_x \in C(K)$ such that $|f_x(y)| \geq 1$ for $y \in B(x)$. By compactness finitely many of these neighborhoods 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 7.26. 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 $\operatorname{Ker}(m_{x_0}) = I_{x_0}$.

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 (5.31)) as the Gelfand–Naimark[6] theorem below will show that the Gelfand transform is bijective for commutative C^* algebras. ⋄

Example 7.27. Consider the commutative algebra of complex matrices of the form $\left(\begin{smallmatrix} a & b \\ 0 & a \end{smallmatrix}\right)$. The only nontrivial ideal is given by matrices of the form $\left(\begin{smallmatrix} 0 & b \\ 0 & 0 \end{smallmatrix}\right)$ (show this). In particular, this algebra is not semi-simpel. ⋄

---

[5] Nathan Jacobson (1910–1999), American mathematician

[6] Mark Naimark (1909–1978), Soviet mathematician

Since $0 \notin \sigma(x)$ implies that x is invertible, the Gelfand representation theorem also contains a useful criterion for invertibility.

Corollary 7.21. *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 7.22 (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 7.23. *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 5.9) 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 7.24 (Gelfand–Naimark)**.** *Suppose X is a unital commutative C^* algebra. Then the Gelfand transform is an isometric isomorphism between C^* algebras.*

Proof. As in a commutative C^* algebra all elements are normal, Lemma 5.8 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.42). □

The first moral from this theorem is that from an abstract point of view every unital commutative C^* algebra is of the form $C(K)$ with K some compact Hausdorff space. Moreover, the formulation also very much resembles 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. (5.34)). 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 7.25 (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 $(y-\alpha)^*(y-\alpha)$. But this element is self-adjoint and hence has 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. Whence $(y-\alpha)^{-1} = ((y-\alpha)^*(y-\alpha))^{-1}(y-\alpha)^* \in Y$. □

Now we can extend Theorem 5.10 to the case of normal operators:

Theorem 7.26 (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

$$\sigma(f(x)) = f(\sigma(x)), \tag{7.39}$$

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 7.28. 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 7.15. *Show that $C^1[a,b]$ is a Banach algebra. What are the characters? Is it semi-simple?*

Problem 7.16. *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[7] coefficients via $f(z) = \sum_{k=0}^{\infty} \hat{f}_k z^k$ known as the* ***Hardy space***[8] *H^1 of the disc.)*

[7]Brook Taylor (1685–1731), English mathematician

[8]G. H. Hardy (1877–1947), English mathematician

Chapter 8

Unbounded operators

Unbounded operators in Banach spaces play an important role in applications to partial differential equations. In addition, the special case of (unbounded) self-adjoint operators in Hilbert spaces forms the foundation of quantum mechanics. We will, however, not pursue this latter case but we refer to the specialized literature (e.g. **[34]**, where you can start in Chapter 2) instead.

8.1. Closed operators

Recall from Section 4.1 that an operator $A : \mathfrak{D}(A) \subseteq X \to Y$ between Banach spaces X and Y is called closed if its graph $\Gamma(A) = \{(x, Ax) | x \in \mathfrak{D}(A)\} \subset X \oplus Y$ is closed. In this section we want to investigate which properties of bounded operators carry over to closed operators.

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 $y = Ax$.

Please observe that the domain $\mathfrak{D}(A)$ is an intrinsic part of the definition of A and that we cannot assume $\mathfrak{D}(A) = X$ unless A is bounded (which is the content of the closed graph theorem, Theorem 4.9). Hence, if we want an unbounded operator to be closed, we have to live with domains. We will however typically assume that $\mathfrak{D}(A)$ is dense and set

$$\mathscr{C}(X,Y) := \{A : \mathfrak{D}(A) \subseteq X \to Y | A \text{ is densely defined and closed}\}. \tag{8.1}$$

One writes $B \subseteq A$ if $\mathfrak{D}(B) \subseteq \mathfrak{D}(A)$ and $Bx = Ax$ for $x \in \mathfrak{D}(B)$. In this case A is called an **extension** of B.

Example 8.1. Two operators having the same prescription but different domains are different. For example

$$\mathfrak{D}(A) := C^1[0,1], \qquad Af := f'$$

and

$$\mathfrak{D}(B) := \{f \in C^1[0,1] | f(0) = f(1) = 0\}, \qquad Bf := f'$$

are two different operators in $X := C[0,1]$. Clearly A is an extension of B. Moreover, both are closed since $f_n \to f$ and $f_n' \to g$ implies that f is differentiable and $f' = g$. Note that A is densely defined while B is not. $\diamond$

Be aware that taking sums or products of unbounded operators is tricky due to the possible different domains. Indeed, if A and B are two operators between Banach spaces X and Y, so is $A + B$ defined on $\mathfrak{D}(A + B) := \mathfrak{D}(A) \cap \mathfrak{D}(B)$. The problem is that $\mathfrak{D}(A + B)$ might contain nothing more than zero. Similarly, if $A : \mathfrak{D}(A) \subseteq X \to Y$ and $B : \mathfrak{D}(B) \subseteq Y \to Z$, then the composition BA is defined on $\mathfrak{D}(BA) := \{x \in \mathfrak{D}(A) | Ax \in \mathfrak{D}(B)\}$.

Example 8.2. Consider $X := C_0[0,1]$. Let M be the subspace of trigonometric polynomials in X and N be the subspace of piecewise linear functions in X. Then both M and N are dense with $M \cap N = \{0\}$. $\diamond$

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 1.16). 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

$$\overline{\Gamma(A)} \cap \{(0,y) | y \in Y\} = \{(0,0)\}. \tag{8.2}$$

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. Any linear subset $\mathfrak{D} \subseteq \mathfrak{D}(A)$ with the property that A restricted to $\mathfrak{D}$ has the same closure, $\overline{A|_{\mathfrak{D}}} = \overline{A}$, is called a **core** for 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

$$\begin{aligned} \mathfrak{D}(\overline{A}) &= \{x \in X | \exists x_n \in \mathfrak{D}(A),\ y \in Y : x_n \to x \text{ and } Ax_n \to y\}, \\ \overline{A}x &= y. \end{aligned} \tag{8.3}$$

There is yet another way of defining the closure: Define the **graph norm** associated with A by

$$\|x\|_A := \|x\|_X + \|Ax\|_Y, \quad x \in \mathfrak{D}(A). \tag{8.4}$$

Since we have $\|Ax\| \le \|x\|_A$ we see that $A : \mathfrak{D}(A) \to Y$ is bounded with norm at most one. Thus far $(\mathfrak{D}(A), \|.\|_A)$ is a normed space and it suggests itself to consider its completion X_A. Then one can check (Problem 8.5) that X_A can be regarded as a subset of X if and only if A is closable. In this case the completion can be identified with $\mathfrak{D}(\overline{A})$ and the closure of A in X coincides with the extension from Theorem 1.16 of A in X_A. In particular, A is closed if and only if $(\mathfrak{D}(A), \|.\|_A)$ is complete.

Example 8.3. Consider the multiplication operator A in $\ell^p(\mathbb{N})$ defined by $Aa_j := ja_j$ on $\mathfrak{D}(A) := \ell_c(\mathbb{N})$, where $\ell_c(\mathbb{N})$ denotes the sequences with compact support (i.e., finitely many nonzero terms).

(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

$$\mathfrak{D}(\overline{A}) = \begin{cases} \{a \in \ell^p(\mathbb{N}) | (ja_j)_{j=1}^\infty \in \ell^p(\mathbb{N})\}, & 1 \le p < \infty, \\ \{a \in c_0(\mathbb{N}) | (ja_j)_{j=1}^\infty \in c_0(\mathbb{N})\}, & p = \infty, \end{cases}$$

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

$$a_j^n := \begin{cases} a_j, & j \le n, \\ 0, & j > n. \end{cases}$$

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 1.10) and setting $Aa = 0$ for every new element from this Hamel basis we obtain a (still unbounded) operator which is everywhere defined. However, this extension cannot be closed! ⋄

Example 8.4. Here is a simple example of a nonclosable operator: Let $X = Y := \ell^2(\mathbb{N})$ 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 $1 \le j \le n$ 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 8.5 (Sobolev[1] spaces)**.** Let $X := L^p(0,1)$, $1 \le p < \infty$, and consider $Af := f'$ on $\mathfrak{D}(A) := C^1[0,1]$. Then it is not hard to see that A is *not* closed (take a sequence g_n of continuous functions which converges in L^p

[1] Sergei Sobolev (1908–1989), Soviet mathematician

to a non-continuous function, cf. Example 1.12, and consider its primitive $f_n(x) = \int_0^x g_n(y)dy)$. It is however closable. To see this suppose $f_n \to 0$ and $f_n' \to g$ in L^p. Then $f_n(0) = f_n(x) - \int_0^x f_n'(y)dy \to -\int_0^x g(y)dy$. But a sequence of constant functions can only have a constant function as a limit implying $g \equiv 0$ as required. The domain of the closure is the **Sobolev space** $W^{1,p}(0,1)$ and this is one way of defining Sobolev spaces.

Note, that for $f \in W^{1,p}(0,1)$ there is a sequence $f_n \in C^1[0,1]$ such that $(f_n, f_n') \to (f,g)$ in $L^p(0,1) \oplus L^p(0,1)$. Moreover, one can show (Problem 8.6) that the sequence f_n converges uniformly implying $W^{1,p}(0,1) \subset C[0,1]$. Taking the limit in $f_n(x) = f_n(0) + \int_0^x f'(y)dy$ further shows $f(x) = f(0) + \int_0^x g(y)dy$. Since the fact that A is closable implies that g is uniquely determined, this justifies the notation $f' := g$. In the case $p = 1$ these functions are also known as the absolutely continuous functions $AC[0,1] := W^{1,1}(0,1)$.

Instead of the usual graph norm $W^{1,p}(0,1)$ one chooses the p-norm for the direct sum $X \oplus_p X$ such that the graph norm reads

$$\|f\|_{1,p} := \big(\|f\|_p^p + \|f'\|_p^p\big)^{1/p}.$$

In particular, the case $p = 2$ will lead to a Hilbert space usually denoted by $H^1(0,1) := W^{1,2}(0,1)$. ⋄

Example 8.6. Another example are point evaluations in $L^p(0,1)$, $1 \le p < \infty$: 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 $\mathfrak{D}(\ell_{x_0}) := C[0,1] \subseteq L^p(0,1)$. 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 8.2). ⋄

For the closure of sums and products see Problem 8.11 and Problem 8.12, respectively.

Given a subset $\Gamma \subseteq X \oplus Y$ we can define

$$\Gamma^{-1} := \{(y,x) | (x,y) \in \Gamma\} \subseteq Y \oplus X. \tag{8.5}$$

In particular, applying this to the graph of an operator A, we will obtain the graph of its inverse (provided A is invertible). Hence we see that an invertible operator is closed if and only if its inverse is closed. Slightly more general, we have:

Lemma 8.1. *Suppose an operator A between Banach spaces is closable and $\overline{A}$ is injective. Then $\overline{A}^{-1} = \overline{A^{-1}}$.*

Proof. This follows from $\Gamma(A^{-1}) = \Gamma^{-1}(A)$ and

$$\overline{\Gamma(A^{-1})} = \overline{\Gamma^{-1}(A)} = \overline{\Gamma(A)}^{-1} = \Gamma^{-1}(\overline{A}) = \Gamma(\overline{A}^{-1}). \qquad \square$$

Note that A injective does not imply $\overline{A}$ injective in general.

Example 8.7. 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 8.2. *Suppose $A \in \mathscr{C}(X,Y)$ is 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.*

Example 8.8. Note that in Example 8.3 the inverse $B := A^{-1}$ is the bounded operator $Ba_j := \frac{1}{j}a_j$ defined on $\mathfrak{D}(B) = \operatorname{Ran}(A) = \ell_c(\mathbb{N})$. Hence, the closure is $\overline{B}a_j := \frac{1}{j}a_j$ with

$$\mathfrak{D}(\overline{B}) = \overline{\operatorname{Ran}(A)} = \begin{cases} \ell^p(\mathbb{N}), & 1 \le p < \infty, \\ c_0(\mathbb{N}), & p = \infty. \end{cases}$$

Hence $\overline{B} \in \mathscr{L}(X)$ for $1 \le p < \infty$ but not for $p = \infty$. $\diamond$

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

$$\|Ax\| \ge c\|x\|, \qquad x \in \mathfrak{D}(A). \tag{8.6}$$

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. To this end we note:

Lemma 8.3. *Suppose $A \in \mathscr{C}(X,Y)$ is a closed operator between Banach spaces. Then $\operatorname{Ker}(A)$ is closed and the quotient space $\tilde{X} := X/\operatorname{Ker}(A)$ is a Banach space. Moreover, $\tilde{A} : \mathfrak{D}(\tilde{A}) \to Y$ where $\mathfrak{D}(\tilde{A}) = \mathfrak{D}(A)/\operatorname{Ker}(A) \subseteq \tilde{X}$, is a closed operator with $\operatorname{Ker}(\tilde{A}) = \{0\}$ and $\operatorname{Ran}(\tilde{A}) = \operatorname{Ran}(A)$.*

Proof. It is easy to check that $\operatorname{Ker}(A)$ is closed (Problem 8.1). Then $\tilde{A}$ defined via $\tilde{A}[x] := Ax$ is well defined and injective (cf. Problem 1.71). 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 $\tilde{X} \times Y \setminus \Gamma(\tilde{A})$ is open. $\square$

Applying Corollary 8.2 to $\tilde{A}$ gives:

Corollary 8.4. *Suppose $A \in \mathscr{C}(X,Y)$. Then $\operatorname{Ran}(A)$ is closed if and only if*

$$\|Ax\| \ge c \operatorname{dist}(x, \operatorname{Ker}(A)), \qquad x \in \mathfrak{D}(A), \tag{8.7}$$

for some $c > 0$. The sup over all possible c is $\|\tilde{A}^{-1}\|^{-1}$ and is known as the (reduced) ***minimum modulus*** *of A.*

There is also another variant which allows Ax to deviate from y:

Corollary 8.5. *Suppose $A : \mathfrak{D}(A) \subseteq X \to Y$ is closed. Then $\operatorname{Ran}(A)$ is closed if there exists some $0 \le \delta < 1$ and some $\varepsilon > 0$ such that for every $y \in \overline{\operatorname{Ran}(A)}$ we can find a corresponding $x \in \mathfrak{D}(X)$ satisfying*

$$\varepsilon\|x\| + \|y - Ax\| \le \delta\|y\|. \tag{8.8}$$

Conversely, if $\operatorname{Ran}(A)$ is closed this can be done for all $0 \le \delta < 1$ 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 $\|Ax\| \ge \frac{\varepsilon}{\delta}\|x\|$ after maybe adding an element from the kernel to x. This x satisfies $\varepsilon\|x\| + \|y - Ax\| = \varepsilon\|x\| \le \delta\|y\|$ as required.

Conversely, fix $y \in \overline{\operatorname{Ran}(A)}$ and recursively choose a sequence x_n (starting with $x_0 = 0$) such that

$$\varepsilon\|x_n\| + \|(y - A\tilde{x}_{n-1}) - Ax_n\| \le \delta\|y - A\tilde{x}_{n-1}\|, \quad \tilde{x}_n := \sum_{m\le n} x_m.$$

In particular, $\|y - A\tilde{x}_n\| \le \delta^n\|y\|$ as well as $\varepsilon\|x_n\| \le \delta^n\|y\|$, which shows $\tilde{x}_n \to x$ and $A\tilde{x}_n \to y$. Hence $x \in \mathfrak{D}(A)$ and $y = Ax \in \operatorname{Ran}(A)$. □

As for bounded operators we can define the **adjoint operator** using $(A'\ell)(x) := \ell(Ax)$. Of course the linear functional $A'\ell$ will only be defined on $\mathfrak{D}(A)$ and it will be unbounded in general. Hence we define the domain $\mathfrak{D}(A')$ as the set

$$\mathfrak{D}(A') := \{\ell \in Y^* | \ell \circ A \text{ is bounded}\}. \tag{8.9}$$

If we assume $\mathfrak{D}(A)$ to be dense, then Theorem 1.16 implies that $\ell \circ A$ has a unique extension to an element of X^* and we can set

$$A'\ell := \overline{\ell \circ A}, \qquad \ell \in \mathfrak{D}(A'). \tag{8.10}$$

However, note that even if A is densely defined, it can happen that $\mathfrak{D}(A') = \{0\}$ (Problem 8.14). Clearly this extends our previous definition for bounded operators and it is not hard to see that $A \in \mathscr{C}(X,Y)$ is bounded if and only if A' is bounded (Problem 8.13). Finally note that for densely defined operators we have

$$A \subseteq B \quad \Rightarrow \quad B' \subseteq A'. \tag{8.11}$$

Example 8.9. Consider again A from Example 8.3 with $1 \le p < \infty$. Then $X^* \cong \ell^q(\mathbb{N})$ with $\frac{1}{q} + \frac{1}{p} = 1$. Now if $b \in \mathfrak{D}(A')$ there is some $c \in \ell^q(\mathbb{N})$ such that $\ell_c(a) = \ell_b(Aa)$ for all $a \in \ell_c(\mathbb{N})$. That is,

$$\sum_{j=1}^{\infty} c_j a_j = \sum_{j=1}^{\infty} b_j (j a_j), \qquad a \in \ell_c(\mathbb{N}).$$

Choosing $a = \delta^j$ shows $c_j = j b_j$ and hence $A' b_j = j b_j$ with $\mathfrak{D}(A') = \{b \in \ell^q(\mathbb{N}) | (j b_j)_{j=1}^{\infty} \in \ell^q(\mathbb{N})\}$. Note that for $1 < p < \infty$ the domain is dense, while for $p = 1$ we have $\overline{\mathfrak{D}(A')} = c_0(\mathbb{N}) \subset \ell^\infty(\mathbb{N})$. ⋄

Example 8.10. Let us compute the adjoint of B from Example 8.4. Proceeding as in the previous example we get $c_j = b_1$ which is in ℓ^2 if and only if $b_1 = 0$. Thus $\mathfrak{D}(B') = \{b \in \ell^2(\mathbb{N}) | b_1 = 0\}$ and $B' b = 0$. So don't expect the adjoint of a noncloseable operator to contain much information about the operator. ⋄

For the adjoint of sums and products see Problem 8.16 and Problem 8.17, respectively.

There are two other ways of introducing the adjoint operator which are worth while mentioning. First of all an operator $B : \mathfrak{D}(B) \subseteq Y^* \to X^*$ is adjoint to $A : \mathfrak{D}(A) \subseteq X \to Y$ if

$$(By')(x) = y'(Ax), \qquad x \in \mathfrak{D}(A), y' \in \mathfrak{D}(B). \tag{8.12}$$

Then, if $\mathfrak{D}(A)$ is dense, there is a unique maximally adjoint operator, which is precisely the adjoint A'. The second way is using graphs. To this end recall that

$$\Gamma^\perp = \{(x', y') \in X^* \oplus Y^* | x'(x) + y'(y) = 0 \, \forall (x, y) \in \Gamma\} \tag{8.13}$$

and hence the last equality can be rephrased as

$$\Gamma(B) \subseteq \Gamma^{-1}(-A)^\perp = (\Gamma(-A)^\perp)^{-1} \tag{8.14}$$

with equality if $B = A'$. Hence

$$\begin{aligned} \Gamma(A') &= \Gamma^{-1}(-A)^\perp = (\Gamma(-A)^\perp)^{-1} \\ &= \{(y', x') \in X^* \oplus Y^* | x'(x) - y'(y) = 0 \, \forall (x, y) \in \Gamma(A)\} \\ &= \{(y', x') \in X^* \oplus Y^* | x'(x) - y'(Ax) = 0 \, \forall x \in \mathfrak{D}(A)\}. \end{aligned} \tag{8.15}$$

In particular, we conclude that A' is closed. Note that $\Gamma^{-1}(-A)^\perp$ is the graph of an operator if and only if $\mathfrak{D}(A)^\perp = \{0\}$, that is, if and only if $\mathfrak{D}(A)$ is dense (recall (4.7)). Moreover, since $\Gamma^\perp = \overline{\Gamma}^\perp$ we have

$$A' = \overline{A}' \tag{8.16}$$

if A is closable.

If A' is densely defined we can compute the closure by computing the doubly adjoint operator.

Theorem 8.6. *Suppose X and Y are Banach spaces and $A : \mathfrak{D}(A) \subseteq X \to Y$ is densely defined. Then A is closable if and only if $\mathfrak{D}(A')_\perp = \{0\}$. Moreover, if A' is densely defined, then $\overline{A} = J_Y^{-1} A'' J_X$ with $\mathfrak{D}(\overline{A}) = J_X^{-1}\mathfrak{D}(A'')$ and $\operatorname{Ran}(A''|_{J_X(X)}) \subseteq J_Y(Y)$.*

Proof. By Lemma 4.20 we have $\Gamma^{-1}(-A')_\perp = (\Gamma(A)^\perp)_\perp = \overline{\Gamma(A)}$. Hence $\overline{\Gamma(A)}$ is a graph if and only if $\mathfrak{D}(A')_\perp = \{0\}$. Hence A is closable if $\mathfrak{D}(A')$ is dense and the converse holds if Y is reflexive.

Moreover, if $\mathfrak{D}(A')$ is dense, A'' exists and $J\Gamma(\overline{A}) = J\Gamma^{-1}(-A')_\perp \subseteq \Gamma^{-1}(-A')^\perp = \Gamma(A'')$ which shows $\overline{A} = J_Y^{-1} A'' J_X$. □

Of course if Y is reflexive, then the last theorem says that A is closable if and only if A' is densely defined. If Y is not reflexive, this formulation still holds if dense is replaced by weak-$*$ dense — see Problem 6.20. Without this replacement it can fail as we have seen in Example 8.9.

Finally we want to establish the closed range theorem for closed operators. We begin by extending Lemma 4.26.

Lemma 8.7. *Let X, Y be Banach spaces and $A \in \mathscr{C}(X, Y)$, then $\operatorname{Ran}(A)^\perp = \operatorname{Ker}(A')$ and $\operatorname{Ran}(A')_\perp = \operatorname{Ker}(A)$.*

Proof. For the first claim observe: $y' \in \operatorname{Ker}(A')$ implies $0 = (A'y')(x) = y'(Ax)$ for all $x \in \mathfrak{D}(A)$ and hence $y' \in \operatorname{Ran}(A)^\perp$. Conversely, $y' \in \operatorname{Ran}(A)^\perp$ implies $y'(Ax) = 0$ for all $x \in \mathfrak{D}(A)$ and hence $y' \in \mathfrak{D}(A')$ with $A'y' = 0$.

For the second claim observe: $x \in \operatorname{Ker}(A)$ implies $(A'y')(x) = y'(Ax) = 0$ for all $y' \in \mathfrak{D}(A')$ and hence $x \in \operatorname{Ran}(A')_\perp$. Conversely, $x \in \operatorname{Ran}(A')_\perp$ implies $(A'y')(x) = 0$ for all $y' \in \mathfrak{D}(A')$. If we had $(x, 0) \notin \Gamma(A)$, we could find (Corollary 4.15) a linear functional $(x', y') \in X^* \oplus Y^*$ which vanishes on $\Gamma(A)$ and satisfies $(x', y')(x, 0) = 1$. The fact that it vanishes on $\Gamma(A)$ implies $x'(x) + y'(Ax) = 0$ for all $x \in \mathfrak{D}(A)$ implying $y' \in \mathfrak{D}(A')$ with $A'y' = -x'$. But this gives the contradiction $0 = (A'y')(x) = -x'(x) = -1$ and hence $(x, 0) \in \Gamma(A)$, that is, $x \in \operatorname{Ker}(A)$. □

Taking annihilators in these formulas we again obtain

$$\operatorname{Ker}(A')_\perp = (\operatorname{Ran}(A)^\perp)_\perp = \overline{\operatorname{Ran}(A)} \tag{8.17}$$

and

$$\operatorname{Ker}(A)^\perp = (\operatorname{Ran}(A')_\perp)^\perp \supseteq \overline{\operatorname{Ran}(A')}. \tag{8.18}$$

The analog of Theorem 4.25 is

Theorem 8.8. *Suppose X, Y are Banach spaces. If $A \in \mathscr{C}(X, Y)$, then A^{-1} exists and is in $\mathscr{L}(Y, X)$ if and only if $(A')^{-1}$ exists and is in $\mathscr{L}(X^*, Y^*)$. Moreover, in this case we have*

$$(A')^{-1} = (A^{-1})'. \tag{8.19}$$

Proof. If $A^{-1} \in \mathscr{L}(Y, X)$, then Theorem 4.23 implies $(A^{-1})' \in \mathscr{L}(X^*, Y^*)$ and $\Gamma((A^{-1})') = \Gamma^{-1}(-A^{-1})^{\perp} = (\Gamma^{-1}(-A)^{\perp})^{-1} = \Gamma(A')^{-1} = \Gamma((A')^{-1})$ which establishes (8.19).

Conversely, let $(A')^{-1} \in \mathscr{L}(X^*, Y^*)$. Then for any $\ell \in X^*$ and $x \in \mathfrak{D}(A)$ we have $((A')^{-1}\ell)(Ax) = (A'(A')^{-1}\ell)(x) = \ell(x)$ and choosing ℓ normalized such that $\ell(x) = \|x\|$ (Corollary 4.13) we obtain $\|x\| = ((A')^{-1}\ell)(Ax) \leq \|(A')^{-1}\|\|Ax\|$. This shows that A has a bounded inverse with $\|A^{-1}\| \leq \|(A')^{-1}\|$ and hence $\operatorname{Ran}(A)$ is closed by Corollary 8.2. Finally $\operatorname{Ran}(A)^{\perp} = \operatorname{Ker}(A')$ shows $\operatorname{Ran}(A) = Y$. □

Now we are ready to show:

Theorem 8.9 (Closed range)**.** *Suppose X, Y are Banach spaces and $A \in \mathscr{C}(X, Y)$. Then the following items are equivalent:*

(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. Consider $\tilde{X} = X/\operatorname{Ker}(A)$ and $\tilde{Y} = \overline{\operatorname{Ran}(A)}$ and the corresponding operator $\tilde{A}$ as in Lemma 8.3. Then $\tilde{A}$ is a closed injective operator whose range is dense. Now you can literally follow the proof of Theorem 4.28 replacing the used results for bounded operators by the corresponding ones for closed operators. □

We end this section with the remark that one could try to define the concept of a weakly closed operator by replacing the norm topology in the definition of a closed operator by the weak topology. However, Theorem 6.12 implies that this gives nothing new:

Lemma 8.10. *For an operator $A : \mathfrak{D}(A) \subseteq X \to Y$ between Banach spaces the following are equivalent:*

- $\Gamma(A)$ *is closed.*
- $x_n \in \mathfrak{D}(A)$ *with* $x_n \to x$ *and* $Ax_n \to y$ *implies* $x \in \mathfrak{D}(A)$ *and* $y = Ax$.
- $\Gamma(A)$ *is weakly closed.*
- $x_n \in \mathfrak{D}(A)$ *with* $x_n \rightharpoonup x$ *and* $Ax_n \rightharpoonup y$ *implies* $x \in \mathfrak{D}(A)$ *and* $Ax = y$.

Problem* 8.1. *Show that the kernel* $\operatorname{Ker}(A)$ *of a closed operator* A *is closed.*

Problem* 8.2. *A linear functional defined on a dense subspace is closable if and only if it is bounded.*

Problem 8.3. *Let $(m_j)_{j\in\mathbb{N}}$ be a sequence of complex numbers and consider the multiplication operator M in $\ell^p(\mathbb{N})$ defined by $Ma_j := m_j a_j$ on $\mathfrak{D}(A) := \ell_c(\mathbb{N})$. Under which conditions on m is M closable and what is its closure?*

Problem 8.4. *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* 8.5. *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 1.16 of A in X_A. In particular, A is closed if and only if $(\mathfrak{D}(A), \|.\|_A)$ is complete.*

Problem* 8.6. *Consider the Sobolev spaces $W^{1,p}(0,1)$. Show that*

$$\|f\|_\infty \le 2^{1-1/p}\|f\|_{1,p}$$

and conclude that functions from $W^{1,p}(0,1)$ are continuous. (Hint: Start with estimating $f(a) = f(x) - \int_a^x f'(y)dy$ and integrate the result.)

Problem 8.7. *Show that $f \in W^{1,p}(0,1)$, $1 < p < \infty$ is Hölder continuous with*

$$|f(x) - f(y)| \le \|f'\|_p |x-y|^{1-\frac{1}{p}}.$$

Conclude that the embedding $W^{1,p}(0,1) \hookrightarrow L^p(0,1)$ is compact.

Problem 8.8. *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.*

Problem 8.9. *Let $X := C[0,1]$. Show that the operator given by $Ax(t) := \frac{x(t)}{t}$ with $\mathfrak{D}(A) := \{x \in X | \exists \lim_{x\to 0} \frac{x(t)}{t}\}$ is closed.*

Problem 8.10. *Let $X := C[0,1]$. Show that the operator given by $Ax := x'' + x$ with $\mathfrak{D}(A) := \{x(t) \in C^2[0,1] | x(0) = x'(0) = 0\}$ is closed.*

Problem* 8.11. *Show that if $A : \mathfrak{D}(A) \subseteq X \to Y$ is closable and $B \in \mathscr{L}(X,Y)$, then $\overline{A} + B = \overline{A+B}$. Moreover, $A+B$ is closable if and only if A is.*

Give an example where A and B are closable but $A+B$ is not. (Hint: For the very last part see Problem 8.8.)

Problem* 8.12. *Let $A : \mathfrak{D}(A) \subseteq X \to Y$ and $B : \mathfrak{D}(B) \subseteq Y \to Z$ be closable. If $A \in \mathscr{L}(X,Y)$, then BA is closable and $\overline{BA} = \overline{B}A$. Similarly, if $B^{-1} \in \mathscr{L}(Z,Y)$, then BA is closable and $\overline{BA} = B\overline{A}$.*

Problem* 8.13. *Show that for $A \in \mathscr{C}(X,Y)$ we have $A \in \mathscr{L}(X,Y)$ if and only if $A' \in \mathscr{L}(Y^*, X^*)$.*

Problem 8.14. *Show that $\mathfrak{D}(A')$ is trivial if and only if $\Gamma(A) \subset X \times Y$ is dense. Let $X = Y$ be a separable Hilbert space and choose an ONB $\{u_n\}_{n\in\mathbb{N}}$ plus a dense set $\{x_n\}_{n\in\mathbb{N}}$. Define $Au_n := x_n$ and extend A by linearity to $\mathfrak{D}(A) = \operatorname{span}\{u_n\}_{n\in\mathbb{N}}$. Show that $\Gamma(A)^\perp = \{(0,0)\}$ and conclude $\mathfrak{D}(A') = \{0\}$.*

Problem 8.15. *Compute the adjoint of the embedding $I : \ell^1(\mathbb{N}) \hookrightarrow \ell^2(\mathbb{N})$.*

Problem* 8.16. *Show that $A' + B' \subseteq (A+B)'$ if $A+B$ is densely defined. Show that we have equality if one of the operators is bounded. Give an example where equality fails.*

Problem* 8.17. *Let $A : \mathfrak{D}(A) \subseteq X \to Y$ and $B : \mathfrak{D}(B) \subseteq Y \to Z$ be densely defined. If $B \in \mathscr{L}(Y,Z)$, then BA is densely defined and $(BA)' = A'B'$. Similarly, if $A^{-1} \in \mathscr{L}(Y,X)$, then BA is densely defined and $(BA)' = A'B'$.*

Problem 8.18. *Let A be a closed operator. Show that for every $\alpha \in \rho(A)$ the expression $\|f\|_\alpha := \|(A-\alpha)x\|$ defines a norm which is equivalent to the graph norm.*

Problem 8.19. *Let X_j be Banach spaces. A sequence of operators $A_j \in \mathscr{C}(X_j, X_{j+1})$*

$$X_1 \xrightarrow{A_1} X_2 \xrightarrow{A_2} X_3 \cdots X_n \xrightarrow{A_n} X_{n+1}$$

*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* 8.20. *Let $A : \mathfrak{D}(A) \subseteq X \to X$ be an unbounded operator and $B \in \mathscr{L}(X)$ bounded. Then A and B are said to **commute** if*

$$BA \subseteq AB.$$

Of course if $A \in \mathscr{L}(X)$, then we have equality and this definition reduces to the usual one. Note that the definition implies in particular, that B leaves $\mathfrak{D}(A)$ invariant, $B\mathfrak{D}(A) \subseteq \mathfrak{D}(A)$.

Show that if A is invertible, then A commutes with B if and only if A^{-1} commutes with B. Conclude that if A has a nonempty resolvent set, then A commutes with B if and only if $R_A(\alpha)$ commutes with B for one $\alpha \in \rho(A)$. Moreover, in this case this holds for all $\alpha \in \rho(A)$.

Problem* 8.21. *Let $A : \mathfrak{D}(A) \subseteq X \to X$ be closable and $B \in \mathscr{L}(X)$ bounded. Then if A commutes with B so does $\overline{A}$.*

8.2. Spectral theory for unbounded operators

Let $A \in \mathscr{C}(X)$ be a closed operator on a Banach space X. As in the case of bounded operators (cf. Section 5.1) we define the **resolvent set** via

$$\rho(A) := \{\alpha \in \mathbb{C} | A - \alpha : \mathfrak{D}(A) \to X \text{ is bijective with a bounded inverse}\} \tag{8.20}$$

and call

$$R_A(\alpha) := (A - \alpha)^{-1}, \qquad \alpha \in \rho(A) \tag{8.21}$$

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:

Lemma 8.11. *Let $A \in \mathscr{C}(X)$ be a closed operator on a Banach space X. Then the resolvent set is open and if $\alpha_0 \in \rho(A)$ we have*

$$R_A(\alpha) = \sum_{n=0}^{\infty} (\alpha - \alpha_0)^n R_A(\alpha_0)^{n+1}, \quad |\alpha - \alpha_0| < \|R_A(\alpha_0)\|^{-1}. \tag{8.22}$$

In particular, the resolvent is analytic and

$$\|(A - \alpha)^{-1}\| \geq \frac{1}{\operatorname{dist}(\alpha, \sigma(A))}. \tag{8.23}$$

Proof. By comparing with the geometric series one sees that the series for $R_A(\alpha)$ converges and hence defines a bounded operator. Denote this operator by R. Now fix $x \in X$ and consider $R_N := \sum_{n=0}^{N} (\alpha - \alpha_0)^n R_A(\alpha_0)^{n+1}$. Then $x_N := R_N x \in \mathfrak{D}(A)$ and

$$\begin{aligned}
(A - \alpha)x_N &= (A - \alpha_0)x_N - (\alpha - \alpha_0)x_N \\
&= \sum_{n=0}^{N} (\alpha - \alpha_0)^n R_A(\alpha_0)^n x - \sum_{n=0}^{N} (\alpha - \alpha_0)^{n+1} R_A(\alpha_0)^{n+1} \\
&= x - (\alpha - \alpha_0)^{N+1} R_A(\alpha_0)^{N+1}.
\end{aligned}$$

Hence $(A - \alpha)x_N \to x$ implying $x \in \mathfrak{D}(A)$ and $(A - \alpha)Rx = x$. Similarly one shows $R(A - \alpha)x = x$ for $x \in \mathfrak{D}(A)$. $\square$

It is also straightforward to verify the **first resolvent identity**

$$\begin{aligned}
R_A(\alpha_0) - R_A(\alpha_1) &= (\alpha_0 - \alpha_1) R_A(\alpha_0) R_A(\alpha_1) \\
&= (\alpha_0 - \alpha_1) R_A(\alpha_1) R_A(\alpha_0),
\end{aligned} \tag{8.24}$$

for $\alpha_0, \alpha_1 \in \rho(A)$.

As a consequence of the previous lemma we obtain the useful

Lemma 8.12. *Let $A \in \mathscr{C}(X)$ be a closed operator on a Banach space X. We have $\alpha \in \sigma(A)$ if there is a sequence $x_n \in \mathfrak{D}(A)$ such that $\|x_n\| = 1$ and $\|(A-\alpha)x_n\| \to 0$. If α is a boundary point of $\rho(A)$, then the converse is also true. Such a sequence is called a* ***Weyl sequence****.*

Proof. Let x_n be a Weyl sequence. Then $\alpha \in \rho(A)$ is impossible by $1 = \|x_n\| = \|R_A(\alpha)(A-\alpha)x_n\| \le \|R_A(\alpha)\|\|(A-\alpha)x_n\| \to 0$. Conversely, by (8.23), there is a sequence $\alpha_n \to \alpha$ and corresponding nonzero vectors $y_n \in X$ such that $\|R_A(\alpha_n)y_n\|\|y_n\|^{-1} \to \infty$. Let $x_n := R_A(\alpha_n)y_n$ and rescale y_n such that $\|x_n\| = 1$. Then $\|y_n\| \to 0$ and hence

$$\|(A-\alpha)x_n\| = \|y_n + (\alpha_n - \alpha)x_n\| \le \|y_n\| + |\alpha - \alpha_n| \to 0$$

shows that x_n is a Weyl sequence. □

The set of all $\alpha \in \mathbb{C}$ for which a Weyl sequence exists is called the **approximate point spectrum**

$$\sigma_{ap}(A) := \{\alpha \in \mathbb{C} | \exists x_n \in \mathfrak{D}(A) : \|x_n\| = 1, \|(A-\alpha)x_n\| \to 0\} \tag{8.25}$$

and the above lemma could be phrased as

$$\partial\sigma(A) \subseteq \sigma_{ap}(A) \subseteq \sigma(A). \tag{8.26}$$

Note that there are two possibilities if $\alpha \in \sigma_{ap}(A)$: Either α is an eigenvalue or $(A-\alpha)^{-1}$ exists but is unbounded. By the closed graph theorem the latter case is equivalent to the fact that $\operatorname{Ran}(A-\alpha)$ is not closed. In particular, we have

$$\sigma_p(A) \cup \sigma_c(A) \subseteq \sigma_{ap}(A) \subseteq \sigma(A). \tag{8.27}$$

Example 8.11. If α is an eigenvalue and x a corresponding normalized eigenfunction, then $x_n := x$ will be a Weyl sequence. Hence you should think of x_n as approximate eigenfunctions. Conversely, if a Weyl sequence converges, $x_n \to x$, then we also have $Ax_n \to \alpha x$ which shows that x is an eigenfunction (assuming A is closed). However, in general a Weyl sequence does not have to converge. Indeed, consider the multiplication operator $(Aa)_j = \frac{1}{j}a_j$ in $\ell^p(\mathbb{N})$. Then $x_n := \delta^n$ is a Weyl sequence for $\alpha = 0$, but it clearly has no convergent subsequence. ⋄

However, note that for unbounded operators the spectrum will no longer be bounded in general and both $\sigma(A) = \emptyset$ as well as $\sigma(A) = \mathbb{C}$ are possible.

Example 8.12. Consider $X := C[0,1]$ and $A = \frac{d}{dx}$ with $\mathfrak{D}(A) = C^1[0,1]$. We obtain the eigenvalues by solving the ordinary differential equation $x'(t) = \alpha x(t)$ which gives $x(t) = \mathrm{e}^{\alpha t}$. Hence every $\alpha \in \mathbb{C}$ is an eigenvalue, that is, $\sigma(A) = \mathbb{C}$.

Now let us modify the domain and look at $A_0 = \frac{d}{dx}$ with $\mathfrak{D}(A_0) = \{x \in C^1[0,1] | x(0) = 0\}$ and $X_0 := \{x \in C[0,1] | x(0) = 0\}$. Then the previous eigenfunctions do not satisfy the boundary condition $x(0) = 0$ and hence A_0 has no eigenvalues. Moreover, the solution of the inhomogeneous ordinary differential equation $x'(t) - \alpha x(t) = y(t)$ is given by $x(t) = x(0)\mathrm{e}^{\alpha t} + \int_0^t \mathrm{e}^{\alpha(t-s)} y(s)ds$. Hence $R_{A_0}(\alpha)y(t) = \int_0^t \mathrm{e}^{\alpha(t-s)} y(s)ds$ is the resolvent of A_0. Consequently $\sigma(A_0) = \emptyset$. ⋄

Note that if A is closed, then bijectivity implies boundedness of the inverse (see Corollary 8.2). Moreover, by Lemma 8.1 an operator with nonempty resolvent set must be closed.

The point, continuous, and residual spectrum can be defined as in Section 7.1 and Lemma 7.1 holds for closed operators.

Lemma 8.13. *Suppose $A \in \mathscr{C}(X)$ is a closed operator on some Banach space X. Then*

$$\sigma(A) = \sigma(A') \tag{8.28}$$

and

$$R_A(\alpha)' = R_{A'}(\alpha), \qquad \alpha \in \rho(A) = \rho(A'). \tag{8.29}$$

Moreover,

$$\begin{aligned} &\sigma_p(A') \subseteq \sigma_p(A) \cup \sigma_r(A), && \sigma_p(A) \subseteq \sigma_p(A') \cup \sigma_r(A'),\\ &\sigma_r(A') \subseteq \sigma_p(A) \cup \sigma_c(A), && \sigma_r(A) \subseteq \sigma_p(A'),\\ &\sigma_c(A') \subseteq \sigma_c(A), && \sigma_c(A) \subseteq \sigma_r(A') \cup \sigma_c(A'). \end{aligned} \tag{8.30}$$

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. Literally follow the proof of Lemma 7.1 using the corresponding results for closed operators: Theorem 8.8, Lemma 8.7, and Theorem 8.6. □

Let us also note the following spectral mapping result.

Lemma 8.14. *Suppose $A \in \mathscr{C}(X)$ is an injective and closed operator on some Banach space X with $\mathrm{Ran}(A)$ dense. Then*

$$\sigma(A^{-1}) \setminus \{0\} = (\sigma(A) \setminus \{0\})^{-1} \tag{8.31}$$

and

$$R_{A^{-1}}(\alpha^{-1}) = -\alpha A R_A(\alpha) = -\alpha - \alpha^2 R_A(\alpha), \qquad \alpha \in \rho(A) \setminus \{0\}. \tag{8.32}$$

In addition, for $\alpha \neq 0$ we have $\mathrm{Ker}((A-\alpha))^n = \mathrm{Ker}((A^{-1} - \alpha^{-1})^n)$ as well as $\mathrm{Ran}((A-\alpha))^n = \mathrm{Ran}((A^{-1} - \alpha^{-1})^n)$ for any $n \in \mathbb{N}$.

Proof. Throughout this proof we assume $\alpha \neq 0$. We first show the formula for the resolvent of A^{-1}. To this end choose some $\alpha \in \rho(A) \setminus \{0\}$ and note that $-\alpha A R_A(\alpha) = -\alpha(A-\alpha+\alpha)R_A(\alpha) = -\alpha - \alpha^2 R_A(\alpha)$ is a bounded operator from $X \to \operatorname{Ran}(A) = \mathfrak{D}(A^{-1})$ and

$$(A^{-1}-\alpha^{-1})(-\alpha A R_A(\alpha))x = (-\alpha + A)R_A(\alpha)x = x, \quad x \in X.$$

Conversely, if $y \in \mathfrak{D}(A^{-1}) = \operatorname{Ran}(A)$, we have $y = Ax$ and hence

$$(-\alpha A R_A(\alpha))(A^{-1}-\alpha^{-1})y = A R_A(\alpha)((A-\alpha)x) = Ax = y.$$

Thus (8.32) holds and $\alpha^{-1} \in \rho(A^{-1})$. Interchanging the roles of A and A^{-1} establishes the first part.

Next note that for $x \in \mathfrak{D}(A)$ the equation $(A-\alpha)x = y$ is equivalent to $x = \frac{1}{\alpha}(Ax - y)$ and hence $(A-\alpha)x \in \operatorname{Ran}(A^m)$ implies $x \in \operatorname{Ran}(A^m)$ for any $m \in \mathbb{N}$. Consequently we get that $x \in \operatorname{Ker}((A-\alpha)^n)$ implies $x \in \operatorname{Ran}(A^m) = \mathfrak{D}(A^{-m})$ for any $m \in \mathbb{N}$ and $0 = A^{-n}(A-\alpha)^n\alpha = (-\alpha)^n(A^{-1}-\alpha^{-1})^n x$. So $\operatorname{Ker}((A-\alpha)^n) \subseteq \operatorname{Ker}((A^{-1}-\alpha^{-1})^n)$ and equality follows by reversing the roles of A and A^{-1}.

Similarly, $x \in \operatorname{Ran}((A-\alpha)^n)$ implies $x = (A-\alpha)^n y$ for some $y \in \mathfrak{D}(A^n) = \operatorname{Ran}(A^{-n})$. Consequently $y = A^{-n}z$ for some $z \in \mathfrak{D}(A^{-n}) = \operatorname{Ran}(A^n)$ and $x = (A-\alpha)^n y = (A-\alpha)^n A^{-n} z = (-\alpha)^n (A^{-1}-\alpha^{-1})^n z \in \operatorname{Ran}((A^{-1}-\alpha^{-1})^n)$. So $\operatorname{Ran}((A-\alpha)^n) \subseteq \operatorname{Ran}((A^{-1}-\alpha^{-1})^n)$ and equality follows again by reversing the roles of A and A^{-1}. □

Concerning $\alpha = 0$ note that $0 \in \sigma(A^{-1})$ if and only if A is unbounded and vice versa.

Example 8.13. If A is bounded with $\sigma(A) = \sigma_c(A) = \{0\}$, then $(A^{-1} - \alpha^{-1}) = A(\mathbb{I} - \alpha A)^{-1}$ for all $\alpha \in \mathbb{C}$ and $\sigma(A^{-1}) = \emptyset$ in this case. ⋄

In particular we can apply this lemma to the resolvent in case $\alpha_0 \in \rho(A)$ which shows

$$\sigma(A) = \alpha_0 + (\sigma(R_A(\alpha_0)) \setminus \{0\})^{-1} \tag{8.33}$$

and $\operatorname{Ker}(R_A(\alpha_0)-\alpha)^n = \operatorname{Ker}(A-\alpha_0-\frac{1}{\alpha})^n$ as well as $\operatorname{Ran}(R_A(\alpha_0)-\alpha)^n = \operatorname{Ran}(A-\alpha_0-\frac{1}{\alpha})^n$ for $\alpha \neq 0$ and $n \in \mathbb{N}$.

For example, this can be used to apply Theorem 7.7 to unbounded operators in case they have a compact resolvent. To this end note that if we have $R_A(\alpha) \in \mathscr{K}(X)$ for one $\alpha \in \rho(A)$, then this holds in fact for all $\alpha \in \rho(A)$ by the first resolvent identity (8.24) since compact operators form an ideal.

Theorem 8.15. *Let $A \in \mathscr{C}(X)$ be a closed operator on some Banach space X. Suppose $R_A(\alpha) \in \mathscr{K}(X)$ for one $\alpha \in \rho(A)$. Then the spectrum of A consists only of discrete eigenvalues with finite (geometric and algebraic) multiplicity and we have the splitting into closed invariant subspaces*

$$X = \operatorname{Ker}(A-\alpha)^n \dotplus \operatorname{Ran}(A-\alpha)^n, \qquad \alpha \in \sigma(A), \tag{8.34}$$

where n is the index of α.

Proof. The claim follows by combining the previous lemma with the spectral theorem for compact operators (Theorem 7.7 and Lemma 7.5). □

Example 8.14. Consider again A_0 from the previous problem. Then

$$A_0^{-1} : X_0 \to X_0, \quad x \mapsto x(t) = \int_0^t x(s) ds$$

is compact by Problem 3.7 (cf. also Problem 7.4). Hence we get again $\sigma(A_0) = \emptyset$ without the need of computing the resolvent. ⋄

Example 8.15. Of course another example of unbounded operators with compact resolvent are regular Sturm–Liouville operators as shown in Section 3.3. ⋄

This result says in particular, that for $\alpha \in \sigma_p(A)$ we can split $X = X_0 \oplus X_1$ where both $X_0 := \mathrm{Ker}(A-\alpha)^n$ and $X_1 := \mathrm{Ran}(A-\alpha)^n$ are invariant subspaces for A. Consequently we can split $A = A_0 \oplus A_1$, where A_0 is the restriction of A to the finite dimensional subspace X_0 (with $A_0 - \alpha$ a nilpotent matrix) and A_1 is the restriction of A to X_1. Moreover, $\mathrm{Ker}(A_1 - \alpha) = \{0\}$ by construction. Now note that for $\beta \in \rho(A)$ we must have $\beta \in \rho(A_0) \cap \rho(A_1)$ with $R_A(\beta) = R_{A_0}(\beta) \oplus R_{A_1}(\beta)$ (cf. Problem 8.25) which shows that $R_{A_1}(\beta) \in \mathscr{K}(X_1)$ for $\beta \in \rho(A)$. Now since $\alpha \notin \sigma_p(A_1)$ this tells us $\alpha \in \rho(A_1)$ and hence α is an isolated eigenvalue (in particular, it is normal).

Problem* 8.22. *Let A be a closed operator. Show* (8.24)*. Moreover, conclude*

$$\frac{d^n}{d\alpha^n} R_A(\alpha) = n! R_A(\alpha)^{n+1}, \qquad \frac{d}{d\alpha} R_A(\alpha)^n = n R_A(\alpha)^{n+1}.$$

Problem 8.23. *Suppose $A = \overline{A_0}$. If $x_n \in \mathfrak{D}(A)$ is a Weyl sequence for $\alpha \in \sigma(A)$, then there is also one with $\tilde{x}_n \in \mathfrak{D}(A_0)$.*

Problem* 8.24. *Suppose $A_j \mathfrak{D}(A_j) \subseteq X_j \to Y_j)$, $j = 1, 2$. Then $A_1 \oplus A_2$ is closable if and only if both A_1 and A_2 are closable. Moreover, in this case we have $\overline{A_1 \oplus A_2} = \overline{A_1} \oplus \overline{A_1}$.*

Problem* 8.25. *Suppose $A_j \in \mathscr{C}(X_j)$, $j = 1, 2$. Then $A_1 \oplus A_2 \in \mathscr{C}(X_1 \oplus X_2)$ and $\sigma(A_1 \oplus A_2) = \sigma(A_1) \cup \sigma(A_2)$ with*

$$R_{A_1 \oplus A_2}(\alpha) = R_{A_1}(\alpha) \oplus R_{A_2}(\alpha), \quad \alpha \in \rho(A_1) \cap \rho(A_2).$$

8.3. Reducing subspaces and spectral projections

In the previous section we have seen that if A has a compact resolvent, then we can split X into two invariant subspaces such that all the spectral information pertaining to an eigenvalue is captured by restricting A to the first subspace, while the restriction to the second subspace has a bounded inverse. In this section we want to discuss such a splitting in more detail.

To begin with, recall that for a given closed subspace $M \subseteq X$ there might not be a corresponding closed subspace N such that $X = M \dotplus N$. Subspaces with this property are called complemented and M is complemented if and only if there is a projection $P = P^2 \in \mathscr{L}(X)$ such that $M = \operatorname{Ran}(P)$. In this case we have $N = \operatorname{Ker}(P)$ and $1 - P$ is a projection with $N = \operatorname{Ran}(1 - P)$. Recall that these problems do not arise if M is finite dimensional (or finite codimensional) or if we are in a Hilbert space (where we can choose $N = M^{\perp}$). Moreover, in this case we have $X \cong M \oplus N$.

So we need to assume that we have a complemented subspace such that both M and N are invariant with respect to A. Moreover, if A is unbounded we also need to assume that the domain is compatible with this splitting, that is, if we split a vector from the domain, both pieces will be again in the domain. In this case M is said to **reduce** $A : \mathfrak{D}(A) \subseteq X \to X$ and this is equivalent to the requirement that the corresponding projection P commutes with A (cf. Problem 8.20) in the sense that

$$PA \subseteq AP. \tag{8.35}$$

Note that the definition implies in particular, that P leaves $\mathfrak{D}(A)$ invariant, $P\mathfrak{D}(A) \subseteq \mathfrak{D}(A)$. Clearly, in this case A also commutes with the complementary projection $Q := 1 - P$. If A has nonempty resolvent set, then A will commute with P if and only if it commutes with $R_A(\alpha)$ for one (and hence for all) $\alpha \in \rho(A)$. Moreover, if A commutes with P, the same is true for $\overline{A}$ (Problem 8.21).

Lemma 8.16. *Let $A \in \mathscr{C}(X)$ be a closed operator on some Banach space X. Suppose $P \in \mathscr{L}(X)$ is a projection which commutes with A. Let $M :=$ $\operatorname{Ran}(P)$ and $N := \operatorname{Ker}(P)$. Then we have $A = A_0 \oplus A_1$ with $A_0 = A|_M \in \mathscr{C}(M)$ and $A_1 = A|_N \in \mathscr{C}(N)$. Here $\mathfrak{D}(A|_M) := \mathfrak{D}(A) \cap M$.*

Proof. First of all note that $A_0 : \mathfrak{D}(A_1) \subseteq M \to M$ is well defined as $A_0 x = Ax$ for $x \in \mathfrak{D}(A_0) := M \cap \mathfrak{D}(A) = P\mathfrak{D}(A)$. Moreover, A_0 is densely defined since if $x \in M$ we can find a sequence $x_n \in \mathfrak{D}(A)$ with $x_n \to x$ and hence also $Px_n \in \mathfrak{D}(A_0) \to Px = x$. Similarly we see that A_0 is closed since $\Gamma(A_0) = \Gamma(A) \cap (M \oplus M)$.

Exchanging P and Q shows that the same conclusions hold for $A_1 : \mathfrak{D}(A_1) \subseteq N \to N$ defined as $A_1 x = Ax$ for $x \in \mathfrak{D}(A_1) := N \cap \mathfrak{D}(A) = Q\mathfrak{D}(A)$. □

Hence in such a situation the investigation of A can be reduced to the investigation of A_0 and A_1 (cf. Problems 8.24 and 8.25).

In a finite dimensional case the projection onto a generalized eigenspace is given as the negative residue of the resolvent.

Example 8.16. Indeed consider

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

such that

$$(A-\alpha)^{-1} = \frac{1}{1-\alpha}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \frac{1}{-\alpha}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \frac{1}{\alpha^2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \qquad \diamond$$

This motivates the following definition: Suppose $\gamma : [0,1] \to \rho(A) \subset \mathbb{C}$ is a smooth Jordan curve. Recall that a Jordan curve is a closed curve without any self intersections and the Jordan curve theorem states that it splits the complex plane into two connected regions which both have the curve as a boundary. In particular, the interior region will be bounded. While this seems quite obvious, a proof turns out to be quite delicate. However, in the typical application of our result γ will just be a circle and hence one can always think of a circle.

Then we define the corresponding **Riesz projector** (after F. Riesz) as

$$P_\gamma := -\frac{1}{2\pi\mathrm{i}} \oint_\gamma R_A(\alpha) d\alpha. \tag{8.36}$$

Here the integral is defined as a Riemann integral (cf. Section 9.1 for details), explicitly

$$\oint_\gamma R_A(\alpha) d\alpha := \lim_{n\to\infty} \sum_{j=1}^n R_A\big(\gamma(j/n)\big)\gamma'(j/n)/n. \tag{8.37}$$

Choosing some $x \in X$ and some $\ell \in X^*$ we get

$$\ell(P_\gamma x) = -\frac{1}{2\pi\mathrm{i}} \oint_\gamma \ell(R_A(\alpha)x) d\alpha, \tag{8.38}$$

where now the integral on the right is an ordinary path integral in the complex plane. In particular, this shows that all the convenient facts from complex analysis about line integrals of holomorphic functions are at our disposal. For example, we can continuously deform the curve γ within $\rho(A)$ without changing the integral and the Cauchy integral theorem holds.

Lemma 8.17. *Let $A \in \mathscr{C}(X)$ be a closed operator on some Banach space X and let $\gamma : [0,1] \to \rho(A)$ be a Jordan curve. Then P_γ from* (8.36) *is a projection which reduces A.*

Proof. First of all note that $P_\gamma \in \mathscr{L}(X)$ since (see again Section 9.1 for details) the right-hand side of (8.37) converges in the operator norm. Moreover, since the integral is independent of γ we can take a slightly deformed path γ' in the exterior of γ to compute:

$$\begin{aligned}
P_\gamma^2 &= -\frac{1}{4\pi^2}\oint_{\gamma'}\oint_\gamma R_A(\alpha)R_A(\beta)d\alpha\, d\beta \\
&= -\frac{1}{4\pi^2}\oint_{\gamma'}\left(\oint_\gamma R_A(\alpha)\frac{d\alpha}{\alpha-\beta}\right)d\beta + \frac{1}{4\pi^2}\oint_{\gamma'} R_A(\beta)\left(\oint_\gamma \frac{d\alpha}{\alpha-\beta}\right)d\beta \\
&= -\frac{1}{4\pi^2}\oint_\gamma R_A(\alpha)\left(\oint_{\gamma'}\frac{d\beta}{\alpha-\beta}\right)d\alpha = -\frac{1}{2\pi\mathrm{i}}\oint_\gamma R_A(\alpha)d\alpha = P_\gamma
\end{aligned}$$

Here we have used the first resolvent identity (8.24) to obtain the second line. To obtain the third line we have used Fubini for the first double integral and the fact that in the second integral the inner integral vanishes by the Cauchy integral theorem since β lies in the exterior of γ. To obtain the last line observe that the inner integral equals $-2\pi\mathrm{i}$ since α lies in the interior of γ'.

Finally, since P_γ commutes with $R_A(\alpha)$ it also commutes with A (Problem 8.20), that is, P_γ reduces A. □

Example 8.17. Corollary 3.14 shows that the resolvent of a Sturm–Liouville operator has a simple pole at every eigenvalue whose negative residue is precisely the projector onto the corresponding eigenspace. In particular, if we choose $\gamma(t) := E_j + \varepsilon \mathrm{e}^{2\pi\mathrm{i}t}$ with ε so small that the interior of γ contains no eigenvalues other than E_j, then $P_\gamma = \langle u_j, .\rangle u_j$. ⋄

Theorem 8.18 (F. Riesz)**.** *Let $A \in \mathscr{C}(X)$ be a closed operator on some Banach space X. Let $\gamma : [0,1] \to \rho(A)$ be a Jordan curve and denote by S_0, S_1 its interior, exterior, respectively. Then the operators in the splitting from Lemma 8.16 with respect to P_γ satisfy*

$$\sigma(A_0) = \sigma(A)\cap S_0, \quad \textit{and} \quad \sigma(A_1) = \sigma(A)\cap S_1. \tag{8.39}$$

Moreover, A_0 is bounded.

Proof. Note that for α in the exterior of γ an analogous calculation (using the first resolvent identity) as in the previous lemma shows

$$R_A(\alpha)P_\gamma = \frac{1}{2\pi\mathrm{i}}\oint_\gamma R_A(\alpha)\frac{d\beta}{\alpha-\beta},$$

where the right-hand side is analytic in the exterior of γ. Similarly, for α in the interior of γ we have

$$R_A(\alpha)(1-P_\gamma) = -\frac{1}{2\pi\mathrm{i}}\oint_\gamma R_A(\alpha)\frac{d\beta}{\alpha-\beta},$$

where the right-hand side is analytic in the interior of γ. Hence Problem 8.26 implies $S_1 \subset \rho(A_0)$ and $S_0 \subset \rho(A_1)$. Since one also has $\rho(A_j) \subset \rho(A)$ by Problem 8.25, the claim about the spectra follows.

To see that AP_γ is bounded note that for $x \in X$ and $\ell \in \mathfrak{D}(A')$ we have

$$(A'\ell)(P_\gamma x) = -\frac{1}{2\pi\mathrm{i}} \oint_\gamma \ell(AR_A(\alpha)x) d\alpha,$$

which shows that AP is bounded since $AR_A(\alpha) = 1 + \alpha R_A(\alpha)$ is. □

Of course this result applies to isolated eigenvalues. Thus if α is an isolated eigenvalue of A we can split $A = A_0 \oplus A_1$ with $\sigma(A_0) = \{\alpha\}$ and $\sigma(A_1) = \sigma(A) \setminus \{\alpha\}$. Moreover, since $\mathrm{Ker}((A-\alpha)^n) = \mathrm{Ker}((A_0-\alpha)^n)$ the geometric and algebraic multiplicity of α as an eigenvalue of A and as an eigenvalue of A_1 agree. In particular, the algebraic multiplicity equals $\dim(\mathrm{Ran}(P_\gamma))$.

Problem* 8.26. *Let $U \subseteq \mathbb{C}$ be open and connected and $R : U \to \mathscr{L}(X)$ be weakly holomorphic (i.e. $\alpha \to \ell(R(\alpha)x)$ is holomorphic for every $\ell \in X^*$ and $x \in X$). If $A \in \mathscr{C}(X)$ satisfies $U_0 \subseteq \rho(A)$, where $U_0 \subseteq U$ is nonempty and open, and if $R(\alpha) = R_A(\alpha)$ for $\alpha \in U_0$, then we have $U \subseteq \rho(A)$ and $R(\alpha) = R_A(\alpha)$ for $\alpha \in U$.*

8.4. Relatively bounded and relatively compact operators

In many applications operators have the structure $A + B$, where A is well-understood and B can be considered *small* with respect to A. This raises the questions about conditions on B which ensure that certain nice properties of A persist when adding B. In this section we discuss two basic conditions.

Let $A : \mathfrak{D}(A) \subseteq X \to Y$ be an operator between Banach spaces X and Y. An operator $B : \mathfrak{D}(B) \subseteq X \to Y$ is called **A-bounded** or **relatively bounded** with respect to A if $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$ and if there are constants $a, b \geq 0$ such that

$$\|Bx\| \leq a\|Ax\| + b\|x\|, \qquad x \in \mathfrak{D}(A). \tag{8.40}$$

The infimum of all constants a for which a corresponding b exists such that (8.40) holds is called the **A-bound** of B. The set of all A-bounded operators forms a vector space denoted by $\mathscr{L}_A(X,Y)$ (Problem 8.27). Note that B is relatively bounded if and only if it is bounded with respect to the graph norm, that is $\tilde{\mathscr{L}}_A(X,Y) \cong \mathscr{L}([\mathfrak{D}(A)], Y)$, where $\tilde{\mathscr{L}}_A(X,Y) = \{B \in \mathscr{L}_A(X,Y) | \mathfrak{D}(B) = \mathfrak{D}(A)\}$ and we have written $[\mathfrak{D}(A)]$ to indicate that $\mathfrak{D}(A)$ is equipped with the graph norm of A.

It is important to emphasize that the A-bound will in general not be attained, since b will typically increase as a approaches the A-bound. Moreover, this concept is only of interest for unbounded operators, since if A is bounded, then the relatively bounded operators are just the bounded operators.

Example 8.18. 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 (Example 8.3) while B is not closable (Example 8.4). (Compare also Problem 8.8.)

Now

$$\sum_{j\in\mathbb{N}} |a_j| = \sum_{j\le n} |a_j| + \sum_{j>n} \frac{1}{j} \cdot |ja_j| \le \sqrt{n}\|a\|_2 + \sqrt{\sum_{j>n} \frac{1}{j^2}}\|Aa\|_2$$

shows that B is A-bounded with A-bound zero. ◇

There is also an important equivalent characterization in terms of the resolvent.

Lemma 8.19. *Suppose $A \in \mathscr{C}(X)$ is closed with nonempty resolvent set. Then $B \in \mathscr{L}_A(X)$ if and only if $BR_A(\alpha) \in \mathscr{L}(X)$ for one (and hence for all) $\alpha \in \rho(A)$.*

Moreover, the A-bound of B is no larger than $\inf_{\alpha\in\rho(A)} \|BR_A(\alpha)\|$.

Proof. This follows since $R_A(\alpha) : X \to [\mathfrak{D}(A)]$ is a homeomorphism by the inverse mapping theorem. In fact, this can also be seen directly since $y = (A-\alpha)x$ implies $\|R_A(\alpha)y\| \le \|R_A(\alpha)\|\|Ax\| + |\alpha|\|x\|$.

Moreover, the same idea shows that for $x \in \mathfrak{D}(A)$ we have

$$\|Bx\| = \|BR_A(\alpha)(A-\alpha)x\| \le a\|(A-\alpha)x\| \le a\|Ax\| + (a|\alpha|)\|x\|,$$

where $a := \|BR_A(\alpha)\|$. Finally, note that if $BR_A(\alpha)$ is bounded for one $\alpha \in \rho(A)$, it is bounded for all $\alpha \in \rho(A)$ by the first resolvent formula. □

Note that if B is closable and A has nonempty resolvent set, then it is relatively bounded if and only if $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$. In fact, in this case Problem 8.12 implies that $BR_A(\alpha)$ is closed and hence bounded by the closed graph theorem (Theorem 4.9).

Now we come to our first application.

Lemma 8.20. *Suppose $A : \mathfrak{D}(A) \subseteq X \to Y$ and $B \in \mathscr{L}_A(X,Y)$ with A-bound less than one. Then the graph norms of A and $A+B$ are equivalent. In particular, in this case $A+B$ is closable/closed if and only if A is.*

Proof. From $\|Ax\| \le \|(A+B)x\| + \|Bx\| \le \|(A+B)x\| + a\|Ax\| + b\|x\|$ we obtain

$$\|Ax\| \le \frac{1}{1-a}\|(A+B)x\| + \frac{b}{1-a}\|x\|$$

which shows that the graph norms of A and $A+B$ are equivalent. □

If A is closable, then the fact that B is A-bounded always implies that there is an extension $\bar{B}$ to $\mathfrak{D}(\overline{A})$ given by extending B with respect to the graph norm using continuity (i.e., as in Theorem 1.16). Moreover, note that (8.40) continues to hold for $\overline{A}$ and $\bar{B}$ (with the same constants). This implies that we can assume A to be closed without loss of generality and that it suffices to verify (8.40) on a core of A. However, while $\overline{A}+\bar{B}$ is bounded (and hence closed) on $[\mathfrak{D}(\overline{A})]$, it might not be closed on X since the graph norm of $\overline{A}+\bar{B}$ might be weaker than the graph norm of $\overline{A}$. Conversely, if the graph norms are equivalent, there is no need that B has a closed extension in X. The following two examples illustrate these observations.

Example 8.19. If A is unbounded and $B = -A$, then $A + B = 0$ and hence the condition that the A-bound is less than one cannot be dropped in general. ⋄

Example 8.20. Consider the operators from Example 8.18. Then, since A is closed, so is $A + B$ by our lemma (which can also be verified directly, cf. Problem 8.8). Nevertheless, note that B alone is not closable. ⋄

Example 8.21. Let $1 \le p < \infty$. Consider the differentiation operator $A : W^{1,p}(0,1) \subset L^p(0,1) \to L^p(0,1)$ from Example 8.5 and let $B : C[0,1] \subset L^p(0,1) \to L^p(0,1)$ be the point evaluation at 0, that is $(Bf)(x) = f(0)$. Then by Problem 8.6 B is relatively bounded and satisfies (8.40) with $a = b = 1$ (note that Hölder's inequality (Problem 1.38) implies $\|f\|_1 \le \|f\|_p$). This is clearly not good enough to apply our lemma and hence we need to work a bit harder in this case. If $1 < p < \infty$ we can invoke Hölder's inequality to obtain

$$|f(0)| = \left| f(x) - \int_0^x f'(y)dy \right| \le |f(x)| + x^{1/q}\|f'\|_p,$$

where $q = \frac{p}{p-1}$. Integrating form 0 to ε with respect to x further shows (using $\|f\|_1 \le \|f\|_p$)

$$|f(0)| \le \frac{q\varepsilon^{1/q}}{1+q}\|f'\|_p + \frac{1}{\varepsilon}\|f\|_p.$$

Hence the relative bound is 0 for $1 < p < \infty$ and our lemma applies. In the case $p = 1$ this argument breaks down. In fact, it turns out that for $p = 1$ the A-bound is one. To see this let $f_n(x) := \max(1 - nx, 0)$ such that $\|f_n\|_1 = \frac{1}{2n}$ and $\|f_n'\|_1 = |f_n(0)| = 1$. Then letting $n \to \infty$ in

$$1 = |f_n(0)| = \|Bf_n\|_1 \le a\|Af_n\|_1 + b\|f_n\| = a + \frac{b}{2n}$$

shows $a \ge 1$ and hence the A bound of B is one and our lemma does not apply. We will come back to this case below. ⋄

Next, there is also a convenient formula for the resolvent of $A + B$.

Lemma 8.21. *Let A, B be two given operators with $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$ such that A and $A+B$ are closed. Then we have the* ***second resolvent formula***

$$R_{A+B}(\alpha) - R_A(\alpha) = -R_A(\alpha)BR_{A+B}(\alpha) = -R_{A+B}(\alpha)BR_A(\alpha) \tag{8.41}$$

for $\alpha \in \rho(A) \cap \rho(A+B)$.

Proof. We abbreviate $C := A + B$ and compute

$$R_C(\alpha) + R_A(\alpha)BR_C(\alpha) = R_A(\alpha)(C - \alpha)R_C(\alpha) = R_A(\alpha). \qquad \square$$

Moreover, we also give a criterion for a point to be in the resolvent set of $A + B$.

Lemma 8.22. *Let $A \in \mathscr{C}(X)$ and $B \in \mathscr{L}_A(X)$ satisfying* (8.40). *If $\alpha \in \rho(A)$ and $\|BR_A(\alpha)\| < 1$, then $\alpha \in \rho(A+B)$ with*

$$R_{A+B}(\alpha) = R_A(\alpha)(1 - BR_A(\alpha))^{-1}. \tag{8.42}$$

This condition holds in particular if

$$a\|AR_A(\alpha)\| + b\|R_A(\alpha)\| < 1. \tag{8.43}$$

Proof. Since we can write $A+B+\alpha = (1+BR_A(\alpha))(A-\alpha)$ the first claim follows from Problem 8.28. The second follows from

$$\|BR_A(\alpha)x\| \le a\|AR_A(\alpha)x\| + b\|R_A(\alpha)x\|. \qquad \square$$

Now we turn to the second condition: Let $A : \mathfrak{D}(A) \subseteq X \to Y$. An operator $B : \mathfrak{D}(B) \subseteq X \to Y$ is called A**-compact** or **relatively compact** with respect to A if $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$ and whenever x_n and Ax_n are bounded sequences, then Bx_n has a convergent subsequence.

The set of all A-compact operators forms a vector space denoted by $\mathscr{K}_A(X,Y)$. As before, note that B is relatively compact if and only if it is compact with respect to the graph norm, that is $\tilde{\mathscr{K}}_A(X,Y) \cong \mathscr{K}([\mathfrak{D}(A)], Y)$, where $\tilde{\mathscr{K}}_A(X,Y) = \{B \in \mathscr{K}_A(X,Y) | \mathfrak{D}(B) = \mathfrak{D}(A)\}$. Of course every compact operator is relatively bounded and if the embedding $[\mathfrak{D}(A)] \hookrightarrow X$ is compact, then every bounded operator is relatively compact.

Example 8.22. Consider the operators from Example 8.18 or 8.21. Then, B is relatively compact since its range is one-dimensional. In fact, the same is true for any relatively bounded operator B whose range is finite dimensional (cf. Problem 8.30). ⋄

Example 8.23. Consider the differential operator $A : W^{1,p}(0,1) \subseteq L^p(0,1) \to L^p(0,1)$, $1 < p < \infty$ from Example 8.5. Then since the embedding $W^{1,p}(0,1) \subseteq L^p(0,1)$ is compact (Problem 8.7), every bounded operator $B \in \mathscr{L}(L^p(0,1))$ is relatively compact. For example one can choose B to be a multiplication operator by a bounded function. ⋄

Again there is an equivalent characterization in terms of resolvents.

Lemma 8.23. *Suppose $A \in \mathscr{C}(X)$ is closed with nonempty resolvent set. Then $B \in \mathscr{K}_A(X)$ if and only if $BR_A(\alpha) \in \mathscr{K}(X)$ for one (and hence for all) $\alpha \in \rho(A)$.*

Proof. Again this follows since $R_A(\alpha) : X \to [\mathfrak{D}(A)]$ is a homeomorphism. Moreover, if $BR_A(\alpha)$ is compact for one $\alpha \in \rho(A)$, it is compact for all $\alpha \in \rho(A)$ by the first resolvent formula. $\square$

As already pointed out before, if A is closable, there is a unique extension $\bar{B}$ of B to $\mathfrak{D}(\overline{A})$ using the graph norm. Moreover, if B is relatively compact with respect to A, then $\bar{B}$ is relatively compact with respect to $\overline{A}$. This follows from Lemma 3.3 upon using the graph norm on $\mathfrak{D}(A)$. Moreover, in this case $\overline{A} + \bar{B}$ is closed without restriction on the A-bound.

Lemma 8.24. *Suppose $A \in \mathscr{C}(X,Y)$ and $B \in \mathscr{K}_A(X,Y)$. Then $A+B$ is closed and B is also relatively compact with respect to $A+B$.*

Proof. Abbreviate $C := A + B$. We first show that B is relatively compact with respect to C. If it were not, we could find a bounded sequence $x_n \in \mathfrak{D}(A)$ such that Cx_n is bounded but Ax_n is unbounded. After dropping some terms we can assume $\|Ax_n\| \to \infty$. Then $\tilde{x}_n := \|Ax_n\|^{-1}x_n$ is a null sequence and so is $C\tilde{x}_n$. Moreover, since B is A-compact we can assume that $B\tilde{x}_n \to y$ converges. Hence $A\tilde{x}_n = (C-B)\tilde{x}_n \to -y$ implying $y = 0$ as A is closed, contradicting $\|A\tilde{x}_n\| = 1$.

Now let $x_n \in \mathfrak{D}(A)$ be a sequence with $x_n \to x$ and $Cx_n \to y$. Since by the first part B is C-compact, we can pass to a subsequence such that also $Bx_n \to z$. Consequently $Ax_n = (C-B)x_n \to y - z$ implying $x \in \mathfrak{D}(A)$ and $Ax = y - z$ since A is closed. Moreover, since B is A-bounded we also have $Bx = z$. Thus we have $x \in \mathfrak{D}(C) = \mathfrak{D}(A)$ and $Cx = y = Ax + Bx$ which shows that C is closed. $\square$

Example 8.24. Consider once more the operators from Example 8.18 or 8.21. Again our lemma implies that $A+B$ is closed and this time we don't even need the computation of the A-bound. Moreover, in the case of Example 8.21 this now covers the full range $1 \le p < \infty$. $\diamond$

Applications of these notions will be given in the next section.

Problem 8.27. *Suppose B_j, $j = 1, 2$, are A bounded with respective A-bounds a_i, $i = 1, 2$. Show that $\alpha_1 B_1 + \alpha_2 B_2$ is also A bounded with A-bound less than $|\alpha_1|a_1 + |\alpha_2|a_2$.*

Problem* 8.28. *Suppose A is closed and B satisfies $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$:*

- *Show that $1 + B$ has a bounded inverse if $\|B\| < 1$.*

- *Suppose A has a bounded inverse. Then so does $A+B$ if $\|BA^{-1}\| < 1$. In this case we have $\|(A+B)^{-1}\| \le \frac{\|A^{-1}\|}{1-\|BA^{-1}\|}$.*

Problem 8.29. *Suppose A, B are linear operators with $\mathfrak{D}(A) \subseteq \mathfrak{D}(B)$ and $\|Bx\| \le a\|Ax\|^{\alpha}\|x\|^{1-\alpha}$ for all $x \in \mathfrak{D}(A)$ and some $a > 0$, $\alpha \in (0,1)$. Then B is relatively bounded with A-bound 0. (Hint: Young's inequality (1.25).)*

Problem 8.30. *Show that a relatively bounded operator $B \in \mathscr{L}_A(X,Y)$ with finite dimensional range is of the form*

$$Bx = \sum_{j=1}^{n} (x_j'(x) + y_j'(Ax))y_j$$

for some $x_j' \in X^*$, $y_j' \in Y^*$, and $y_j \in Y$, $1 \le j \le n$. Moreover, any such operator is relatively compact. (Hint: Start with the finite rank operator $\widehat{B} \in \mathscr{L}([\mathfrak{D}(A)], Y)$ and use $[\mathfrak{D}(A)] \cong \Gamma(A) \subset X \oplus Y$.)

Problem 8.31. *Suppose $B : \mathfrak{D}(B) \subseteq X \to Y$ is relatively compact with respect to $A : \mathfrak{D}(A) \subseteq X \to Y$. Then the A-bound of B is zero if either*

(i) *B is closable, or*

(ii) *A is closed and Y is reflexive.*

(Hint: Argue by contradiction. If the claim were wrong, then for any $a > 0$ one could find a sequence $x_n \in \mathfrak{D}(A)$ such that $\|x_n\|_A = 1$ and $\|Bx_n\| \ge a\|Ax_n\| + n\|x_n\|$. Since $B \in \mathscr{L}([\mathfrak{D}(A)], Y)$ we must have $x_n \to 0$. Moreover, we can assume $Bx_n \to y$ and we can get a contradiction if we can show $y = 0$. In the first case this is immediate. In the second case use that $\Gamma(A)$ is weakly closed.)

8.5. Unbounded Fredholm operators

The definition of a Fredholm operator from Section 7.2 can be extended to closed operators in a straightforward manner: A closed operator $A \in \mathscr{C}(X,Y)$ between Banach spaces X and Y is called a **Fredholm operator** if both its kernel and cokernel are finite dimensional. In this case we define its **index** as

$$\operatorname{ind}(A) := \dim \operatorname{Ker}(A) - \dim \operatorname{Coker}(A). \tag{8.44}$$

In fact, many results for bounded Fredholm operators carry over to the unbounded case using the following observation: Let us denote A regarded as an operator from $\mathfrak{D}(A)$ equipped with the graph norm $\|.\|_A$ to Y by $\widehat{A}$. Recall that $\widehat{A}$ is bounded (cf. Problem 8.5). Moreover, $\operatorname{Ker}(\widehat{A}) = \operatorname{Ker}(A)$ and $\operatorname{Ran}(\widehat{A}) = \operatorname{Ran}(A)$. Consequently, $A \in \mathscr{C}(X,Y)$ is Fredholm if and only if $\widehat{A}$ is.

For example, applying Lemma 7.8 to $\widehat{A}$ shows that a closed operator with finite cokernel has closed range. In particular, a Fredholm operator has closed range.

Example 8.25. Consider the operator A from Example 8.12. There we have seen $\dim \operatorname{Ker}(A-\alpha) = 1$ for every $\alpha \in \mathbb{C}$. Moreover, the solution of the inhomogeneous differential equation $f' - \alpha f = g$ is given by

$$f(x) = f(0)\mathrm{e}^{\alpha x} + \int_0^x \mathrm{e}^{\alpha(x-y)} g(y)dy,$$

which shows $\operatorname{Ran}(A-\alpha) = X$. Thus $A-\alpha$ is Fredholm with $\operatorname{ind}(A-\alpha) = 1$. ⋄

Literally following the proof of Lemma 7.9 (using Lemma 8.7 and Theorem 8.9) gives

Lemma 8.25. *Suppose* $A \in \mathscr{C}(X,Y)$ *with* $\operatorname{Ran}(A)$ *closed. Then*

$$\operatorname{Ker}(A') \cong \operatorname{Coker}(A)^*, \qquad \operatorname{Coker}(A') \cong \operatorname{Ker}(A)^*. \tag{8.45}$$

In particular, we have

$$\dim \operatorname{Ker}(A') = \dim \operatorname{Coker}(A), \qquad \dim \operatorname{Ker}(A) = \dim \operatorname{Coker}(A'). \tag{8.46}$$

And hence Riesz' theorem continues to hold:

Theorem 8.26. *A closed operator A is Fredholm if and only if A' is and in this case*

$$\operatorname{ind}(A') = -\operatorname{ind}(A). \tag{8.47}$$

Applying Dieudonné's and Yood's theorem to $\widehat{A}$ gives:

Theorem 8.27. *Suppose* $A \in \mathscr{C}(X,Y)$ *is Fredholm. Then there is an* $\varepsilon > 0$ *such that for every* $B \in \mathscr{L}_A(X,Y)$ *with norm smaller than* ε*, we have that* $A+B$ *is Fredholm with index* $\operatorname{ind}(A+B) = \operatorname{ind}(A)$*. Similarly, for every* $B \in \mathscr{K}_A(X,Y)$ *the operator* $A+B$ *is Fredholm with index* $\operatorname{ind}(A+B) = \operatorname{ind}(A)$*.*

The structure of Fredholm operators is essentially the same as in the bounded case. As discussed in Section 7.2 there are closed subspaces $X_0 \subseteq X$ and $Y_0 \subseteq Y$ such that $X = \operatorname{Ker}(A) \dotplus X_0$ and $Y = Y_0 \dotplus \operatorname{Ran}(A)$, respectively. Moreover, we have corresponding projections $P \in \mathscr{L}(X)$ with $\operatorname{Ran}(P) = \operatorname{Ker}(A)$ and $Q \in \mathscr{L}(Y)$ with $\operatorname{Ran}(Q) = Y_0$. With respect to the decomposition $\operatorname{Ker}(A) \dotplus X_0 \to Y_0 \dotplus \operatorname{Ran}(A)$ our Fredholm operator is given by

$$A = \begin{pmatrix} 0 & 0 \\ 0 & A_0 \end{pmatrix}, \tag{8.48}$$

where A_0 is the restriction of A to $\mathfrak{D}(A) \cap X_0 \to \operatorname{Ran}(A)$. By construction A_0 is closed and surjective implying that it has a bounded inverse (Corollary 8.2). Defining

$$B := \begin{pmatrix} 0 & 0 \\ 0 & A_0^{-1} \end{pmatrix} \tag{8.49}$$

we get

$$AB = \mathbb{I} - Q, \qquad BA = (\mathbb{I} - P)|_{\mathfrak{D}(A)} \tag{8.50}$$

and hence A is invertible up to finite rank operators. The converse can be shown as in the bounded case which gives the analog of Theorem 7.14:

Theorem 8.28. *An operator $A \in \mathscr{C}(X,Y)$ is Fredholm if and only if there exist $B_1, B_2 \in \mathscr{L}(Y,X)$ such that $\overline{B_1 A - \mathbb{I}} \in \mathscr{K}(X)$ and $AB_2 - \mathbb{I} \in \mathscr{K}(Y)$.*

To establish the analog of Lemma 7.11 we need

Lemma 8.29. *Let X_0 be a closed subspace of a normed space X with finite codimension. Then for any dense subspace $D \subseteq X$ there is a finite-dimensional space $N \subseteq D$ with $X = X_0 \dotplus N$.*

Proof. Suppose that $\tilde{N} = \operatorname{span}\{\tilde{x}_j\}_{j=1}^n$ complements X_0. Then we can write every $x \in X$ uniquely as $x = \sum_{j=1}^n \alpha_j \tilde{x}_j + x_0$ with $x_0 \in X_0$. Choosing $\|x_0\| + \sum_{j=1}^n |\alpha_j|$ as a norm on $X_0 \oplus \tilde{N}$ (recall that all norms on the finite dimensional part are equivalent), there is a constant C such that

$$\|x_0\| + \sum_{j=1}^n |\alpha_j| \le C\|x\|$$

since $X_0 \oplus \tilde{N} \cong X$. Now we use that D is dense and replace $\tilde{x}_j$ by some $x_j \in D$ with $\|\tilde{x}_j - x_j\| \le \frac{C}{2}$. Then $N = \operatorname{span}\{x_j\}_{j=1}^n$ is the required subspace. Indeed for $x = \sum_{j=1}^n \alpha_j x_j + x_0$ the above estimate implies

$$\|x_0\| + \frac{1}{2}\sum_{j=1}^n |\alpha_j| \le C\|x\|.$$

In particular, choosing $x_0 = 0$ this shows that the elements x_j are linearly independent. Moreover, we still have $X_0 \cap N = \{0\}$ since if $x_0 = \sum_{j=1}^n \alpha_j x_j \in X_0$ we can apply the above estimate to $x = -\sum_{j=1}^n \alpha_j x_j + x_0 = 0$ to obtain $\|x_0\| + \frac{1}{2}\sum_{j=1}^n |\alpha_j| = 0$.

Finally the quotient map $X \to X/X_0$ gives an injective map from $N \to X/X_0$ and since both spaces have the same dimension this map must also be surjective, that is, $X_0 \dotplus N = X$. □

Lemma 8.30. *Let X, Y and Z be Banach spaces. Suppose $A \in \mathscr{C}(X,Y)$ and $B \in \mathscr{C}(Y,Z)$ are Fredholm. Then $BA \in \mathscr{C}(X,Z)$ is Fredholm with*

$$\operatorname{ind}(BA) = \operatorname{ind}(A) + \operatorname{ind}(B). \tag{8.51}$$

Proof. We first need to show $BA \in \mathscr{C}(X,Z)$.

To see that $\mathfrak{D}(BA) = \{x \in \mathfrak{D}(A) | Ax \in \mathfrak{D}(B)\}$ is dense we use that $Y = \operatorname{Ran}(A) + Y_0$, where we can assume $Y_0 \subseteq \mathfrak{D}(B)$ by Lemma 8.29. Hence $\operatorname{Ran}(A) \cap \mathfrak{D}(B)$ is dense in $\operatorname{Ran}(A)$ and thus $\tilde{X}_0 := A_0^{-1} \operatorname{Ran}(A) \cap \mathfrak{D}(B)$ is dense in $\mathfrak{D}(A) \cap X_0$, which in turn is dense in X_0. In summary, $\mathfrak{D}(BA) = \operatorname{Ker}(A) \dotplus \tilde{X}_0$ is dense.

To see that BA is closed we consider a sequence $x_n \in \mathfrak{D}(BA)$ with $x_n \to x$ and $BAx_n \to z$. Write $Y = \operatorname{Ker}(B) \dotplus Y_0$ and let P be the corresponding projection onto Y_0. Then $BPAx_n = BAx_n \to z$ and hence $PAx_n \to y_2 := B_0^{-1} z$. Now if $(\mathbb{I} - P)Ax_n \in \operatorname{Ker}(B)$ is bounded, we can assume that $(\mathbb{I} - P)Ax_n \to y_2$ after passing to a subsequence. Then $Ax_n \to y := y_1 + y_2$, which shows that $x \in \mathfrak{D}(A)$ with $Ax = y$ and $y \in \mathfrak{D}(B)$ with $By = z$, that is, $x \in \mathfrak{D}(BA)$ with $BAx = z$. Thus it remains to show that $(\mathbb{I} - P)Ax_n \in \operatorname{Ker}(B)$ is bounded. If not, we can pass to a subsequence such that $c_n := \|(\mathbb{I} - P)Ax_n\| \to \infty$. Now, after passing to another subsequence, we can assume that $\tilde{y}_n := c_n^{-1}(\mathbb{I} - P)Ax_n \to \tilde{y}$. Moreover, $c_n^{-1}Ax_n - \tilde{y}_n = c_n^{-1}PAx_n \to 0$ since $PAx_n \to y_2$. Hence $A(c_n^{-1}x_n) \to \tilde{y}$ and $c_n^{-1}x_n \to 0$ implying $\tilde{y} = A0 = 0$ and contradicting $\|\tilde{y}\| = 1$.

Now we can proceed as in the proof of Lemma 7.11. In this respect note that by definition $\operatorname{Ker}(BA) \subseteq \mathfrak{D}(A)$ and hence A is well defined as a map from $\operatorname{Ker}(A) \to \operatorname{Ker}(BA)$. Moreover, as noted before, we can choose $Y = Y_0 \dotplus \operatorname{Ran}(A)$ with $Y_0 \subseteq \mathfrak{D}(B)$ and hence also $\tilde{B} : Y/\operatorname{Ran}(A) \to Z/\operatorname{Ran}(BA)$ is well defined. Thus the rest follows as in the proof of Lemma 7.11. □

Moreover, the essential and Fredholm spectrum can be defined as in the bounded case and the same argument shows that

$$\sigma_{ess}(A) = \bigcap_{K \in \mathscr{K}(X)} \sigma(A + K). \tag{8.52}$$

Furthermore, Theorem 7.16 continues to hold.

Theorem 8.31 (Weyl)**.** *Let $A \in \mathscr{C}(X)$, then*

$$\sigma_\Phi(A + K) = \sigma_\Phi(A), \quad \sigma_{ess}(A + K) = \sigma_{ess}(A), \quad K \in \mathscr{K}_A(X). \tag{8.53}$$

Example 8.26. Consider again the operator A from Example 8.12. Since $A - \alpha$ is Fredholm for every α we have $\sigma_{ess}(A) = \emptyset$. Moreover, since the embedding $C^1[0,1] \hookrightarrow C[0,1]$ is compact, every bounded operator $B \in \mathscr{L}(C[0,1])$ is relatively compact and we have $\sigma_{ess}(A + B) = \sigma_{ess}(A) = \emptyset$ in this case. ⋄

We end this section with a useful criterion extending Lemma 8.12. To this end we call a Weyl sequence x_n a **singular Weyl sequence** if in addition x_n has no convergent subsequence.

Lemma 8.32 (Weyl's criterion)**.** *We have $\alpha \in \sigma_\Phi(A)$ if there is a singular Weyl sequence $x_n \in \mathfrak{D}(A)$ such that $\|x_n\| = 1$, $\|(A-\alpha)x_n\| \to 0$, and x_n has no convergent subsequence.*

Proof. By Lemma 8.12 we have $\alpha \in \sigma(A)$ and hence we need to show that $A - \alpha$ is not Fredholm. If $\dim(\operatorname{Ker}(A-\alpha)) = \infty$ there is nothing to do. Hence we can assume $\dim(\operatorname{Ker}(A-\alpha)) < \infty$ and there is a projector $P = P^2 \in \mathscr{L}(X)$ with $\operatorname{Ker}(P) = \operatorname{Ker}(A-\alpha)$. Consider $y_n := (1-P)x_n$. Then $(A-\alpha)y_n = (A-\alpha)x_n \to 0$ and y_n has no convergent subsequence. In indeed, if it had a convergent subsequence, then by compactness of P, there is a further subsequence which works for both y_n, Px_n, and hence also for x_n, a contradiction. Moreover, $\|y_n\| = \|x_n - Px_n\| \to 1$ and hence we can assume $\|y_n\| = 1$ after rescaling. Moreover, since $X \cong \operatorname{Ker}(A-\alpha) \oplus X_0$, where $X_0 := \operatorname{Ran}(P)$ and $X/\operatorname{Ker}(A-\alpha) \cong X_0$ we see that $\operatorname{dist}(y_n, \operatorname{Ker}(A-\alpha)) \ge \varepsilon > 0$ and hence the $\operatorname{Ran}(A-\alpha)$ is not closed by Corollary 8.4. $\square$

Note that one way of ensuring that x_n has no convergent subsequence is to require $x_n \rightharpoonup 0$. For example, in a Hilbert space one could try to choose the elements orthogonal.

Example 8.27. Let us take $X := L^p(\mathbb{R})$, $1 \le p < \infty$, and start with $A_0 = \frac{1}{\mathrm{i}}\frac{d}{dx}$ on $\mathfrak{D}(A_0) = C_c^\infty(\mathbb{R})$. Then the closure $A := \overline{A_0}$ is defined on $\mathfrak{D}(A) = W^{1,p}(\mathbb{R})$, where the Sobolev space $W^{1,p}(\mathbb{R})$ is defined as the closure of $C_c^\infty(\mathbb{R})$ with respect to the norm

$$\|f\|_{1,p}^p = \int_{\mathbb{R}} \big(|f'(x)|^p + |f(x)|^p\big) dx.$$

(Similarly $L^p(\mathbb{R})$ can be understood as the closure of $C_c^\infty(\mathbb{R})$ with respect to the norm $\|f\|_p^p = \int_{\mathbb{R}} |f(x)|^p dx$.) Then any eigenfunction must be of the form

$$f(x) = f(0)\mathrm{e}^{-\mathrm{i}\alpha x}$$

and such a function will never be square integrable unless $f(0) = 0$. However, for $\alpha \in \mathbb{R}$ this function is at least bounded and we could try to get approximating eigenfunctions by restricting $\mathrm{e}^{-\mathrm{i}\alpha x}$ to compact intervals. More precisely, choose $\varphi_n \in C_c^\infty(\mathbb{R})$ such that φ_n is symmetric with $\varphi_n(x) = 1$ for $0 \le x \le n$, $\varphi_n(x) = 1$ for $x \ge n+1$ and such that the piece on $[n, n+1]$ is independent of n. Set $u_n(x) := \varphi_n(x)\mathrm{e}^{-\mathrm{i}\alpha x}$. Then $\|u_n\|_p = (2(C_0 + n))^{1/p}$ while $\|(A-\alpha)u_n\|_p = (2C_1)^{1/p}$ and thus $u_n/\|u_n\|_p$ is a Weyl sequence. Since the same is true for $u_n(x + r_n)$ we get a singular Weyl sequence by choosing r_n such that the supports of φ_n and φ_m are disjoint for $n \ne m$. Hence we conclude $\mathbb{R} \subset \sigma_\Phi(A)$.

If $\alpha \in \mathbb{C} \setminus \mathbb{R}$ we can write down the resolvent of A. To this end note that the solution of the inhomogeneous equation $-\mathrm{i}f' - \alpha f = g$ is given by

$$f(x) = f(a)\mathrm{e}^{\mathrm{i}\alpha(x-a)} + \mathrm{i}\int_a^x \mathrm{e}^{\mathrm{i}\alpha(x-y)} g(y)dy.$$

If g has compact support, we can shift a beyond the support of g and hence we expect

$$R_A(\alpha)g(x) := \mathrm{i}\int_{\pm\infty}^x \mathrm{e}^{\mathrm{i}\alpha(x-y)} g(y)dy, \qquad \mp\mathrm{Im}(\alpha) > 0.$$

Note that here we have shifted a to the left/right of the support depending on the sign of $\mathrm{Im}(\alpha)$ such that the above expression at least formally makes sense without the compact support assumption. Then Young's inequality for convolutions implies

$$\|R_A(\alpha)g\|_p \le \frac{1}{|\mathrm{Im}(\alpha)|}\|g\|_p,$$

which shows $\alpha \in \rho(A)$ with $\|R_A(\alpha)\| \le |\mathrm{Im}(\alpha)|^{-1}$.

In summary, we have $\sigma(A) = \sigma_\Phi(A) = \sigma_{ess}(A) = \mathbb{R}$.

Moreover, since the embedding $W^{1,p}(a,b) \hookrightarrow L^p(a,b)$ is compact (Problem 8.7) and the embedding $L^p(a,b) \hookrightarrow L^p(\mathbb{R})$ is bounded for every bounded interval (a,b) every multiplication operator Q with a function $q \in L^\infty_c(\mathbb{R})$ (bounded with compact support) is relatively compact. Moreover, every bounded function $q \in L^\infty_0(\mathbb{R})$ which vanishes as $|x| \to \infty$ can be approximated by bounded functions with compact support q_n in the sup norm. Hence we also have $Q_n \to Q$ in the operator norm, and for any $q \in L^\infty_0(\mathbb{R})$ we have $\sigma_\Phi(A+Q) = \sigma_{ess}(A+Q) = \mathbb{R}$.

However, note that while the spectrum of $A+Q$ could be different from the spectrum of A, this is not the case here. Indeed one can verify that the resolvent is given by

$$R_{A+Q}(\alpha)g(x) := \mathrm{i}\int_{\pm\infty}^x \mathrm{e}^{\mathrm{i}\alpha(x-y)-\mathrm{i}\int_y^x q(t)dt} g(y)dy, \qquad \mp\mathrm{Im}(\alpha) > 0. \qquad \diamond$$

Problem 8.32. *Consider $X := C[0,1]$ and $A = \frac{d}{dx}$ with $\mathfrak{D}(A) = \{f \in C^1[0,1] | f(0) = f(1) = 0\}$. Investigate when $A - \alpha$ is Fredholm and compute the essential spectrum of A.*

Part 3

Nonlinear Functional Analysis

Chapter 9

Analysis in Banach spaces

9.1. Single variable calculus in Banach spaces

As a warmup we will look at mappings from an interval to a Banach space. This case is somewhat simpler than the case of mappings between Banach spaces, but nevertheless is sufficient for many applications.

Let X be a Banach spaces. 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

$$\dot{f}(t) := \lim_{\varepsilon \to 0} \frac{f(t+\varepsilon) - f(t)}{\varepsilon} \tag{9.1}$$

exists. If t is a boundary point, the limit/derivative is understood as the corresponding onesided limit/derivative.

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 $A \in \mathscr{L}(X, Y)$ and $f \in C^k(I, X)$, then $Af \in C^k(I, Y)$ and $\frac{d}{dt} Af = A\dot{f}$.

The following version of the mean value theorem will be crucial.

Theorem 9.1 (Mean value theorem)**.** *Let X be a Banach space and $I \subseteq \mathbb{R}$ an interval. Suppose $f \in C^1(I, X)$. Then*

$$\|f(t) - f(s)\| \le M|t - s|, \qquad M := \sup_{\tau \in [s,t]} \|\dot{f}(\tau)\|, \tag{9.2}$$

for $s \le t \in I$.

Proof. Fix $\tilde{M} > M$ and consider $d(\tau) := \|f(\tau) - f(s)\| - \tilde{M}(\tau - s)$ for $\tau \in [s,t]$. Suppose $[s,\tau_0]$ is the largest interval on which $d(\tau) \le 0$ holds. If $\tau_0 < t$ there must be a sequence $\varepsilon_n \downarrow 0$ such that

$$\begin{aligned} 0 < d(\tau_0 + \varepsilon_n) &\le \|f(\tau_0+\varepsilon_n) - f(\tau_0)\| - \tilde{M}\varepsilon_n + d(\tau_0) \\ &= \|\dot{f}(\tau_0)\varepsilon_n + o(\varepsilon_n)\| - \tilde{M}\varepsilon_n \le (M - \tilde{M} + o(1))\varepsilon_n < 0. \end{aligned}$$

Taking $n \to \infty$ contradicts our assumption. □

In particular,

Corollary 9.2. *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

$$t_0 = a < t_1 < t_2 < \cdots < t_{n-1} < t_n = b \tag{9.3}$$

such that $f(t)$ is constant on each of the open intervals (t_{j-1}, t_j). 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 step functions. Moreover, since the step functions form an algebra, the regulated functions form a Banach algebra.

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. Slightly more general, note that piecewise continuous functions are regulated since every piecewise continuous function is the sum of a continuous function and a step function.

For a step function $f \in S(I,X)$ we can define a linear map $\int : S(I,X) \to X$ by

$$\int_a^b f(t)dt := \sum_{j=1}^n f_j (t_j - t_{j-1}), \tag{9.4}$$

where f_j is the value of f on (t_{j-1}, t_j). This map satisfies

$$\left\| \int_a^b f(t)dt \right\| \le \|f\|_\infty (b-a) \tag{9.5}$$

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

$$\left\| \int_a^b f(t)dt \right\| \le \int_a^b \|f(t)\| dt \tag{9.6}$$

since this holds for step functions by the triangle inequality and hence for all functions by approximation.

We remark that it is possible to extend the integral to a larger class of functions in various ways. The first generalization is to replace step functions by simple functions (and at the same time one could also replace the Lebesgue measure on I by an arbitrary finite measure). Then the same approach defines the integral for uniform limits of simple functions. However, things only get interesting when you also replace the sup norm by an L^1 type seminorm: $\|f\|_1 := \int \|f(x)\|\, d\mu(x)$. As before the integral can be extended to all functions which can be approximated by simple functions with respect to this seminorm. This is know as the Bochner[1] integral and we refer to Section 5.5 from [**40**] for details. For our purpose the present approach will be sufficient.

If $A \in \mathscr{L}(X,Y)$, then $f \in R(I,X)$ implies $Af \in R(I,Y)$ and

$$A \int_a^b f(t)dt = \int_a^b A\, f(t)dt. \tag{9.7}$$

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

$$\ell(\int_a^b f(t)dt) = \int_a^b \ell(f(t))dt, \qquad f \in R(I,X). \tag{9.8}$$

Moreover, 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 $\int_{t_2}^{t_1} f(s)ds := -\int_{t_1}^{t_2} f(s)ds$. Note that we could replace (t_1,t_2) by a closed or half-open 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 9.3 (Fundamental theorem of calculus)**.** *Let X be a Banach space and $I \subseteq \mathbb{R}$ an interval. Suppose $F \in C^1(I,X)$, then*

$$F(t) = F(a) + \int_a^t \dot{F}(s)ds. \tag{9.9}$$

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

$$\begin{aligned}\|\int_a^{t+\varepsilon} f(s)ds - \int_a^t f(s)ds - f(t)\varepsilon\| &= \|\int_t^{t+\varepsilon} (f(s)-f(t))ds\| \\ &\le |\varepsilon| \sup_{s\in[t,t+\varepsilon]} \|f(s)-f(t)\|.\end{aligned}$$

[1]Salomon Bochner (1899–1982), Austrian–Hungarian mathematician

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 9.2. Choosing $t = a$ finally shows $F(a) = C$. $\square$

Problem* 9.1 (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 9.2. *Let X be a Banach algebra and $\mathcal{G}(X)$ the group of invertible elements. Show that if $f \in C^1(I, \mathcal{G}(X))$, then $f^{-1} \in C^1(I,X)$ with*

$$\frac{d}{dt} f^{-1}(t) = -f^{-1}(t)\dot{f}(t)f^{-1}(t).$$

(Hint: Corollary 5.2)

Problem* 9.3. *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* 9.4. *Let $A : \mathfrak{D}(A) \subseteq X \to X$ be a closed operator. Show that* (9.7) *holds for $f \in C(I,X)$ with* $\operatorname{Ran}(f) \subseteq \mathfrak{D}(A)$ *and $Af \in C(I,X)$.*

Problem 9.5. *Let $I = [a,b]$ and $J = [c,d]$ be two compact intervals. Suppose $f(s,t) : I \times J \to X$ is regulated in the sense that it is a uniform limit of step functions being constant on disjoint open rectangles $(s_{j-1}, s_j) \times (t_{k-1}, t_k)$ whose closure cover $I \times J$. Show that*

$$\int_J \left(\int_I f(s,t)ds \right) dt = \int_I \left(\int_J f(s,t)dt \right) ds.$$

(Hint: One way is to use linear functionals and reduce it to the classical Fubini theorem.)

9.2. Multivariable calculus in Banach spaces

We now turn to 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 $U \subseteq X$ be open. 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

$$\|L\| := \sup_{|u|=1} |Lu|. \tag{9.10}$$

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

$$F(x+u) = F(x) + dF(x)\, u + o(u), \tag{9.11}$$

where o, O are the **Landau symbols** (after Edmund Landau). Explicitly

$$\lim_{u\to 0}\frac{|F(x+u)-F(x)-dF(x)\,u|}{|u|}=0. \tag{9.12}$$

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.

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. Clearly this is consistent with the definition (9.1) from the previous section.

Example 9.1. 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. ◇

The previous example emphasizes that for $F:U\subseteq X\to Y$ it makes a big difference whether X is a real or a complex Banach space. In fact, in case of a complex Banach space X, we obtain a version of complex differentiability which of course is much stronger than real differentiability. Note that in this respect it makes no difference whether Y is real or complex.

Example 9.2. Suppose $f\in C^1(\mathbb{R})$ with $f(0)=0$. Let $X:=\ell^p_{\mathbb{R}}(\mathbb{N})$, 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

$$\delta F(x,u) := \lim_{\varepsilon \to 0} \frac{F(x+\varepsilon u) - F(x)}{\varepsilon}, \qquad \varepsilon \in \mathbb{R} \setminus \{0\}, \tag{9.13}$$

which are also known as **Gâteaux derivative**[2] or **variational derivative**. Indeed, if F is differentiable at x, then (9.11) implies

$$\delta F(x,u) = dF(x)u. \tag{9.14}$$

In particular, we call F Gâteaux differentiable at $x \in U$ if the limit on the right-hand side in (9.13) 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

$$\delta F(x,\lambda u) = \lambda\, \delta F(x,u), \qquad \lambda \in \mathbb{R}. \tag{9.15}$$

Example 9.3. 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 $F(x,y) := 0$ else is Gâteaux differentiable at 0 with Gâteaux derivative $\delta F(0) = 0$, which is clearly linear. However, F is not differentiable.

[2]René Gâteaux (1889–1914), French mathematician

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. $\diamond$

Example 9.4. Let $X := L^2_{\mathbb{R}}(0,1)$ and consider

$$F : X \to X, \qquad x \mapsto \sin(x).$$

First of all note that by $|\sin(t)| \le |t|$ our map is indeed from X to X and since sine is Lipschitz continuous we get the same for F: $\|F(x)-F(y)\|_2 \le \|x-y\|_2$. 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 9.7). $\diamond$

Example 9.5. Consider $L^p(X, d\mu)$, $1 \le p < \infty$ and let $G : \mathbb{C} \to \mathbb{R}$ be (real) differentiable with

$$|G(z)| \le C|z|^p, \qquad \sqrt{|\partial_x G(z)|^2 + |\partial_y G(z)|^2} \le C|z|^{p-1}, \qquad z = x + \mathrm{i}y,$$

or, if μ is finite,

$$|G(z)| \le C(1 + |z|^p), \qquad \sqrt{|\partial_x G(z)|^2 + |\partial_y G(z)|^2} \le C(1 + |z|^{p-1}).$$

Note that the first condition comes for free from the second in the finite case and also in the general case if $G(0) = 0$. We only write down the estimates in the first case and leave the easy adaptions for the second case as an exercise.

Then

$$N(f) := \int_X G(f) d\mu$$

is Gâteaux differentiable and we have

$$\delta N(f)g = \int_X \big((\partial_x G)(f)\mathrm{Re}(g) + (\partial_y G)(f)\mathrm{Im}(g)\big)d\mu.$$

In fact, by the chain rule $h(\varepsilon) := G(f+\varepsilon g)$ is differentiable with $h'(0) = (\partial_x G)(f)\mathrm{Re}(g) + (\partial_y G)(f)\mathrm{Im}(g)$. Moreover, by the mean value theorem

$$\begin{aligned}\left|\frac{h(\varepsilon)-h(0)}{\varepsilon}\right| &\le \sup_{0\le\tau\le\varepsilon} \sqrt{(\partial_x G)(f+\tau g)^2 + (\partial_y G)(f+\tau g)^2}\,|g| \\ &\le C2^{p-1}(|f|^{p-1}+|g|^{p-1})|g|\end{aligned}$$

and hence we can invoke dominated convergence to interchange differentiation and integration. Note that using Hölder's inequality this last estimate also shows Lipschitz continuity on bounded sets:

$$|N(f)-N(g)| \le C(\|f\|_p + \|g\|_p)^{p-1}\|f-g\|_p.$$

In particular, for $1 < p < \infty$ the norm

$$N(f) := \int_X |f|^p d\mu$$

is Gâteaux differentiable with

$$\delta N(f)g = p\int_X |f|^{p-2}\mathrm{Re}(fg^*)d\mu. \qquad \diamond$$

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 have linearity (which is easy to check):

Lemma 9.4. *Let X and Y be two Banach spaces and let $U \subseteq X$ be open. 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 9.5. *Let X and Y be two Banach spaces and let $U \subseteq X$ be open. 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*

$$|F(x+u) - F(x)| \le M|u|, \qquad |u| \le \delta. \tag{9.16}$$

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 9.6. 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 $F(x, y) = 0$ else shows: It is Gâteaux differentiable at 0 with $\delta F(0) = 0$ but it is not continuous since $\lim_{\varepsilon\to 0} F(\varepsilon, \varepsilon^2) = 1 \neq 0 = F(0, 0)$. $\diamond$

If F is differentiable for all $x \in U$ we call F **differentiable**. In this case we get a map

$$dF: \begin{array}{ccc} U & \to & \mathscr{L}(X,Y) \\ x & \mapsto & dF(x) \end{array}. \tag{9.17}$$

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 9.7. 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. $\diamond$

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

$$dF(x + v) = dF(x) + d^2F(x)v + o(v). \tag{9.18}$$

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

$$\|d^2F(x)\| = \sup_{|v|=1} \|d^2F(x)v\| = \sup_{|u|=|v|=1} \|(d^2F(x)v)u\|. \tag{9.19}$$

Example 9.8. Note that if $F \in \mathscr{L}(X, Y)$, then $dF(x) = F$ (independent of x) and $d^2F(x) = 0$. $\diamond$

Example 9.9. 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 9.10. Suppose $f \in C^2(\mathbb{R})$ with $f(0) = 0$ and continue Example 9.2. 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 $F \in C^r(U, Y)$, $r \ge 1$, 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 9.11. 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

$$dM(x, y)(u_1, u_2) = u_1 y + x u_2.$$

Consequently dM is linear in (x, y) and hence

$$(d^2M(x, y)(v_1, v_2))(u_1, u_2) = u_1 v_2 + v_1 u_2.$$

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, $r \ge 1$, then F is called a **diffeomorphism** of class C^r.

For the composition of mappings we have the usual **chain rule**.

Lemma 9.6 (Chain rule). *Let X, Y and Z be Banach spaces and let $U \subseteq X$, $V \subseteq Y$ be open. If $F \in C^r(U, V)$ and $G \in C^r(V, Z)$, $r \ge 1$, then $G \circ F \in C^r(U, Z)$ and*

$$d(G \circ F)(x) = dG(F(x)) \circ dF(x), \qquad x \in X. \tag{9.20}$$

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 $|\tilde{v}| \le \|dF(x)\||u| + |o(u)|$ 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 $\ell \in Y^*$ is a bounded linear functional, then $d(\ell \circ F) = d\ell \circ dF = \ell \circ dF$. As an application of this result we obtain

Theorem 9.7 (Schwarz[3]). *Let Y be a Banach space and $U \subseteq \mathbb{R}^n$ open. Suppose $F \in C^2(U, Y)$. Then*

$$\partial_i \partial_j F = \partial_j \partial_i F$$

for any $1 \le i, j \le n$.

Proof. Since this is a local result we can assume $U = \mathbb{R}^n$ for notational simplicity without loss of generality. 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 $\ell \in Y^*$ be a bounded linear functional, then $\ell \circ F \in C^2(\mathbb{R}^2, \mathbb{R})$ and hence $\partial_i \partial_j (\ell \circ F) = \partial_j \partial_i (\ell \circ F)$ by the classical theorem of Schwarz. Moreover, by our remark preceding this lemma $\partial_i \partial_j (\ell \circ F) = \partial_i \ell(\partial_j F) = \ell(\partial_i \partial_j F)$ and hence $\ell(\partial_i \partial_j F) = \ell(\partial_j \partial_i F)$ for every $\ell \in Y^*$ 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

$$\partial_s \partial_t G(t, s)\Big|_{(s,t)=0} = \partial_s dF(x + sv)u\Big|_{s=0} = (d^2 F(x)u)v.$$

Since by the previous theorem the oder of the derivatives is irrelevant, we obtain

$$d^2 F(u, v) = d^2 F(v, u), \tag{9.21}$$

that is, $d^2 F$ is a symmetric bilinear form. This result easily generalizes to higher derivatives. To this end we introduce some notation first.

[3] Hermann Schwarz (1843–1921), German mathematician

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

$$\|L\| := \sup_{x:|x_1|=\cdots=|x_n|=1} |L(x_1, \dots, x_n)| < \infty. \tag{9.22}$$

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 $L : (x_1, \dots, x_n) \mapsto L(x_1, \dots, x_n)$ maps to $x_1 \mapsto L(x_1, .)$.

Lemma 9.8. *Suppose $F \in C^r(X, Y)$, where X and Y are Banach spaces. Then for every $x \in X$ we have that*

$$d^r F(x)(u_1, \dots, u_r) = \partial_{t_1} \cdots \partial_{t_r} F(x + \sum_{i=1}^r t_i u_i)|_{t_1=\cdots=t_r=0}. \tag{9.23}$$

Moreover, $d^r F(x) \in \mathscr{L}^r(X, Y)$ is a bounded symmetric multilinear form.

Proof. The representation (9.23) follows using induction as before. Symmetry follows since the order of the partial derivatives can be interchanged by Lemma 9.7. □

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 9.9):

$$L(u_1, \dots, u_n) = \frac{1}{n!} \partial_{t_1} \cdots \partial_{t_n} L(\sum_{i=1}^n t_i u_i). \tag{9.24}$$

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

$$dL(x)u = L(u_1, x_2) + L(x_1, u_2) \tag{9.25}$$

since

$$\begin{aligned} L(x_1 + u_1, x_2 + u_2) - L(x_1, x_2) &= L(u_1, x_2) + L(x_1, u_2) + L(u_1, u_2) \\ &= L(u_1, x_2) + L(x_1, u_2) + O(|u|^2) \end{aligned} \tag{9.26}$$

as $|L(u_1, u_2)| \le \|L\| |u_1| |u_2| = O(|u|^2)$. If X is a Banach algebra and $L(x_1, x_2) = x_1 x_2$ we obtain the usual form of the product rule.

Next we have the following mean value theorem.

Theorem 9.9 (Mean value). *Let X and Y be Banach spaces. Suppose $U \subseteq X$ is open and $F : U \to Y$ is Gâteaux differentiable at every $x \in U$. If U is convex, then*

$$|F(x)-F(y)| \le M|x-y|, \qquad M := \sup_{0\le t\le 1} |\delta F((1-t)x+ty, \frac{x-y}{|x-y|})|. \tag{9.27}$$

Conversely, (for any open U) if

$$|F(x) - F(y)| \le M|x-y|, \qquad x, y \in U, \tag{9.28}$$

then

$$\sup_{x\in U, |e|=1} |\delta F(x,e)| \le M. \tag{9.29}$$

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 (9.15). Hence the first part follows from Theorem 9.1.

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 9.10. *Suppose $U \subseteq X$ is open and $F \in C^1(U, Y)$. Then F is locally Lipschitz continuous.*

Proof. Fix $x_0 \in U$ 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 9.12. 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 9.11. *Suppose U is an open connected subset of a Banach space X. A Gâteaux 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.*

As an application of the fundamental theorem of calculus (Theorem 9.3) we obtain a generalization of the well-known fact that continuity of the directional derivatives implies continuous differentiability.

Lemma 9.12. *Let X and Y be two Banach spaces and let $U \subseteq X$ be open. Suppose $F : U \to Y$ is Gâteaux differentiable such that the Gâteaux derivative is linear and continuous, $\delta F \in C(U, \mathscr{L}(X,Y))$. 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

$$\begin{aligned} |F(x+u) - F(x) - \delta F(x)u| &= \left| \left(\int_0^1 \big(\delta F(x+tu) - \delta F(x)\big)dt \right) u \right| \\ &\le \left(\int_0^1 \|\delta F(x+tu) - \delta F(x)\| dt \right) |u| \\ &\le \max_{t\in[0,1]} \|\delta F(x+tu) - \delta F(x)\|\, |u|. \end{aligned}$$

By the continuity assumption on δF, the right-hand side is $o(u)$ as required. □

As another consequence we obtain **Taylor's theorem**.

Theorem 9.13 (Taylor)**.** *Let X and Y be two Banach spaces and let $U \subseteq X$ be open. If $F \in C^{r+1}(U,Y)$, then*

$$\begin{aligned} F(x+u) =& F(x) + dF(x)u + \frac{1}{2} d^2F(x)u^2 + \cdots + \frac{1}{r!} d^rF(x)u^r \\ &+ \left(\frac{1}{r!} \int_0^1 (1-t)^r d^{r+1}F(x+tu)dt \right) u^{r+1}, \end{aligned} \tag{9.30}$$

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 (9.25) 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. $\square$

Of course this also gives the Peano form for the remainder:

Corollary 9.14. *Suppose $U \subseteq X$ is open and $F \in C^r(U,Y)$. Then*

$$F(x+u) = F(x) + dF(x)u + \frac{1}{2}d^2F(x)u^2 + \cdots + \frac{1}{r!}d^rF(x)u^r + o(|u|^r). \tag{9.31}$$

Proof. Just estimate

$$\begin{aligned} &\left|\left(\frac{1}{(r-1)!}\int_0^1 (1-t)^{r-1} d^rF(x+tu)dt - \frac{1}{r!}d^rF(x)\right)u^r\right| \\ &\leq \frac{|u|^r}{(r-1)!}\int_0^1 (1-t)^{r-1}\|d^rF(x+tu) - d^rF(x)\|dt \\ &\leq \frac{|u|^r}{r!}\sup_{0\leq t\leq 1}\|d^rF(x+tu) - d^rF(x)\|. \end{aligned}$$

$\square$

Finally we remark that it is often necessary to equip $C^r(U,Y)$ with a norm. A suitable choice is

$$\|F\| = \sum_{0\leq j\leq r}\sup_{x\in U}\|d^jF(x)\|. \tag{9.32}$$

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 (9.12) 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). If U is open we avoid 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 9.6. *Let X be a real Hilbert space, $A \in \mathscr{L}(X)$ and $F(x) := \langle x, Ax\rangle$. Compute d^nF.*

Problem* 9.7. *Let $X := C([0,1],\mathbb{R})$ and suppose $f \in C^1(\mathbb{R})$. Show that*

$$F : X \to X, \qquad x \mapsto f \circ x$$

is differentiable for every $x \in X$ with derivative given by

$$(dF(x)y)(t) = f'(x(t))y(t).$$

Problem 9.8. *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* 9.9. *Show* (9.24).

9.3. Minimizing nonlinear functionals via calculus

Many problems in applications lead to finding the minimum (or maximum) of a given (nonlinear) functional $F : X \to \mathbb{R}$ on some Banach space X. For example, many physical problems can be described by an energy functional and one seeks a solution which minimizes this energy. 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 minimum or maximum. In fact, by virtue of our version of Taylor's theorem (9.31) 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 $u \in \partial B_1(0)$. 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 through x, that is, we look at the behavior of the function

$$f(t) := F(x + tu) \tag{9.33}$$

for a fixed direction $u \in B_1(0)$. Hence this approach is also known as **calculus of variations**. 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

$$\delta^n F(x,u) := \left(\frac{d}{dt}\right)^n F(x + tu)\Big|_{t=0} \tag{9.34}$$

with the derivative defined as a limit as in (9.13). 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

$$\delta^n F(x,\lambda u) = \lambda^n \delta^n F(x,u), \qquad \lambda\in\mathbb{R}. \tag{9.35}$$

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 through a fixed point. This is demonstrated by the following example:

Example 9.13. 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 $B_n \subset \{(x,y)|x\ge 0, \frac{x^2}{2}\le y\le 2x^2\}$. 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 $F(0,0)=0$, $\delta F((0,0),(u,v))=0$, $\delta^2F((0,0),(u,v)) = 2(u^2+v^2)$, and $\delta^k F((0,0),(u,v))=0$ for $k\ge 3$.

In particular, $F(ut,vt)$ has a strict local minimum at $t=0$ for every $(u,v)\in\mathbb{R}^2\backslash\{0\}$, 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 9.15. *Let X be a Banach spaces and let $U\subseteq X$ be open. 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 minimum is that $\delta F(x,u)=0$ and $\delta^2F(x,u)\ge 0$ for all $u\in X$. A sufficient condition for a strict local minimum is if in addition $\delta^2F(x,u)\ge c>0$ for all $u\in\partial B_1(0)$ and δ^2F 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 δ^2F imply that there is some $\varepsilon>0$ such that $\delta^2F(y,u)\ge\frac{c}{2}$ for all $y\in B_\varepsilon(x)$ and all $u\in\partial B_1(0)$. Equivalently, $\delta^2F(y,u)\ge\frac{c}{2}|u|^2$ for all $y\in B_\varepsilon(x)$ and all $u\in X$. Hence applying Taylor's theorem to $f(t)$ using $\ddot f(t) = \delta^2F(x+tu,u)$ gives

$$F(x+u) = f(1) = f(0) + \int_0^1 (1-s)\ddot f(s)ds \ge F(x) + \frac{c}{4}|u|^2$$

for $u\in B_\varepsilon(0)$. □

Note that if $F\in C^2(U,\mathbb{R})$ then $\delta^2F(x,u) = d^2F(x)u^2$ and we obtain

Corollary 9.16. *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 $d^2F(x)u^2\ge c|u|^2$ for all $u\in X$.*

Proof. Observe that by $|\delta^2 F(x,u) - \delta^2 F(y,u)| \le \|d^2F(x) - d^2F(y)\| |u|^2$ the continuity requirement from the previous lemma is satisfied. □

Example 9.14. If X is a real 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 $d^2F(x)u^2 \ge c|u|^2$ 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 9.15. 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 $d^2F(0)u^2 = \sum_n n^{-2}u_n^2 > 0$ for $u \ne 0$. However, $F(\delta^m/m) < 0$ shows that 0 is not a 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 $d^2F(x,u) \ge c > 0$ for all $u \in \partial B_1(0)$. ◇

Example 9.16. 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**[4] 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

$$q(t) := q(a) + \frac{t-a}{b-a}(q(b) - q(a)) + x(t), \qquad x \in X.$$

[4]Joseph-Louis Lagrange (1736–1813), Italian mathematician and astronomer

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

$$\begin{aligned}\delta F(x,u) &= \frac{d}{dt}\Big|_{t=0} \int_a^b L(s, q(s) + t\, u(s), \dot q(s) + t\, \dot u(s)) ds \\ &= \int_a^b \big(L_q(s,q(s),\dot q(s)) u(s) + L_{\dot q}(s,q(s),\dot q(s))\dot u(s)\big) ds \\ &= \int_a^b \big(L_q(s,q(s),\dot q(s)) u(s) - \frac{d}{ds} L_{\dot q}(s,q(s),\dot q(s))\big) u(s) ds,\end{aligned}$$

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**

$$\frac{d}{ds} L_{\dot q}(s,q(s),\dot q(s)) = L_q(s,q(s),\dot q(s)).$$

There are two special cases worth while noticing: First of all, if L does not depend on q, then the Euler–Lagrange equation simplifies to

$$L_{\dot q}(s,\dot q(s)) = C$$

and this identity can be derived directly from $\delta F(x,u)$ without performing the integration by parts (Lemma 3.24 from [**40**]). In particular, it suffices to assume $L \in C^1$ and $x \in C^1$ in this case.

Secondly, if L does not depend on s, we can eliminate $L_q \dot q$ from

$$\frac{d}{ds} L(q(s),\dot q(s)) = L_q(q(s),\dot q(s))\dot q(s) + L_{\dot q}(q(s),\dot q(s))\ddot q(s)$$

using the Euler–Lagrange identity which gives

$$\frac{d}{ds} L(q(s),\dot q(s)) = \dot q(s) \frac{d}{ds} L_{\dot q}(q(s),\dot q(s)) + L_{\dot q}(q(s),\dot q(s))\ddot q(s).$$

Rearranging this last equation gives the **Beltrami identity**

$$L(q(s),\dot q(s)) - \dot q(s) L_{\dot q}(q(s),\dot q(s)) = C$$

which must be satisfied by any solution of the Euler–Lagrange equation. Conversely, if $\dot q \neq 0$, a solution of the Beltrami equation also solves the Euler–Lagrange equation. ⋄

Example 9.17. 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 equations 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 9.17. *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 9.18. *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 the fact that a convex function is always above its tangent (cf. Lemma 3.2 from [**40**]) implies $F(y) = f(1) \ge f(0) = F(x)$. □

As in the one-dimensional case, convexity can be read off from the second derivative.

Lemma 9.19. *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 $\delta^2 F(x,u) \ge 0$ 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

$$F(\lambda y + (1-\lambda)x) = f(\lambda) \le \lambda f(1) - (1-\lambda) f(0) = \lambda F(y) - (1-\lambda)F(x)$$

with strict inequality if f is strictly convex. The rest follows from Problem 9.10. □

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 12.

Lemma 9.20. *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 9.10 since $f'(t) = \delta F(x+tu)u$ which shows that δF is (strictly) monotone if and only if f' is (strictly) increasing. □

Example 9.18. The length of a curve $q : [a, b] \to \mathbb{R}^n$ is given by

$$\int_a^b |q'(s)| ds.$$

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, 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 $|q'| \geq c$. 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 $C := \{x \in X | |x'| < \frac{|q_1 - q_0|}{2(b-a)}\}$. We compute

$$\delta F(x, u) = \int_a^b \frac{q'(s) u'(s)}{|q'(s)|} ds$$

which shows by virtue of the du Bois-Reymond[5] Lemma (Lemma 3.24 from [**40**]) that $q'/|q'|$ 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 9.19. Let us try to find a curve $q(t)$ from $q(0) = 0$ to $q(t_1) = q_1$ which minimizes

$$F(q) := \int_0^{t_1} \sqrt{\frac{1 + \dot{q}(t)^2}{t}} dt.$$

[5]Paul du Bois-Reymond (1831–1889), German mathematician

Note that since the function $v \mapsto \sqrt{1+v^2}$ is convex, we obtain that F is convex. Hence it suffices to find a zero of

$$\delta F(q,u) = \int_0^b \frac{\dot{q}(t)\dot{u}(t)}{\sqrt{t(1+\dot{q}(t)^2)}} dt,$$

which shows by virtue of the du Bois-Reymond Lemma (Lemma 3.24 from [**40**]) that $\frac{\dot{q}}{\sqrt{t(1+\dot{q}^2)}} = C^{-1/2}$ is constant or equivalently

$$\dot{q}(t) = \sqrt{\frac{t}{C-t}}$$

and hence

$$q(t) = C \arctan\left(\sqrt{\frac{t}{C-t}}\right) - \sqrt{t(C-t)}.$$

The constant C has to be chosen such that $q(t_1)$ matches the given value q_1. Note that $C \mapsto q(t_1)$ decreases from $\frac{\pi t_1}{2}$ to 0 and hence there will be a unique $C > t_1$ for $0 < q_1 < \frac{\pi t_1}{2}$. ⋄

Example 9.20. The classical **brachistochrone problem**, as posed by Johann Bernoulli, reads: Given two points A and B in a vertical plane, what is the curve traced out by a point acted on only by gravity, which starts at A and reaches B in the shortest time.

This leads to the problem of finding a curve $q(t)$ from $q(0) = q_0 > 0$ to $q(t_1) = 0$ which minimizes

$$F(q) := \int_0^{t_1} \sqrt{\frac{1+\dot{q}(t)^2}{q(t)}} dt.$$

Since the Lagrangian is independent of t, the solution of the Euler–Lagrange equation must satisfy the Beltrami identity (cf. Problem 9.16)

$$\frac{1}{\sqrt{q(t)(1+\dot{q}(t)^2)}} = C$$

Rearranging we obtain

$$\dot{q}(t)^2 = \frac{2r}{q(t)} - 1$$

(with $r := 1/(2C)$) the equation of the cycloid.

One can check that the function $f(q,v) := \sqrt{\frac{1+v^2}{q}}$ is quasi-convex on $\mathbb{R}\times(0,\infty)$ and hence so is F and thus our solution is the unique minimum. ⋄

Problem* 9.10. *Show that a differentiable function $\varphi : (a,b) \to \mathbb{R}$ is (strictly) convex if and only if φ' is (strictly) increasing. Moreover, if φ is twice differentiable it is (strictly) convex if and only if $\varphi'' \geq 0$ ($\varphi'' > 0$ a.e.).*

Problem 9.11. *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 $V'' \le 0$.

Problem 9.12. *Let x be a Hilbert space an $A : \mathfrak{D}(A) \subseteq X \to X$ a (densely defined) symmetric operator. Show that if*

$$\lambda_0 := \inf_{u\in\mathfrak{D}(A): \|u\|=1} \langle u, Au\rangle$$

is attained for some $u_0 \in \mathfrak{D}(A)$, then u_0 is an eigenvector corresponding to the eigenvalue λ_0.

9.4. Minimizing nonlinear functionals via compactness

Another approach for minimizing a nonlinear functional $F : M \subseteq X \to \mathbb{R}$ is based on compactness. 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 9.21. By Lemma 4.29 (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 4.44). ◇

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 B.34) will already be sufficient. This is frequently referred to as the **direct method in the calculus of variations** due to Zaremba[6] and Hilbert:

Theorem 9.21 (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 $F(x_0) = \inf_M F$.*

[6] Stanisław Zaremba (1863–1942), Polish mathematician and engineer

If F is Gâteaux differentiable, then

$$\delta F(x_0, u) = 0 \tag{9.36}$$

for every $u \in X$ with $x_0 + \varepsilon u \in M$ for sufficiently small ε.

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 $F(x_n) \to \inf_M F < \infty$. If M is unbounded, then the fact that F is coercive implies that x_n is bounded. Otherwise, if M is bounded, it is obviously bounded. Hence by Theorem 4.32 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 $\inf_M F = \lim_{n\to\infty} F(x_n) = \liminf_{n\to\infty} F(x_n) \ge F(x_0)$. □

Note that looking for a maximum of F is the same as looking for a minimum of $-F$. However, in this case lower semicontinuous turns into upper semicontinuous, so the conditions on F are different in this case.

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 9.22. *Let X be a Banach space and 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 6.4 we can find a linear functional ℓ which separates $\{x\}$ and M $\mathrm{Re}(\ell(x)) < c \le \mathrm{Re}(\ell(y))$, $y \in M$. But this contradicts $\mathrm{Re}(\ell(x)) < c \le \mathrm{Re}(\ell(x_n)) \to \mathrm{Re}(\ell(x))$. □

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

$$F(\lambda x + (1-\lambda)y) \le \max\{F(x), F(y)\}, \qquad \lambda \in (0,1), \quad x, y \in M. \tag{9.37}$$

It is called **strictly quasiconvex** if the inequality is strict for $x \ne y$. By $\lambda F(x) + (1-\lambda)F(y) \le \max\{F(x), F(y)\}$ every (strictly) convex function is (strictly) quasiconvex. The converse is not true as the following example shows.

Example 9.22. 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 $M = \mathbb{R}$. ⋄

Note that we can extend a (quasi-)convex function $F : M \to \overline{\mathbb{R}}$ to all of X by setting $F(x) = \infty$ for $x \in X \setminus M$ and the resulting function will still be (quasi-)convex and will have the same infimum.

Now we are ready for the next

Lemma 9.23. *Suppose $M \subseteq X$ is a closed convex set of some Banach space X 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 $F(x_n) \le a < F(x_0)$. Then $x_n \in F^{-1}((-\infty, a])$ implying $x_0 \in F^{-1}((-\infty, a])$ as this set is convex (Problem 9.15) and closed (Problem B.34). But this gives the contradiction $a < F(x_0) \le a$. □

Example 9.23. Let $U \subseteq \mathbb{R}^n$ and $K : U \times \mathbb{C} \to [0, \infty)$ measurable. Suppose $u \mapsto K(x, u)$ is convex and continuous for fixed $x \in U$. Then

$$F(u) := \int_U K(x, u(x)) d^n x$$

is weakly sequentially lower semicontinuous on $L^p(U)$ for $1 \le p \le \infty$. Since F is convex, it suffices to show lower semicontinuity. Assume the contrary, then we can find some $u \in L^p$ and a sequence $u_n \to u$ such that $F(u) > \liminf F(u_n)$. After passing to a subsequence we can assume that $u_n(x) \to u(x)$ a.e. and hence $K(x, u_n(x)) \to K(x, u(x))$ a.e. Finally applying Fatou's lemma (Theorem 2.4 from [**40**]) gives the contradiction $F(u) \le \liminf F(u_n)$.

Moreover, if $u \mapsto K(x, u)$ is strictly convex for a.e. $x \in U$, then F is strictly convex. Indeed, in this case

$$\begin{aligned} F(\lambda u + (1-\lambda) v) &= \int_U K(x, \lambda u(x) + (1-\lambda) v(x)) d^n x \\ &\le \int_U \big(\lambda K(x, u(x)) + (1-\lambda) K(x, v(x))\big) d^n x = \lambda F(u) + (1-\lambda) F(v) \end{aligned}$$

and equality would imply $K(x, \lambda u(x) + (1-\lambda) v(x)) = K(x, \lambda u(x) + (1-\lambda) v(x))$ for a.e. x and hence $u(x) = v(x)$ for a.e. x.

Note that this result generalizes to $\mathbb{C}^n$-valued functions in a straightforward manner. ◇

Moreover, in this case our variational principle reads as follows:

Corollary 9.24. *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 9.24. 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 9.24. Note that the assumption that X is reflexive is crucial (Problem 9.13). Moreover, we also get that x_0 is unique if X is strictly convex (see Problem 1.19). ⋄

Example 9.25. 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 9.26. Let $\mathfrak{H}$ be a Hilbert space and let us consider the problem of finding the lowest eigenvalue of a positive operator $A \ge 0$. 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 1.34) 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 9.21. 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 4.29 (ii) versus Lemma 4.30). If we make the additional assumption that A is compact, then q_A is weakly sequentially continuous as can be seen from Theorem 4.33. 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. ⋄

Example 9.27. The classical **Poisson problem**[7] asks for a solution of

$$-\Delta u = f$$

in a bounded domain $U \subset \mathbb{R}^n$ attaining given boundary values g on ∂U. We are going to look for weak solutions, that is, solutions $u \in H^1_{\mathbb{R}}(U)$ satisfying

$$\int_U (\nabla u \cdot \nabla\phi - f\phi)\, d^n x = 0, \qquad \phi \in C_c^\infty(U).$$

We start by introducing the functional

$$F(u) := \int_U \left(\frac{1}{2}|\nabla u|^2 - uf\right) d^n x$$

on $H^1_{\mathbb{R}}(U)$. To incorporate the boundary values we introduce

$$M := \{v \in H^1_{\mathbb{R}}(U) | v|_{\partial U} = g\}.$$

Here the equality $v|_{\partial U} = g$ has to be understood in the sense of traces and hence we need to require U to have a C^1 boundary such that the trace operator is well-defined. Moreover, we assume $f \in L^2_{\mathbb{R}}(U)$ and g in the range of the trace operator, such that M is nonempty. In particular, there is some $\bar{g} \in H^1_{\mathbb{R}}(U)$ with $\bar{g}|_{\partial U} = g$. For example, we can assume that g is Lipschitz continuous in which case it has a Lipschitz continuous extension $\bar{g}$ to $\overline{U}$ (Lemma B.29). In this case $\bar{g} \in W^{1,\infty}_{\mathbb{R}}(U) \subset H^1_{\mathbb{R}}(U)$. Moreover, by continuity of the trace operator, M is closed and convexity is obvious.

To see that F is weakly coercive, let $u = \bar{g} + v$, where $v \in H^1_0(U)$ vanishes on the boundary, then

$$F(u) \geq \frac{1}{2}\|\partial v\|_2^2 - \|\partial\bar{g}\|_2\|\partial v\|_2 - \|f\|_2\|v\|_2 - C,$$

[7] Siméon Denis Poisson (1781–1840), French mathematician, engineer, and physicist

with C depending on f and g only. Now the Poincaré inequality (Theorem 7.38 from [**40**]) $\|v\|_2 \le C_0 \|\partial v\|_2$ implies that $F(u) \to \infty$ if $\|v\|_{1,2} \to \infty$. Finally, since F is convex and continuous, Corollary 9.24 implies existence of a unique minimizer.

Finally, F is Gâteaux differentiable with

$$\delta F(u)v = \int_U (\nabla u \cdot \nabla v - fv)\, d^n x$$

and we have $u + \varepsilon v \in M$ whenever $u \in M$ and $v \in H_0^1(U, \mathbb{R})$. Hence this minimizer solves the weak formulation of our boundary value problem. ⋄

Example 9.28. Let us consider the following nonlinear elliptic problem

$$-\Delta u + u|u| + u = f$$

in $L^2_{\mathbb{R}}(\mathbb{R}^n)$ for a given function $f \in L^2_{\mathbb{R}}(\mathbb{R}^n)$. We are going to look for weak solutions, that is, solutions $u \in H^1_{\mathbb{R}}(\mathbb{R}^n)$ satisfying

$$\int_{\mathbb{R}^n} (\partial u \cdot \partial \phi + (|u|u + u - f)\phi)\, d^n x = 0, \qquad \phi \in C_c^\infty(\mathbb{R}^n).$$

We start by introducing the functional

$$F(u) := \int_{\mathbb{R}^n} \left(\frac{1}{2}|\partial u|^2 + \frac{1}{3}|u|^3 + \frac{1}{2}u^2 - uf \right) d^n x$$

on $X := L^2_{\mathbb{R}}(\mathbb{R}^n)$ and set $F(u) = \infty$ if $u \notin H^1_{\mathbb{R}}(\mathbb{R}^n) \cap L^3_{\mathbb{R}}(\mathbb{R}^n)$. We also choose $M := X$. One checks that for $u \in H^1_{\mathbb{R}}(\mathbb{R}^n) \cap L^3_{\mathbb{R}}(\mathbb{R}^n)$ and $\phi \in C_c^\infty(\mathbb{R}^n)$ this functional has a variational derivative

$$\delta F(u, \phi) = \int_{\mathbb{R}^n} (\partial u \cdot \partial \phi + (|u|u + u - f)\phi)\, d^n x = 0$$

which coincides with the weak formulation of our problem. Hence a minimizer (which is necessarily in $H^1_{\mathbb{R}}(\mathbb{R}^n) \cap L^3_{\mathbb{R}}(\mathbb{R}^n)$) is a weak solution of our nonlinear elliptic problem and it remains to show existence of a minimizer.

First of all note that

$$F(u) \ge \frac{1}{2}\|u\|_2^2 - \|u\|_2 \|f\|_2 \ge \frac{1}{4}\|u\|_2^2 - \|f\|_2^2$$

and hence F is coercive. To see that it is weakly sequentially lower continuous, observe that for the first term this follows from convexity, for the second term this follows from Example 9.23 and the last two are easy. Hence the claim follows. ⋄

If we look at Example 9.28 in the case $f = 0$, our approach will only give us the trivial solution. In fact, for a linear problem one has nontrivial solutions for the homogenous problem only at an eigenvalue. Since the Laplace operator has no eigenvalues on $\mathbb{R}^n$ (as is not hard to see using the Fourier transform), we look at a bounded domain U instead. To avoid the trivial solution we will add a constraint. Of course the natural constraint is to require admissible elements to be normalized. However, since the unit sphere is not weakly closed (its weak closure is the unit ball — see Example 6.10), we cannot simply add this requirement to M. To overcome this problem we will use that another way of getting weak sequential closedness is via compactness:

Lemma 9.25. *Let X, Y be Banach spaces such that X is compactly embedded into Y and let $N : Y \to \mathbb{R}$ be continuous. Then $M := \{x \in X | N(x) = N_0\} \subseteq X$ is weakly sequentially closed for any $N_0 \in \mathbb{R}$. The same holds for $M := \{x \in X | N(x) \le N_0\}$.*

Proof. This follows from Theorem 4.33 since every weakly convergent sequence in X is convergent in Y. □

Theorem 9.26 (Variational principle with constraints)**.** *Let X be a reflexive Banach space and let $F : X \to \mathbb{R}$ be weakly sequentially lower semicontinuous and weakly coercive. Let Y be another Banach space such that X is compactly embedded into Y and let $N : Y \to \mathbb{R}$ be continuous. Fix $N_0 \in \mathbb{R}$ and suppose that $M := \{x \in X | N(x) = N_0\}$ is nonempty. Then there exists some $x_0 \in M$ with $F(x_0) = \inf_M F$.*

If in addition F and N are Gâteaux differentiable and δN does not vanish on M, then there is a constant $\lambda \in \mathbb{R}$ (the ***Lagrange multiplier****) such that*

$$\delta F(x_0) = \lambda \delta N(x_0). \tag{9.38}$$

Proof. Existence follows from Theorem 9.21 which is applicable thanks to our previous lemma. Now choose some $x_1 \in X$ such that $\delta N(x_0)x_1 \neq 0$ and $x \in X$ arbitrary. Then the function

$$f(t, s) := N(x_0 + t\, x + x_1 s)$$

is $C^1(\mathbb{R}^2)$ and satisfies

$$\partial_t f(t, s) = \delta N(x_0 + t\, x + s\, x_1)x, \qquad \partial_s f(t, s) = \delta N(x_0 + t\, x + x_1 s)x_1$$

and since $\partial_s f(0, 0) \neq 0$ the implicit function theorem implies existence of a function $\sigma \in C^1(-\varepsilon, \varepsilon)$ such that $\sigma(0) = 0$ and $f(t, \sigma(t)) = f(0, 0)$, that is, $x(t) := x_0 + t\, x + \sigma(t)x_1 \in M$ for $|t| < \varepsilon$. Moreover,

$$\sigma'(0) = -\frac{\partial_t f(0, 0)}{\partial_s f(0, 0)} = -\frac{\delta N(x_0)x}{\delta N(x_0)x_1}.$$

Hence, as before together with the chain rule

$$\frac{d}{dt} F(x_0 + t\, x + \sigma(t)x_1)|_{t=0} = \delta F(x_0)(x + \sigma'(0)x_1) = \delta F(x_0)x - \lambda \delta N(x_0)x = 0,$$

where

$$\lambda := \frac{\delta F(x_0)x_1}{\delta N(x_0)x_1}.$$

□

Example 9.29. Let $U \subset \mathbb{R}^n$ be a bounded domain and consider

$$F(u) := \frac{1}{2}\int_U |\partial u|^2 d^n x, \qquad u \in H^1_0(U,\mathbb{R})$$

subject to the constraint

$$N(u) := \int_U G(u) d^n x = N_0,$$

where $G : \mathbb{R} \to \mathbb{R}$ is differentiable and satisfies

$$|G'(x)| \le C(1+|x|).$$

This condition implies

$$|G(x)| \le \tilde{C}(1+|x|^2)$$

and ensures that $N(u)$ is well-defined for all $u \in L^2(U)$.

In order to apply the theorem we set $X := H^1_0(U,\mathbb{R})$ and $Y := L^2(U)$. That X is compactly embedded into Y is the Rellich–Kondrachov theorem (Theorem 7.35 from [**40**]). Moreover, by the Poincaré inequality (Theorem 7.38 from [**40**]) we can choose $\|x\|^2 := F(x)$ as an equivalent norm on X. In particular, F satisfies the requirements of our theorem and so does N by Example 9.5. Consequently, if N_0 is such that

$$M := \{u \in X | N(u) = N_0\}$$

is nonempty, there is a minimizer u_0. By Example 9.1 and Example 9.5

$$dF(u_0)u = \int_U (\nabla u_0)(\nabla u)\, d^n x, \quad \delta N(u_0)u = \int_U G'(u_0)u\, d^n x$$

and if we can find some $u \in H^1_0(U)$ such that this derivative is nonzero, then u_0 satisfies

$$\int_U \big(\partial u_0 \cdot \partial u - \lambda G'(u_0)u\big) d^n x = 0, \qquad u \in H^1_0(U,\mathbb{R}),$$

and hence is a weak solution of the nonlinear eigenvalue problem

$$-\Delta u_0 = \lambda G'(u_0).$$

Note that this last condition is for example satisfied if $G(0) = 0$, $G'(x)x > 0$ for $x \neq 0$, and $N_0 > 0$. Indeed, in this case $\delta N(u_0)u_0 = \int_U G'(u_0)u_0 d^n x > 0$ since otherwise we would have $u_0 = 0$ contradicting $0 < N_0 = N(u_0) = N(0) = 0$.

Of course in the case $G(x) = \frac{1}{2}|x|^2$ and $N_0 = 1$ this gives us the lowest eigenvalue of the Laplacian on U with Dirichlet boundary conditions.

Note that using continuous embeddings $L^2 \hookrightarrow L^p$ with $2 \le p \le \infty$ for $n = 1$, $2 \le p < \infty$ for $n = 2$, and $2 \le p \le \frac{2n}{n-2}$ for $n \ge 3$ one can improve this result to the case

$$|G'(x)| \le C(1+|x|^{p-1}). \qquad \diamond$$

Problem 9.13. *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 9.14. *Let $M \subseteq X$ be a convex subset of some vector space X. Show that $F : M \to \overline{\mathbb{R}}$ is convex if and only if its **epigraph** $\operatorname{epi} F := \{(x,a) \in M \times \overline{\mathbb{R}} | F(x) \le a\} \subset X \times \overline{\mathbb{R}}$ is convex.*

Problem* 9.15. *Let $M \subseteq X$ be a convex subset of some vector space X. 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 $a \in \mathbb{R}$.*

9.5. 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

$$|F(x) - F(\tilde{x})| \le \theta |x - \tilde{x}|, \qquad x, \tilde{x} \in C. \tag{9.39}$$

Note that a contraction is continuous. We also recall the notation $F^n(x) = F(F^{n-1}(x))$, $F^0(x) = x$.

Theorem 9.27 (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*

$$|F^n(x) - \overline{x}| \le \frac{\theta^n}{1-\theta} |F(x) - x|, \qquad x \in C. \tag{9.40}$$

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 $n > m$)

$$\begin{aligned} |x_n - x_m| &\le \sum_{j=m+1}^{n} |x_j - x_{j-1}| \le \theta^m \sum_{j=0}^{n-m-1} \theta^j |x_1 - x_0| \\ &\le \frac{\theta^m}{1-\theta} |x_1 - x_0|. \end{aligned} \tag{9.41}$$

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 (9.40) follows after taking the limit $m \to \infty$ in (9.41). $\square$

Example 9.30. Consider $X := C[0,1]$ and let $C := \{x \in X | x(0) = 0, x(1) = 1,\ 0 \le x \le 1\}$. Then the map $F(x)(t) := t\,x(t)$ maps C to C and satisfies $\|F(x) - F(\tilde{x})\|_\infty < \|x - \tilde{x}\|_\infty$ for $x \neq \tilde{x}$ and $x, \tilde{x} \in C$. Nevertheless it has no fixed point in C. $\diamond$

Note that we can replace θ^n by any other summable sequence θ_n (Problem 9.17):

Theorem 9.28 (Weissinger)**.** *Let C be a nonempty closed subset of a Banach space X. Suppose $F : C \to C$ satisfies*

$$|F^n(x) - F^n(y)| \le \theta_n |x - y|, \qquad x, y \in C, \tag{9.42}$$

with $\sum_{n=1}^\infty \theta_n < \infty$. Then F has a unique fixed point $\overline{x}$ such that

$$|F^n(x) - \overline{x}| \le \left(\sum_{j=n}^\infty \theta_j \right) |F(x) - x|, \qquad x \in C. \tag{9.43}$$

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

$$|F(x,y) - F(\tilde{x},y)| \le \theta |x - \tilde{x}|, \quad x, \tilde{x} \in \overline{U},\ y \in V, \tag{9.44}$$

that is, the contraction constant θ is independent of y.

Theorem 9.29 (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 $\overline{x}(.) \in C^r(V, U)$. If F is Lipschitz with respect to the paramter, so is the fixed point.*

Proof. Let us first show that $\overline{x}(y)$ is continuous. From

$$\begin{aligned} |\overline{x}(y+v) - \overline{x}(y)| &= |F(\overline{x}(y+v), y+v) - F(\overline{x}(y), y+v) \\ &\quad + F(\overline{x}(y), y+v) - F(\overline{x}(y), y)| \\ &\le \theta |\overline{x}(y+v) - \overline{x}(y)| + |F(\overline{x}(y), y+v) - F(\overline{x}(y), y)| \end{aligned}$$

we infer

$$|\overline{x}(y+v) - \overline{x}(y)| \le \frac{1}{1-\theta} |F(\overline{x}(y), y+v) - F(\overline{x}(y), y)| \tag{9.45}$$

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,

$$d\,\overline{x}(y) = \partial_x F(\overline{x}(y), y) d\,\overline{x}(y) + \partial_y F(\overline{x}(y), y). \tag{9.46}$$

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 9.9. Hence we get a unique continuous solution $\overline{x}'(y)$. It remains to show

$$\overline{x}(y + v) - \overline{x}(y) - \overline{x}'(y)v = o(v).$$

Let us abbreviate $u := \overline{x}(y+v) - \overline{x}(y)$, then using (9.46) and the fixed point property of $\overline{x}(y)$ we see

$$\begin{aligned}
(1 - \partial_x F(\overline{x}(y), y))(u - \overline{x}'(y)v) = \\
&= F(\overline{x}(y) + u, y + v) - F(\overline{x}(y), y) - \partial_x F(\overline{x}(y), y)u - \partial_y F(\overline{x}(y), y)v \\
&= o(u) + o(v)
\end{aligned}$$

since $F \in C^1(U \times V, U)$ by assumption. Moreover, $\|(1 - \partial_x F(\overline{x}(y), y))^{-1}\| \le (1 - \theta)^{-1}$ and $u = O(v)$ (by (9.45)) implying $u - \overline{x}'(y)v = o(v)$ as desired.

Finally, suppose that the result holds for some $r - 1 \ge 1$. 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 (9.46) shows $d\,\overline{x}(y) \in C^{r-1}(V, U)$ and hence $\overline{x}(y) \in C^r(V, U)$. □

As an important consequence we obtain the implicit function theorem.

Theorem 9.30 (Implicit function). *Let X, Y, and Z be Banach spaces and let U, V be open subsets of X, Y, respectively. Let $F \in C^r(U \times V, Z)$, $r \ge 0$, 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 $r \ge 1$)*

$$d\xi(y) = -(\partial_x F(\xi(y), y))^{-1} \circ \partial_y F(\xi(y), y). \tag{9.47}$$

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 $\|\partial_x G(x, y)\| \le \theta < 1$ for $(x, y) \in U_1 \times V_1$. Thus by the mean value theorem (Theorem 9.9) $G(., y)$ 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

$$|G(x, y) - x_0| = |G(x, y) - G(x_0, y_0)| \le \theta|x - x_0| + (1 - \theta)r < r$$

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 (9.47) 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 9.31 (Inverse function)**.** *Let X, Y be Banach spaces and $U \subseteq X$ open. Suppose $F \in C^r(U, Y)$, $r \geq 1$, 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 9.31. 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: $F(x,y) = (\mathrm{e}^{2x} - y^2 + 3, 4\mathrm{e}^{2x} y - y^3)$. Note that $\det \frac{\partial F}{\partial(x,y)} = 8\mathrm{e}^{4x} + 10\mathrm{e}^{2x} y^2$ and $F(0,2) = (0,0) = F(0,-2)$. ⋄

Example 9.32. 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 9.16. *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 9.27 are satisfied.

Problem* 9.17. *Prove Theorem 9.28. 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 9.28 (except for the estimate) can also be considered as a special case of Theorem 9.27 since the assumption implies that F^n is a contraction for n sufficiently large.*

9.6. 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 9.32. *Let X, Y be Banach spaces and $U \subseteq X$ open. 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*

$$(f_*x)(t) := f(x(t)). \tag{9.48}$$

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 9.14) 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

$$\begin{array}{rlcl} \lambda: & C(I,\mathscr{L}(X,Y)) & \to & \mathscr{L}(C(I,X),C(I,Y)) \\ & T & \mapsto & T_* \end{array},$$

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. $\square$

Now we come to our existence and uniqueness result for the initial value problem in Banach spaces.

Theorem 9.33. *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*

$$\dot{x} = F(t,x,\lambda), \qquad x(t_0) = x_0, \quad (t_0,x_0,\lambda) \in I \times U \times \Lambda, \tag{9.49}$$

has a unique solution $x(t, t_0, x_0, \lambda) \in C^r(I_1 \times I_2 \times U_1 \times \Lambda_1, U)$*, 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

$$\dot{x} = \varepsilon F(x_0 + x), \quad x \in D^1 := \{x \in C^1([-1,1], B_\delta(0)) | x(0) = 0\}, \tag{9.50}$$

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 (9.50) is equivalent to looking for zeros of the function

$$\begin{array}{rlcl} G: & D^1 \times U_0 \times \mathbb{R} & \to & C([-1,1], X), \\ & (x, x_0, \varepsilon) & \mapsto & \dot{x} - \varepsilon F(x_0 + x), \end{array} \tag{9.51}$$

where U_0 is a neighborhood of x_0 and δ sufficiently small such that $U_0 + B_\delta(0) \subseteq U$. Lemma 9.32 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 9.33. 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 1.59) and solves the differential equation (Problem 9.18). ⋄

Example 9.34. 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 (9.49) can be defined.

Suppose that solutions of the initial value problem (9.49) exist locally and are unique (as guaranteed by Theorem 9.33). Let ϕ_1, ϕ_2 be two solutions of (9.49) 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

$$\phi(t) := \begin{cases} \phi_1(t), & t \in I_1, \\ \phi_2(t), & t \in I_2, \end{cases} \tag{9.52}$$

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 9.34. *Suppose the initial value problem* (9.49) *has a unique local solution (e.g. the conditions of Theorem 9.33 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 (9.49) 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 9.35. *Suppose $F \in C^1(\mathbb{R}\times X, X)$ and let $x(t)$ be a maximal solution of the initial value problem* (9.49)*. 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 9.35. Finally, we want 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 through 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 $Aq_n = -k(q_{n+1} - 2q_n + q_{n-1})$ is a bounded operator in $X := \ell^p_{\mathbb{R}}(\mathbb{Z})$ 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 9.18) 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

$$\cos(tA^{1/2}) := \sum_{k=0}^{\infty} \frac{t^{2k}}{(2k)!}A^k, \qquad \frac{\sin(tA^{1/2})}{A^{1/2}} := \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}A^k.$$

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 $\ell^p_{\mathbb{R}}(\mathbb{Z})$ (see Example 9.2). 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 9.33 gives us existence of a local solution in $\ell^p_{\mathbb{R}}(\mathbb{Z}) \oplus \ell^p_{\mathbb{R}}(\mathbb{Z})$. 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 $X := \ell^2_{\mathbb{R}}(\mathbb{Z}) \oplus \ell^2_{\mathbb{R}}(\mathbb{Z})$ as our underlying Banach (in this case even Hilbert) space. Recall that since we assume V to have a minimum at 0, we have $|V(r)| \le C_R r^2$ 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 $V \geq 0$,

$$\|p(t)\|_2^2 \leq 2H(p(t), q(t)) = 2H(p(0), q(0)).$$

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

$$\|q(t)\|_2 \leq \|q(0)\|_2 + \int_0^t \|p_n(s)\|_2 ds \leq \|q(0)\|_2 + \sqrt{2H(p(0), q(0))}t.$$

So Lemma 9.35 ensures that solutions are global in X. Of course every solution from X is also a solution from $Y = \ell^p_{\mathbb{R}}(\mathbb{Z}) \oplus \ell^p_{\mathbb{R}}(\mathbb{Z})$ for all $p \geq 2$ (since the $\|.\|_2$ norm is stronger than the $\|.\|_p$ norm for $p \geq 2$).

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 $V(r) := \mathrm{e}^{-r} + r - 1$. $\diamond$

Example 9.36. 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 $X = \ell^2(\mathbb{Z}) \cong \ell^2_{\mathbb{R}}(\mathbb{Z}) \oplus \ell^2_{\mathbb{R}}(\mathbb{Z})$. 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 discrete Schrödinger operator. Clearly the right-hand side is C^1 for $p \geq 0$ and hence there is a unique local solution by Theorem 9.33. Please note that even though X is a complex Banach space we consider it as a real Banach space

Moreover, for a solution we have

$$\frac{d}{dt}\|u(t)\|_2^2 = 2\mathrm{Re}\langle \dot{u}(t), u(t)\rangle = 2\mathrm{Im}\left(\langle Hu, u\rangle \pm \langle |u(t)|^{2p}u(t), u(t)\rangle\right) = 0$$

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 $H \in \mathscr{L}(X)$. $\diamond$

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* 9.18. *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 1.59.)

Problem 9.19. *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* $V_\alpha(r)$ *has a maximum at* $r = -\frac{1}{\alpha}$. *Now use conservation of energy to conclude that the solution cannot escape the region* $|r| < \frac{1}{|\alpha|}$.*)*

9.7. Bifurcation theory

One of the most basic tasks is finding the zeros of a given function $F \in C^k(U, Y)$, where $U \subseteq X$ is open 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 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

$$F(\mu, x) = 0, \tag{9.53}$$

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

$$\partial_x F(\mu_0, x_0) \tag{9.54}$$

is not invertible at some point x_0 satisfying the equation $F(\mu_0, x_0) = 0$.

Example 9.37. 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 $f(\mu, x) = x^3 - \mu = 0$, $\partial_x f(\mu, x) = 3x^2 = 0$ 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 7.2) continuous linear projections

$$P : X = \operatorname{Ker}(A) \dotplus X_0 \to \operatorname{Ker}(A), \qquad Q : X = X_1 \dotplus \operatorname{Ran}(A) \to X_1. \tag{9.55}$$

Now split our equation into a system of two equations according to the above splitting of the underlying Banach space:

$$F(\mu, x) = 0 \qquad \Leftrightarrow \quad F_1(\mu, u, v) = 0,\ F_2(\mu, u, v) = 0, \tag{9.56}$$

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 (7.23))

$$\begin{aligned} &\partial_u F_1(\mu_0, 0, 0) = 0, \qquad &&\partial_v F_1(\mu_0, 0, 0) = 0,\\ &\partial_u F_2(\mu_0, 0, 0) = 0, \qquad &&\partial_v F_2(\mu_0, 0, 0) = A_0. \end{aligned} \tag{9.57}$$

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

$$F_2(\mu, u, \psi(\mu, u)) = 0, \qquad (\mu, u) \in U. \tag{9.58}$$

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 the original system to the finite dimensional system

$$\tilde{F}_1(\mu, u) = F_1(\mu, u, \psi(\mu, u)) = 0. \tag{9.59}$$

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

$$\partial_u \tilde{F}_1(\mu_0, 0) = \partial_u F_1(\mu_0, 0, 0) + \partial_v F_1(\mu_0, 0, 0) \partial_u \psi(\mu_0, 0) = 0. \tag{9.60}$$

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

$$\dim \mathrm{Ker}(A) = \dim \mathrm{Coker}(A) = 1 \tag{9.61}$$

such that we actually have a problem in $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

Explicitly, let u_0 span $\mathrm{Ker}(A)$ and let u_1 span X_1. Then we can write

$$\tilde{F}_1(\mu, \lambda u_0) = f(\mu, \lambda) u_1, \tag{9.62}$$

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

$$\partial_\lambda f(\mu_0, 0) u_1 = \partial_u \tilde{F}_1(\mu_0, 0) u_0 = 0. \tag{9.63}$$

It remains to investigate f. To split off the trivial solution it suggests itself to write

$$f(\mu, \lambda) = \lambda\, g(\mu, \lambda) \tag{9.64}$$

We already have

$$g(\mu_0, 0) = \partial_\lambda f(\mu_0, 0) = 0 \tag{9.65}$$

and hence if

$$0 \neq \partial_\mu g(\mu_0, 0) = \partial_\mu \partial_\lambda f(\mu_0, 0) \neq 0 \tag{9.66}$$

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

$$\begin{aligned} \partial_\lambda^2 f(\mu_0, 0) u_1 &= \partial_\lambda^2 \tilde{F}_1(\mu_0, \lambda u_0)\Big|_{\lambda=0} = \partial_\lambda^2 F_1(\mu_0, \lambda u_0, \psi(\mu_0, \lambda u_0))\Big|_{\lambda=0} \\ &= \partial_u^2 F_1(\mu_0, 0, 0)(u_0, u_0) = Q \partial_x^2 F(\mu_0, 0)(u_0, u_0) \end{aligned}$$

(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 9.36 (Crandall–Rabinowitz)**.** *Let X be a Banach space. Assume $F \in C^2(\mathbb{R} \times X, X)$ with $F(\mu, 0) = 0$ 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*

$$\partial_\mu \partial_x F(\mu_0, 0) u_0 \notin \mathrm{Ran}(\partial_x F(\mu_0, 0)) \tag{9.67}$$

there are open neighborhoods $I \subseteq \mathbb{R}$ of 0, $J \subseteq \mathbb{R}$ of μ_0, and $U \subseteq \mathrm{span}\{u_0\}$ of 0 plus corresponding functions $\mu \in C^1(I, J)$ and $\psi \in C^2(J \times U, X_0)$ such *that every nontrivial solution of* $F(\mu, x) = 0$ *in a neighborhood of* $(\mu_0, 0)$ *is given by*

$$x(\lambda) = \lambda u_0 + \psi(\mu(\lambda), \lambda u_0). \tag{9.68}$$

Moreover,

$$\mu(\lambda) = \mu_0 - \frac{\ell_1(\partial_x^2 F(\mu_0, 0)(u_0, u_0))}{2\ell_1(\partial_\mu \partial_x F(\mu_0, 0)u_0)}\lambda + o(\lambda), \quad x(\lambda) = \lambda u_0 + o(\lambda). \tag{9.69}$$

where ℓ_1 *is any nontrivial functional which vanishes on* $\mathrm{Ran}(\partial_x F(\mu_0, 0))$.

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 9.38. 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 (with respect to $\phi = \phi_r + \mathrm{i}\phi_i \cong (\phi_r, \phi_i)$)

$$\partial_\phi F(\omega, \phi)u = (H - \omega)u \pm 2p|\phi|^{2(p-1)} \begin{pmatrix} \phi_r^2 & \phi_r\phi_i \\ \phi_r\phi_i & \phi_i^2 \end{pmatrix} u \pm |\phi|^{2p}u$$

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

$$\partial_\omega \partial_\phi F(\omega_0, 0) u_0 = -u_0 \notin \mathrm{Ran}(H - \omega_0) = \mathrm{Ker}(H - \omega_0)^\perp$$

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 a 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 7.16). Hence every point in $\sigma(H) \setminus [-2, 2]$ will be an isolated eigenvalue. $\diamond$

Problem 9.20. *Show that if* $F(\mu, -x) = -F(\mu, x)$*, then* $\psi(\mu, -u) = -\psi(\mu, u)$ *and* $\mu(-\lambda) = \mu(\lambda)$*.*

Chapter 10

The Brouwer mapping degree

10.1. 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

$$f(x) = 0, \quad x \in U. \tag{10.1}$$

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

$$n(\gamma, z_0) := \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{dz}{z - z_0} \in \mathbb{Z}. \tag{10.2}$$

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 $0 \notin f(\gamma)$)

$$n(f(\gamma), 0) = \frac{1}{2\pi \mathrm{i}} \int_\gamma \frac{f'(z)}{f(z)} dz = \sum_k n(\gamma, z_k)\alpha_k, \tag{10.3}$$

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 $f : \gamma \to \mathbb{C}\backslash\{0\}$ and $g : \gamma \to \mathbb{C}\backslash\{0\}$ 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** $H : [0, 1] \times \gamma \to \mathbb{C}\backslash\{0\}$ 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.

10.2. 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 (nonempty) 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 $df(x) = (\partial_{x_j} f_i(x))_{1 \le i,j \le n}$ and the Jacobi determinant of f at $x \in U$ is

$$J_f(x) := \det df(x). \tag{10.4}$$

The set of **regular values** is

$$\mathrm{RV}(f) := \{y \in \mathbb{R}^n | \forall x \in f^{-1}(y) : J_f(x) \neq 0\}. \tag{10.5}$$

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

$$\bar{C}^k(U, \mathbb{R}^n) := C^k(U, \mathbb{R}^n) \cap C(\overline{U}, \mathbb{R}^n) \tag{10.6}$$

and regard them as subspaces of the Banach space $C(\overline{U}, \mathbb{R}^n)$ (cf. Section 9.2). 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

$$\bar{C}^k_y(\overline{U}, \mathbb{R}^n) := \{f \in \bar{C}^k(U, \mathbb{R}^n) | y \notin f(\partial U)\} \tag{10.7}$$

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,

$$\mathrm{dist}(y, f(\partial U)) = \min_{x \in \partial U} |y - f(x)| \tag{10.8}$$

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 10.1. *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, $1 \le i \le N$, 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 $N = 1$ and the rest follows from induction. For (ii) use $N = 1$ and $U_1 = \emptyset$ in (i). For (iii) note that $H(t, x) = (1-t)f(x) + t\,g(x)$ satisfies $|H(t, x) - y| \ge \operatorname{dist}(y, f(\partial U)) - |f(x) - g(x)|$ 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 10.2. *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) 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 $|H(t) - y| \ge |f - y| - |g - f| > 0$, $H(t) = (1-t)f + t\,g$. 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.23 (vi)).

Moreover, note that this result also shows why $\deg(f, U, y)$ cannot be defined meaningful for $y \in f(\partial U)$. 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 $y \notin \mathrm{CV}(f)$. 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

$$\deg(f, U, 0) = \sum_{i=1}^N \deg(f, U(x^i), 0). \tag{10.9}$$

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

$$f(x) = df(0)x + |x| r(x), \qquad r \in C(B_\delta(0), \mathbb{R}^n), \quad r(0) = 0. \tag{10.10}$$

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

$$\deg(f, U, 0) = \sum_{i=1}^N \deg(df(x^i), B_\delta(0), 0) \tag{10.11}$$

and it remains to compute the degree of a nonsingular matrix. To this end we need the following lemma.

Lemma 10.3. *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 10.4. *Suppose $f \in \bar{C}^1_y(U, \mathbb{R}^n)$ and $y \notin \mathrm{CV}(f)$, then a degree satisfying (D1)–(D4) satisfies*

$$\deg(f, U, y) = \sum_{x \in f^{-1}(\{y\})} \operatorname{sign} J_f(x), \tag{10.12}$$

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 $f_1(x) = (1 - g(x_1)h(x_2)\cdots h(x_n), x_2, \dots, x_n)$ and $f_2(x) = (1, x_2, \dots, x_n)$. Clearly $f_1^{-1}(\{0\}) = \{x^1, x^2\}$ with $x^1 = 0$, $x^2 = (2, 0, \dots, 0)$ 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 (10.12). Once we have shown that regular values are dense, it will follow that the degree is uniquely determined by (10.12) since the remaining values follow from point (iii) of Theorem 10.1. On the other hand, we don't even know whether a degree exists since it is unclear whether (10.12) satisfies (D4). Hence we need to show that (10.12) can be extended to $f \in \bar{C}_y(U, \mathbb{R}^n)$ and that this extension satisfies our requirements (D1)–(D4).

10.3. Extension of the determinant formula

Our present objective is to show that the determinant formula (10.12) can be extended to all $f \in \bar{C}_y(U, \mathbb{R}^n)$, where $U \subset \mathbb{R}^n$ is open and bounded. As a preparation we prove that the set of regular values is dense. This is a consequence of a special case of Sard's theorem which says that $\mathrm{CV}(f)$ has zero measure.

Lemma 10.5 (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

$$|f(x) - f(\tilde{x}) - df(\tilde{x})(x - \tilde{x})| \le \int_0^1 |df(\tilde{x} + t(x - \tilde{x})) - df(\tilde{x})||\tilde{x} - x| dt \le \frac{\varepsilon\rho}{N} \tag{10.13}$$

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 (10.13) 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 10.2, $\deg(f, U, y)$ 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 10.1. 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 (10.12) 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 10.6 (Heinz)**.** *Suppose $f \in \bar{C}^1_y(U, \mathbb{R}^n)$ and $y \notin \mathrm{CV}(f)$. Then the degree defined as in* (10.12) *satisfies*

$$\deg(f, U, y) = \int_U \phi_\varepsilon(f(x) - y) J_f(x) d^n x \tag{10.14}$$

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 $f^{-1}(\{y\}) = \{x^i\}_{1\le i\le N}$, the inverse function theorem ensures that 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 with $f|_{U(x^i)}$ a bijection and $J_f(x)$ 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 show that the integral representation is independent of ε as long as $\varepsilon < \operatorname{dist}(y, f(\partial U))$. To this end we will rewrite the difference 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 10.7. *Suppose $f \in C^2(U,\mathbb{R}^n)$ and $u \in C^1(\mathbb{R}^n,\mathbb{R}^n)$, then*

$$(\operatorname{div} u)(f)J_f = \operatorname{div} D_f(u), \tag{10.15}$$

where $D_f(u)_j$ is the determinant of the matrix obtained from df by replacing the j-th column by $u(f)$. Here $\operatorname{div} u = \sum_{j=1}^n \partial_j u_j$ is the ***divergence*** *of a vector field.*

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 10.8. *Let $U \subset \mathbb{R}^n$ be (nonempty) bounded and open. 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*

$$\deg(f,U,y) = \sum_{x \in \tilde{f}^{-1}(y)} \operatorname{sign} J_{\tilde{f}}(x), \tag{10.16}$$

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 $\|f - \tilde{f}\|_\infty < \operatorname{dist}(y, f(\partial U))$, such that $y \in \mathrm{RV}(\tilde{f})$.

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 $t \mapsto I_\varepsilon(f + t\,g,U,y)$, $t \in [0,1]$, is continuous and it is integer valued (since it is equal to our determinant formula) on a dense set. Consequently it must be constant and we can extend I_ε to a function $\bar{I}_\varepsilon$ on all of $B_\delta(f)$ and hence on all of $\bar{C}_y(U,\mathbb{R}^n)$. 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 $\varepsilon < \min_{t\in[0,1]} \operatorname{dist}(H(t)(\partial U),y)$ 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 10.2. 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 $|h(t)| \ge |g| - t|f-g| > 0$ 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 (10.12) 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 10.9. *Suppose $U \subset \mathbb{R}^n$ 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 an $x \in \partial U$ and a $\lambda \ne 0$ such that $f(x) = \lambda x$.*

Proof. By Theorem 11.10 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 $\deg(f,U,0) \ne -1$, 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 $H(t,x) = (1-t)f(x) + tx$. $\square$

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 10.3. 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. $\diamond$

Another illustration 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 10.2).

One more useful observation is that odd functions have odd degree:

Theorem 10.10 (Borsuk)**.** *Let $0 \in U \subset \mathbb{R}^n$ be open, bounded and symmetric with respect to the origin (i.e., $U = -U$). Let $f \in \bar{C}_0(U, \mathbb{R}^n)$ be odd (i.e., $f(-x) = -f(x)$). Then $\deg(f, U, 0)$ is odd.*

Proof. If $f \in \bar{C}_0^1(U)$ 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 $f \in \bar{C}_0(U, \mathbb{R}^n)$ we can choose an approximating $f_0 \in \bar{C}_0^1(U, \mathbb{R}^n)$ and replacing f_0 by its odd part $\frac{1}{2}(f_0(x) - f_0(-x))$ 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 $\varphi(t) \neq 0$ for $t \neq 0$. 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 $x \in U_1 := \{x \in U | x_1 \neq 0\}$ 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 and $df_2(x) = df_1(x)$ by our assumption $\phi'(0) = 0$. After n steps we reach $V_n = U \setminus \{0\}$ and f_n is the approximation we are looking for. $\square$

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 10.11 (Borsuk–Ulam). *Let $0 \in U \subset \mathbb{R}^n$ 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 $\mathbb{R}^m \times \{0\} \subset \mathbb{R}^n$. $\square$

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 10.12 (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 $H(1) = f(\frac{1}{2}x) - f(-\frac{1}{2}x)$ is odd. But then we also have $\deg(f, B_r(0), y) \neq 0$ for $y \in B_\varepsilon(0)$ and thus $B_\varepsilon(0) \subseteq f(B_r(0))$. $\square$

An easy consequence worth while noting is the topological invariance of dimension:

Corollary 10.13. *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 10.1. *Suppose $U = (a, b) \subset \mathbb{R}^1$. 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* 10.2. *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 10.3. *Show that in Borsuk's theorem the condition f is odd can be replaced by $f(x) \neq t f(-x)$ 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 $\operatorname{sign}(f(x)) := \frac{f(x)}{|f(x)|}$).*

10.4. 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 10.14 (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 11.10 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 $x = \tilde{f}(x) \in \mathrm{conv}(f(K)) \subseteq K$. □

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 10.15 (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., **[29]** for details).

10.5. 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 10.16 (Kakutani)**.** *Suppose K is a compact convex subset of $\mathbb{R}^n$ and $f : K \to \mathrm{CS}(K)$. If the set*

$$\Gamma := \{(x,y)|y \in f(x)\} \subseteq K^2 \tag{10.17}$$

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

$$K = \mathrm{conv}(v_1, \dots, v_m), \qquad m \le n+1,$$

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., $\lambda_i \ge 0$, $\sum_{i=1}^m \lambda_i(x) = 1$ 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 $v_{\sigma_1}, \dots, v_{\sigma_m}$ of the vertices you consider the simplex $\mathrm{conv}(v_{\sigma_1}, \frac{1}{2}(v_{\sigma_1}+v_{\sigma_2}), \dots, \frac{1}{m}(v_{\sigma_1}+\cdots+v_{\sigma_m}))$. 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 (see Figure 10.1). 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 by $\mathrm{conv}(v_1^k, \dots, v_m^k)$ such that

$$x^k = \sum_{i=1}^m \lambda_i^k v_i^k = \sum_{i=1}^m \lambda_i^k y_i^k, \quad y_i^k = f^k(v_i^k). \tag{10.18}$$

Since $(x^k, \lambda_1^k, \dots, \lambda_m^k, y_1^k, \dots, y_m^k) \in K \times [0,1]^m \times K^m$ we can assume that this sequence converges to some limit $(x^0, \lambda_1^0, \dots, \lambda_m^0, y_1^0, \dots, y_m^0)$ after passing to a subsequence. Since the subsimplices shrink to a point, this implies $v_i^k \to x^0$

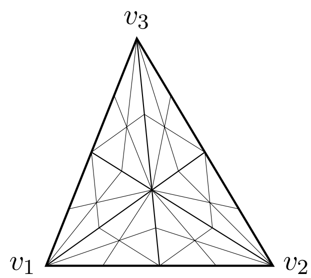


Figure 10.1. Iterated barycentric subdivision of a simplex

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 (10.18) 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

$$R_i(\varphi) \in \mathbb{R}, \qquad \varphi = (\varphi_1, \dots, \varphi_n) \in \Phi = \bigtimes_{i=1}^{n} \Phi_i \tag{10.19}$$

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

$$s(\varphi) = \prod_{i=1}^{n} s_i(\varphi), \qquad s_i(\varphi) = s_i^{k_i},\ \varphi = (k_1, \dots, k_n) \in \Phi \tag{10.20}$$

(assuming all players make their choice independently) and the expected payoff for player i is

$$R_i(s) = \sum_{\varphi \in \Phi} s(\varphi) R_i(\varphi). \tag{10.21}$$

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

$$R_i(s \setminus \overline{s}_i) = \max_{\tilde{s}_i \in S_i} R_i(s \setminus \tilde{s}_i) \tag{10.22}$$

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

$$R_i(s) = \sum_{k=1}^{m_i} s_i^k R_i(s/k) \tag{10.23}$$

we have $\overline{s}_i \in B_i(s)$ if and only if $\overline{s}_i^k = 0$ whenever $R_i(s \setminus k) < \max_{1 \le l \le m_i} R_i(s \setminus l)$. 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,

$$\overline{s} \in B(\overline{s}). \tag{10.24}$$

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

$$\begin{array}{c|cc} R_1 & d_2 & c_2 \\ \hline d_1 & 0 & 2 \\ c_1 & -1 & 1 \end{array} \qquad \begin{array}{c|cc} R_2 & d_2 & c_2 \\ \hline d_1 & 0 & -1 \\ c_1 & 2 & 1 \end{array} \tag{10.25}$$

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 10.17 (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 (10.22) for s^m, $\overline{s}^m$

$$R_i(s^m \setminus \tilde{s}_i) \le R_i(s^m \setminus \overline{s}_i^m), \qquad \tilde{s}_i \in S_i,\ 1 \le i \le n,$$

implies (10.22) for the limits s, $\overline{s}$

$$R_i(s \setminus \tilde{s}_i) \le R_i(s \setminus \overline{s}_i), \qquad \tilde{s}_i \in S_i,\ 1 \le i \le n,$$

by continuity of $R_i(s)$. □

10.6. 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 10.18 (Reduction property)**.** *Let $U \subset \mathbb{R}^n$ 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*

$$\deg(\mathbb{I} + f, U, y) = \deg(\mathbb{I} + f_m, U_m, y), \tag{10.26}$$

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 10.2 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 10.19 (Product formula)**.** *Let $U \subset \mathbb{R}^n$ 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*

$$\deg(g \circ f, U, y) = \sum_j \deg(f, U, G_j) \deg(g, G_j, y), \tag{10.27}$$

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, $g^{-1}(\{y\}) \subset \bigcup_j G_j$. 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 $g^{-1}(\{y\}) \subset \bigcup_{j=1}^m G_j$. 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 $y\in \mathrm{CV}(g\circ f)$ 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 10.1 (iii) and hence $\deg(g,\tilde L_l,y)=\deg(g,L_l,y)$ by Theorem 10.1 (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. □

10.7. The Jordan curve theorem

In this section we want to show how the product formula (10.27) for the Brouwer degree can be used to prove the famous **Jordan curve theorem**[1] 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 10.20. *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 11.10 we can extend both to $\mathbb{R}^n$. Then Theorem 10.1 (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 10.1 (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. □

[1] Camille Jordan (1838–1922), French mathematician

Chapter 11

The Leray–Schauder mapping degree

11.1. 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

$$\deg(f, U, y) := \deg(\phi \circ f \circ \phi^{-1}, \phi(U), \phi(y)) \tag{11.1}$$

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 in 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

$$J_{A\circ\tilde{f}^*\circ A^{-1}}(Ay^*) = \det(A) J_{\tilde{f}^*}(y^*) \det(A^{-1}) = J_{\tilde{f}^*}(y^*) \tag{11.2}$$

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 11.1. 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.

11.2. 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 $\mathcal{K}(U, Y)$ 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 $\mathcal{K}(U, Y) \subseteq C_b(U, Y)$ if U is bounded (recall that compact sets are automatically bounded).

If K is compact, then $\mathcal{K}(K, Y) = C(K, Y)$ (since the continuous image of a compact set is compact) and if $\dim(Y) < \infty$, then $\mathcal{F}(U, Y) = \mathcal{K}(U, Y)$. In particular, if $U \subset \mathbb{R}^n$ is bounded, then $\mathcal{F}(\overline{U}, \mathbb{R}^n) = \mathcal{K}(\overline{U}, \mathbb{R}^n) = C(\overline{U}, \mathbb{R}^n)$.

Example 11.2. 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 $\{x_j\}_{j\in\mathbb{N}}$, 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 11.1. *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 $|P_\varepsilon(x) - x| \le \varepsilon$ 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 $\mathrm{supp}(\phi_i) \subset B_\varepsilon(x_i)$ 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

$$|P_\varepsilon(x) - x| = |\sum_{i=1}^n \phi_i(x) x - \sum_{i=1}^n \phi_i(x) x_i| \le \sum_{i=1}^n \phi_i(x) |x - x_i| \le \varepsilon. \qquad \square$$

This lemma enables us to prove the following important result.

Theorem 11.2. *Let $U \subset X$ be bounded, then the closure of $\mathcal{F}(U,Y)$ in $C_b(U,Y)$ is $\mathcal{K}(U,Y)$.*

Proof. Suppose $F_N \in \mathcal{K}(U,Y)$ converges to F. If $F \notin \mathcal{K}(U,Y)$ then we can find a sequence $x_n \in U$ such that $|F(x_n) - F(x_m)| \ge \rho > 0$ for $n \ne m$. If N is so large that $|F - F_N| \le \rho/4$, then

$$\begin{aligned} |F_N(x_n) - F_N(x_m)| &\ge |F(x_n) - F(x_m)| - |F_N(x_n) - F(x_n)| \\ &\qquad - |F_N(x_m) - F(x_m)| \\ &\ge \rho - 2\frac{\rho}{4} = \frac{\rho}{2} \end{aligned}$$

This contradiction shows $\overline{\mathcal{F}(U,Y)} \subseteq \mathcal{K}(U,Y)$. Conversely, let $F \in \mathcal{K}(U,Y)$, set $K := \overline{F(U)}$ and choose P_ε according to Lemma 11.1. Then $F_\varepsilon = P_\varepsilon \circ F \in \mathcal{F}(U,Y)$ converges to F. Hence $\mathcal{K}(U,Y) \subseteq \overline{\mathcal{F}(U,Y)}$ and we are done. $\square$

Finally, let us show some interesting properties of mappings $\mathbb{I} + F$, where $F \in \mathcal{K}(U,Y)$.

Lemma 11.3. *Let $\overline{U} \subseteq X$ be bounded and closed. Suppose $F \in \mathcal{K}(\overline{U}, X)$, 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 $F \in \mathcal{K}(\overline{U}, Y)$ and $G \in C(Y, Z)$, then $G \circ F \in \mathcal{K}(\overline{U}, Z)$ and similarly, if $G \in C_b(\overline{V}, \overline{U})$, then $F \circ G \in \mathcal{K}(\overline{V}, Y)$.

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.

11.3. The Leray–Schauder mapping degree

For an open set $U \subset X$ of a Banach space X we set

$$\bar{\mathcal{K}}_y(U, X) := \{F \in \mathcal{K}(\overline{U}, X) | y \notin (\mathbb{I} + F)(\partial U)\} \tag{11.3}$$

and $\bar{\mathcal{F}}_y(U, X) := \{F \in \mathcal{F}(\overline{U}, X) | y \notin (\mathbb{I} + F)(\partial U)\}$. Note that for $F \in \bar{\mathcal{K}}_y(U, X)$ we have $\mathrm{dist}(y, (\mathbb{I} + F)(\partial U)) > 0$ since $\mathbb{I} + F$ maps closed sets to closed sets (cf. Problem B.52).

Abbreviate $\rho := \mathrm{dist}(y, (\mathbb{I} + F)(\partial U))$ 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

$$\deg(\mathbb{I} + F, U, y) := \deg(\mathbb{I} + F_1, U_1, y) \tag{11.4}$$

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

$$\deg(\mathbb{I} + F_i, U_0, y) = \deg(\mathbb{I} + F_i, U_i, y), \quad i = 1, 2, \tag{11.5}$$

by the reduction property. Moreover, set $H(t) = \mathbb{I} + (1-t)F_1 + t\,F_2$ implying $H(t) \in \bar{\mathcal{K}}_y(U_0, X_0)$, $t \in [0, 1]$, since $|H(t) - (\mathbb{I} + F)| < \rho$ for $t \in [0, 1]$. Hence homotopy invariance

$$\deg(\mathbb{I} + F_1, U_0, y) = \deg(\mathbb{I} + F_2, U_0, y) \tag{11.6}$$

shows that (11.4) is independent of F_1, X_1.

Theorem 11.4. *Let U be a bounded open subset of a (real) Banach space X and let $F \in \bar{\mathcal{K}}_y(U, X)$, $y \in X$. 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* $H(t) \in \bar{\mathcal{K}}_{y(t)}(U,X)$, $t \in [0,1]$, *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 11.2 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 10.2. □

In addition, Theorem 10.1 and Theorem 10.2 hold for the new situation as well (no changes are needed in the proofs).

Theorem 11.5. *Let* U *be a bounded open subset of a (real) Banach space* X *and let* $F \in \bar{\mathcal{K}}_y(U,X)$, $y \in X$. *If* $F, G \in \bar{\mathcal{K}}_y(U,X)$, *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* $\bar{\mathcal{K}}_y(U,X)$.

(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 11.6 (Borsuk)**.** *Let* $U \subseteq X$ *be open, bounded and symmetric with respect to the origin (i.e.,* $U = -U$*). Let* $F \in \bar{\mathcal{K}}_0(U,X)$ *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 11.7 (Brouwer)**.** *Let* $U \subseteq X$ *be open and let* $F \in \mathcal{K}(U,X)$ *be compact with* $\mathbb{I} + F$ *locally injective. Then* $\mathbb{I} + F$ *is also open.*

11.4. 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 11.8 (Schaefer fixed point or Leray–Schauder principle)**.** *Let X be a Banach space. Suppose $F \in \mathcal{K}(X, X)$ 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 $H(t) \in \bar{\mathcal{K}}_0(B_\rho(0), X)$ due to the a priori bound. □

Now we can extend the Brouwer fixed point theorem to infinite dimensional spaces as well.

Theorem 11.9 (Schauder fixed point)**.** *Let K be a closed, convex, and bounded subset of a Banach space X. If $F \in \mathcal{K}(K, K)$, 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 11.10 below we can find a continuous retraction $R : X \to K$ (i.e., $R(x) = x$ for $x \in K$) and consider $\tilde{F} = F \circ R \in \mathcal{K}(\overline{B_\rho(0)}, \overline{B_\rho(0)})$. Now either $t\tilde{F}$ has a fixed point on the boundary $\partial B_\rho(0)$ or the compact homotopy $H(t, x) := -t\tilde{F}(x)$ satisfies $0 \notin (\mathbb{I} - t\tilde{F})(\partial B_\rho(0))$ 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 11.10. *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 refinement $\{O_\lambda\}_{\lambda \in \Lambda}$ of this cover (see Lemma B.31) and define

$$\phi_\lambda(x) := \frac{\operatorname{dist}(x, X \setminus O_\lambda)}{\sum_{\mu \in \Lambda} \operatorname{dist}(x, X \setminus O_\mu)}.$$

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 $\operatorname{dist}(x_\lambda, O_\lambda) \le 2\operatorname{dist}(K, O_\lambda)$. 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 $|\bar{F}(x) - F(x_0)| \le \varepsilon$ for all $x \in X$ with $|x - x_0| < \delta$. Suppose $x \notin K$, then $|\bar{F}(x) - F(x_0)| \le \sum_{\lambda\in\Lambda} \phi_\lambda(x)|F(x_\lambda) - F(x_0)|$. By our construction, x_λ should be close to x for all λ with $x \in \overline{O_\lambda}$ since x is close to K. In fact, if $x \in \overline{O_\lambda}$ we have

$$\begin{aligned} |x - x_\lambda| &\le \operatorname{dist}(x_\lambda, O_\lambda) + \operatorname{diam}(O_\lambda) \\ &\le 2\operatorname{dist}(K, O_\lambda) + \operatorname{diam}(O_\lambda), \end{aligned}$$

where $\operatorname{diam}(O_\lambda) := \sup_{x,y\in O_\lambda} |x - y|$. 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 $O_\lambda \subset B_{\rho(\tilde{x})}(\tilde{x})$ and hence $\operatorname{diam}(O_\lambda) \le 2\rho(\tilde{x}) \le \operatorname{dist}(K, B_{\rho(\tilde{x})}(\tilde{x})) \le \operatorname{dist}(K, O_\lambda)$. Putting it all together implies that we have $|x - x_\lambda| \le 3\operatorname{dist}(K, O_\lambda) \le 3|x_0 - x|$ whenever $x \in \overline{O_\lambda}$ and thus

$$|x_0 - x_\lambda| \le |x_0 - x| + |x - x_\lambda| \le 4|x_0 - x| \le 4\delta$$

as expected. By our choice of δ we have $|F(x_\lambda) - F(x_0)| \le \varepsilon$ for all λ with $\phi_\lambda(x) \ne 0$. Hence $|F(x) - F(x_0)| \le \varepsilon$ whenever $|x - x_0| \le \delta$ and we are done. □

Example 11.3. 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

$$\begin{aligned} |F(x)(t) - F(y)(t)| &\le \int_0^1 \mathrm{e}^{-ts} |\cos(\lambda x(s)) - \cos(\lambda y(s))| ds \\ &\le \int_0^1 \mathrm{e}^{-ts} \lambda |x(s) - y(s)| ds \le \lambda \|x - y\|_\infty. \end{aligned}$$

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:

$$\|F(x)\|_\infty \le 1, \qquad \|F(x)'\|_\infty \le 1.$$

Hence the Arzelà–Ascoli theorem (Theorem B.40) implies that the image of F is a compact subset of the unit ball and hence $F \in \mathcal{K}(\bar{B}_1(0), \bar{B}_1(0))$. 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 11.11. *Let $U \subset X$ be open and bounded and let $F \in \mathcal{K}(\overline{U}, X)$. Suppose there is an $x_0 \in U$ such that*

$$F(x) - x_0 \neq \alpha(x - x_0), \qquad x \in \partial U,\ \alpha \in (1, \infty). \tag{11.7}$$

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 11.12. *Let $F \in \mathcal{K}(\bar{B}_\rho(0), X)$. 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 (11.7) 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 (11.7) 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 $|F(x) - x|^2 = |F(x)|^2 - 2\langle F(x), x\rangle + |x|^2$. □

11.5. 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 11.13. *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*

$$\begin{array}{rccl} F: & C(I, \mathbb{R}^n) & \to & C(I, \mathbb{R}^n) \\ & x(t) & \mapsto & F(x)(t) = \int_a^{\tau(t)} f(t, s, x(s)) ds \end{array} \tag{11.8}$$

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 $|f(t,s,x) - f(t,s,y)| \le \varepsilon_1$ for $|x-y| < \delta$. But this implies

$$\begin{aligned}\|F(x) - F(x_0)\|_\infty &\le \sup_{t\in I} \int_a^{\tau(t)} |f(t,s,x(s)) - f(t,s,x_0(s))| ds \\ &\le \sup_{t\in I} (b-a)\varepsilon_1 = \varepsilon,\end{aligned}$$

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.40). Thus F is compact. □

As a first application we use this result to show existence of solutions to integral equations.

Theorem 11.14. *Let F be as in the previous lemma. Then the integral equation*

$$x - \lambda F(x) = y, \qquad \lambda \in \mathbb{R},\ y \in C(I,\mathbb{R}^n) \tag{11.9}$$

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 11.15 (Peano[1]). *Consider the initial value problem*

$$\dot{x} = f(t,x), \quad x(t_0) = x_0, \tag{11.10}$$

where $f \in C(I \times U, \mathbb{R}^n)$ and $I \subset \mathbb{R}$ is an interval containing t_0. Then (11.10) 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 $t \ge 0$ since $t \to -t$ amounts to $f \to -f$.

Now observe, that (11.10) 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 $M(\varepsilon,\rho) \le \tilde{M}(\varepsilon)(1+\rho)$. Then

$$|x(t)| \le \int_0^t |f(s,x(s))|ds \le \tilde{M}(\varepsilon) \int_0^t (1+|x(s)|)ds$$

implies $|x(t)| \le \exp(\tilde{M}(\varepsilon)\varepsilon)$ by Gronwall's inequality (Problem 11.1). 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 through 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

$$\rho \partial_t v = \eta \Delta v - \rho(v \cdot \nabla)v - \nabla p + K, \tag{11.11}$$

[1] Giuseppe Peano (1858–1932), Italian mathematician

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

$$0 = \eta\Delta v - \rho(v\cdot\nabla)v - \nabla p + K. \tag{11.12}$$

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 $H^1_0(U,\mathbb{R}^3)$ with scalar product

$$\langle u, v\rangle = \sum_{i,j=1}^{3} \int_U (\partial_j u_i)(\partial_j v_i)dx. \tag{11.13}$$

Recall that by the Poincaré inequality (Theorem 7.38 from [**40**]) the corresponding norm is equivalent to the usual one. In order to take care of the incompressibility condition we will choose

$$\mathcal{X} := \{v \in H^1_0(U,\mathbb{R}^3) | \nabla\cdot v = 0\}. \tag{11.14}$$

as our configuration space (check that this is a closed subspace of $H^1_0(U,\mathbb{R}^3)$).

Now we multiply (11.12) by $w \in \mathcal{X}$ and integrate over U

$$\begin{aligned}\int_U \Big(\eta\Delta v - \rho(v\cdot\nabla)v - K\Big)\cdot w\, d^3x &= \int_U (\nabla p)\cdot w\, d^3x \\ &= \int_U p(\nabla w) d^3x = 0,\end{aligned} \tag{11.15}$$

where we have used integration by parts (Lemma 7.11 from [**40**] (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

$$\eta\langle v, w\rangle - a(v,v,w) - \int_U K\cdot w\, d^3x = 0, \quad \text{for all } w \in \mathcal{X}, \tag{11.16}$$

where

$$a(u,v,w) := \sum_{j,k=1}^{3} \int_U u_k v_j (\partial_k w_j)\, d^3x. \tag{11.17}$$

In other words, (11.16) represents a necessary solubility condition for the Navier–Stokes equations and a solution of (11.16) 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 (11.15). 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 $p \in L^2(U,\mathbb{R})$ 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

$$a(v,v,v) = \sum_{j,k} \int_U v_k v_j (\partial_k v_j)\, d^3x = \frac{1}{2} \sum_{j,k} \int_U v_k \partial_k (v_j v_j)\, d^3x$$
$$= -\frac{1}{2} \sum_{j,k} \int_U (v_j v_j) \partial_k v_k\, d^3x = 0, \qquad v \in \mathcal{X}. \tag{11.18}$$

We proceed by studying (11.16). 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

$$\int_U K \cdot w\, d^3x = \langle \tilde{K}, w\rangle, \qquad w \in \mathcal{X}. \tag{11.19}$$

Moreover, applying the Cauchy–Schwarz inequality twice to each summand in $a(u,v,w)$ we see

$$|a(u,v,w)| \le \sum_{j,k} \Big(\int_U (u_k v_j)^2 dx \Big)^{1/2} \Big(\int_U (\partial_k w_j)^2 dx \Big)^{1/2}$$
$$\le \|w\| \sum_{j,k} \Big(\int_U (u_k)^4 dx \Big)^{1/4} \Big(\int_U (v_j)^4 dx \Big)^{1/4} = \|u\|_4 \|v\|_4 \|w\|. \tag{11.20}$$

Since by the Gagliardo–Nierenberg–Sobolev inequality (Theorem 7.26 from [**40**]) there is a continuous embedding $H^1(U,\mathbb{R}^3) \hookrightarrow L^4(U,\mathbb{R}^3)$ (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

$$a(u,v,w) = \langle B(u,v), w\rangle, \qquad w \in \mathcal{X}. \tag{11.21}$$

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

$$\langle \eta v - B(v,v) - \tilde{K}, w\rangle = 0, \qquad w \in \mathcal{X}, \tag{11.22}$$

and hence

$$\eta v - B(v,v) = \tilde{K}. \tag{11.23}$$

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 7.35 from [**40**]).

Motivated by this analysis we formulate the following theorem which implies existence of weak solutions and uniqueness for sufficiently small outer forces.

Theorem 11.16. *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*

$$|a(u,v,w)| \le \alpha\|u\|_Y\|v\|_Y\|w\| \tag{11.24}$$

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

$$\eta\langle v,w\rangle - a(v,v,w) = \langle \tilde{K}, w\rangle, \qquad w \in \mathcal{X}. \tag{11.25}$$

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 (11.24) 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* 11.1 (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 $\varphi(t) \le \alpha \mathrm{e}^{\int_0^t \beta(s)ds}$. (Hint: Differentiate $\log\left(\alpha + \int_0^t \beta(s)\varphi(s)ds\right)$*.)*

Chapter 12

Monotone maps

12.1. 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 12.1. *Suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuous and satisfies*

$$\lim_{|x|\to\infty} \frac{F(x)x}{|x|} = \infty. \tag{12.1}$$

Then the equation

$$F(x) = y \tag{12.2}$$

has a solution for every $y \in \mathbb{R}^n$. If F is strictly monotone

$$(F(x) - F(y))(x - y) > 0, \qquad x \neq y, \tag{12.3}$$

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 10.2. The second claim is trivial. $\square$

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

$$\langle F(x) - F(y), x - y \rangle \geq 0, \qquad x, y \in \mathfrak{H}, \tag{12.4}$$

strictly monotone if

$$\langle F(x) - F(y), x - y \rangle > 0, \qquad x \neq y \in \mathfrak{H}, \tag{12.5}$$

and finally **strongly monotone** if there is a constant $C > 0$ such that

$$\langle F(x) - F(y), x - y \rangle \geq C \|x - y\|^2, \qquad x, y \in \mathfrak{H}. \tag{12.6}$$

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**

$$\lim_{\|x\| \to \infty} \frac{\langle F(x), x \rangle}{\|x\|} = \infty. \tag{12.7}$$

(Just take $y = 0$ in the definition of strong monotonicity.) Hence the following result is not surprising.

Theorem 12.2 (Zarantonello). *Let $\mathfrak{H}$ be a Hilbert space. Suppose $F \in C(\mathfrak{H}, \mathfrak{H})$ is (globally) Lipschitz continuous and strongly monotone. Then, for each $y \in \mathfrak{H}$ the equation*

$$F(x) = y \tag{12.8}$$

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 $t \in (0, \frac{2C}{L^2})$, G is a uniform contraction and the rest follows from the uniform contraction principle. $\square$

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

$$x_{n+1} = x_n - \frac{C}{L^2}(F(x_n) - y), \quad x_0 = y, \tag{12.9}$$

converges to the solution.

Example 12.1. Let $A \in \mathscr{L}(\mathfrak{H})$ and consider $F(x) = Ax$. Then the condition

$$\langle Ax, x\rangle \ge C\|x\|^2$$

implies that A has a bounded inverse $A^{-1} : \mathfrak{H} \to \mathfrak{H}$ with $\|A^{-1}\| \le C^{-1}$ (cf. Problem 1.57). ⋄

12.2. 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:

$$a(x, y) = b(y), \qquad \text{for all } y \in \mathfrak{H}, \tag{12.10}$$

where $a : \mathfrak{H}^2 \to \mathbb{R}$ and $b : \mathfrak{H} \to \mathbb{R}$. For this equation the following result holds.

Theorem 12.3 (Nonlinear Lax–Milgram theorem)**.** *Let $\mathfrak{H}$ be a Hilbert space. 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*

$$a(x, x - y) - a(y, x - y) \ge C\|x - y\|^2 \tag{12.11}$$

and

$$|a(x, z) - a(y, z)| \le L\|z\|\|x - y\|. \tag{12.12}$$

Then there is a unique $x \in \mathfrak{H}$ such that (12.10) *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 (12.11) the map F is strongly monotone. Moreover, by (12.12) we infer

$$\|F(x) - F(y)\| = \sup_{\tilde{x} \in \mathfrak{H}, \|\tilde{x}\|=1} |\langle F(x) - F(y), \tilde{x}\rangle| \le L\|x - y\|$$

that F is Lipschitz continuous. Now apply Theorem 12.2. □

The special case where $a \in \mathscr{L}^2(\mathfrak{H}, \mathbb{R})$ is a bounded bilinear form which is strongly coercive, that is,

$$a(x,x) \geq C\|x\|^2, \qquad x \in \mathfrak{H}, \tag{12.13}$$

is usually known as (linear) Lax–Milgram theorem (Theorem 2.17).

Example 12.2. For example, let $U \subset \mathbb{R}^n$ be a bounded domain and consider the Dirichlet problem for the second order nonlinear **elliptic problem**

$$-\sum_{i,j=1}^{n} \partial_i A_{ij}(x)\partial_j u(x) + F(x, u(x)) = w(x)$$

with $A_{i,j} \in L^\infty(U, \mathbb{R})$ and $F : U \times \mathbb{R}^n \to \mathbb{R}$. If we impose Dirichlet boundary conditions, then it is natural to look for solutions from the class $\mathfrak{D} = \{u \in H^1_0(U, \mathbb{R}) | A_{ij}\partial_j u \in H^1(U, \mathbb{R}), 1 \leq i, j \leq n\}$. Multiplying this equation with a function $v \in H^1_0(U, \mathbb{R})$ and integrating over U gives the associated weak formulation

$$a(u, v) = w(v),$$

where

$$a(u, v) := \sum_{i,j=1}^{n} \int_U (A_{ij}(x)(\partial_j u(x))(\partial_i v(x)) + F(x, u(x))v(x))\, d^n x,$$

$$w(v) := \int_U w(x)v(x) d^n x.$$

Here we have assumed $w \in L^2(U, \mathbb{R})$ but, somewhat more general, $w \in H^1(U, \mathbb{R})^*$ would also suffice.

If we require

$$C := \inf_{e \in S^n, x \in U} e_i A_{ij}(x) e_j > 0$$

as well as

$$|F(x, u_1) - F(x, u_2)| \leq L|u_1 - u_2| \text{ and } (F(x, u_1) - F(x, u_2))(u_1 - u_2) \geq 0,$$

then the assumption of the nonlinear Lax–Milgram theorem are satisfied on $H^1_0(U, \mathbb{R})$. ⋄

12.3. The main theorem of monotone maps

Now we return to the investigation of $F(x) = y$ and weaken the conditions of Theorem 12.2. 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**

$$\lim_{n\to\infty} \langle F(x_n), y\rangle = \langle F(x), y\rangle, \qquad \text{for all } y \in \mathfrak{H} \tag{12.14}$$

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

$$F_n(x_n) = y_n, \qquad x_n, y_n \in \mathfrak{H}_n. \tag{12.15}$$

More precisely, let P_n be the (linear) projection onto $\mathfrak{H}_n$ and set $F_n(x_n) = P_n F(x_n)$, $y_n = P_n y$ (verify that F_n is continuous and monotone!).

Now Lemma 12.1 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

$$\langle F_n(x), x\rangle > \|y_n\| \|x\|, \qquad \|x\| \ge R_n,\ x \in \mathfrak{H}_n. \tag{12.16}$$

Since $\langle F_n(x), x\rangle = \langle P_n F(x), x\rangle = \langle F(x), P_n x\rangle = \langle F(x), x\rangle$ 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

$$\|x_n\| \le R. \tag{12.17}$$

Now by Theorem 4.32 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,

$$\langle x_n, z\rangle \to \langle x, z\rangle, \qquad \text{for every } z \in \mathfrak{H}. \tag{12.18}$$

And it remains to show that x is indeed a solution. This follows from

Lemma 12.4. *Suppose $F : \mathfrak{H} \to \mathfrak{H}$ is demicontinuous and monotone, then*

$$\langle y - F(z), x - z\rangle \ge 0 \quad \textit{for every } z \in \mathfrak{H} \tag{12.19}$$

implies $F(x) = y$.

Proof. Choose $z = x \pm tw$, then $\mp\langle y - F(x \pm tw), w\rangle \ge 0$ and by continuity $\mp\langle y - F(x), w\rangle \ge 0$. Thus $\langle y - F(x), w\rangle = 0$ for every w implying $y - F(x) = 0$. □

Now we can show

Theorem 12.5 (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*

$$F(x) = y \tag{12.20}$$

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 (Problem 4.42) 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

$$t \mapsto \langle F(x + ty), z\rangle \tag{12.21}$$

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.

Some set theory

At the beginning of the 20th century Russell showed with his famous paradox "Is $\{x|x \notin x\}$ a set?" that naive set theory can lead to contradictions. Hence it was replaced by **axiomatic set theory**, more specific we will take the **Zermelo–Fraenkel set theory** (**ZF**)[1], 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 only A and B as elements and nothing else. 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 $\mathcal{F}$ and nothing else. In particular, given two sets A, B there exists a set $A \cup B = \bigcup\{A, B\}$ consisting precisely

[1]Ernst Zermelo (1871–1953), German logician and mathematician
[1]Abraham Fraenkel (1891–1965), German-born Israeli mathematician

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** $\mathfrak{P}(A)$ which contains precisely the 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 $0 = \emptyset$, $1 = \{\emptyset\}$, $2 = \{\emptyset, \{\emptyset\}\}$, ...

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 $\mathfrak{P}(X)$ be the collections of all subsets of a set X. Then $\mathfrak{P}(X)$ is partially ordered by inclusion $\subseteq$. $\diamond$

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 $A(k)$ 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[2])**.** *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.

[2]Max Zorn (1906–1993), German mathematician

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 $|\mathbb{N}| \leq |A|$, countable if $|A| \leq |\mathbb{N}|$, and countably infinite if $|A| = |\mathbb{N}|$.

Theorem A.3 (Bernstein–Cantor–Schröder)**.** *$|A| \leq |B|$ and $|B| \leq |A|$ implies $|A| = |B|$.*

Proof. Assume without loss of generality that A and B are disjoint and suppose $\varphi : A \to B$ and $\psi : B \to A$ are two injective maps. For any start value $a \in A$ or $b \in B$ we can form a unique two-sided sequence of elements by alternately applying φ and ψ. This sequence is uniquely defined for both positive and negative indices but might terminate at a negative integer since our maps are not surjective. In any case, if an element appears in two such 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 $A \uplus B$ and it suffices to find a bijection between the elements from A and B in the range of each sequence. If the sequence terminates at an element in A we can take φ. If the sequence

[2]Felix Bernstein (1878–1956), German mathematician

[2]Georg Cantor (1845–1918), German mathematician and founder of set theory

[2]Ernst Schröder (1841–1902), German mathematician

terminates at an element in B we can take ψ^{-1}. If the sequence is doubly infinite either of the previous choices will do. $\square$

Theorem A.4 (Zermelo)**.** *Either* $|A| \le |B|$ *or* $|B| \le |A|$.

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 $|A| \le |B|$ and if $\varphi_m(A_m) = B$ we have $|B| \le |A|$. $\square$

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. $\square$

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. $\square$

Corollary A.7. *Any infinite set can be written as a union of two disjoint subsets having the same cardinality as the original set.*

Proof. By the lemma we can write $A = \dot\bigcup A_\alpha$, where all A_α are countably infinite. Now split $A_\alpha = B_\alpha \dot\cup C_\alpha$ into two disjoint countably infinite sets (map A_α bijective to $\mathbb{N}$ and then split into even and odd elements). Then the desired splitting is $A = B \dot\cup C$ with $B = \dot\bigcup B_\alpha$ and $C = \dot\bigcup C_\alpha$. $\square$

Theorem A.8. *Suppose* A *or* B *is infinite. Then* $|A \cup B| = \max\{|A|, |B|\}$.

Proof. Suppose A is infinite and $|B| \le |A|$. Then $|A| \le |A \cup B| \le |A \uplus B| \le |A \uplus A| = |A|$ by the previous corollary. Here $\uplus$ denotes the disjoint union. $\square$

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 (Hessenberg[3]). *Suppose A is infinite and $B \neq \emptyset$. Then $|A \times B| = \max\{|A|, |B|\}$.*

Proof. Without loss of generality we can assume $|B| \le |A|$ (otherwise exchange both sets). Then $|A| \le |A \times B| \le |A \times A|$ 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. $\square$

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: $|[0,1)| \le |\mathfrak{P}(\mathbb{N})|$. 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 $|\mathfrak{P}(\mathbb{N})| \le |[0,1)|$. $\diamond$

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[4] and Cohen[5] 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.)*

---

[3]Gerhard Hessenberg (1874–1925), German mathematician

[4]Kurt Gödel (1906–1978), Austrian logician, mathematician, and philosopher

[5]Paul Cohen (1934–2007), American mathematician

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 [**19**] or the nice book by Kaplansky [**17**]. 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[1] 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 [**36**].

B.1. Basics

One of the key concepts in analysis is approximation which boils down to 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 set 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) $d(x, z) \le d(x, y) + d(y, z)$ (**triangle inequality**).

[1] `https://doi.org/10.2307/3614400`

If (i) does not hold, d is called a **pseudometric**. Note that if one requires the triangle inequality in the form $d(x,z) \le d(x,y) + d(z,y)$, then item (ii) could be dropped as it would follow from the other two (show this). Also nonnegativity $d \ge 0$ comes for free from (i) and (iii).

As a straightforward consequence we record the **inverse triangle inequality** (Problem B.1)

$$|d(x,y) - d(z,y)| \le d(x,z). \tag{B.1}$$

Example B.1. The role model for a metric space is of course Euclidean space $\mathbb{R}^n$ (or $\mathbb{C}^n$) together with $d(x,y) := (\sum_{k=1}^n |x_k - y_k|^2)^{1/2}$. ⋄

Example B.2. For any set X we can define the **discrete metric** as $d(x,x) = 0$ and $d(x,y) = 1$ if $x \ne y$. ⋄

Several notions from $\mathbb{R}^n$ carry over to (pseudo)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 open ball around x. If x is an interior point of U, then U is also called a **neighborhood** of x. 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. A point which is an interior point of the complement $X \setminus U$ is known as **exterior point**. Note that every point is either an interior, an exterior, or a boundary point of U.

Example B.3. 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 boundary points of U and the points $(-\infty,-1) \cup (1,\infty)$ are the exterior points of U.

Let $U := \mathbb{Q}$, the set of rational numbers. Then U has no interior points and $\partial U = \mathbb{R}$. ⋄

A point x is called a **limit point** of U (also **accumulation** or **cluster point**) if for any open ball $B_r(x)$ around 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 $x \in U$ which is not a limit point of U is called an **isolated point** of U. In other words, $x \in U$ is isolated 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**.

Example B.4. Consider $\mathbb{R}$ with the usual metric and let $U := (-1, 1)$. The points $[-1, 1]$ are the limit points of U. In the set $\mathbb{N} \subset \mathbb{R}$ every point is isolated and thus $\mathbb{N}$ is a discrete set in $\mathbb{R}$. ⋄

A set all of whose points are interior points is called **open**. The family of open sets $\mathcal{O}$ satisfies the properties

(O1) $\emptyset, X \in \mathcal{O}$,

(O2) $O_1, O_2 \in \mathcal{O}$ implies $O_1 \cap O_2 \in \mathcal{O}$,

(O3) $\{O_\alpha\} \subseteq \mathcal{O}$ implies $\bigcup_\alpha O_\alpha \in \mathcal{O}$.

That is, $\mathcal{O}$ is closed under finite intersections and arbitrary unions. Indeed, (O1) is obvious, (O2) follows since the intersection of two open balls centered at x is again an open ball centered at x (explicitly $B_{r_1}(x) \cap B_{r_2}(x) = B_{\min(r_1,r_2)}(x)$), and (O3) follows since every ball contained in one of the sets is also contained in the union.

Example B.5. Of course every open ball $B_r(x)$ is an open set since $y \in B_r(x)$ implies $B_s(y) \subseteq B_r(x)$ for $s < r - d(x, y)$. ⋄

Example B.6. Property (O2) clearly extends to finite intersections, but not to infinite intersections: The sets $U_n := (-\frac{1}{n}, \frac{1}{n}) \subset \mathbb{R}$ are open, but $\bigcap_{n\in\mathbb{N}} U_n = \{0\}$ is not open. ⋄

In various situations it tuns out beneficial to relax the quantitative notion of distance provided by a metric to a qualitative one. The key observation is that open balls can be replaced by open sets such that it suffices to know when a set is open. This motivates the following definition:

A space X together with a family of subsets $\mathcal{O}$, the open sets, satisfying (O1)–(O3), is called a **topological space**. The notions of interior point, limit point, neighborhood, etc. carry over to topological spaces if we replace "open ball around x" by "open set containing x".

The complement of an open set is called a **closed set**. It follows from **De Morgan's laws**

$$X \setminus \big(\bigcup_\alpha U_\alpha\big) = \bigcap_\alpha (X \setminus U_\alpha), \quad X \setminus \big(\bigcap_\alpha U_\alpha\big) = \bigcup_\alpha (X \setminus U_\alpha) \tag{B.3}$$

that the family of closed sets $\mathcal{C}$ satisfies

(C1) $\emptyset, X \in \mathcal{C}$,

(C2) $C_1, C_2 \in \mathcal{C}$ implies $C_1 \cup C_2 \in \mathcal{C}$,

(C3) $\{C_\alpha\} \subseteq \mathcal{C}$ 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**

$$\overline{U} := \bigcap_{C\in\mathcal{C}, U\subseteq C} C, \tag{B.4}$$

and the largest open set contained in a given set U is called the **interior**

$$U^\circ := \bigcup_{O\in\mathcal{O}, O\subseteq U} O. \tag{B.5}$$

Lemma B.1. *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. Clearly U° contains all interior points of U. Conversely, since U° is open, every point is an interior point of U° and hence also of U.

For the second claim observe that by Problem B.12 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. But these are precisely the limit points as noted before.

The last claim is left as Problem B.13. □

Example B.7. For any $x \in X$ in a metric space X the **closed ball**

$$\bar{B}_r(x) := \{y \in X | d(x,y) \le r\} \tag{B.6}$$

is a closed set (check that its complement is open). But in general we have only

$$\overline{B_r(x)} \subseteq \bar{B}_r(x) \tag{B.7}$$

since an isolated point y with $d(x,y) = r$ will not be a limit point of $B_r(x)$. In $\mathbb{R}^n$ (or $\mathbb{C}^n$) we have of course equality. ⋄

It is not hard to see that the closure satisfies the following axioms (**Kuratowski closure axioms**):

(K1) $\overline{\emptyset} = \emptyset$,

(K2) $U \subseteq \overline{U}$,

(K3) $\overline{\overline{U}} = \overline{U}$,

(K4) $\overline{U \cup V} = \overline{U} \cup \overline{V}$.

In fact, one can show that an operator satisfying these axioms can equivalently be used to define the topology by observing that the closed sets are precisely those which satisfy $\overline{U} = U$ (cf. Problems B.10, B.11). Similarly, the open sets are precisely those which satisfy $U^\circ = U$ and taking complements of the Kuratowski closure axioms (Problem B.12) gives:

(K1') $X^\circ = X$,

(K2') $U^\circ \subseteq U$,

(K3') $(U^\circ)^\circ = U^\circ$,

(K4') $(U \cap V)^\circ = U^\circ \cap V^\circ$.

Example B.8. Considering $X := \mathbb{R}$ and $U := \mathbb{Q}$, $V := \mathbb{R} \setminus \mathbb{Q}$, we see that (K4) fails for intersections and (K4') fails for unions. $\diamond$

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). Intuitively, the trivial topology says that every point is arbitrarily close to any other point, while the discrete topology says, that all points are well seperated.

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 $\mathcal{O}_2 \subseteq \mathcal{O}_1$.

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 one can define the weakest topology with a certain property to be the intersection of all topologies with this property (provided there is one at all).

Example B.9. 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

$$\tilde{d}(x, y) := \sum_{k=1}^{n} |x_k - y_k|. \tag{B.8}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} |x_k| \le \sqrt{\sum_{k=1}^{n} |x_k|^2} \le \sum_{k=1}^{n} |x_k| \tag{B.9}$$

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. $\diamond$

Example B.10. We can always replace a metric d by the bounded metric

$$\tilde{d}(x, y) := \frac{d(x, y)}{1 + d(x, y)} \tag{B.10}$$

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). $\diamond$

Every subset 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**). It follows that a set $C \subseteq Y$ is closed if and only if there is some closed set $\tilde{C} \subseteq X$ such that $C = \tilde{C} \cap Y$. Note that if Y is open, then $O \subseteq Y$ will be open in the relative topology if and only if it is open in X. Similarly for closed sets if Y is closed.

Example B.11. 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 $Y := [-1, 1]$. ⋄

Example B.12. If X is a metric space, then so is any subset $Y \subseteq X$ by simply restricting the metric to Y. In this case the relative topology on Y coincides with the metric topology. ⋄

The family of all open sets will typically be quite large and it hence will be often be helpful to consider subfamilies of open sets which still contain the full information about the topology. To this end, 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 $B \in \mathcal{B}$ with $x \in B \subseteq U(x)$.

Example B.13. In a metric space the balls form a base for the topology. ⋄

Example B.14. If $\mathcal{B}$ is a base for X and $Y \subseteq X$, then $\mathcal{B} \cap Y$ is a base for Y with respect to the relative topology. ⋄

Since an open set O is a neighborhood of every one of its points, it can be written as $O = \bigcup_{O \supseteq B \in \mathcal{B}} B$ and we have

Lemma B.2. *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. □

Frequently one wants to generate a topology by requiring certain sets to be open. More precisely, given a collection $\mathcal{M}$ of subsets of X we can define the topology generated by $\mathcal{M}$ as the weakest topology (i.e., the intersection of all topologies) containing $\mathcal{M}$. In general $\mathcal{M}$ will not be a base (Problem B.14). However, if we require $\mathcal{M}$ to be closed under finite intersections and $\bigcup \mathcal{M} = X$, then it will be a base for the topology generated by $\mathcal{M}$ (Problem B.15). Accordingly 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.15. The intervals form a base for the topology on $\mathbb{R}$. Intervals of the form (α, ∞) or $(-\infty, \alpha)$ with $\alpha \in \mathbb{R}$ are a subbase for topology of $\mathbb{R}$. ◇

Example B.16. 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 $[-1, 1]$. ◇

Given two bases we can use them to check if the corresponding topologies are equal.

Lemma B.3. *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 $O_1 \in \mathcal{O}_2$. □

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. Note that if x has a countable neighborhood base $\{B_n\}_{n\in\mathbb{N}}$, we can assume the base to be **nested**, $B_{n+1} \subseteq B_n$, without loss of generality. Indeed, simply replace B_n by $\tilde{B}_n := \bigcap_{m\le n} B_m$.

Example B.17. In a metric space the open balls $\{B_{1/m}(x)\}_{m\in\mathbb{N}}$ are a nested 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. ◇

Example B.18. Let $X := \mathbb{R}$ (or any other uncountable set) and call $O \subseteq X$ open if either O is empty or $X \setminus O$ is countable. This defines a topology (check this), known as the **cocountable topology**. Then X is not first countable. Indeed, suppose there were a countable neighborhood base B_j

for $0 \in X$. Without loss of generality we can assume $B_j \neq X$ for all j and hence there is some $x_j \in X \setminus B_j$ for every j. Now $N := X \setminus \bigcup_j \{x_j\}$ is a neighborhood of 0 for which $B_j \not\subseteq N$ for all j, a contradiction. ◇

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. *Let* $X := \mathbb{N}$. *Which of the following families of subsets form the open sets of a topology when augmented with the empty set and* X*?*

(i) $U_n := \{j \in \mathbb{N} | j \leq n\}$, $n \in \mathbb{N}$.

(ii) $U_n := \{j \in \mathbb{N} | j \geq n\}$, $n \in \mathbb{N}$.

(iii) *all finite subsets of* $\mathbb{N}$.

(iv) *all infinite subsets of* $\mathbb{N}$.

Problem B.5. *Consider* $X := \{a, b\}$ *with the open sets given by* $\mathcal{O} := \{\emptyset, \{a\}, X\}$. *Show that this is a topology and compute the closed sets and the closure of* $\{a\}$.

Problem B.6. *Consider* $X := \mathbb{Z}$ *with the topology generated by the pseudometric* $d(x,y) := ||x| - |y||$. *Compute the interior and the closure of* $\{1\}$.

Problem B.7. *Show De Morgan's laws.*

Problem B.8. *Show that the closure satisfies the Kuratowski closure axioms.*

Problem B.9. *Let* X *be a topological space and let* $A, A_\alpha, B \subset X$, *where* α *is any index. Show:*

(i) $A \subseteq B \quad \Rightarrow \quad A^\circ \subseteq B^\circ$

(ii) $A \subseteq B \quad \Rightarrow \quad \overline{A} \subseteq \overline{B}$

(iii) $\left(\bigcap_\alpha A_\alpha\right)^\circ \subseteq \bigcap_\alpha (A_\alpha)^\circ$

(iv) $\bigcup_\alpha \overline{A_\alpha} \subseteq \overline{\bigcup_\alpha A_\alpha}$

Problem B.10. *Let* $X = \{1,2,3\}$ *and define* $c : \mathfrak{P}(X) \to \mathfrak{P}(X)$ *via* $c(\{1\}) = \{1\}$, $c(\{2\}) = \{1,2\}$, $c(\{3\}) = \{2,3\}$ *and*

$$c(U) = \bigcup_{x \in U} c(\{x\}), \quad U \subset X,$$

with the convention $c(\emptyset) = \emptyset$.

Which of the Kuratowski closure axioms does c satisfy? What are the closed sets according to c. Do they give raise to a topology? If yes, do you have $\overline{U} = c(U)$ with respect to this topology?

Problem B.11. *Let $c : \mathfrak{P}(X) \to \mathfrak{P}(X)$ be an operator satisfying the Kuratowski closure axioms (K1,K2,K4). Show that the sets $U \in \mathfrak{P}(X)$ satisfying $\overline{U} = U$ satisfy the axioms for closed sets. Show that $c(U) = \overline{U}$ provided (K3) holds in addition. (Hint: First show that $U \subseteq V$ implies $\overline{U} \subseteq V$.)*

Problem* B.12. *Let X be a topological space and let $U \subset X$. 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.13. *Show that the boundary of U is given by $\partial U = \overline{U} \setminus U^\circ$.*

Problem B.14. *Let $X := \{1, 2, 3\}$ and $A := \{1, 2\}$, $B := \{2, 3\}$. What is the topology generated by $\mathcal{M} := \{A, B\}$? Is $\mathcal{M}$ a base for this topology?*

Problem B.15. *Suppose $\mathcal{M}$ is a collection of sets closed under finite intersections containing $\emptyset$ and let $X := \bigcup \mathcal{M}$. Then the topology generated by $\mathcal{M}$ is given by $\mathcal{O}(\mathcal{M}) := \{\bigcup_\alpha M_\alpha | M_\alpha \in \mathcal{M}\}$.*

Problem B.16. *Equip $X := \mathbb{R}$ with the* ***lower limit topology*** *generated by half-open intervals of the form $[a, b)$. Show that this topology is finer than the metric topology. Also show that the sets $[a, b)$ are both open and closed with respect to this topology, and the same holds true for the intervals $[a, \infty)$ and $(-\infty, b)$.*

Problem B.17. *Show that the lower limit topology from Problem B.16 is first countable but not second countable.*

Problem B.18. *Every subset $Y \subseteq X$ of a first/second countable topological space is first/second countable with respect to the relative topology.*

B.2. Convergence and completeness

A sequence $(x_n)_{n=1}^\infty$ from a metric space X 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. Explicitly we have $x_n \to x$ if for every $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n \geq N$. 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 $r > 0$.

Of course 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 $n \geq N$.

Example B.19. According to the discrete topology a sequence will converge if and only if it is eventually constant. According to the trivial topology every sequence will converge to every point. In particular, the last example shows that the limit is not unique if the space contains more than one point. $\diamond$

Ensuring uniqueness of limits requires another property: A topological space is called a **Hausdorff space** (sometimes also T_2) if for any two different points there are always two disjoint neighborhoods.

Example B.20. 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. $\diamond$

It is straightforward to check that in a Hausdorff space the limit is unique (Problem B.22). Note also that in a Hausdorff space sets consisting of one point are closed (Problem B.20).

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.21. For example, we can call a set $U \subseteq X$ **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 interior 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 nested neighborhood base (Problem B.23). Hence the converse is also true in this case. $\diamond$

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. Problem 5.11 from [**40**]).

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 $(x_n)_{n=1}^\infty \in X^{\mathbb{N}}$ in a metric space X 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.22. Both $\mathbb{R}^n$ and $\mathbb{C}^n$ are complete metric spaces. ⋄

Example B.23. 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.16). ⋄

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** (in X) if its closure is all of X, that is, if $\overline{U} = X$. A topological space is called **separable** if it contains a countable dense set.

Lemma B.7. *A second countable space is separable and the converse holds for metric spaces.*

Proof. From every countable base we obtain a dense set by selecting an element from each set in the base. Indeed, let $x \in X$ be given and let S be the selected points from some countable base. From the base we can take those sets containing x to obtain a countable neighborhood base $\{B_n\}$. Considering $\tilde{B}_n := \bigcap_{m \le n} B_m$ we can choose a point $x_n \in \tilde{B}_n \cap S$ (since each $\tilde{B}_n$ contains some set from our base). By construction $x_n \to x$ and hence S is dense.

Conversely, from every dense set we get a countable base by considering open balls with radii $1/n$ and centers in the dense set. □

Example B.24. Consider the lower limit topology on $X := \mathbb{R}$ generated by half-open intervals of the form $[a, b)$ (Problem B.16). Then the set of rationals $\mathbb{Q}$ is dense, since for every $x \notin \overline{\mathbb{Q}}$ there must be an open neighborhood O of x not intersection $\mathbb{Q}$. But O must contain a set of the form $[x, x + 1/n)$ for some n contradicting the assumption that O does not intersect $\mathbb{Q}$. On the other hand, X is not second countable (Problem B.17). ⋄

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 $d(y_{n_k,k}, y) \le d(y_{n_k,k}, x_{n_k}) + d(x_{n_k}, y) \le 2/k \to 0$. □

If X is an (incomplete) metric space, consider the set of all Cauchy sequences $\mathcal{X}$ from X. Call two Cauchy sequences $x = (x_n)_{n\in\mathbb{N}}$ and $y = (y_n)_{n\in\mathbb{N}}$ equivalent if $d_X(x_n, y_n) \to 0$ and denote by $\bar{X}$ the set of all equivalence classes $[x]$. 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 of the limits whenever both Cauchy sequences converge in X.

Theorem B.9. *The space $\bar{X}$ is a complete 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 $\xi_n = [(x_{n,j})_{j=1}^\infty]$ be a Cauchy sequence in $\bar{X}$. Without loss of generality (by dropping terms) we can choose the representatives $x_{n,j}$ such that $d_X(x_{n,j}, x_{n,k}) \le \frac{1}{n}$ for $j, k \ge n$.

To see that $\xi = [(x_{j,j})_{j=1}^\infty]$ is its limit, let $\varepsilon > 0$ be given and choose an N such that $d_{\bar{X}}(\xi_n, \xi_m) \le \varepsilon/3$ for $n, m \ge N$. By definition this means that

$$\lim_{j\to\infty} d_X(x_{n,j}, x_{m,j}) \le \frac{\varepsilon}{3}.$$

Moreover, by increasing N if necessary, we can also assume $\frac{2}{N} \le \frac{\varepsilon}{3}$. Then

$$\begin{aligned} d_X(x_{n,k}, x_{m,m}) &\le d_X(x_{n,m}, x_{n,j}) + d_X(x_{n,j}, x_{m,j}) + d_X(x_{m,j}, x_{m,m}) \\ &\le \frac{2}{N} + d_X(x_{n,j}, x_{m,j}), \qquad n, m, k, j \ge N, \end{aligned}$$

and taking the limit $j \to \infty$ we conclude

$$d_X(x_{n,k}, x_{m,m}) \le \varepsilon, \qquad n, m, k \ge N.$$

Now choosing $k = n$ we see that $x_{n,n}$ is a Cauchy sequence and hence $\xi \in \bar{X}$. Moreover, choosing $k = m$ shows $d_{\bar{X}}(\xi_n - \xi) \le \varepsilon$ for $n \ge N$ and hence $\xi_n \to \xi$. □

Notice 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.39 below). In particular, it is no restriction to assume that a metric space is complete.

Problem B.19. *Let X be some nonempty set and define $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$ if $x \neq y$. Show that (X, d) is a metric space. When are sequences convergent? When is X separable?*

Problem* B.20. *Show that singeltons are closed if and only if X is a T_1* ***space****, that is, a topological space where for any points $x, y \in X$ there is a neighborhood of x disjoint from y. Evidently every Hausdorff space is T_1.*

Problem B.21. *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.22. *Show that in a Hausdorff space limits are unique. Show that the converse is true for a first countable space.*

Problem* B.23. *Let X be a first countable topological space. Show that x is a limit point of $U \subseteq X$ if and only if there is a sequence $x_n \in U \setminus \{x\}$ converging to x.*

Problem B.24. *Let X be a metric space and denote by $B(X)$ the set of all bounded functions $X \to \mathbb{C}$. Introduce the metric*

$$d(f,g) = \sup_{x \in X} |f(x) - g(x)|, \quad f, g \in B(X).$$

Show that $B(X)$ is complete.

Problem B.25. *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 $\operatorname{Ran}(f) := f(X)$ 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 $\operatorname{Ran}(f) = Y$. A function f which is both injective and surjective is called **bijective**.

Recall that we always have

$$f^{-1}(\bigcup_\alpha V_\alpha) = \bigcup_\alpha f^{-1}(V_\alpha), \qquad f^{-1}(\bigcap_\alpha V_\alpha) = \bigcap_\alpha f^{-1}(V_\alpha),$$
$$f^{-1}(Y \setminus V) = X \setminus f^{-1}(V) \tag{B.15}$$

as well as

$$f(\bigcup_\alpha U_\alpha) = \bigcup_\alpha f(U_\alpha), \qquad f(\bigcap_\alpha U_\alpha) \subseteq \bigcap_\alpha f(U_\alpha),$$
$$f(X) \setminus f(U) \subseteq f(X \setminus U) \tag{B.16}$$

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

$$d_Y(f(x), f(y)) \le \varepsilon \quad \text{if} \quad d_X(x, y) \le \delta. \tag{B.17}$$

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. Slightly more general, f is called **Lipschitz** if there is constant $L \ge 0$ such that $d_Y(f(x), f(y)) \le L\, d_X(x, y)$. Clearly, Lipschitz functions are continuous since we can choose $\delta := \varepsilon/L$.

Lemma B.10. *Let $f : X \to Y$ be a function between metric spaces X, Y. 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 $x_n \notin f^{-1}(V)$. 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.28).

In particular, (iii) implies that f is continuous if and only if the preimage of every open set is again open (equivalently, by (B.15), 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**. Similarly, if the image of every closed set is closed, then f is called **closed**. Note that due to the lack of equality in (B.16) these two conditions are not equivalent in general unless f is injective.

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.

In a topological space X a function $f : X \to \overline{\mathbb{R}}$ is **lower semicontinuous** if the set $f^{-1}((a,\infty])$ is open for every $a \in \mathbb{R}$. Similarly, f is **upper semicontinuous** if the set $f^{-1}([-\infty,a))$ is open for every $a \in \mathbb{R}$. Clearly f is lower semicontinuous if and only if $-f$ is upper semicontinuous.

Finally, 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.26. *Let X, Y be topological spaces and $U \subseteq X$, $V \subseteq Y$. Show that if $f : X \to Y$ is continuous and* $\operatorname{Ran}(f) \subseteq V$*, then the restriction $f : U \to V$ is continuous when U, V are equipped with the relative topology.*

Problem B.27 (Gluing lemma)**.** *Let X, Y be topological spaces and $\{U_\alpha\}_{\alpha\in A}$ an cover of X (i.e. $\bigcup_\alpha U_\alpha = X$). Suppose $f_\alpha : U_\alpha \to Y$ are continuous functions with respect to the relative topology on U_α such that $f_\alpha = f_\beta$ on $U_\alpha \cap U_\beta$ and hence $f : X \to Y$ is well defined by $f(x) := f_\alpha(x)$ if $x \in U_\alpha$. Show that f is continuous if either all sets U_α are open or if the index set A is finite and all sets U_α are closed.*

Problem* B.28. *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 at $x \in X$. Show that the converse holds if X is first countable.*

Problem B.29. *Let X, Y be topological spaces and let $f : X \to Y$ be continuous. Show that $f(\overline{A}) \subseteq \overline{f(A)}$ for any $A \subset X$. Show that equality fails in general.*

Problem B.30. *Let X, Y be topological spaces and let $f : X \to Y$ be continuous. Show that if X is separable, then so is $f(X)$.*

Problem B.31. *Let X, Y be topological spaces and let $f : X \to Y$. Show that f is open if and only if the images from all sets of a base are open.*

Problem B.32. *Let X be a topological space and $f : X \to \overline{\mathbb{R}}$. Let $x_0 \in X$ and let $\mathcal{B}(x_0)$ be a neighborhood base for x_0. Define*

$$\liminf_{x\to x_0} f(x) := \sup_{U\in\mathcal{B}(x_0)} \inf_U f, \qquad \limsup_{x\to x_0} f(x) := \inf_{U\in\mathcal{B}(x_0)} \sup_U f.$$

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) $\liminf_{x\to x_0}(f(x)+g(x)) \ge \liminf_{x\to x_0} f(x) + \liminf_{x\to x_0} g(x)$.

Here in (iii) $g : X \to \overline{\mathbb{R}}$ *is assumed to be such that no undefined expressions of the from* $\infty - \infty$ *appear.*

Moreover, show that

$$\liminf_{n\to\infty} f(x_n) \ge \liminf_{x\to x_0} f(x), \qquad \limsup_{n\to\infty} f(x_n) \le \limsup_{x\to x_0} f(x)$$

for every sequence $x_n \to x_0$ *and there exists a sequence attaining equality if* X *is a metric space.*

Problem B.33. *Show that the supremum over lower semicontinuous functions is again lower semicontinuous.*

Problem* B.34. *Let* X *be a topological space and* $f : X \to \overline{\mathbb{R}}$. *Show that* f *is lower semicontinuous if and only if*

$$\liminf_{x\to x_0} f(x) \ge f(x_0), \qquad x_0 \in X.$$

Similarly, f *is upper semicontinuous if and only if*

$$\limsup_{x\to x_0} f(x) \le f(x_0), \qquad x_0 \in X.$$

Show that a lower semicontinuous function is also sequentially lower semicontinuous

$$\liminf_{n\to\infty} f(x_n) \ge f(x_0), \qquad x_n \to x_0,\ x_0 \in X.$$

Show the converse if X *is a metric space. (Hint: Problem B.32.)*

Problem B.35. *Show that the sum of two lower semicontinuous functions is again lower semicontinuous (provided the sum is well-defined, that is, no terms of the form* $\infty - \infty$ *appear).*

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 $X \times Y$.

Example B.25. 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:

$$\tilde{d}((x_1,y_1),(x_2,y_2)) := \sqrt{d_X(x_1,x_2)^2 + d_Y(y_1,y_2)^2}. \tag{B.20}$$

As noted in Example B.9, the topology (and thus also convergence/continuity) is independent of this choice. $\diamond$

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. In fact, products of open sets from a base for X and Y form a base of the product topology. Again the projections onto the first and second component are continuous and a sequence (x_n, y_n) will converge with respect to the product topology if and only if both components converge. In the case of metric spaces this of course agrees with the topology defined via the product metric (B.19).

Example B.26. Note that the projections onto the components are open maps since the projection of the product of two open sets is open and these sets form a base for the product topology (Problem B.31).

However, they are not closed, as the example $\{(x, y) \in \mathbb{R}^2 | xy = 1\}$ shows. ⋄

Clearly this easily extends to any finite number of topological spaces. The case of an arbitrary number of topological spaces will be discussed in Section B.8.

Problem B.36. *Let X, Y be topological spaces and $A \subseteq X$, $B \subseteq Y$. Show that $(A \times B)^\circ = A^\circ \times B^\circ$ and $\overline{A \times B} = \overline{A} \times \overline{B}$.*

Problem B.37. *Let X, Y be topological spaces. Show that $X \times Y$ has the following properties if both X and Y have.*

(i) *Hausdorff*
(ii) *separable*
(iii) *first/second countable*

Problem B.38. *Show that X is Hausdorff if and only if the diagonal $\Delta := \{(x, x) | x \in X\} \subseteq X \times X$ is closed.*

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.11 (Lindelöf[2]). *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.

[2]Ernst Leonard Lindelöf (1870–1946), Finnish mathematician

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 subset $K \subset X$ is called **compact** if every open cover of K has a finite subcover. Note that $K \subset X$ is compact if and only if it is compact with respect to the relative topology (Problem B.39). A set is called **relatively compact** if its closure is compact.

Example B.27. Consider $X := \mathbb{R}$ with the lower semicontinuity topology generated by sets of the form $[a, \infty)$ for $a \in \mathbb{R}$. Then a set is compact if and only if it has a minimum — Problem B.41. ⋄

There is also a dual formulation in terms of closed sets. To this end, a (nonempty) family of subsets is said to have the **finite intersection property** if the intersection over any finite subfamily is nonempty.

Lemma B.12. *A topological space is compact if and only if every family of closed subsets with the finite intersection property has a nonempty intersection.*

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.13. *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.*

(vi) *The product of finitely many compact sets is compact.*

Proof. (i) If $\{O_\alpha\}_{\alpha\in A}$ is an open cover for $f(K)$, then $\{f^{-1}(O_\alpha)\}_{\alpha\in A}$ is one for K and by assumption there is a finite subset $A_1 \subseteq A$ such that $\{f^{-1}(O_\alpha)\}_{\alpha\in A_1}$ still covers YK. Hence $\{O_\alpha\}_{\alpha\in A_1}$ covers $f(K)$.

(ii) Let $\{O_\alpha\}$ be an open cover for the closed subset $Y \subseteq K$. Then $\{O_\alpha\} \cup \{X \setminus Y\}$ is an open cover for K 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 a Hausdorff space, 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 $y_1, \dots, y_n$ 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.

(vi) Let $K \subseteq X$ and $L \subseteq Y$ be compact. Let $\{O_\alpha\}$ be an open cover for $K \times L$. For every $(x,y) \in K \times L$ there is some $\alpha(x,y)$ such that $(x,y) \in O_{\alpha(x,y)}$. By definition of the product topology, there is some open rectangle $U(x,y) \times V(x,y) \subseteq O_{\alpha(x,y)}$. Hence for fixed x, $\{V(x,y)\}_{y\in L}$ is an open cover of L. Hence there are finitely many points $y_k(x)$ such that the $V(x, y_k(x))$ cover L. Set $U(x) = \bigcap_k U(x, y_k(x))$. Since finite intersections of open sets are open, $\{U(x)\}_{x\in K}$ is an open cover and there are finitely many points x_j such that the $U(x_j)$ cover K. By construction, the $U(x_j) \times V(x_j, y_k(x_j)) \subseteq O_{\alpha(x_j, y_k(x_j))}$ cover $K \times L$. □

Example B.28. The Hausdorff property is crucial for (iii) and (v) to hold: Consider again $X := \mathbb{R}$ with the lower semicontinuity topology. Then the sets $[0,1) \cup (2,\infty)$ and $[2,\infty)$ are both compact but neither is closed and their intersection $(2,\infty)$ is not compact. ⋄

As a consequence we obtain a simple criterion when a continuous function is a homeomorphism.

Corollary B.14. *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. □

In many situations an important fact is that compact sets have a limit point.

Lemma B.15. *Let X be a topological space and $K \subseteq X$ be relatively compact. Then every infinite subset of K has a limit point.*

Proof. Let $J \subseteq K$ be infinite without limit points. Then J is discrete and every point of J has an open neighborhood containing no other point from J. These neighborhoods together with $X \setminus J$ cover $\overline{K}$ but contain no finite subcover, a contradiction. □

This is of particular importance in connection with sequences. To this end we call a subset $K \subseteq X$ **sequentially compact** if every sequence from K has a convergent subsequence whose limit is in K.

Corollary B.16. *In a first countable Hausdorff space every compact set is sequentially compact.*

Proof. A sequence x_n has either finitely many values or infinitely many values. In the first case, one of the values appears infinitely many times and gives rise to a convergent subsequence. In the second case we can choose a countable neighborhood base $\{B_n\}$ for the limit point x. Now choose x_{n_1} from $B_1 \setminus \{x\}$. Set $\tilde{B}_2 := (B_1 \cap B_2) \setminus \{x_1, \dots, x_{n_1}\}$ (by Problem B.20 $\tilde{B}_2$ is a neighborhood of x) and choose x_{n_2} from $\tilde{B}_2 \setminus \{x\}$. Continuing like this gives a subsequence which converges to x. □

Corollary B.17. *A compact metric space X is complete and separable.*

Proof. Completeness is immediate since every Cauchy sequence has a convergent subsequence by the previous corollary.

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 the converse of Corollary B.16 is also true. In fact we will show equivalence to another property which is frequently easier to establish.

A subset U of a 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.

Note that we can assume that the centers of these balls are all in U. Indeed, given $\varepsilon > 0$ we can choose an $\varepsilon/2$-cover $\{B_{\varepsilon/2}(x_j)\}_{j=1}^n$ and (after removing redundant balls which have empty intersection with U) points $y_j \in B_{\varepsilon/2}(x_j) \cap U$. Then $B_{\varepsilon/2}(x_j) \subseteq B_\varepsilon(y_j)$ and hence $\{B_\varepsilon(y_j)\}_{j=1}^n$ is an ε-cover for U with centers from U.

Lemma B.18. *Let X be a complete metric space. Then the following are equivalent for $K \subseteq X$:*

(i) *K is relatively compact.*

(ii) *Every sequence from K has a convergent subsequence.*

(iii) *K is totally bounded.*

Proof. (i) $\Rightarrow$ (ii) This follows since $\overline{K}$ is sequentially compact by Corollary B.16.

(ii) $\Rightarrow$ (iii) Let $\varepsilon > 0$ be given and suppose $x_1, \dots, x_n \in K$ are chosen such that $d(x_j, x_k) \geq \varepsilon$ for $j \neq k$. Then this process must stop at some finite index n since otherwise the resulting set of points would not have a limit point. But then the balls $B_\varepsilon(x_j)$, $1 \leq j \leq n$, cover K as required.

(iii) ⇒ (i) We need to show that $\overline{K}$ is compact. Since K is totally bounded, so is $\overline{K}$ (Problem B.45) and we can assume that K is closed without loss of generality.

Now assume we had a cover U_α for K which has no finite subcover. By (iii) K can be covered by finitely many balls of radius 1 and the intersection of at least one of these balls with K, call it K_1, cannot be covered by finitely many U_α. Now repeat this process choosing radii $1/2$, $1/3$, ... to obtain a sequence of nested closed sets $K_n \supseteq K_{n+1}$ such that K_n is contained in a ball of radius $1/n$ and no K_n can be covered by finitely many U_α.

Now choose $x_n \in K_n$ and note that this sequence is Cauchy and hence has a limit $x_n \to x \in \bigcap_n K_n$. In particular, $x \in U_{\alpha_0}$ for some α_0 and thus $K_n \subset U_{\alpha_0}$ for n sufficiently large. This contradicts our construction and hence K is compact. □

Recall that a set in a metric space 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 they can be covered by finitely many balls. In $\mathbb{R}^n$ (or $\mathbb{C}^n$) the converse also holds.

Theorem B.19 (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.13 (ii), (iii), and (vi) it suffices to show that a closed interval in $I \subseteq \mathbb{R}$ is compact. Moreover, by Lemma B.18, 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 $|y_n - y_m| \le \frac{b-a}{2n}$ for $m \ge n$), which converges by completeness of $\mathbb{R}$. □

By Lemma B.15 this is equivalent to

Theorem B.20 (Bolzano–Weierstraß[3])**.** *Every bounded infinite subset of $\mathbb{R}^n$ (or $\mathbb{C}^n$) has at least one limit point.*

Using Lemma B.13 (i) we also obtain the **extreme value theorem**.

Theorem B.21 (Weierstraß)**.** *Let X be nonempty and compact. Every continuous function $f : X \to \mathbb{R}$ attains its maximum and minimum.*

[3]Bernard Bolzano (1781–1848), Bohemian mathematician, logician, philosopher, theologian and Catholic priest

Proof. By Lemma B.13 (i) the range of f is compact. In particular, the range contains a maximum and a minimum. To see the last claim, pick a sequence converging to the supremum (or infimum) and extract a convergent subsequence. $\square$

See Problem B.47 for an extension.

A metric space X for which the Heine–Borel theorem holds is called **proper**. Lemma B.13 (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. Moreover, by Corollary B.17 a σ-compact metric space is complete and separable.

Lemma B.22. *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.*

The implications (ii) $\Rightarrow$ (iii) $\Leftrightarrow$ (vi) hold for topological spaces.

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) $\Rightarrow$ (iv): Let $X = \bigcup_n K_n$ with K_n compact. Without loss $K_n \subseteq K_{n+1}$ (consider $\bigcup_{m\le n} K_m$). 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 sets and choose a finite subcover). Now define $U_1 := V(K_1)$ and $U_{n+1} := V(\overline{U_n} \cup K_n)$. (iv) $\Rightarrow$ (i): Each of the sets $\overline{U_n}$ has a countable dense subset by Corollary B.17. The union gives a countable dense set for X. Since every $x \in U_n$ for some n, X is also locally compact. $\square$

A sequence of compact sets $\overline{U_n}$ as in (iv) is known as a **compact exhaustion** of X. Since every point of X is an interior point of some U_n, every compact set $K \subseteq X$ is contained in U_n for n sufficiently large (cover K by taking an open neighborhood contained in some U_n for every point).

Example B.29. Consider the set of bounded sequences $\ell^\infty(\mathbb{N})$ with the metric $d(x,y) = \sup_{n\in\mathbb{N}} |x_n - y_n|$. Denote by δ^j the sequence which is 1 at the jth place and 0 else. Let $X_j := \{\lambda\delta^j | \lambda \in [0,1]\}$ and note that the metric on X_j inherited from $\ell^\infty(\mathbb{N})$ is the same as the usual metric from $\mathbb{R}$. Then $X := \bigcup_{j\in\mathbb{N}} X_j$ is a 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 $d((\varepsilon/2)\delta^j, (\varepsilon/2)\delta^k) = \varepsilon/2$ for $k \neq j$. ◇

Example B.30. $X := (0,1)$ with the usual metric is locally compact and σ-compact but not proper. ◇

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.56) there is a new metric with the claimed properties (Problem B.46).

Problem B.39. *Show that a subset K of a topological space X is compact if and only if it is compact with respect to the relative topology.*

Problem B.40. *Let X, Y be topological spaces, Y Hausdorff, and let $f : X \to Y$ be continuous. Show that $f(\overline{A}) = \overline{f(A)}$ for any $A \subset X$.*

Problem* B.41. *Consider $X := \mathbb{R}$ with the lower semicontinuity topology $\mathcal{O} := \{(a,\infty) | a \in \mathbb{R}\} \cup \{\emptyset, \mathbb{R}\}$. Show that this is indeed a topology. Show that a set $C \subset X$ is compact if and only if it has a minimum.*

Problem* B.42 (Alexandroff[4] 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.43. *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.44. *Let X be a metric space. Show that a totally bounded set $U \subseteq X$ is bounded.*

Problem B.45. *Let X be a metric space. Show that if $U \subseteq X$ is totally bounded, so is $\overline{U}$.*

[4]Pavel Alexandroff (1896–1982), Soviet mathematician

Problem B.46. *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.

Problem B.47. *Let X be a nonempty sequentially compact topological space. Show that a sequentially lower (upper) semicontinuous function $f: X \to \overline{\mathbb{R}}$ attains its infimum (supremum) on X.*

B.6. 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 continuous 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 nonempty open sets U and V which split A according to $A = (U \cap A) \cup (V \cap A)$ with $U \cap A = V \cap A = \emptyset$.

Example B.31. In $\mathbb{R}$ the nonempty connected sets are precisely the intervals (Problem B.48). 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.32. 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). $\diamond$

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.33. Every normed vector space is (locally) path-connected since every ball is path-connected (consider straight lines). In fact this also holds for locally convex spaces. Every open subset of a locally (path-)connected space is locally (path-)connected. $\diamond$

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.34. The spaces $\mathbb{R}$ and $\mathbb{R}^n$, $n > 1$, are not homeomorphic. In fact, removing any point from $\mathbb{R}$ gives a disconnected space while removing a point form $\mathbb{R}^n$ still leaves it (path-)connected. $\diamond$

We collect a few simple but useful properties below.

Lemma B.23. *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 no two sets are disjoint. 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 $X_1, \dots, X_n$ are (path-)connected, then so is $\bigtimes_{j=1}^n X_j$.*

(v) *Suppose $A \subseteq X$ is connected, then any B with $A \subseteq B \subseteq \overline{A}$ 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'). We first prove (ii) under the weaker assumption $\bigcap_\alpha A_\alpha \neq \emptyset$.

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 $0 \leq t \leq \frac{1}{2}$ and $\gamma_\alpha \odot \gamma_\beta(t) = \gamma_\beta(2t-1)$ for $\frac{1}{2} \leq t \leq 1$ (cf. Problem B.27).

(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 (ii').

(ii). Follows combining (ii') with (iii).

(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. Indeed, a splitting $\{x\} \cup A = U \cup V$ with $x \in U$ (w.l.o.g.) would give a splitting $A = (U \setminus \{x\}) \cup V$ and hence $U = \{x\}$ and $V = A$ or $U = A$ and $V = \emptyset$, both of which are impossible. 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.35. Consider the graph of the function $f : (0, 1] \to \mathbb{R}$, $x \mapsto \sin(\frac{1}{x})$. Then its graph $\Gamma(f) \subseteq \mathbb{R}^2$ is path-connected and the 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.

Note that this example shows that item (v) does not extend to the path-connected case. ⋄

Theorem B.24 (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.48. *Show that a nonempty subset of $\mathbb{R}$ is connected if and only if it is an interval.*

Problem B.49. *Let $U = \dot{\bigcup}_j U_j \subseteq \mathbb{R}^n$ be an open set with U_j its connected components. Show $\partial U_j \subseteq \partial U$. Show that in the case of finitely many components we have $\bigcup_{j=1}^n \partial U_j = \partial U$. Show that this fails for infinitely many components in general.*

Problem B.50. *Show that $\mathbb{R}$ with the lower limit topology (Problem B.16) is* ***totally disconnected****, that is, the only (nonempty) connected subsets are singeltons.*

Problem B.51. *Let X be a topological space and suppose $A_\alpha \subseteq X$ such that no two sets are seperated. Show that $\bigcup_\alpha A_\alpha$ is connected. (Hint: By Lemma B.23 (iii) it suffices to consider the case of two sets.)*

B.7. Constructing continuous functions

In many applications one needs to construct continuous functions with prescribed values. A typical problem would be to find a cut-off function which is one on a given set C_1 and vanishes on some other set C_0. By continuity we can assume both sets to be closed. In a metric space X this problem is relatively easy. To this end we define the **distance** between a point $x \in X$ and a subset $Y \subseteq X$ as

$$\operatorname{dist}(x, Y) := \inf_{y \in Y} d(x, y). \tag{B.21}$$

Note that $x \in \overline{Y}$ if and only if $\operatorname{dist}(x, Y) = 0$ (Problem B.52).

Lemma B.25. *Let X be a metric space and $Y \subseteq X$ nonempty. Then*

$$|\operatorname{dist}(x, Y) - \operatorname{dist}(z, Y)| \le d(x, z). \tag{B.22}$$

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 $\operatorname{dist}(z, Y) - \operatorname{dist}(x, Y) \le d(x, z)$. □

Now the function we are looking for is

$$f(x) := \frac{\operatorname{dist}(x, C_0)}{\operatorname{dist}(x, C_1) + \operatorname{dist}(x, C_0)}.$$

In a general topological space this questions is more delicate. As a first observation note, that if we have found such a function, then we can separate our closed sets via open neighborhoods by considering $f^{-1}((-\infty, \frac{1}{2}))$ and $f^{-1}((\frac{1}{2}, \infty))$. This motivates the following definition:

A topological space is called **normal** if for any two disjoint closed sets C_0 and C_1, there are disjoint open sets O_0 and O_1 such that $C_j \subseteq O_j$, $j = 0, 1$.

Observe that X is normal if and only if for every closed set $\tilde{C}$ and every open neighborhood $\tilde{O}$ of $\tilde{C}$, there exists an open set $\tilde{O}_1$ and a closed set $\tilde{C}_1$ such that $\tilde{C} \subset \tilde{O}_1 \subset \tilde{C}_1 \subset \tilde{O}$ (just use the identification $C_1 \leftrightarrow \tilde{C}$, $C_0 \leftrightarrow X \setminus \tilde{O}$ and $O_1 \leftrightarrow \tilde{O}_1$, $O_0 \leftrightarrow X \setminus \tilde{C}_1$).

Example B.36. By our preceding considerations, every metric space is normal. In fact, this still holds for a pseudo metric space. Hence, note that a normal space might not be Hausdorff (unless we additionally require singeltons to be closed). ⋄

Lemma B.26 (Urysohn[5]). *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_0, there exists a continuous function $f : X \to [0, 1]$ which is one on C_1 and zero on C_0.*

[5]Pavel Urysohn (1898–1924), Soviet mathematician

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. $O_0 := X \setminus C_0$). 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.

Using our second characterization of normal 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 $C_q \subseteq O_p$ for $p < q$. Now set $f_n^s(x) := \max\{q = k/2^n | x \in O_q,\ 0 \le k < 2^n\}$ for $x \in O_0$ and $f_n^s(x) := 0$ else as well as $f_n^i(x) := \min\{q = k/2^n | x \notin C_q,\ 0 < k \le 2^n\}$ for $x \notin C_1$ and $f_n^i(x) := 1$ else. Then $f_n^s(x) \nearrow f^s(x) := \sup\{q | x \in O_q\}$ 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, $0 \le f_n^i(x) - f_n^s(x) - 2^{-n} \le 2^{-n}$ and thus $f^s(x) = f^i(x)$. 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 $O_0 := f^{-1}([0,1/2))$ and $O_1 := f^{-1}((1/2,1])$.

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. $\square$

An important class of normal spaces are compact Hausdorff spaces.

Lemma B.27. *Every compact Hausdorff space is normal.*

Proof. We first show that we can separate a point $x \in X$ and a closed subset $C \subset X$ provided $x \notin C$. Indeed, since X is Hausdorff, for every $y \in C$ there are disjoint open sets U_y containing x and V_y containing y. Moreover, since X is compact so is C (Lemma B.13 (ii)) and hence finitely many sets $V_{y_1}, \dots, V_{y_n}$ cover C. Hence $U := \bigcap U_{y_j}$ is an open set containing x which is disjoint from the open set $V := \bigcup V_{y_j}$ containing C.

Now given two disjoint closed sets $C_1, C_2 \subset X$, the first parts implies that for every $y \in C_1$ there are disjoint open sets U_y containing C_1 and V_y containing y. Hence we can proceed as before. $\square$

Another important result is the **Tietze extension theorem**:

Theorem B.28 (Tietze[6]). *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 $\bar{f} : X \to [-1,1]$. If in addition X is locally compact and C is compact, then f can be chosen to have compact support.*

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 $\bar{f} := \sum_{j=1}^\infty f_j$.

To see the last claim multiply f with a continuous function which equals 1 on C and has compact support (which exists by Urysohn's lemma). □

Note that by extending each component we can also handle functions with values in $\mathbb{R}^n$.

There is also a corresponding result for Lipschitz continuous functions.

Lemma B.29 (McShane[7]). *Let X be a metric space and $Y \subseteq X$ nonempty. Then a Lipschitz continuous functions $f : Y \to \mathbb{R}$*

$$|f(x) - f(y)| \le L\, d(x,y), \qquad x, y \in Y, \tag{B.23}$$

has a continuous extensions

$$\bar{f}(x) = \sup_{y\in Y} \big(f(y) - L\, d(x,y)\big) \tag{B.24}$$

such that $|\bar{f}(x) - \bar{f}(y)| \le L\, d(x,y)$ for all $x, y \in X$.

Proof. Since $f(y) - L\, d(x,y) \le \bar{f}(x)$ with equality for $x = y$ in case $x \in Y$ we see that $\bar{f}(x) = f(x)$ in this case. To see the Lipschitz condition choose $x, x' \in X$ and suppose $\bar{f}(x') \le \bar{f}(x)$ without loss of generality. Then, if

[6] Heinrich Tietze (1880–1964), Austrian mathematician

[7] Edward J. McShane (1904–1989), American mathematician

$\bar{f}(x') < \infty$ we have

$$\begin{aligned}\bar{f}(x) - \bar{f}(x') &= \sup_{y\in Y}\big(f(y) - L\,d(x,y)\big) - \sup_{y\in Y}\big(f(y) - L\,d(x',y)\big)\\ &\le \sup_{y\in Y}\Big(\big(f(y) - L\,d(x,y)\big) - \big(f(y) - L\,d(x',y)\big)\Big)\\ &= L\sup_{y\in Y}\big(d(x',y) - d(x,y)\big) \le L\,d(x',x).\end{aligned}$$

In particular, for $x' \in Y$ we conclude that $\bar{f}(x)$ is finite for all $x \in X$. □

Note that the proof easily extends to (e.g.) Hölder continuous functions by replacing d by d^γ in the definition of $\bar{f}$.

A **partition of unity** is a collection of functions $h_\alpha : X \to [0,1]$ such that $\sum_\alpha h_\alpha(x) = 1$. We will only consider the case when the partition of unity is **locally finite**, that is, when every x has a neighborhood where all but a finite number of the functions h_α vanish. Moreover, given a cover $\{O_\beta\}$ of X a partition of unity is called **subordinate** to this cover if every h_α has support contained in some set O_β from this cover.

In the case of subsets of $\mathbb{R}^n$ we are interested in the existence of smooth partitions 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.37. 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.30. *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.*

Proof. Let U_j be as in Lemma B.22 (iv). For the compact set $\overline{U}_j$ 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. □

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.31 (Stone[8]). *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 $d(y,z) > 2^{-i} \ge 2^{-n-m+1}$. □

Problem B.52. *Let X be a metric space and $Y \subseteq X$. 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.53. *Let X be a metric space and $Y, Z \subseteq X$. 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.54. *Let X be some normed vector space and $Y \subseteq Z \subseteq X$ with Z open. Show* $\operatorname{dist}(Y,\partial Z) = \operatorname{dist}(Y, X\setminus Z)$.

Problem B.55. *Let X be a metric space. Let $K \subseteq U \subseteq X$ with K compact and U open. Show that there is some $\varepsilon > 0$ such that $K_\varepsilon := \{x \in X | \operatorname{dist}(x,K) < \varepsilon\} \subseteq U$.*

[8] Arthur Harold Stone (1916–2000), British mathematician

Problem B.56. *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.22 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.8. Initial and final topologies

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 in Section B.4.

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. Explicitly, sets of the form

$$U = \bigcap_{j=1}^{n} \pi_{\alpha_j}^{-1}(O_{\alpha_j}) = \{x \in X | x_{\alpha_1} \in O_{\alpha_1}, \dots, x_{\alpha_n} \in O_{\alpha_n}\}, \tag{B.25}$$

where $O_{\alpha_j} \subseteq X_{\alpha_j}$ are open, are a base for the topology. Of course we could choose a base from each X_α and restrict O_{α_j} to sets from the corresponding base for X_{α_j}.

Example B.38. Note that an infinite product $\bigtimes_{\alpha \in A} O_\alpha$ of open sets $O_\alpha \subseteq X_\alpha$ will not be open in the product topology unless $O_\alpha = X_\alpha$ except for finitely many α (or one of them, and hence the entire product, equals the empty set). In particular, if infinitely many $O_\alpha \subsetneq X_\alpha$, then no open neighborhood fits into A and hence $A^\circ = \emptyset$. The corresponding statement for closed sets is true (Problem B.58). ⋄

Example B.39. 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 are close to x at a given finite number of points from A. 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 from A, 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). $\diamond$

Concerning products of compact sets we have

Theorem B.32 (Tychonoff[9]). *The product $K := \times_{\alpha\in A} K_\alpha$ of an arbitrary collection of compact topological spaces $\{K_\alpha\}_{\alpha\in A}$ is compact with respect to the product topology.*

Proof. By Lemma B.12 it suffices to show that any family $\mathcal{F}$ of closed subsets of K having the finite intersection property has nonempty intersection. Moreover, we can extend such a family to a maximal family of closed sets having the finite intersection property. To this end observe that 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 since we can add any finite intersection without affecting the finite intersection property).

Denote by $\pi_\alpha : K \to K_\alpha$ the projection onto K_α. 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 $\pi_\alpha(F)\cap F_\alpha\neq\emptyset$). 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 every $F\in\mathcal{F}_M$. Hence $x\in\overline{F}=F$ for every $F\in\mathcal{F}_M$, that is, $x\in\bigcap_{\mathcal{F}_M} F$. $\square$

The product topology 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

$$f_\alpha^{-1}(O_\alpha) = \{x\in X | x_\alpha\in O_\alpha\},$$

where $O_\alpha\subseteq Y_\alpha$ is open, known as open **cylinders**. Finite intersections of such sets,

$$\bigcap_{j=1}^{n} f_{\alpha_j}^{-1}(O_{\alpha_j}) = \{x\in X | f_{\alpha_1}(x)\in O_{\alpha_1},\dots,f_{\alpha_n}(x)\in O_{\alpha_n}\},$$

[9]Andrey Tychonoff (1906–1993), Soviet mathematician (official transliteration: Andrey Tikhonov)

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. Note also that the image of a cylinder set under f_α is open and hence the maps f_α are open (Problem B.31).

Example B.40. If X is a topological space and $Y \subseteq X$, then the relative topology on Y is the initial topology with respect to the inclusion map $j : Y \hookrightarrow X$ since $j^{-1}(A) = A \cap Y$. ⋄

The initial topology has the following characteristic property:

Lemma B.33. *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.41. 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 $\pi : X \to X/\sim$. ⋄

Example B.42. Let X_α be a collection of topological spaces. The **disjoint union**

$$X := \dot{\bigcup_{\alpha\in A}} X_\alpha$$

is usually given the final topology from the canonical injections $i_\alpha : X_\alpha \hookrightarrow X$ such that $O \subseteq X$ is open if and only if $O \cap X_\alpha$ is open for all $\alpha \in A$. ⋄

Lemma B.34. *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.57. *Let X_α be topological spaces and let $X := \bigtimes_{\alpha\in A} X_\alpha$ with the product topology. Show that the projection maps are open.*

Problem B.58. *Let X_α be topological spaces and let $X := \bigtimes_{\alpha\in A} X_\alpha$ with the product topology. Show that the product $\bigtimes_{\alpha\in A} C_\alpha$ of closed sets $C_\alpha \subseteq X_\alpha$ is closed.*

Problem B.59. *Let X_α be topological spaces and $U_\alpha \subseteq X_\alpha$. Show that $\overline{\bigtimes_{\alpha\in A} U_\alpha} = \bigtimes_{\alpha\in A} \overline{U_\alpha}$. Show that $(\bigtimes_{\alpha\in A} U_\alpha)^\circ = \bigtimes_{\alpha\in A} U_\alpha^\circ$ holds if $U_\alpha = X_\alpha$ for all but finitely many $\alpha \in A$.*

Problem B.60. *Let X_α be topological spaces. Show that $X := \bigtimes_{\alpha\in A} X_\alpha$ has the following properties if all X_α have.*

(i) *Hausdorff*

(ii) *separable*

(iii) *first/second countable*

Problem B.61. *Let $X_\alpha = [0,1]$ with the usual topology and consider $X := [0,1]^{[0,1]} := \bigtimes_{\alpha\in[0,1]} X_\alpha$ with the product topology. Show that X is Hausdorff and compact but not sequentially compact. (Hint: Consider the function $x_n(\alpha)$ with maps $\alpha \in [0,1]$ to the n'th digit of its binary expansion.)*

Problem B.62. *Let $\{(X_j, d_j)\}_{j\in\mathbb{N}}$ be a sequence of metric spaces. Show that*

$$d(x,y) := \sum_{j\in\mathbb{N}} \frac{1}{2^j} \frac{d_j(x_j,y_j)}{1+d_j(x_j,y_j)} \quad \text{or} \quad d(x,y) := \sup_{j\in\mathbb{N}} \frac{1}{2^j} \frac{d_j(x_j,y_j)}{1+d_j(x_j,y_j)}$$

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.9. 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. On $C_b(X,Y)$ we can introduce a metric via

$$d(f,g) := \sup_{x\in X} d_Y(f(x),g(x)). \tag{B.26}$$

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.43. 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 we have $d_Y(f_n(x),f(x)) < \varepsilon$ for $n \ge N$. 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.35. *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 $n \ge N$ implying $f_n^{-1}(B_{\varepsilon/2}(y)) \subseteq f^{-1}(B_\varepsilon(y))$ since $d(f_n(z),y) < \frac{\varepsilon}{2}$ implies $d(f(z),y) \le d(f(z),f_n(z)) + d(f_n(z),y) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$ for $n \ge N$. □

Corollary B.36. *Let X be a topological space and Y a complete metric space. Then $C_b(X,Y)$ is a complete metric space.*

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 $d_Y(f(x),f_n(x)) = \lim_{m\to\infty} d_Y(f_m(x),f_n(x)) \le \sup_{m\ge n} d(f_m,f_n)$ 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 $f^{-1}(V)$ will be a compact neighborhood of x. Hence without this assumption $f \in C_c(X,Y)$

must vanish at every point which does not have a compact neighborhood and $C_c(X,Y)$ will not be sufficiently rich.

Example B.44. Let X be a separable and locally compact metric space and $Y := \mathbb{C}^n$. Then

$$C_0(X,\mathbb{C}^n) = \{f \in C_b(X,\mathbb{C}^n) | \ \forall\varepsilon > 0, \exists K \subseteq X \text{ compact}: \\ |f(x)| < \varepsilon,\ x \in X \setminus K\}. \tag{B.27}$$

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.22) and let φ_m be a corresponding sequence as in Urysohn's lemma (Lemma B.26). 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 functions which vanish at the boundary (including the case as $|x| \to \infty$ if X is unbounded). ◇

Lemma B.37. *If X is a separable and locally compact metric 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 $f(x), g(x) \in I_j$. □

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

$$d_Y(f(x), f(y)) \le \varepsilon \quad \text{whenever} \quad d_X(x,y) < \delta. \tag{B.28}$$

Note that with the usual definition of continuity one fixes x and then chooses δ depending on x. Here δ has to be independent of x. Note that the uniform limit of uniformly continuous functions is again uniformly continuous (Problem B.63) and hence $C_{buc}(X,Y) \subseteq C_b(X,Y)$ is a closed subspace. If the domain is compact, this extra condition comes for free.

Theorem B.38. *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 $d_Y(f(x_n), f(y_n)) \geq \varepsilon$. 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. □

Example B.45. If X is not compact, there are bounded continuous functions which are not uniformly continuous. For example, $f(x) := \sin(x^2)$ is in $C_b(\mathbb{R})$ but not in $C_{buc}(\mathbb{R})$. ⋄

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.39. *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 $C(X,Y)$, generalizing Theorem 1.13. 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

$$d_Y(f(x), f(y)) \leq \varepsilon \quad \text{whenever} \quad y \in U(x), \quad \forall f \in F. \tag{B.29}$$

Theorem B.40 (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 bounded. If X is connected it suffices that the set $\{f(x_0) | f \in F\}$ is bounded for one $x_0 \in X$.*

Proof. Suppose F is equicontinuous and bounded for a fixed x_0. Fix $\varepsilon > 0$. By compactness of X there are finitely many points $x_1, \dots, x_n \in X$ such that the neighborhoods $U(x_j)$ (from the definition of equicontinuity) cover X. Let $U_1 = U(x_j)$ be a neighborhood which contains x_0. Then, for some fixed $y \in Y$, we have $d_Y(f(x), y) \le d_Y(f(x), f(x_j)) + d_Y(f(x_j), f(x_0)) + d_Y(f(x_0), y) \le \sup_{f\in F} d_Y(f(x_0), y) + 2\varepsilon$ for every $x \in U_1$ and every $f \in F$. Now define U_k recursively by starting with U_1 and adding to U_{k+1} all remaining neighborhoods which intersect U_k. Then we have $d_Y(f(x), y) \le \sup_{f\in F} d_Y(f(x_0), y) + 2k\varepsilon$ for every $x \in U_k$ and every $f \in F$. If X is connected, some U_k for $k \le n$ will cover all of X. Hence F is bounded with $d_Y(f(x), y) \le \sup_{f\in F} d_Y(f(x_0), y) + 2n\varepsilon$ for every $x \in X$ and every $f \in F$.

Next consider $P : C(X, Y) \to Y^n$, $P(f) = (f(x_1), \dots, f(x_n))$. Then $P(F)$ is bounded and $d(f, g) \le 3\varepsilon$ 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 $d_Y(f(x), g(x)) \le d_Y(f(x), f(x_j)) + d_Y(f(x_j), g(x_j)) + d_Y(g(x_j), g(x)) \le 3\varepsilon$. Hence F is relatively compact by Lemma 1.11.

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 $1 \le j \le n$. Then $f \in B_\varepsilon(f_j)$ for some j and hence $d_Y(f(y), f(x)) \le d_Y(f(y), f_j(y)) + d_Y(f_j(y), f_j(x)) + d_Y(f_j(x), f(x)) \le 3\varepsilon$, 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

$$p_1(t) := 0, \qquad p_{n+1}(t) := p_n(t) + \frac{t^2 - p_n(t)^2}{2}. \tag{B.30}$$

Then this sequence of polynomials satisfies $p_n(t) \le p_{n+1}(t) \le |t|$ and converges pointwise to $|t|$ for $t \in [-1,1]$ (show this). Hence by Dini's theorem (Problem B.66) it converges uniformly. By scaling we get the corresponding result for arbitrary compact subsets of the real line.

Theorem B.41 (Stone[10]–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 $x_1 \ne x_2$ there is some function $f \in F$ such that $f(x_1) \ne f(x_2)$), 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. □

Example B.46. The special case where K is a compact interval and $F := \{1, x\}$ recovers the classical Weierstraß theorem (Theorem 1.3). ⋄

Example B.47. The set $\{f \in C(K,\mathbb{R}) | f(x_0) = 0\}$ for some $x_0 \in K$ is a closed algebra which, in particular, is not dense. The same is true for the set $\{f \in C(K,\mathbb{R}) | f(x_1) = f(x_2)\}$ for some $x_1, x_2 \in K$. These examples show that the above two conditions are also necessary. ⋄

[10] Marshall H. Stone (1903–1989), American mathematician

Theorem B.42 (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.43. *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 $\{f \in C(K) | f(x_0) = 0\}$ for some $x_0 \in K$.*

Proof. There are two possibilities: either all $f \in F$ vanish at one point $x_0 \in K$ (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 point x_0 we have $\overline{A} \subseteq \{f \in C(K) | f(x_0) = 0\}$ 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 $f(x_0) = 0$ we get $f = \tilde{f} + \alpha$ with $\tilde{f} \in \overline{A}$ and $\alpha \in \mathbb{C}$. But $0 = f(x_0) = \tilde{f}(x_0) + \alpha = \alpha$ implies $f = \tilde{f} \in \overline{A}$, that is, $\overline{A} = \{f \in C(K) | f(x_0) = 0\}$. □

Problem B.63. *Show that the uniform limit of uniformly continuous functions is again uniformly continuous.*

Problem B.64. *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.65. *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*

$$d_Y(f(x), f(y)) \le \varepsilon \quad \text{whenever} \quad d_X(x,y) < \delta, \quad \forall f \in F. \tag{B.31}$$

Show that if X is compact, then a family F is pointwise equicontinuous if and only if it is uniformly equicontinuous.

Problem* B.66 (Dini[11] 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.67. *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 $z_0 \in \mathbb{C} \setminus \mathbb{R}$ 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.68. *Let $U \subseteq \mathbb{C} \setminus \mathbb{R}$ 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.69. *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 $C(K)$.*

[11] Ulisse Dini (1845–1918), Italian mathematician and politician

Bibliography

[1] H. W. Alt, *Lineare Funktionalanalysis*, 4th ed., Springer, Berlin, 2002.

[2] M. Berger and M. Berger, *Perspectives in Nonlinearity*, Benjamin, New York, 1968.

[3] A. Bowers and N. Kalton, *An Introductory Course in Functional Analysis*, Springer, New York, 2014.

[4] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, 2011.

[5] S.-N. Chow and J. K. Hale, *Methods of Bifurcation Theory*, Springer, New York, 1982.

[6] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer, New York, 1994.

[7] K. Deimling, *Nichtlineare Gleichungen und Abbildungsgrade*, Springer, Berlin, 1974.

[8] K. Deimling, *Nonlinear Functional Analysis*, Springer, Berlin, 1985.

[9] L. C. Evans, *Weak Convergence Methods for nonlinear Partial Differential Equations*, CBMS 74, American Mathematical Society, Providence, 1990.

[10] L. C. Evans, *Partial Differential Equations*, 2nd ed., American Mathematical Society, Providence, 2010.

[11] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd. ed., Wiley, Hoboken NJ, 1999.

[12] J. Franklin, *Methods of Mathematical Economics*, Springer, New York, 1980.

[13] I. Gohberg, S. Goldberg, and M.A. Kaashoek, *Basic Classes of Linear Opeartors*, Springer, Basel, 2003.

[14] G. Grubb, *Distributions and Operators*, Springer, New York, 2009.

[15] E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer, Berlin, 1965.

[16] K. Jänich, *Toplogy*, Springer, New York, 1995.

[17] I. Kaplansky, *Set Theory and Metric Spaces*, AMS Chelsea, Providence, 1972.

[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer, New York, 1966.

[19] J. L. Kelley, *General Topology*, Springer, New York, 1955.

[20] O. A. Ladyzhenskaya, *The Boundary Values Problems of Mathematical Physics*, Springer, New York, 1985.

[21] P. D. Lax, *Functional Analysis*, Wiley, New York, 2002.

[22] B. M. Levitan and V. V. Zhikov, *Almost Periodic Functions and Differential Equations*, Cambridge University Press, Cambridge, 1982.

[23] E. Lieb and M. Loss, *Analysis*, 2nd ed., Amer. Math. Soc., Providence, 2000.

[24] N. Lloyd, *Degree Theory*, Cambridge University Press, London, 1978.

[25] R. Meise and D. Vogt, *Introduction to Functional Analysis*, Oxford University Press, Oxford, 2007.

[26] F. W. J. Olver et al., *NIST Handbook of Mathematical Functions*, Cambridge University Press, Cambridge, 2010.

[27] M. Reed and B. Simon, *Methods of Modern Mathematical Physics I. Functional Analysis*, rev. and enl. edition, Academic Press, San Diego, 1980.

[28] J. R. Retherford, *Hilbert Space: Compact Operators and the Trace Theorem*, Cambridge University Press, Cambridge, 1993.

[29] J.J. Rotman, *Introduction to Algebraic Topology*, Springer, New York, 1988.

[30] H. Royden, *Real Analysis*, Prencite Hall, New Jersey, 1988.

[31] W. Rudin, *Real and Complex Analysis*, 3rd edition, McGraw-Hill, New York, 1987.

[32] M. Růžička, *Nichtlineare Funktionalanalysis*, Springer, Berlin, 2004.

[33] M. Schechter, *Principles of Functional Analysis*, 2nd ed., Amer. Math. Soc., Providence, 2002.

[34] H. Schröder, *Funktionalanalysis*, 2nd ed., Harri Deutsch Verlag, Frankfurt am Main, 2000.

[35] B. Simon, *A Comprehensive Course in Analysis*, Amer. Math. Soc., Providence, 2015.

[36] L. A. Steen and J. A. Seebach, Jr., *Counterexamples in Topology*, Springer, New York, 1978.

[37] M. E. Taylor, *Measure Theory and Integration*, Amer. Math. Soc., Providence, 2006.

[38] G. Teschl, *Mathematical Methods in Quantum Mechanics; With Applications to Schrödinger Operators*, Amer. Math. Soc., Providence, 2009.

[39] G. Teschl, *Ordinary Differential Equations and Dynamical Systems*, Amer. Math. Soc., Providence, 2012.

[40] G. Teschl, *Topics in Real Analysis*, Amer. Math. Soc., Providence, to appear.

[41] J. Weidmann, *Lineare Operatoren in Hilberträumen I: Grundlagen*, B.G.Teubner, Stuttgart, 2000.

[42] D. Werner, *Funktionalanalysis*, 7th edition, Springer, Berlin, 2011.

[43] M. Willem, *Functional Analysis*, Birkhäuser, Basel, 2013.

[44] E. Zeidler, *Applied Functional Analysis: Applications to Mathematical Physics*, Springer, New York 1995.

[45] E. Zeidler, *Applied Functional Analysis: Main Principles and Their Applications*, Springer, New York 1995.

[46] A. Zygmund, *Trigonometric series*, 3rd ed., Cambridge University Press, Cambridge, 2002.

Glossary of notation

$\arg(z)$	... argument of $z \in \mathbb{C}$; $\arg(z) \in (-\pi, \pi]$, $\arg(0) = 0$
$B_r(x)$	... open ball of radius r around x, 362
$B(X)$	... Banach space of bounded measurable functions
$\mathfrak{B}$	$= \mathfrak{B}^1$
$\mathfrak{B}^n$	... Borel σ-algebra of $\mathbb{R}^n$, see [**40**]
$\mathbb{C}$	... 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, 399
$C_c(U)$	... 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
$C_c^\infty(U)$	... set of compactly supported smooth functions
$C(U,Y)$	... set of continuous functions from U to Y, 268
$C^r(U,Y)$	... set of r times continuously differentiable functions, 274
$C_b^r(U,Y)$	... functions in C^r with derivatives bounded, 46, 279
$C_c^r(U,Y)$	... functions in C^r with compact support
$c_0(\mathbb{N})$	... set of sequences converging to zero, 12
$\mathscr{C}(X,Y)$	... set of closed linear operators from X to Y, 234
$\mathrm{CP}(f)$	... critical points of f, 313
$\mathrm{CS}(K)$	... nonempty convex subsets of K, 326
$\mathrm{CV}(f)$	... critical values of f, 313
$\chi_\Omega(.)$	... characteristic function of the set Ω
$\mathfrak{D}(.)$	... domain of an operator
$\delta_{n,m}$	... Kronecker delta, 14

$\deg(D,f,y)$	... mapping degree, 313, 320
det	... determinant
dim	... dimension of a linear space
div	... divergence of a vector filed, 319
$\mathrm{diam}(U)$	$= \sup_{(x,y)\in U^2} d(x,y)$ diameter of a set
$\mathrm{dist}(U,V)$	$= \inf_{(x,y)\in U\times V} d(x,y)$ distance of two sets
$D_y^r(\overline{U},Y)$	... functions in $C^r(\overline{U},Y)$ which do not attain y on the boundary, 313
e	... Napier's constant, $\mathrm{e}^z = \exp(z)$
$\mathrm{epi}\, F$	... epigraph of F, 295
dF	... derivative of F, 268
$\mathcal{F}(X,Y)$	... set of compact finite dimensional functions, 334
$\Phi(X,Y)$	... set of all linear Fredholm operators from X to Y, 218
$\Phi_0(X,Y)$	... set of all linear Fredholm operators of index 0, 218
$\mathrm{GL}(n)$	... general linear group in n dimensions
$\Gamma(z)$	... gamma function, see [**40**]
$\Gamma(A)$	... graph of an operator, 125
$\Gamma(f_1,\dots,f_n)$	... Gram determinant, 58
$\mathfrak{H}$	... a Hilbert space
conv(.)	... convex hull
$\mathcal{H}(U)$	... set of holomorphic functions on a domain $U\subseteq\mathbb{C}$, 311
$H^k(U)$	$= W^{k,2}(U)$, Sobolev space
$H_0^k(U)$	$= W_0^{k,2}(U)$, Sobolev space
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, 313
$\mathrm{Ker}(A)$	... kernel of an operator A, 34
$\mathscr{K}(X,Y)$	... set of compact linear operators from X to Y, 79
$\mathcal{K}(U,Y)$	... set of compact maps from U to Y, 334
$\bar{\mathcal{K}}_y(U,Y)$	... functions in $\mathcal{K}(\overline{U},Y)$ which do not attain y on the boundary, 336
λ^n	... Lebesgue measure in $\mathbb{R}^n$, see [**40**]
$\mathscr{L}(X,Y)$	... set of all bounded linear operators from X to Y, 36
$\mathscr{L}(X)$	$= \mathscr{L}(X,X)$
$L^p(X,d\mu)$	... Lebesgue space of p integrable functions, see [**40**]
$L^\infty(X,d\mu)$	... Lebesgue space of bounded functions, see [**40**]
$L^p_{loc}(X,d\mu)$	... locally p integrable functions, see [**40**]
$\mathcal{L}^2_{cont}(I)$	... space of continuous square integrable functions, 23
$\ell^p(\mathbb{N})$	... Banach space of p summable sequences, 10
$\ell^2(\mathbb{N})$	... Hilbert space of square summable sequences, 21
$\ell^\infty(\mathbb{N})$	... Banach space of bounded sequences, 11
$\ell_c(\mathbb{N})$	... sequences with finitely many nonzero terms
max	... maximum

$\mathbb{N}$	... 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, 313
$\mathrm{Ran}(A)$	... range of an operator A, 34
$\mathrm{Re}(.)$	... real part of a complex number
$R(I, X)$	... set of regulated functions, 266
$\sigma(A)$	... spectrum of an operator A, 160 ,209
S^{n-1}	$= \{x \in \mathbb{R}^n \mid \lvert x \rvert = 1\}$ unit sphere in $\mathbb{R}^n$
$\mathrm{sign}(z)$	$= z/\lvert z \rvert$ for $z \neq 0$ and 1 for $z = 0$; complex sign function
$S(I, X)$	... step functions $f : I \to X$, 266
sup	... supremum
$\mathrm{supp}(f)$	... support of a function f, 376
$\mathrm{span}(M)$	... set of finite linear combinations from M, 13
$W^{k,p}(U)$	... Sobolev space, see [**40**]
$W_0^{k,p}(U)$	... Sobolev space, see [**40**]
$\mathbb{Z}$	... 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 the operator A, 65
$\overline{A}$	... closure of the operator A, 234
$\hat{f}$	$= \mathcal{F}f$, Fourier coefficients/transform of f, 73
$\check{f}$	$= \mathcal{F}^{-1}f$, inverse Fourier transform of f
$\lvert x \rvert$	$= \sqrt{\sum_{j=1}^n \lvert x_j \rvert^2}$ Euclidean norm in $\mathbb{R}^n$ or $\mathbb{C}^n$
$\lvert \Omega \rvert$	... Lebesgue measure of a Borel set Ω
$\lVert . \rVert$	... norm, 20
$\lVert . \rVert_p$	... norm in the Banach space ℓ^p and L^p, 10, 29
$\lVert . \rVert_{k,\infty}$	... sup norm for derivatives, 46
$\lVert . \rVert_{k,\gamma,\infty}$	... Hölder norm for derivatives, 47

$\langle .,..\rangle$	... scalar product in $\mathfrak{H}$, 20
$\oplus$	... direct/orthogonal sum of vector spaces or operators, 41, 70
$\widehat{\oplus}$	... direct sum of operators with the same image space, 42
$\otimes$	... tensor product, 71
$\mathbin{\dot\cup}$	... union of disjoint sets
$\lfloor x\rfloor$	$= \max\{n \in \mathbb{Z} \vert n \le x\}$, floor function
$\lceil x\rceil$	$= \min\{n \in \mathbb{Z} \vert n \ge x\}$, ceiling function
∂	$= (\partial_1 f, \dots, \partial_m f)$ gradient in $\mathbb{R}^m$
∂_α	... partial derivative in multi-index notation
$\partial_x F(x,y)$	... partial derivative with respect to x, 273
∂U	$= \overline{U} \setminus U^\circ$ boundary of the set U, 362
$\overline{U}$	... closure of the set U, 364
U°	... interior of the set U, 364
$M^\perp$	... orthogonal complement, 59
(λ_1, λ_2)	$= \{\lambda \in \mathbb{R} \,\vert\, \lambda_1 < \lambda < \lambda_2\}$, open interval
$[\lambda_1, \lambda_2]$	$= \{\lambda \in \mathbb{R} \,\vert\, \lambda_1 \le \lambda \le \lambda_2\}$, closed interval
$x_n \to x$	... norm convergence, 9
$x_n \rightharpoonup x$	... weak convergence, 145
$x_n \overset{*}{\rightharpoonup} x$	... weak-$*$ convergence, 153
$A_n \to A$	... norm convergence of operators
$A_n \xrightarrow{s} A$	... strong convergence of operators, 150
$A_n \rightharpoonup A$	... weak convergence of operators, 150

Index