JOINT INSTITUTE
交大密西根学院

Vv556 Methods of Applied Mathematics I
Theory of Linear Operators

Horst Hohberger

University of Michigan - Shanghai Jiaotong University
Joint Institute

Fall Term 2018

Office Hours, Email, TAs

- Please read the Course Profile, which has been uploaded to the Resources section on the Canvas course site.
- My office is Room 441c in in the Longbin Building.
- My email is horst@sjtu.edu.cn and I'll try to answer email queries within 24 hours.
- Office hours will be announced on Canvas.
- Please also make use of the Discussion tab on Canvas for asking questions, making comments or giving feedback on the course.

Coursework

- There will be weekly coursework (assignments) throughout the term.
- You will be randomly assigned into ***assignment groups*** of three students; you are expected to collaborate within each group and hand in a single, common solution paper to each coursework.
- Each student must achieve 60% of the total coursework points by the end of the term in order to obtain a passing grade for the course. However, the assignment points have no effect on the course grade.
- Each member of an assignment group will receive the same number of points for each submission. However, there will be an opportunity for team members to anonymously evaluate each others' contributions to the assignments. In cases where one or more group members consistently do not contribute a commensurate share of the work, a TA will investigate the situation and individual group members may lose some or all of their marks.

Coursework

- Please hand in your coursework on time, by the date given on each set of course work. Late work will not be accepted unless you come to me personally and I find your explanation for the lateness acceptable.
- You can be deducted up to 10% of the awarded marks for an assignment if you fail to write neatly and legibly.
- You are encouraged to compose your coursework solutions in LaTeX. While this is optional, there will be a 10% bonus to the awarded marks for those assignment handed in as typed LaTeX manuscripts.

  LaTeX is open-source software for mathematical typesetting, and there are various implementations available. I suggest that you use Baidu or Google to find a suitable implementation for your computer and OS. LaTeX is widely used for writing theses and scientific papers, so it may be quite useful for you to learn it.
- Further details can be found in the course description.

Use of Wikipedia and Other Sources; Honor Code Policy

- The correct way of using outside sources is to understand the contents of your source and then to write in your own words and without referring back to the source the solution of the problem. Your solution should differ in style significantly from the published solution. If you are not sure whether you are incorporating too much material from your source in your solutions, then you must cite the source that you used.
- You may and are required to collaborate freely with other students in your assignment group. However, you may not communicate at all about concrete coursework with students from other groups. However, discussing general questions regarding the lecture contents with any other student is of course fine and encouraged.

  Do not show or explain your solutions to any student outside your assignment group.

Use of Wikipedia and Other Sources; Honor Code Policy

In this course, the following actions are examples of violations of the Honor Code ("another student" means a student outside your assignment group):

- Showing another student your written solution to a problem.
- Sending a screenshot of your solution via QQ, email or other means to another student.
- Showing another student the written solution of a third student; distributing some student's solution to other students.
- Viewing another student's written solution.
- Copying your solution in electronic form (LaTeX source, PDF, JPG image etc.) to the computer hardware (flash drive, hard disk etc.) of another student. Having another student's solution in electronic form on your computer hardware.

If you have any questions regarding the application of the Honor Code, please contact me or any of the TAs.

Grading Policy

The grade will be composed of the course work and the exams as follows:

- First midterm exam: 30 points
- Second midterm exam: 30 points
- Final exam: 40 points

Class Attendance and Absence for Medical Reasons

If you are unable to attend an exam or a class, you should notify me. The following rules apply:

- Absence for illness should be supported by a hospital/doctor's certificate. A note that a student visited a medical facility is not sufficient excuse for missing a Tuesday class or an exam. The note must specifically indicate that the student was incapable of attending a class or taking the exam due to medical problems.
- **Late** medical excuses must satisfy the following criteria to be valid:
  - (i) The problem must be confirmed by the doctor to be so severe that the student could not participate in the exam.
  - (ii) The problem must have occurred so suddenly that it was impractical to contact me in advance.
  - (iii) The student must be in contact with me immediately after the exam or Tuesday class with the required documentation.

Literature

We will use various textbook sources for the course. These include

- E. Kreyszig, ***Introductory Functional Analysis with Applications***, Wiley 1989;
- K. Jänich, ***Linear Algebra***, Springer 1994.
- S. Lang, ***Linear Algebra***, 2nd Ed., Addison-Wesley 1972.
- I. Stakgold and M. Holst, ***Green's Functions and Boundary Value Problems***, 3rd Ed., Wiley 2011.

Part I

Infinite-Dimensional Vector Spaces

Introduction

Introduction

Under certain assumptions (small displacements, no external forces) the transversal displacement of a vibrating string of length L satisfies the wave equation

$$c^2 u_{xx} - u_{tt} = 0, \qquad 0 < x < L, \ t \in \mathbb{R}, \tag{1.1.1}$$

where $c > 0$ is related to the tension and the density of the string. If the ends of the string are fixed, the boundary conditions

$$u(0, t) = u(L, t) = 0, \qquad t \in \mathbb{R}$$

may be imposed. Solutions of the wave equation then have the form

$$u_n(x, t) = \big(\alpha_n \sin(n\pi ct/L) + \beta_n \cos(n\pi ct/L)\big) \sin(n\pi x/L), \qquad n \in \mathbb{N},$$

where $\alpha_n, \beta_n \in \mathbb{R}$ are constants determined by the initial displacement and velocity of the string.

Introduction

Hence, there is a (countably) infinite family of solutions. It also turns out that there are no other solutions! From the linearity of (1.1.1) it is clear that any sum of the u_n is again a solution. However, experiments show that not all displacements of strings take the form of finite linear combinations of trigonometric functions. In fact, if the string is initially displaced according to

$$u(x,0) = x(x-L), \qquad u_t(x,0) = 0.$$

can we write

$$u(x,t) = \sum_{n=0}^{\infty} \big(\alpha_n \sin(n\pi ct/L) + \beta_n \cos(n\pi ct/L)\big) \sin(n\pi x/L)$$

for certain coefficients α_n, β_n? If so, we would have

$$u(x,0) = x(x-L) = \sum_{n=0}^{\infty} \beta_n \sin(n\pi x/L)$$

Introduction

In what sense can such an identity hold? What types of functions are permissible for $u(x, 0)$? Any continuous function? Perhaps even discontinuous functions?

The same question arises in other differential equations involving Bessel functions, Legendre polynomials and other exotic functions in place of the trigonometric sine and cosines. Therefore, it makes sense to tackle this question from a fundamental point of view:

Can any continuous function be written as an "infinite linear combination" of a given set of functions? In other words, does there exist an algebraic basis for the vector space of continuous functions? Such a basis would of course need to be infinite in size.

We will extend the known methods of linear algebra to tackle this and other, related questions.

Prerequisites from Linear Algebra

We assume that the following concepts from linear algebra are familiar:

- Vector spaces, norms, scalar products
- Linear independence of vectors, bases, dimension of vector spaces
- Linear maps, matrices, determinants
- Eigenvalue problems for matrices

We may briefly recall some of the definitions, but for details we refer to the literature.

Vector Spaces

1.1.1. Definition. A vector space over a field $\mathbb{F}$ (we only consider $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is a triple $(V, +, \cdot)$ where

(i) V is any set;

(ii) $+ \colon V \times V \to V$ is a map (called addition) with the following properties:
- $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$ (***associativity***),
- $u + v = v + u$ for all $u, v \in V$ (***commutativity***),
- there exists an element $e \in V$ such that $v + e = v$ for all $v \in V$ (***existence of a unit element***),
- for every $v \in V$ there exists an element $-v \in V$ such that $v + (-v) = e$;

(iii) $\cdot \colon \mathbb{F} \times V \to V$ is a map (called scalar multiplication) with the following properties:
- $1 \cdot u = u$ for all $u \in V$,
- $\lambda \cdot (u + v) = \lambda \cdot u + \lambda \cdot v$ for all $\lambda \in \mathbb{F}$, $u, v \in V$,
- $(\lambda + \mu) \cdot u = \lambda \cdot u + \mu \cdot u$ for all $\lambda, \mu \in \mathbb{F}$, $u \in V$,
- $(\lambda\mu) \cdot u = \lambda \cdot (\mu \cdot u)$ for all $\lambda, \mu \in \mathbb{F}$, $u \in V$.

Subspaces

1.1.2. Definition. Suppose that $(V, +, \cdot)$ is a vector space and $U \subset V$. If $(U, +, \cdot)$ is also a vector space, we say that U is a ***linear subspace*** or ***sub-vectorspace*** of V.

1.1.3. Notation.

- If the definition of the addition and scalar multiplication in a vector space is clear from the context, we write simply V instead of $(V, +, \cdot)$.
- If $(U, +, \cdot)$ is a subspace of $(V, +, \cdot)$, we write $(U, +, \cdot) \subset (V, +, \cdot)$ or just $U \subset V$. Hence, "$U \subset V$" can indicate either that U (as a set) is a subset of the set V or that U (as a vector space) is a subspace of the space V. This ambiguity should not cause any difficulty.

Vector Spaces

1.1.4. Examples.

(i) The set of n-tuples

$$\mathbb{F}^n := \{x = (x_1, \dots, x_n) \colon x_1, \dots, x_n \in \mathbb{F}\}$$

is a vector space with the so-called ***component-wise*** addition and scalar multiplication:

$$\begin{aligned} x + y &= (x_1, \dots, x_n) + (y_1, \dots, y_n) \\ &:= (x_1 + y_1, \dots, x_n + y_n), \\ \lambda x &= \lambda(x_1, \dots, x_n) \\ &:= (\lambda x_1, \dots, \lambda x_n) \end{aligned}$$

for all $x, y \in \mathbb{F}^n$ and $\lambda \in \mathbb{F}$. We will use both

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad x = (x_1, \dots, x_n) \quad \text{interchangeably.}$$

Vector Spaces

(ii) The set of polynomials of degree at most $n \in \mathbb{N}$,

$$\mathscr{P}_n = \left\{ f : \mathbb{F} \to \mathbb{F} : f(x) = \sum_{k=0}^{n} a_k x^k, \ \ a_0, a_1 \ldots, a_n \in \mathbb{R} \right\}$$

is a vector space with the so-called ***point-wise*** addition and scalar multiplication: for polynomials p and q given by $p(x) = a_0 + a_1 x + \cdots + a_n x^n$, $q(x) = b_0 + b_1 x + \cdots + b_n x^n$ and $\lambda \in \mathbb{F}$ we define polynomials $p + q$ and λp through

$$(p+q)(x) := p(x) + q(x) = \sum_{k=0}^{n} (a_k + b_k) x^k,$$

$$(\lambda p)(x) := \lambda \cdot p(x) = \sum_{k=0}^{n} (\lambda a_k) x^k.$$

Vector Spaces

(iii) The set of polynomials of any degree,

$$\mathscr{P} = \left\{ f : \mathbb{F} \to \mathbb{F} : f \in \mathscr{P}_n \text{ for some } n \in \mathbb{N} \right\}$$

is a vector space with point-wise addition and scalar multiplication. In particular, the sum of two polynomials of degree m and n, respectively, is a polynomial of degree $\max(n, m)$.

(iv) The set of complex-valued continuous functions on a subset $\Omega \subset \mathbb{F}^n$,

$$C(\Omega, \mathbb{F}) = \{ f : \Omega \to \mathbb{C} : f \text{ is continuous on } \Omega \}$$

is a vector space with point-wise addition and scalar multiplication. We often write $C(\Omega)$ to abbreviate $C(\Omega, \mathbb{C})$.

(v) For any $n \in \mathbb{N}$, $\mathscr{P}_n \subset \mathscr{P} \subset C(\mathbb{F}, \mathbb{F})$. Note that in all three spaces, the same addition and scalar multiplication (i.e., pointwise) is used.

Linear Combinations and Span

1.1.5. Definition. Let $x_1, \dots, x_n \in V$ and $\lambda_1, \dots, \lambda_n \in \mathbb{F}$. Then the expression

$$\sum_{k=1}^{n} \lambda_k x_k = \lambda_1 x_1 + \cdots + \lambda_n x_n$$

is called a ***linear combination*** of the vectors $x_1, \dots, x_n$.

The set

$$\operatorname{span}\{x_1, \dots, x_n\} = \left\{ y \in V : y = \sum_{k=1}^{n} \lambda_k x_k,\ \lambda_1, \dots, \lambda_n \in \mathbb{F} \right\}$$

is called the (***linear***) ***span*** of the vectors $x_1, \dots, x_n$.

Span of Subsets

More generally, if V is a vector space and M is some subset of V, then we can define the ***span of*** M as the set containing all (finite) linear combinations of elements of M, i.e.,

$$\operatorname{span} M := \left\{ v \in V : \exists_{n \in \mathbb{N}} \exists_{\lambda_1, \ldots, \lambda_n \in \mathbb{F}} \exists_{m_1, \ldots, m_n \in M} : v = \sum_{i=1}^{n} \lambda_i m_i \right\}.$$

Note that this definition does not presume that M is a subspace, just an arbitrary subset of V. Furthermore, although only finite linear combinations are considered, the set M may well be infinite in size.

Moreover, even though M is just any set, $\operatorname{span} M$ will be a subspace of V.

1.1.6. Example. Let $M = \{f \in C(\mathbb{R}, \mathbb{R}) : f(x) = x^n,\ n \in \mathbb{N}\}$ denote the set of all monomials in the space of continuous functions on $\mathbb{R}$. Then $\operatorname{span} M = \mathscr{P}$, the space of polynomials.

Linear Independence

1.1.7. Definition. Let V be a real or complex vector space and $x_1, \ldots, x_n \in V$. Then the vectors $x_1, \ldots, x_n$ are said to be ***independent*** if for all $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$,

$$\sum_{k=1}^{n} \lambda_k x_k = 0 \quad \Rightarrow \quad \lambda_1 = \lambda_2 = \cdots = \lambda_n = 0.$$

We say that a finite set $M \subset V$ is an ***independent set*** if the elements of M are independent.

Bases and the Standard Basis of $\mathbb{F}^n$

1.1.8. Definition. Let V be a real or complex vector space. An n-tuple $\mathscr{B} = (b_1, \dots, b_n) \in V^n$ is called an ***(ordered and finite) basis*** of V if every vector v has a unique representation

$$v = \sum_{i=1}^{n} \lambda_i b_i, \qquad \lambda_i \in \mathbb{F}.$$

The numbers λ_i are called the ***coordinates*** of v with respect to $\mathscr{B}$.

1.1.9. Example. The tuple of vectors $(e_1, \dots, e_n)$, $e_i \in \mathbb{R}^n$,

$$e_i = (0, \dots 0, \underset{\substack{\uparrow \\ i\text{th} \\ \text{entry}}}{1}, 0, \dots, 0), \qquad i = 1, \dots, n,$$

is called the ***standard basis*** or ***canonical basis*** of $\mathbb{R}^n$.

Characterization of Bases

Sometimes we are not interested in the order of the elements of a basis, and write $\mathscr{B} = \{b_1, \dots, b_n\}$, replacing the tuple by a set. This is known as an ***unordered basis***.

1.1.10. Theorem. Let V be a real or complex vector space. An n-tuple $\mathscr{B} = (b_1, \dots, b_n) \in V^n$ is a finite basis of V if and only if

(i) the vectors $b_1, \dots, b_n$ are linearly independent, i.e., $\mathscr{B}$ is an independent set, and

(ii) $V = \operatorname{span} \mathscr{B}$.

Finite- and Infinite-Dimensional Spaces

1.1.11. Definition. A vector space V is said to be ***finite-dimensional*** if either

- $V = \{0\}$ or
- V possesses a finite basis.

If V is not finite-dimensional, we say that it is ***infinite-dimensional***.

1.1.12. Example.

(i) The space of polynomials of degree at most n, $\mathscr{P}_n$, is finite-dimensional, because it has the basis $\mathscr{B} = (1, x, x^2, \dots, x^n)$.

(ii) The space of polynomials of any degree, $\mathscr{P}$, is infinite-dimensional.

We omit the proof of the following theorem, which is usually shown in elementary courses on linear algebra:

1.1.13. Theorem. Let V be a real or complex finite-dimensional vector space, $V \neq \{0\}$. Then any basis of V has the same length (number of elements) n.

Dimension

1.1.14. Definition. Let V be a finite-dimensional real or complex vector space.

1. If $V = \{0\}$, we define the dimension of V, to be zero and write $\dim V = 0$.
2. If $V \neq \{0\}$, we define $\dim V = n$, where n is the length of any basis of V.

If V is an infinite-dimensional vector space we write $\dim V = \infty$.

1.1.15. Examples.

1. $\dim \mathbb{R}^n = n$
2. $\dim \mathscr{P}_n = n + 1$
3. $\dim \mathscr{P} = \infty$
4. $\dim\{(x_1, x_2) \in \mathbb{R}^2 : x_2 = 3x_1\} = 1$

Infinite-Dimensional Spaces of Functions

A typical undergraduate course in linear algebra will focus on finite-dimensional spaces; these are particularly simple, because they can be fully characterized by finite bases. However, many vector spaces of functions are of great practical interest and infinite-dimensional. These include ($\Omega \subset \mathbb{F}^n$):

- The space of continuous functions, $C(\Omega, \mathbb{C})$,
- The space of polynomials over the real or complex numbers, $\mathscr{P}(\mathbb{F})$.
- The space of bounded functions,

$$L^\infty(\Omega) := \left\{ f : \Omega \to \mathbb{C} : \sup_{x \in \Omega} |f(x)| < \infty \right\},$$

- The spaces of p-integrable functions, $p \geq 1$,

$$L^p(\Omega) := \left\{ f : \Omega \to \mathbb{C} : \int_\Omega |f(x)|^p < \infty \right\}.$$

Infinite-Dimensional Spaces of Sequences

We use the notation $x = (x_n)_{n\in\mathbb{N}}$ or simply (x_n) to denote a sequence of elements $x_0, x_1, x_2, \ldots$. In the following examples, each $x_n \in \mathbb{C}$:

- The space of null sequences,

$$c_0 = \{(x_n) : x_n \to 0 \text{ as } n \to \infty\}$$

- The space of bounded sequences,

$$\ell^\infty := \left\{(x_n) : \sup_{n\in\mathbb{N}} |x_n| < \infty\right\}$$

- The spaces of p-summable sequences, $p \geq 1$,

$$\ell^p = \left\{(x_n) : \sum_{n\in\mathbb{N}} |x_n|^p < \infty\right\}.$$

Pointwise Addition and Scalar Multiplication

For the sets defined above to become vector spaces, we need to introduce the operations of addition and mutliplication with scalars. The standard definitions are called ***point-wise*** operations, as we simply add the values of the functions or sequences at each point:

$$(f+g)(x) := f(x) + g(x), \qquad (\lambda f)(x) := \lambda \cdot f(x),$$
$$(a+b)_n := a_n + b_n, \qquad (\lambda a)_n := \lambda \cdot a_n.$$

for functions f, g, sequences $(a_n), (b_n)$ and scalars $\lambda \in \mathbb{F}$.

It is not immediately obvious that ℓ^p and $L^p(\Omega)$ are vector spaces; is the sum of two elements again an element of the space? To verify this, we need some preliminary results.

Hölder's Inequality

We begin with an interesting inequality of real numbers:

1.1.16. Lemma. Fix $1 < p < \infty$ and q such that $1/p + 1/q = 1$ and let $a, b \geq 0$. Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{1.1.2}$$

Proof.

To prove (1.1.2), consider the graph of the function

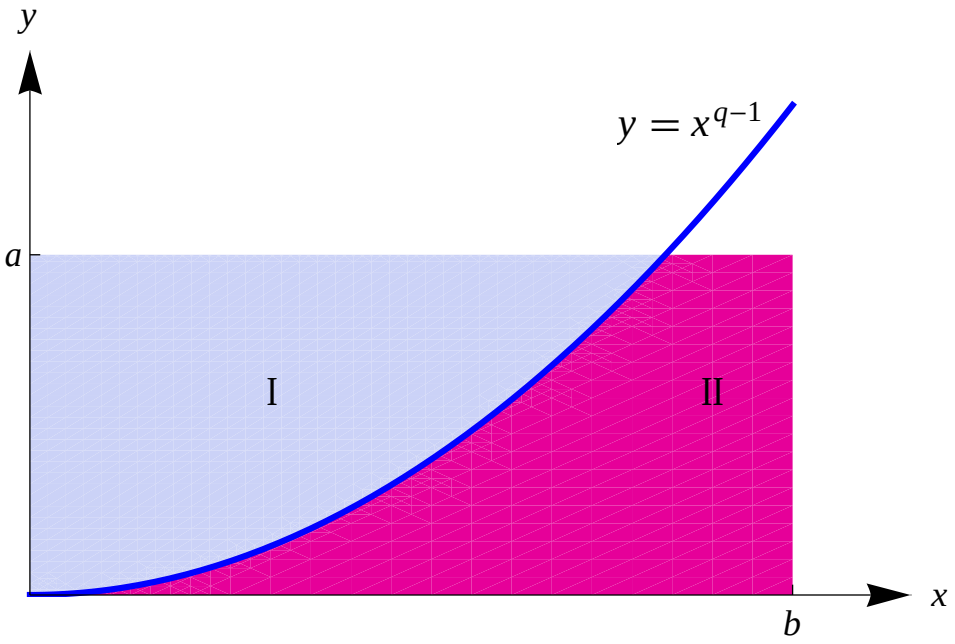
$$y = F(x) = x^{q-1}.$$

Since $(p-1)(q-1) = 1$, this is equivalent to

$$x = F^{-1}(y) = y^{p-1}.$$

Hölder's Inequality

Proof (continued).



From the graph above it is clear that

$$\text{Area(I)} = \int_0^a y^{p-1}\,dy = \frac{a^p}{p}, \qquad \text{Area(II)} \le \int_0^b x^{q-1}\,dx = \frac{b^q}{q},$$

and $\text{Area(I)} + \text{Area(II)} = ab$, so (1.1.2) is proven. □

Hölder's Inequality

We can now prove the ***Hölder inequality for series***:

1.1.17. Hölder's Inequality. Fix $1 < p < \infty$ and q such that $1/p + 1/q = 1$. Suppose that $x \in \ell^p$ and $y \in \ell^q$. Then

$$\sum_{n=0}^{\infty} |x_n y_n| \leq \left(\sum_{n=0}^{\infty} |x_n|^p \right)^{1/p} \left(\sum_{n=0}^{\infty} |y_n|^q \right)^{1/q}, \tag{1.1.3}$$

where $x = (x_n)$, $y = (y_n)$.

For $p = q = 2$ the inequality (1.1.3) is called the ***Cauchy-Schwartz inequality for series***:

$$\sum_{n=0}^{\infty} |x_n y_n| \leq \sqrt{\sum_{n=0}^{\infty} |x_n|^2} \sqrt{\sum_{n=0}^{\infty} |y_n|^2} \tag{1.1.4}$$

Hölder's Inequality

Proof.

Suppose $x \in \ell^p$ and $y \in \ell^q$ are given. Define

$$\widetilde{x} := \frac{1}{\left(\sum_{n=0}^{\infty}|x_n|^p\right)^{1/p}}x, \qquad \widetilde{y} := \frac{1}{\left(\sum_{n=0}^{\infty}|y_n|^q\right)^{1/q}}y,$$

so that $\sum_{n=0}^{\infty}|\widetilde{x}_n|^p = \sum_{n=0}^{\infty}|\widetilde{y}_n|^q = 1$. By (1.1.2),

$$|\widetilde{x}_n\widetilde{y}_n| \leq \frac{|\widetilde{x}_n|^p}{p} + \frac{|\widetilde{y}_n|^q}{q}$$

which implies

$$\sum_{n=0}^{\infty}|\widetilde{x}_n\widetilde{y}_n| \leq \frac{1}{p}\underbrace{\sum_{n=0}^{\infty}|\widetilde{x}_n|^p}_{=1} + \frac{1}{q}\underbrace{\sum_{n=0}^{\infty}|\widetilde{y}_n|^q}_{=1} = \frac{1}{p} + \frac{1}{q} = 1.$$

Minkowski's Inequality

Proof (continued).
Since

$$\sum_{n=0}^{\infty}|\widetilde{x}_n\widetilde{y}_n| = \frac{1}{\left(\sum_{n=0}^{\infty}|x_n|^p\right)^{1/p}}\frac{1}{\left(\sum_{n=0}^{\infty}|y_n|^q\right)^{1/q}}\sum_{n=0}^{\infty}|x_n y_n|$$

the proof is complete. □

We next prove the Minkowski inequality for series:

1.1.18. Minkowski's Inequality. For $1 \le p < \infty$ and $x, y \in \ell^p$ we have

$$\left(\sum_{n=0}^{\infty}|x_n + y_n|^p\right)^{1/p} \le \left(\sum_{n=0}^{\infty}|x_n|^p\right)^{1/p} + \left(\sum_{n=0}^{\infty}|y_n|^p\right)^{1/p}, \tag{1.1.5}$$

where $x = (x_n)$ and $y = (y_n)$.

Minkowski's Inequality

Proof.

We write $z_n := x_n + y_n$ for short and consider $p > 1$ (the case $p = 1$ is trivial). Then

$$|z_n|^p = |x_n + y_n| \cdot |z_n|^{p-1} \leq (|x_n| + |y_n|)|z_n|^{p-1}.$$

We can apply this inequality to finite sums, obtaining

$$\sum_{n=0}^{N} |z_n|^p \leq \sum_{n=0}^{N} |x_n| \cdot |z_n|^{p-1} + \sum_{n=0}^{N} |y_n| \cdot |z_n|^{p-1}$$

for any $N \in \mathbb{N}$. For $q = p/(p-1)$, Hölder's inequality gives

$$\sum_{n=0}^{N} |x_n| \cdot |z_n|^{p-1} \leq \left(\sum_{n=0}^{N} |x_n|^p\right)^{1/p} \left(\sum_{n=0}^{N} |z_n|^p\right)^{1/q}.$$

Minkowski's Inequality

Proof (continued).

Similarly,

$$\sum_{n=0}^{N} |y_n| \cdot |z_n|^{p-1} \leq \left(\sum_{n=0}^{N} |y_n|^p\right)^{1/p} \left(\sum_{n=0}^{N} |z_n|^p\right)^{1/q},$$

so

$$\left(\sum_{n=0}^{N} |z_n|^p\right)^{1-1/q} \leq \left(\sum_{n=0}^{N} |x_n|^p\right)^{1/p} + \left(\sum_{n=0}^{N} |y_n|^p\right)^{1/p}$$

Since $1 - 1/q = 1/p$, we obtain the desired inequality by letting $N \to \infty$. Since both series on the right converge by assumption, so does the series on the left. $\square$

Function and Sequence Spaces

Minkowski's inequality immediately yields that if $x, y \in \ell^p$, then $x + y \in \ell^p$. Since clearly $\lambda x \in \ell^p$ for any $\lambda \in \mathbb{F}$, we see that ℓ^p is indeed a vector space. Analogous inequalities can be used to show that the sets L^p are vector spaces.

In fact, the spaces ℓ^p and L^p are structurally very similar; while we are mostly interested in function spaces in applications, the sequence spaces are easier to discuss technically (the summation of a sequence is simpler than the integration of a function) and we will use the ℓ^p spaces to illustrate many basic ideas.

It is easy to see that the function and sequence spaces we have introduced can not have a finite basis and are infinite-dimensional. For further investigations, it becomes necessary to introduce some more structure: we need the concept of the size of a vector, generalizing the modulus of complex numbers. Such a generalization is called a norm.

Normed Vector Spaces

1.2.1. Definition. Let V be a real or complex vector space. Then a map $\|\cdot\|: V \to \mathbb{R}$ is called a ***norm*** on V if for all $u, v \in V$ and all $\lambda \in \mathbb{F}$,

(i) $\|v\| \geq 0$ for all $v \in V$ and $\|v\| = 0$ if and only if $v = 0$,

(ii) $\|\lambda \cdot v\| = |\lambda| \cdot \|v\|$,

(iii) $\|u + v\| \leq \|u\| + \|v\|$ (triangle inequality).

The pair $(V, \|\cdot\|)$ is called a ***normed vector space*** or a ***normed linear space***.

1.2.2. Examples.

- In $\mathbb{F}^n$, we can define norms by

$$\|x\|_p := \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}, \qquad x = (x_1, \dots, x_n) \in \mathbb{F}^n,$$

for any $p \geq 1$. Minkowski's inequality (1.1.5) is then simply the triangle inequality for the norm. The case $p = 2$ corresponds to the ***euclidean*** or ***canonical*** norm.

Normed Vector Spaces

- More generally, in each space ℓ^p, $p \geq 1$, a nor is defined by

$$\|x\|_p := \left(\sum_{k=1}^{\infty} |x_k|^p\right)^{1/p}, \qquad x = (x_0, x_1, x_2, \ldots) \in \ell^p,$$

- In any ℓ^p, $1 \leq p \leq \infty$, we can define a norm by

$$\|x\|_\infty := \sup_{n \in \mathbb{N}} |x_k|.$$

- Similarly, in $C(\Omega, \mathbb{C})$ for a bounded set $\Omega \in \mathbb{F}^n$ we can define the norms $(1 \leq p < \infty)$

$$\|f\|_p := \left(\int_\Omega |f(x)|^p \, dx\right)^{1/p}, \qquad \|f\|_\infty := \sup_{x \in \Omega} |f(x)|.$$

Open Balls, Open and Closed Sets

Having introduced a norm, we can define open balls and open sets:

1.2.3. Definition. Let $(V, \|\cdot\|)$ be a normed vector space. Then

(i) For $r > 0$ and $x \in V$, the set

$$B_r(x) := \{y \in V : \|x - y\| < r\}$$

is called an ***open ball of radius r centered at x***.

(ii) A set $\Omega \subset V$ is said to be ***open*** if for every $x \in \Omega$ there exists an $\varepsilon > 0$ such that $B_\varepsilon(x) \subset \Omega$.

(iii) A set $\Omega \subset V$ is said to be ***closed*** if $\Omega^c := V \setminus \Omega$ is open.

1.2.4. Examples.

- For any fixed point $x \in V$ and any $r > 0$, the open ball $B_r(x)$ is an open set.
- The empty set $\emptyset$ is (vacuously) open.
- The entire space V is open.

Boundary Points and Closure

- The empty set $\emptyset$ is closed.
- The entire space V is closed.
- For any $x \in V$ the one-element set $\{x\} \subset V$ is closed.

We see that a set can be open, closed, both open and closed, or neither open nor closed.

1.2.5. Definition and Theorem. Let $(V, \|\cdot\|)$ be a normed vector space and $\Omega \subset V$.

(i) A point $x \in V$ is a ***boundary point*** of Ω if for any $\varepsilon > 0$,

$$B_\varepsilon(x) \cap \Omega \neq \emptyset \qquad \text{and} \qquad B_\varepsilon(x) \cap \Omega^c \neq \emptyset.$$

The set of boundary points is denoted by $\partial\Omega$.

(ii) The ***closure*** of Ω is defined as

$$\overline{\Omega} := \Omega \cup \partial\Omega.$$

The closure is the smallest closed set that contains Ω. In particular, Ω is closed if and only if it contains its boundary points.

Boundary Points and Closure

Proof.

Suppose that Ω is closed. Then Ω^c is open and can not contain a boundary point, since any point of the complement must have the property that a sufficiently small ε-ball (ball of radius ε) centered at that point is contained fully in the complement. Therefore, $\partial\Omega \subset \Omega$.

Conversely, suppose $\partial\Omega \subset \Omega$. Then for any point $x \notin \Omega$ it is true that for any $\varepsilon > 0$ the ball $B_\varepsilon(x)$ intersects the complement Ω^c. But since such a point can not be a boundary point, there must be an ε_0 such that $B_{\varepsilon_0}(x) \cap \Omega = \emptyset$, i.e., $B_{\varepsilon_0}(x) \subset \Omega^c$. Since this is true for any $x \in \Omega^c$, the complement of Ω is open and Ω is closed. □

The following lemma is an immediate result of the definitions and the proof is left to the reader:

1.2.6. Corollary. Let $(V, \|\cdot\|)$ be a normed vector space and $\Omega \subset V$. Then $x \in \overline{\Omega}$ if and only if $B_\varepsilon(x) \cap \Omega \neq \emptyset$ for any $\varepsilon > 0$.

Sequences in Vector Spaces

Recall that a sequence of complex numbers (x_n) converges to a limit x if and only if

$$\forall_{\varepsilon>0}\ \exists_{N\in\mathbb{N}}\ \forall_{n>N}\quad |x_n - x| < \varepsilon.$$

An analogous definition can be used for sequences in normed vector spaces:

1.2.7. Definition. A sequence in a normed vector space $(V, \|\cdot\|)$ is a map $(x_n)\colon \mathbb{N} \to V$.

(i) We say that (x_n) ***converges*** to an element $x \in V$ if

$$\forall_{\varepsilon>0}\ \exists_{N\in\mathbb{N}}\ \forall_{n>N}\quad \|x_n - x\| < \varepsilon$$

(ii) We say that (x_n) is a ***Cauchy sequence*** if

$$\forall_{\varepsilon>0}\ \exists_{N\in\mathbb{N}}\ \forall_{m,n>N}\quad \|x_n - x_m\| < \varepsilon$$

Sequences of Continuous Functions

1.2.8. Example. Consider $C([0,1])$, the vector space of continuous functions on the interval $[0,1] \subset \mathbb{R}$, imbued with the norm

$$\|f\|_\infty := \sup_{x\in[0,1]} |f(x)|, \qquad f \in C([0,1]).$$

Then a sequence in $C([0,1])$ is a sequence of functions (f_n) and $f_n \to f$ if

$$\forall_{\varepsilon>0}\ \exists_{N\in\mathbb{N}}\ \forall_{n>N} \quad \underbrace{\sup_{x\in[0,1]} |f_n(x) - f(x)|}_{=\|f_n - f\|_\infty} < \varepsilon.$$

In calculus, one says that (f_n) converges to f ***uniformly***. This is just the ordinary norm convergence in the vector space.

Uniform convergence is in contrast to ***point-wise convergence***: $f_n \to f$ pointwise if $f_n(x) \to f(x)$ as $n \to \infty$ for all x.

Sequences of Continuous Functions

For example, the sequence of functions (f_n) given by

$$f_n(x) = \begin{cases} 1 - 2n|x - 1/2n| & 0 \le x < 1/n, \\ 0 & 1/n \le x \le 1, \end{cases}$$

does not converge to $f(x) = 0$ uniformly, even though $\lim_{n\to\infty} f_n(x) = 0$ for all $x \in [0, 1]$. This is because

$$\|f_n - f\|_\infty = \sup_{x\in[0,1]} |f_n(x) - f(x)| = 1 \not\to 0$$

as $n \to \infty$.

An alternative norm on $C([0, 1])$ is given by

$$\|f\|_1 := \int_0^1 |f(x)|\, dx, \qquad f \in C([0, 1]).$$

Sequences of Continuous Functions

We say that a sequence of functions (f_n) ***converges in the mean*** to a function f if

$$\underset{\varepsilon>0}{\forall}\ \underset{N\in\mathbb{N}}{\exists}\ \underset{n>N}{\forall} \quad \int_0^1 |f_n(x) - f(x)|\, dx < \varepsilon.$$

Continuing our example, we see that

$$\begin{aligned}\|f_n - f\|_1 = \int_0^1 |f_n(x) - f(x)|\, dx &= \int_0^{1/n} (1 - 2n|x - 1/2n|)\, dx \\ = \frac{1}{n} - \frac{2n}{2}\frac{1}{n^2} &= \frac{1}{2n} \xrightarrow{n\to\infty} 0.\end{aligned}$$

We see that the convergence of a sequence in a vector space in general depends on the norm that is used!

Sequences of Continuous Functions

We finally remark that since

$$\|f\|_1 = \int_0^1 |f(x)|\,dx \leq \sup_{x\in[0,1]} |f(x)| = \|f\|_\infty$$

uniform convergence implies convergence in the mean, i.e.,

$$\|f_n - f\|_\infty \to 0 \qquad \Rightarrow \qquad \|f_n - f\|_1 \to 0$$

Our example has shown that the converse is false, i.e., mean convergence does not imply uniform convergence.

In the same way, (f_n) converges to f uniformly only if $f_n \to f$ pointwise, i.e.,

$$\|f_n - f\|_\infty \to 0 \qquad \Rightarrow \qquad |f_n(x) - f(x)| \to 0 \quad \text{for all } x.$$

Sequences and Closed Sets

We may also use sequences to give a more direct description of closed sets:

1.2.9. Theorem. Let $(V, \|\cdot\|)$ be a normed vector space, $\Omega \subset V$ non-empty and $\overline{\Omega}$ the closure of Ω.

(i) Then $x \in \overline{\Omega}$ if and only if there exists a sequence (x_n) with $x_n \in \Omega$ for all $n \in \mathbb{N}$ and $x_n \to x$ as $n \to \infty$.

(ii) The set Ω is closed if and only if the limit $x \in V$ of all convergent sequences (x_n) with $x_n \in \Omega$ for all $n \in \mathbb{N}$ actually lies in Ω.

Proof.

(i) By Corollary 1.2.6, $x \in \overline{\Omega}$ if and only if for every n there exists an $x_n \in \Omega$ such that $x_n \in B_{1/n}(x)$. These points define a sequence (x_n) such that $\|x_n - x\| < 1/n$, i.e., a sequence converging to x.

(ii) By (i), $x \in \overline{\Omega}$. Suppose that Ω is closed. Then $\Omega = \overline{\Omega}$ and $x \in \Omega$. If Ω is not closed, there exists a boundary point y not in Ω. But then we can find a sequence (x_n) with $x_n \in \Omega$ that converges to y. □

An Example of a Subspace that is Not Closed

1.2.10. Example. Consider $(C([0,1]), \|\cdot\|_\infty)$, the space of continuous functions on the interval $[0,1]$ imbued with the supremum-norm. Then $\mathscr{P}([0,1])$, the set of polynomials on the interval $[0,1]$, is a subspace of $C([0,1])$. As a set, $\mathscr{P}([0,1])$ is not closed in $C([0,1])$.

We can see this easily, since the sequence (p_n) of polynomials given by

$$p_n(x) = \sum_{k=0}^{n} \left(-\frac{1}{2}\right)^k x^k$$

converges to the function $f \notin \mathscr{P}([0,1])$ given by $f(x) = \frac{1}{1+x/2}$:

$$\left\| \frac{1}{1+x/2} - \sum_{k=0}^{n} \left(-\frac{1}{2}\right)^k x^k \right\|_\infty = \sup_{x\in[0,1]} \left| \frac{1}{1+x/2} - \frac{1-(-x/2)^{n+1}}{1+x/2} \right|$$

$$\leq \sup_{x\in[0,1]} \left(\frac{x}{2}\right)^n \xrightarrow{n\to\infty} 0.$$

Hence, by Theorem 1.2.9 (ii), $\mathscr{P}([0,1])$ is not closed.

Dense Sets in Normed Vector Spaces

1.2.11. Definition. Let $(V, \|\cdot\|)$ be a normed vector space and $\Omega \subset V$. Then Ω is said to be ***dense*** in V if $\overline{\Omega} = V$.

1.2.12. Remark. From Corollary 1.2.6 it follows that Ω is dense in V if and only if for any $x \in V$ and any $\varepsilon > 0$ there exists an $y \in \Omega$ such that $\|x - y\| < \varepsilon$. We say that for any $x \in V$ there is an arbitrarily close $y \in \Omega$.

1.2.13. Example. The set of rational numbers $\Omega = \mathbb{Q}$ is dense in the set of real numbers $V = (\mathbb{R}, |\cdot|)$: for any real number x and for any $\varepsilon > 0$ we can find a rational number y such that $|x - y| < \varepsilon$. This is achieved, e.g., by truncating the decimal expansion of x at a suitable position and letting y be this truncated number.

The Weierstraß Approximation Theorem

We have seen in Example 1.2.10 that the set of polynomials on a closed interval is not closed in the set of continuous functions, since its boundary points may be continuous, non-polynomial functions. In fact, we now show that all continuous functions are boundary points of the set of polynomials:

1.2.14. Weierstraß Approximation Theorem. The set $\mathscr{P}([a,b])$ is dense in $C([a,b])$, i.e., for every continuous function f and every $\varepsilon > 0$ there exists a polynomial p such that

$$\|f - p\|_\infty = \sup_{x \in [a,b]} |f(x) - p(x)| < \varepsilon.$$

We say that every continuous function can be approximated uniformly and arbitrarily well by a polynomial. Note that this is a stronger statement than usually encountered in calculus: Taylor's Theorem, for example, allows only the uniform approximation of smooth (infinitely often differentiable) functions.

The Weierstraß Approximation Theorem

There are many proofs of the Weierstraß Approximation Theorem, most using advanced mathematical techniques. We will repeat here a beautiful elementary proof given by Lebesgue in his first published paper at age 23 (he received his doctorate four years later).

Proof.

We use the fact that a continuous function f on a closed interval $[a, b]$ is also ***uniformly continuous***, i.e.,

$$\forall_{\varepsilon>0}\ \exists_{\delta>0}\ \forall_{x,y\in[a,b]}\quad |x-y|<\delta \Rightarrow |f(x)-f(y)|<\varepsilon/4. \tag{1.2.1}$$

This fact is usually proven in elementary calculus classes and we won't repeat the proof here. We need to show that for any $\varepsilon > 0$ and any $f \in C([a, b])$ we can find a polynomial $p \in \mathscr{P}$ such that $\varrho(f, p) < \varepsilon$, i.e.,

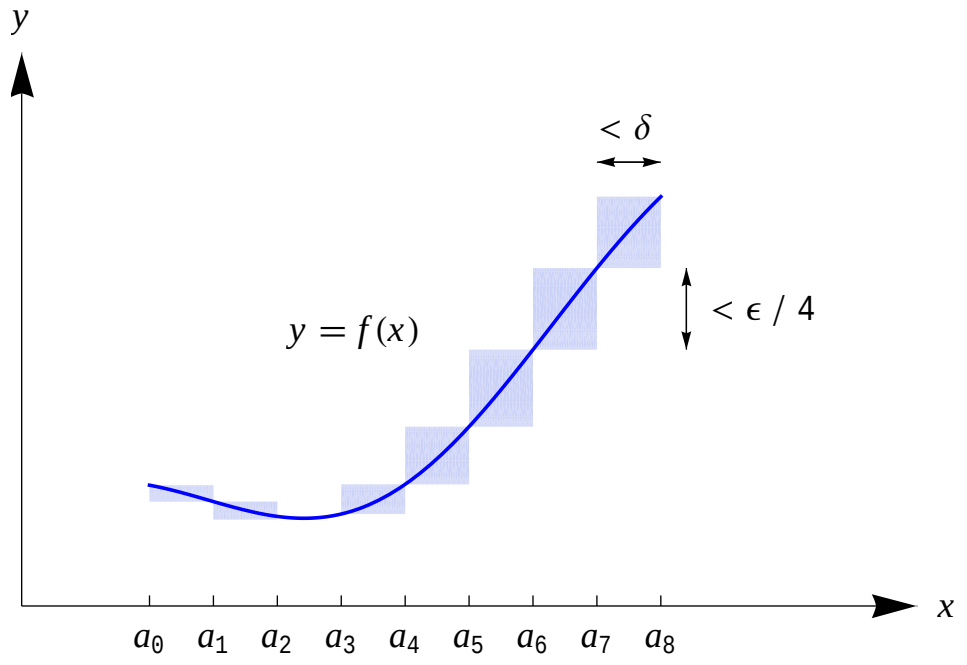
$$\sup_{x\in[a,b]} |f(x)-p(x)|<\varepsilon.$$

The Weierstraß Approximation Theorem

Proof (continued).

Fix $\varepsilon > 0$ and $f \in C([a, b])$. Then we can find some $\delta > 0$ so that (1.2.1) holds. We choose a partition $(a_0, a_1, \ldots, a_n)$ of the interval $[a, b]$, i.e., a set of numbers with the properties

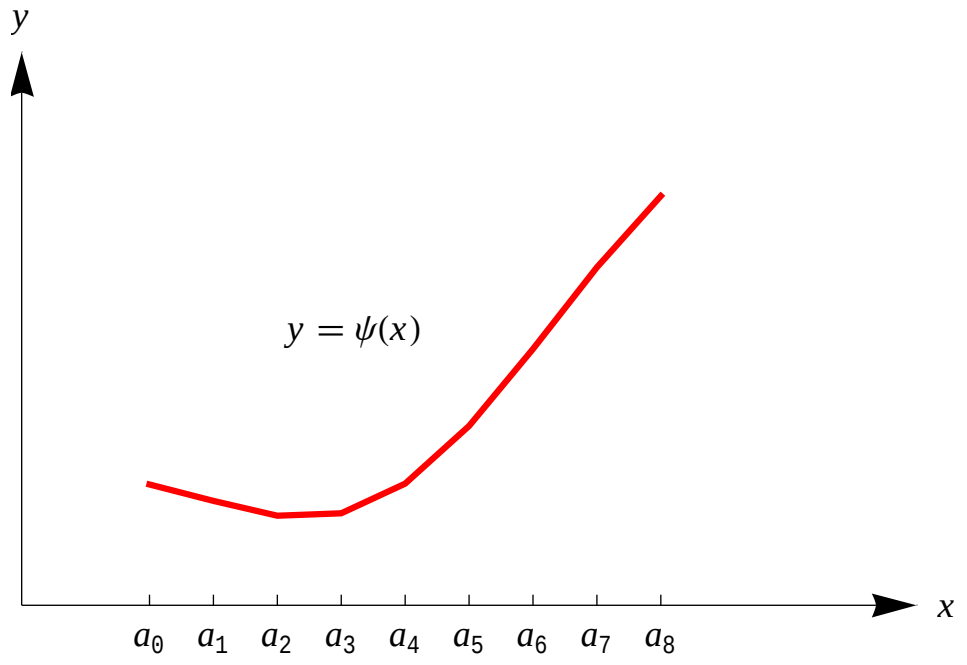
$$a_0 = a, \qquad a_n = b, \qquad 0 < a_k - a_{k-1} < \delta, \qquad k = 1, \ldots, n.$$



The Weierstraß Approximation Theorem

Proof (continued).

We then define a piecewise linear function ψ by joining the points $(a_k, f(a_k))$ with straight lines:



Then $|\psi(a_k) - \psi(x)| < |f(a_k) - f(a_{k-1})| < \dfrac{\varepsilon}{4}$ for $a_{k-1} < x < a_k$.

The Weierstraß Approximation Theorem

Proof (continued).

Furthermore, by (1.2.1),

$$|f(a_k) - f(x)| < \frac{\varepsilon}{4} \qquad \text{for } a_{k-1} < x < a_k.$$

Since $f(a_k) = \psi(a_k)$, the triangle inequality yields

$$|\psi(x) - f(x)| < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2} \qquad \text{for } a_{k-1} < x < a_k.$$

and hence

$$\begin{aligned} \|f - \psi\|_\infty &= \sup_{x \in [a,b]} |\psi(x) - f(x)| \\ &= \max_{1 \le k \le n} \sup_{x \in [a_{k-1}, a_k]} |\psi(x) - f(x)| < \frac{\varepsilon}{2}. \end{aligned}$$

The Weierstraß Approximation Theorem

Proof (continued).

We have therefore reduced the problem to approximating the piecewise linear function ψ by a polynomial p, since if we can find a p such that

$$\|\psi - p\|_\infty < \frac{\varepsilon}{2}, \tag{1.2.2}$$

then

$$\|f - p\|_\infty \leq \|f - \psi\|_\infty + \|\psi - p\|_\infty \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

As a first step, we shift and scale ψ: define

$$\varphi(x) := \psi((b-a)x + a).$$

Then φ is a piecewise linear function $\varphi\colon [0,1] \to \mathbb{R}$.

The Weierstraß Approximation Theorem

Proof (continued).

If we find a polynomial $q\colon [0,1] \to \mathbb{R}$ such that

$$\sup_{x\in[0,1]} |\varphi(x) - q(x)| < \varepsilon,$$

then (1.2.2) holds for

$$p(x) := q\left(\frac{x-a}{b-a}\right).$$

We now discuss how to approximate a piecewise linear function on $[0,1]$. Set

$$\varphi_k(x) = \alpha_k\big((x-\beta_k) + |x-\beta_k|\big) = \begin{cases} 0 & x < \beta_k, \\ 2\alpha_k(x-\beta_k) & x \geq \beta_k, \end{cases} \tag{1.2.3}$$

where $\alpha_k, \beta_k \in \mathbb{R}$.

The Weierstraß Approximation Theorem

Proof (continued).

The piecewise linear function $\varphi\colon [0,1] \to \mathbb{R}$ can then be represented in the form

$$\varphi(x) = f(a) + \varphi_1(x) + \varphi_2(x) + \cdots + \varphi_n(x)$$

for certain functions φ_k as in (1.2.3). It follows that

$$\varphi(x) = \underbrace{f(a) + \sum_{k=1}^{n} \alpha_k (x - \beta_k)}_{\text{polynomial}} + \sum_{k=1}^{n} \alpha_k |x - \beta_k|$$

It is hence sufficient to approximate the modulus $|x|$ by a polynomial, because that can then be shifted and translated to approximate each term $\alpha_k|x - \beta_k|$ in the sum, yielding finally a polynomial approximation to φ.

The Weierstraß Approximation Theorem

Proof (continued).

We will use the binomial series to approximate the square root:

$$\sqrt{1-z} = 1 - \sum_{n=1}^{\infty} \frac{1}{2^{2n-1}} \binom{2n-2}{n-1} \frac{z^n}{n} \tag{1.2.4}$$

In the assignments you will check that (1.2.4) holds and that the series converges whenever $|z| < 1$. It also converges when $|z| = 1$, since

$$\begin{aligned} \sum_{n=1}^{N} \frac{1}{2^{2n-1} n} \binom{2n-2}{n-1} &= \lim_{x \nearrow 1} \sum_{n=1}^{N} \frac{1}{2^{2n-1} n} \binom{2n-2}{n-1} x^n \\ &\leq 1 - \lim_{x \nearrow 1} \sqrt{1-x} = 1 \end{aligned} \tag{1.2.5}$$

The Weierstraß Approximation Theorem

Proof (continued).

We note that it is sufficient to consider $x \in [-1, 1]$, where

$$|x| = \sqrt{x^2} = \sqrt{1 - (1 - x^2)} = 1 - \sum_{n=1}^{\infty} \frac{1}{2^{2n-1}} \binom{2n-2}{n-1} \frac{(1-x^2)^n}{n}.$$

The convergence of the series is even uniform on $[-1, 1]$ (why?), so that truncating the series at a suitable high term yields a polynomial that approximates $|x|$ uniformly. This completes the proof. □

Infinite Bases in Normed Vector Spaces

Colloquially, we might define an "infinite basis" in a normed vector space as consisting of a sequence $(b_n)_{n\in\mathbb{N}}$ such that every vector v has a unique representation in the form

$$v = \sum_{n=0}^{\infty} \lambda_n b_n \tag{1.2.6}$$

for certain numbers λ_n, $n \in \mathbb{N}$. However, this equation is not as simple as it looks, since the infinite sum (series) poses questions of convergence:

- Does the right-hand side of (1.2.6) converge for any sequence of numbers (λ_n)?
- In what sense does the equality in (1.2.6) hold?

Schauder Bases

To highlight these issues, we make the following, more precise, definition:

1.2.15. Definition. Let $(V, \|\cdot\|)$ be a normed vector space. A sequence of vectors (v_n) in V is called a ***Schauder basis*** if for every $v \in V$ there exists a unique sequence of scalars (λ_n) such that

$$v = \lim_{N\to\infty} \sum_{n=0}^{N} \lambda_n v_n,$$

or, equivalently,

$$\left\| v - \lambda_0 v_0 - \lambda_1 v_1 - \cdots - \lambda_N v_N \right\| \xrightarrow{N\to\infty} 0.$$

This definition should be compared with Definition 1.1.8. Since convergence plays an essential role, a Schauder basis can only be defined in a ***normed*** vector space.

Monomials

Let us discuss a first example: in $(C([a,b]), \|\cdot\|_\infty)$ consider the sequence of monomials on an interval,

$$m_n \in C([a,b]), \qquad m_n(x) = x^n, \qquad n \in \mathbb{N}.$$

In order for the sequence (m_k) to be a basis of $C([a,b])$, we need to verify that every continuous function u on $[a,b]$ has a representation

$$u(x) = \sum_{n=0}^{\infty} \lambda_n x^n$$

with uniquely determined coefficients $\lambda_n \in \mathbb{R}$.

The ***existence*** of such a representation follows from the Weierstraß Approximation Theorem: for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ and numbers $\lambda_n \in \mathbb{R}$ such that

$$\left\| u - \sum_{n=0}^{N} \lambda_n m_n \right\|_\infty < \varepsilon.$$

However, the ***uniqueness*** of the coefficients λ_n is far from clear.

Monomials

Recall from Theorem 1.1.10 that a basis in a finite-dimensional space is characterized by the following two conditions:

- the span of the basis vectors is equal to the entire space,
- the basis vectors are linearly independent.

In an infinite-dimensional space, the first condition may be amended to stating that

- the span of the vectors of a Schauder basis is dense in the space

since we are considering limits of linear combinations instead of exact equalities.

However, it is not immediately clear how to translate the condition of linear independence (which guarantees the uniqueness of a basis representation) into the infinite-dimensional case. Linear independence is based on finite linear combinations and extending these to infinite linear combinations is problematical.

A way out is to consider a more restrictive approach, based on orthogonality of vectors. This requires the introduction of a scalar product.

Inner Product Spaces

1.3.1. Definition. Let V be a real or complex vector space. Then a map $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{F}$ is called a ***scalar product*** or ***inner product*** if for all $u, v, w \in V$ and all $\lambda \in \mathbb{F}$

(i) $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$,

(ii) $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$,

(iii) $\langle u, \lambda v \rangle = \lambda \langle u, v \rangle$,

(iv) $\langle u, v \rangle = \overline{\langle v, u \rangle}$.

The pair $(V, \langle \cdot, \cdot \rangle)$ is called an ***inner product space***.

1.3.2. Remark. Properties (iii) and (iv) imply that

$$\langle \lambda u, v \rangle = \overline{\langle v, \lambda u \rangle} = \overline{\lambda \langle v, u \rangle} = \overline{\lambda} \langle u, v \rangle.$$

We say that the inner product is linear in the second component and anti-linear in the first component.

The Induced Norm

1.3.3. Examples.

- In $\mathbb{R}^n$ we define the ***canonical*** or ***standard scalar product***

$$\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i, \qquad x, y \in \mathbb{R}^n. \tag{1.3.1}$$

- In $\mathbb{C}^n$ we can define the inner product

$$\langle x, y \rangle := \sum_{i=1}^{n} \overline{x_i} y_i, \qquad x, y \in \mathbb{C}^n.$$

- In $C([a, b])$, the space of complex-valued, continuous functions on the interval $[a, b]$, we can define an inner product by

$$\langle f, g \rangle := \int_a^b \overline{f(x)} g(x)\, dx, \qquad f, g \in C([a, b]).$$

The Induced Norm

1.3.4. Definition. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. The map

$$\| \cdot \| \colon V \to \mathbb{R}, \qquad \|v\| = \sqrt{\langle v, v \rangle}$$

is called the ***induced norm*** on V.

1.3.5. Examples.

- The induced norm in $\mathbb{R}^n$ and $\mathbb{C}^n$ is given by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{n} |x_i|^2} = \|x\|_2, \tag{1.3.2}$$

  which is the usual euclidean norm.
- The induced norm on $C([a, b])$ is

$$\|f\| = \sqrt{\langle f, f \rangle} = \sqrt{\int_a^b |f(x)|^2 \, dx} = \|f\|_2$$

  which is just the 2-norm.

The Cauchy-Schwartz Inequality

1.3.6. Cauchy-Schwarz Inequality. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product vector space. Then

$$|\langle u, v \rangle| \leq \|u\| \cdot \|v\| \qquad \text{for all } u, v \in V$$

where $\| \cdot \|$ is the induced norm.

Proof.
Let $e := v/\|v\|$. Then $\langle e, e \rangle = \langle v, v \rangle / \|v\|^2 = 1$ and

$$\begin{aligned} 0 \leq \|u - \langle e, u \rangle e\|^2 &= \langle u - \langle e, u \rangle e, u - \langle e, u \rangle e \rangle \\ &= \|u\|^2 - |\langle e, u \rangle|^2 \end{aligned}$$

It follows that

$$|\langle u, v \rangle|^2 = \|v\|^2 \cdot |\langle u, e \rangle|^2 \leq \|u\|^2 \cdot \|v\|^2. \qquad \square$$

The Induced Norm

1.3.7. Corollary. The induced norm is actually a norm, i.e., it satisfies

1. $\|x\| \geq 0$, $\|x\| = 0 \Leftrightarrow x = 0$,
2. $\|\lambda x\| = |\lambda| \cdot \|x\|$,
3. $\|x + y\| \leq \|x\| + \|y\|$

for all $x, y \in V$ and $\lambda \in \mathbb{R}$.

Proof.

All properties except for the triangle inequality are easily checked. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \|x + y\|^2 &= \|x\|^2 + \|y\|^2 + 2\,\mathrm{Re}\langle x, y\rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2|\langle x, y\rangle| \\ &\leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

□

Angle Between Vectors

1.3.8. Definition. Let V be an inner product space and $u, v \in V$. We define the ***angle*** $\alpha(u, v) \in [0, \pi]$ ***between*** u ***and*** v by

$$\cos \alpha(u, v) = \frac{\langle u, v \rangle}{\|u\| \|v\|}. \tag{1.3.3}$$

This definition makes sense, since by the Cauchy-Schwarz inequality

$$\left| \frac{\langle u, v \rangle}{\|u\| \|v\|} \right| = \frac{|\langle u, v \rangle|}{\|u\| \|v\|} \leq 1.$$

In $\mathbb{R}^2$ and $\mathbb{R}^3$ the expression (1.3.3) corresponds to the geometric notion of the (cosine of the) angle between two vectors.

Angle Between Vectors

1.3.9. Example. For $x, y \in \mathbb{R}^2$ we have $\sphericalangle(x, y) = \alpha(x, y)$.

We may assume that $\|x\| = \|y\| = 1$ and we consider the case

$$x = \begin{pmatrix} \cos\varphi_1 \\ \sin\varphi_1 \end{pmatrix}, \qquad y = \begin{pmatrix} \cos\varphi_2 \\ \sin\varphi_2 \end{pmatrix}, \qquad 0 < \varphi_1 < \varphi_2 < \pi.$$

Then $\sphericalangle(x, y) = \varphi_2 - \varphi_1$ and

$$\begin{aligned} \cos\sphericalangle(x, y) &= \cos(\varphi_2 - \varphi_1) = \cos\varphi_2 \cos\varphi_1 + \sin\varphi_2 \sin\varphi_1 \\ &= \langle x, y\rangle = \cos\alpha(x, y) \end{aligned}$$

In a similar manner, one can prove that $\sphericalangle(x, y) = \alpha(x, y)$ for $x, y \in \mathbb{R}^3$.

In applications, we are nearly exclusively interested in whether or not two vectors are at right-angles to each other, i.e., whether they are perpendicular.

Orthogonality and Pythagoras's Theorem

1.3.10. Definition. Let $(V, \langle \cdot\,, \cdot \rangle)$ be an inner product vector space.

(i) A vector $v \in V$ is said to be ***normed*** (or ***normalized***) if $\langle v, v \rangle = 1$. This is equivalent to $\|v\| = 1$.

(ii) Two vectors $u, v \in V$ are said to be ***orthogonal*** or ***perpendicular*** if $\langle u, v \rangle = 0$. We then write $u \perp v$.

1.3.11. Pythagoras's Theorem. Let $(V, \langle \cdot\,, \cdot \rangle)$ be an inner product space, $x, y \in V$ such that $x \perp y$ and $z = x + y$. Then

$$\|z\|^2 = \|x\|^2 + \|y\|^2.$$

Proof.

We see directly that

$$\|z\|^2 = \langle z, z \rangle = \langle x + y, x + y \rangle = \langle x, x \rangle + \underbrace{\langle x, y \rangle}_{=0} + \underbrace{\langle y, x \rangle}_{=0} + \langle y, y \rangle$$

$$= \|x\|^2 + \|y\|^2.$$

□

Orthonormal Systems

1.3.12. Definition. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product vector space. A family of vectors $\{v_n\}_{n \in I} \subset V$, $I \subset \mathbb{N}$, is called an ***orthonormal system*** if

$$\langle v_m, v_n \rangle = \delta_{mn} = \begin{cases} 1 & \text{for } m = n, \\ 0 & \text{for } m \neq n, \end{cases} \qquad j, k \in I,$$

i.e., if $\|v_n\| = 1$ and $v_m \perp v_n$ for $m \neq n$.

1.3.13. Example. The ***standard basis vectors*** in $\mathbb{R}^3$,

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

are an orthonormal system (e_1, e_2, e_3) with respect to the scalar product (1.3.1).

Orthogonal Polynomials

1.3.14. Example. Consider the space $C([-1,1])$ of complex-valued continuous functions imbued with the scalar product given by

$$\langle u, v\rangle = \int_{-1}^{1} \overline{u(x)} v(x)\, dx. \tag{1.3.4}$$

Let us regard the monomials

$$m_0(x) = 1, \qquad m_1(x) = x, \qquad m_2(x) = x^2.$$

Then

$$\langle m_0, m_1\rangle = \int_{-1}^{1} m_0(x) m_1(x)\, dx = \int_{-1}^{1} 1 \cdot x\, dx = 0,$$

$$\langle m_0, m_2\rangle = \int_{-1}^{1} 1 \cdot x^2\, dx = \frac{2}{3},$$

$$\langle m_1, m_2\rangle = \int_{-1}^{1} x \cdot x^2\, dx = 0.$$

Thus, $m_0 \perp m_1$ and $m_1 \perp m_2$ but $m_0 \not\perp m_2$.

Orthonormal Systems

1.3.15. Example. In ℓ^2 a scalar product can be defined by

$$\langle x, y \rangle := \sum_{n=0}^{\infty} \overline{x_n} y_n.$$

Then the set $(e_n)_{n \in \mathbb{N}}$ is an orthonormal system, where

$$e_n = (0, \ldots, 0, 1, 0, \ldots)$$

is a sequence equal to zero in every entry except for the nth entry, where it equals one. In other words,

$$(e_n)_k = \begin{cases} 1 & k = n, \\ 0 & k \neq n. \end{cases}$$

Orthonormal Bases

1.3.16. Definition and Theorem. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product vector space and $\{e_n\}_{n\in I} \subset V$, $I \subset \mathbb{N}$ an orthonormal system in V. If $\operatorname{span}\{e_n\}$ is dense in V, then the ordered set (e_n) is a Schauder basis of V. In particular, any $v \in V$ has the unique representation

$$v = \sum_{n\in I} \langle e_n, v \rangle e_n \tag{1.3.5}$$

We say that $\{e_n\}$ is an ***orthonormal basis*** (ONB) of V.

Proof.
Since the span of $\{e_n\}$ is dense in V, every $v \in V$ has a representation in the form

$$v = \sum_{n\in I} \lambda_n e_n$$

for certain $\lambda_n \in \mathbb{F}$. It remains to show (1.3.5).

Orthonormal Bases

Proof (continued).

We prove only the case $I = \mathbb{N}$. Let

$$v_N := \sum_{n=0}^{N} \lambda_n e_n.$$

Then $\|v_N - v\| \to 0$ as $N \to \infty$. Furthermore, for any $m \in \mathbb{N}$ and any $N > m$,

$$\langle e_m, v_N \rangle = \sum_{n=0}^{N} \lambda_n \langle e_m, e_n \rangle = \lambda_m$$

so we see that $\lambda_n = \langle e_n, v_N \rangle$ for $n \leq N$. Fix $n \in \mathbb{N}$ and choose $N \geq n$. Then

$$|\lambda_n - \langle e_n, v \rangle| = |\langle e_n, v_N - v \rangle| \leq \|v_N - v\| \xrightarrow{N \to \infty} 0$$

by the Cauchy-Schwartz inequality. This shows (1.3.5). □

Fourier-Euler Basis

We now have a useful concept of basis in an infinite-dimensional inner product space: a system of vectors that is orthogonal and whose span is dense. There are two basic examples worth mentioning now:

1.3.17. Example. The functions $b_n \in C([0, 2\pi])$ given by

$$b_n(x) = \frac{1}{\sqrt{2\pi}} e^{inx}, \qquad n \in \mathbb{Z}, \tag{1.3.6}$$

give an orthonormal system $\{b_n\}_{n\in\mathbb{Z}}$ with respect to the scalar product

$$\langle u, v\rangle = \int_0^{2\pi} \overline{u(x)} v(x)\, dx.$$

The span of $\{b_n\}_{n\in\mathbb{Z}}$ is also dense in $C([0, 2\pi])$, but we have not proved this yet. The proof is quite complicated and will have to be postponed. The basis (1.3.6) is called the ***Fourier-Euler basis*** of the continuous functions. We will analyze it more closely in a later section.

A Polynomial Basis?

1.3.18. Example. The span of the functions $m_n \in C([-1,1])$ given by

$$m_n(x) = x^n, \qquad n \in \mathbb{N}, \tag{1.3.7}$$

is dense in $C([-1,1])$ (by the Weierstraß Approximation theorem), but the functions are not orthogonal (see Example 1.3.14. In order to construct an orthonormal basis for $C([1,1])$, we need to orthonormalize them. This is done using the ***Gram-Schmidt process***, which we now describe.

Orthogonal Complement

1.3.19. Definition. If $M \subset V$ is a subspace, the set

$$M^{\perp} := \left\{ v \in V : \underset{u \in M}{\forall} \langle v, u \rangle = 0 \right\}$$

is called the ***orthogonal complement*** of M.

1.3.20. Lemma. $M^{\perp}$ is a subspace of V.

Proof.
If $v_1, v_2 \in M^{\perp}$, then

$$\langle v_1 + v_2, u \rangle = \langle v_1, u \rangle + \langle v_2, u \rangle = 0 + 0 = 0$$

for all $u \in M$, so $v_1 + v_2 \in M^{\perp}$. Similarly, if $v \in M^{\perp}$ and $\lambda \in \mathbb{F}$, then $\langle \lambda v, u \rangle = \overline{\lambda} \langle v, u \rangle = 0$, so $\lambda v \in M^{\perp}$. Thus $M^{\perp}$ is a subspace of V. $\square$

Projection Theorem

1.3.21. Projection Theorem. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and $u_1, \dots, u_n$ a finite orthonormal system in V, i.e.,

$$\langle u_j, u_k \rangle = \delta_{jk} = \begin{cases} 1 & \text{for } j = k, \\ 0 & \text{for } j \neq k, \end{cases} \qquad j, k = 1, \dots, n.$$

Let $U := \operatorname{span}\{u_1, \dots, u_n\}$. Then for every $v \in V$ there exists a unique representation

$$v = u + w \qquad \text{where } u = \sum_{i=1}^{n} \langle u_i, v \rangle u_i \in U \text{ and } w \in U^{\perp}.$$

Projection Theorem

Proof.

We first show the uniqueness of the decomposition: Assume $v = u + w = u' + w'$. Then by Pythagoras's theorem,

$$0 = \|u - u' + (w - w')\|^2 = \|u - u'\|^2 + \|w - w'\|^2,$$

so $\|u - u'\| = \|w - w'\| = 0$. Thus $u = u'$ and $w = w'$. Regarding the existence of such a decomposition, it is clear that u lies in U. We need to show that $w = v - u \in U^\perp$. For this, it is sufficient to show that $\langle w, u_j \rangle = 0$ for $j = 1, \dots, n$, since $(u_1, \dots, u_n)$ is a basis of U. Now for $j = 1, \dots, n$ we have

$$\langle w, u_j \rangle = \langle v, u_j \rangle - \langle u, u_j \rangle = \langle v, u_j \rangle - \sum_{i=1}^{r} \overline{\langle u_i, v \rangle} \underbrace{\langle u_i, u_j \rangle}_{=\delta_{ij}}$$

$$= \langle v, u_j \rangle - \langle v, u_j \rangle = 0.$$

□

Bessel's Inequality

As a consequence of the Projection Theorem 1.3.21 and Pythagoras's Theorem 1.3.11 we obtain the following important result:

1.3.22. Bessel Inequality. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and $\{e_k\}_{n \in I} \subset V$, $I \subset \mathbb{N}$, be an orthonormal system in V. Then, for any $v \in V$,

$$\sum_{n \in I} |\langle e_n, v \rangle|^2 \leq \|v\|^2. \tag{1.3.8}$$

Bessel's Inequality

Proof.

Let us assume first that $I = \{1, \dots, N\}$ is finite. Let $v \in V$ be any vector. Then, by the projection theorem, we can write

$$v = u + w = \sum_{n=0}^{N} \langle e_n, v \rangle e_n + w$$

where $u \perp (v - u)$. Then, by Pythagoras's theorem,

$$0 \leq \|v - u\|^2 = \|v\|^2 - \|u\|^2 = \|v\|^2 - \sum_{n=0}^{N} |\langle e_n, v \rangle|^2. \tag{1.3.9}$$

This proves (1.3.8) for the case of finite N. If $I = \mathbb{N}$, we can let $N \to \infty$ on both sides of the estimate (1.3.9) and again obtain (1.3.8). □

The Riemann-Lebesgue Lemma

An immediate corollary of the Bessel inequality is the followinG:

1.3.23. Riemann-Lebesgue Lemma. Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space and $\{e_k\}_{n \in I} \subset V$, $I \subset \mathbb{N}$, an infinite orthonormal system in V. Then, for any $v \in V$,

$$\langle e_n, v \rangle \xrightarrow{n \to \infty} 0. \tag{1.3.10}$$

The result follows from the fact that the series in (1.3.8) can only converge if the sequence of summands converges to zero.

Gram-Schmidt Orthonormalization

The goal of Gram-Schmidt orthonormalization is to obtain an orthonormal system from any family of vectors.

Assume that we have a family of vectors $\{v_k\}_{k\in I} \subset V$, $I \subset \mathbb{N}$, in an inner product space V. We wish to construct a new family $\{w_k\}_{k\in I} \subset V$, $I \subset \mathbb{N}$, that is orthonormal. We start with v_1 and norm it, defining

$$w_1 := \frac{v_1}{\|v_1\|}$$

Next, we want to obtain from v_2 a vector w_2 such that $w_1 \perp w_2$. By Theorem 1.3.21, v_2 has a unique representation as a sum $v_2 = x + y$, where $x \in \operatorname{span}\{w_1\}$ and $y \in (\operatorname{span}\{w_1\})^\perp$. Now $x = \langle w_1, v_2\rangle w_1$, so

$$y = v_2 - \langle w_1, v_2\rangle w_1 \in (\operatorname{span}\{w_1\})^\perp.$$

(Of course, y is independent and even orthogonal to w_1.)

Gram-Schmidt Orthonormalization

It just remains to normalize y, and we define

$$w_2 := \frac{v_2 - \langle w_1, v_2 \rangle w_1}{\| v_2 - \langle w_1, v_2 \rangle w_1 \|}.$$

Now we can write

$$v_3 = \langle w_1, v_3 \rangle w_1 + \langle w_2, v_3 \rangle w_2 + y,$$

where $y \in (\text{span}\{w_1, w_2\})^{\perp}$. Thus

$$w_3 := \frac{v_3 - \langle w_2, v_3 \rangle w_2 - \langle w_1, v_3 \rangle w_1}{\| v_3 - \langle w_2, v_3 \rangle w_2 - \langle w_1, v_3 \rangle w_1 \|}$$

will be normed and orthogonal to w_1 and w_2.

Gram-Schmidt Orthonormalization

Proceeding in this way, we set

$$w_1 := \frac{v_1}{\|v_1\|}$$

$$w_k := \frac{v_k - \sum_{j=1}^{k-1} \langle w_j, v_k \rangle w_j}{\left\| v_k - \sum_{j=1}^{k-1} \langle w_j, v_k \rangle w_j \right\|}, \qquad k = 2, 3, 4, \dots,$$

and hence obtain an orthonormal system as desired.

1.3.24. Example. Continuing the discussion in Examples 1.3.14 and 1.3.18, we consider the monomials

$$m_0(x) = 1, \qquad m_1(x) = x, \qquad m_2(x) = x^2$$

in $C([-1, 1])$.

Orthogonal Polynomials

We apply the orthonormalization procedure to obtain

$$q_0(x) = \frac{m_0(x)}{\|m_0\|} = \frac{1}{\sqrt{\int_{-1}^{1} |1|^2 \, dx}} = \frac{1}{\sqrt{2}},$$

$$q_1(x) = \frac{m_1(x) - \langle q_0, m_1 \rangle q_0(x)}{\|m_1 - \langle q_0, m_1 \rangle q_0\|} = \frac{x}{\sqrt{\int_{-1}^{1} |x|^2 \, dx}} = \sqrt{\frac{3}{2}} x,$$

$$q_2(x) = \frac{m_2(x) - \langle q_1, m_2 \rangle q_1(x) - \langle q_0, m_2 \rangle q_0(x)}{\|m_2 - \langle q_1, m_2 \rangle q_1 - \langle q_0, m_2 \rangle q_0\|}$$

$$= \frac{x^2 - 1/3}{\sqrt{\int_{-1}^{1} |x^2 - 1/3|^2 \, dx}} = \frac{3\sqrt{5}}{2\sqrt{2}} (x^2 - 1/3).$$

This procedure can be applied to all monomials $m_n(x) = x^n$ to yield a system of orthonormal polynomials in $C([-1, 1])$, called the ***Legendre polynomials***. Since the set of all polynomials is dense, the Legendre polynomials constitute an ONB of $C([-1, 1])$.

Best Approximation

The use of an ***orthonormal*** basis (as opposed to any Schauder basis) also has significant advantages in practical approximation problems. In applications, one usually wants to use only a few basis vectors to approximately represent a given vector. The question is now how to select the coefficients optimally. More precisely:

Let $(V, \langle \cdot\,, \cdot \rangle)$ be an inner product space, $v \in V$ and $\mathscr{B} = \{e_n\}$ an orthonormal system in V. We seek to approximate v using a linear combination of the first $N \in \mathbb{N}$ elements of $\mathscr{B}$,

$$v \approx \sum_{n=1}^{N} \lambda_n e_n, \qquad \lambda_1, \dots, \lambda_N \in \mathbb{F}. \tag{1.3.11}$$

The goal is to choose the coefficients $\lambda_1, \dots, \lambda_N$ in such a way as to minimize the approximation error

$$\left\| v - \sum_{i=1}^{N} \lambda_n e_n \right\|.$$

Best Approximation

Note that

$$\left\| v - \sum_{i=1}^{N} \lambda_n e_n \right\|^2 = \|v\|^2 + \sum_{n=1}^{N} |\lambda_n|^2 - \sum_{n=1}^{N} \lambda_n \langle v, e_n \rangle - \sum_{n=1}^{N} \overline{\lambda}_n \langle e_n, v \rangle$$

$$= \|v\|^2 + \sum_{n=1}^{N} \left| \langle e_n, v \rangle - \lambda_n \right|^2 - \sum_{n=1}^{N} |\langle e_n, v \rangle|^2. \quad (1.3.12)$$

It is clear that (1.3.12) is minimal if $\lambda_n = \langle e_n, v \rangle$, i.e., the coefficients in (1.3.11) are just the coefficients of a basis expansion. We also see that

$$\left\| v - \sum_{i=1}^{n} \langle e_i, v \rangle e_i \right\| \leq \left\| v - \sum_{i=1}^{N} \langle e_i, v \rangle e_i \right\| \qquad \text{for } n > N, \quad (1.3.13)$$

so the approximation can only improve when we add further elements of the orthonormal system $\mathscr{B}$ to the approximation. The previous coefficients do not need to be recalculated when more orthonormal vectors are added.

The Legendre Polynomials

1.4.1. Definition. For $n \in \mathbb{N}$ the function

$$P_n \colon [-1,1] \to \mathbb{R}, \qquad P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}[(x^2-1)^n]. \tag{1.4.1}$$

is said to be the ***nth Legendre polynomial***.

1.4.2. Remark. The expression (1.4.1) is known as ***Rodrigues's formula*** for the Legendre polynomials. The Legendre polynomials can also be defined as the unique bounded solution to the ***Legendre differential equation*** for $x \in (-1,1)$,

$$\frac{d}{dx}\left((x^2-1)\frac{dy}{dx}\right) = n(n+1)y, \qquad \text{with} \qquad y(1) = 1. \tag{1.4.2}$$

This equation occurs naturally in the study of partial differential equations.

The Legendre Polynomials

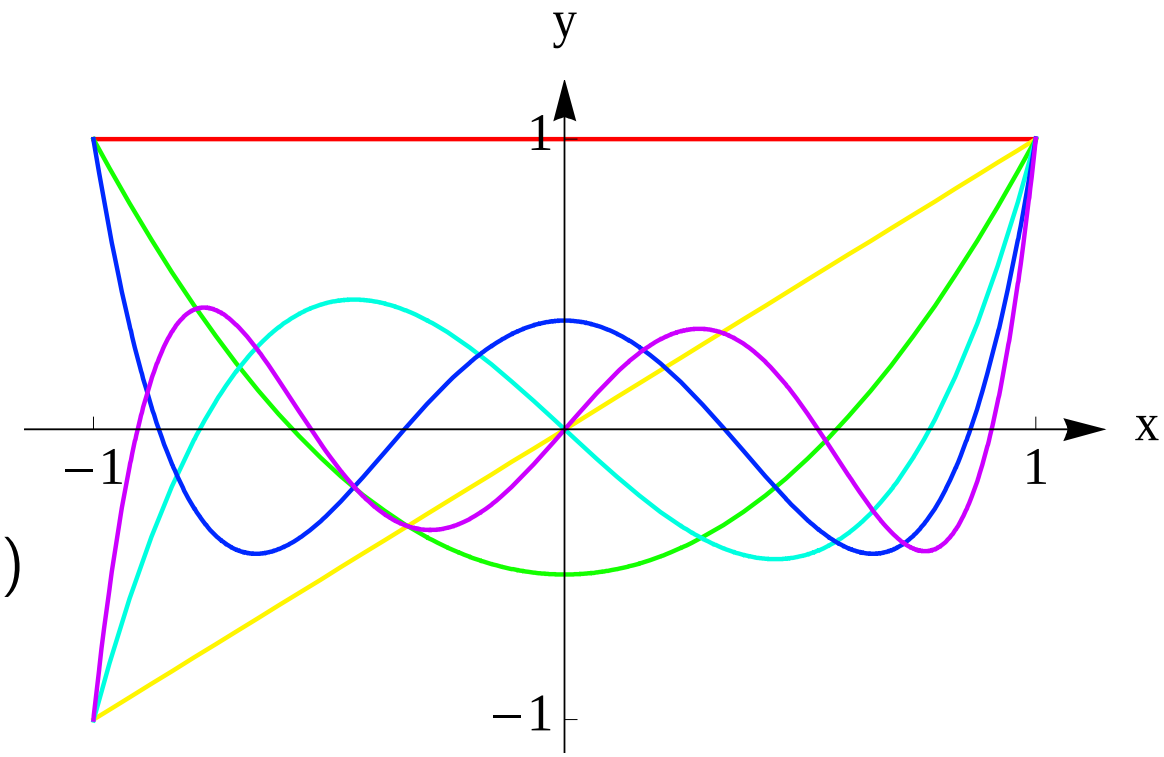
It can be easily seen from Rodriques's formula that each P_n is a polynomial of degree n. We plot the first six Legendre polynomials below:

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

$$P_3(x) = \frac{1}{2}(5x^3 - 3x),$$

$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3),$$

$$P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$$



Orthogonality of the Legendre Polynomials

1.4.3. Theorem. The Legendre polynomials are orthogonal with respect to the scalar product (1.3.4), i.e.,

$$\langle P_n, P_m \rangle = \int_{-1}^{1} P_n(x) P_m(x)\, dx = 0, \qquad n \neq m. \tag{1.4.3}$$

Proof.

Since P_n is a polynomial of degree n, it is sufficient to show that P_n is orthogonal to any monomial

$$m_k : [-1, 1] \to \mathbb{R}, \qquad m_k(x) = x^k, \tag{1.4.4}$$

of degree $k < n$. We define $u_n(x) := (x^2 - 1)^n$ so

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}[(x^2 - 1)^n] = \frac{1}{2^n n!} u_n^{(n)}(x).$$

Orthogonality of the Legendre Polynomials

Proof (continued).

Note that for any $k < n$, $u_n^{(k)}(-1) = u_n^{(k)}(1) = 0$. Let $0 \leq k < n$. Then, integrating by parts, we find

$$\begin{aligned}\langle m_k, P_n \rangle &= \frac{1}{2^n n!} \int_{-1}^{1} x^k u_n^{(n)}(x)\, dx \\ &= \underbrace{\frac{1}{2^n n!} u_n^{(n-1)}(x) x^k \Big|_{-1}^{1}}_{=0} - \frac{k}{2^n n!} \int_{-1}^{1} x^{k-1} u_n^{(n-1)}(x)\, dx.\end{aligned}$$

Repeatedly integrating by parts, we obtain

$$\langle m_k, P_n \rangle = \frac{(-1)^k k!}{2^n n!} \int_{-1}^{1} u_n^{(n-k)}(x)\, dx = \frac{(-1)^k k!}{2^n n!} u_n^{(n-k-1)}(x) \Big|_{-1}^{1} = 0.$$

This completes the proof. □

Legendre Polynomials and Orthonormalized Monomials

It turns out that (up to normalization) the Legendre polynomials are precisely the polynomials obtained from monomials through orthonormalization.

1.4.4. Theorem. Let $(e_n)_{n\in\mathbb{N}}$ denote the sequence of polynomials obtained from the monomials

$$m_k : [-1,1] \to \mathbb{R}, \qquad m_k(x) = x^k,$$

through Gram-Schmidt orthonormalization. Then

$$e_n = \frac{1}{\|P_n\|} P_n. \tag{1.4.5}$$

Legendre Polynomials and Orthonormalized Monomials

Proof.

Fix $n \in \mathbb{N}$ and define the space of all real polynomials of degree less than or equal to n,

$$\mathscr{P}_n := \left\{ p \colon [-1,1] \to \mathbb{R} \colon p(x) = \sum_{k=0}^{n} a_k x^k \right\}.$$

Then $\{e_0, e_1, \dots, e_{n-1}, e_n\}$ is a basis of $\mathscr{P}_n$ and we have the basis representation

$$P_n(x) = \sum_{k=0}^{n-1} \lambda_k e_k(x) + \lambda_n e_n(x), \qquad \lambda_1, \dots, \lambda_n \in \mathbb{R}.$$

Since (e_n) is an orthonormal basis, we have

$$\lambda_k = \langle e_k, P_n \rangle.$$

Legendre Polynomials and Orthonormalized Monomials

Proof (continued).

As P_n is orthogonal to all monomials of degree less than n, we see that

$$P_n(x) = \lambda_n e_n(x)$$

for some $\lambda_n \in \mathbb{R}$. Using $\|e_n\| = 1$,

$$\|P_n\| = \|\lambda_n e_n\| = |\lambda_n|,$$

so

$$\frac{1}{\|P_n\|} P_n(x) = \pm e_n(x).$$

To determine the sign, note that the coefficients of x^n in $P_n(x)$ and $e_n(x)$ are both positive (why?), so (1.4.5) holds. $\square$

The Fourier-Legendre Basis

Thus, any function $f \in C([-1,1])$ can be expanded in a series of the form

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{\|P_n\|^2} \langle P_n, f \rangle P_n(x),$$

where

$$\langle P_n, f \rangle = \int_{-1}^{1} f(x) P_n(x)\, dx \qquad \text{and} \qquad \|P_n\|^2 = \int_{-1}^{1} (P_n(x))^2\, dx.$$

In the assignments, we will establish

$$\|P_n\| = \sqrt{\frac{2}{2n+1}}.$$

Hence, for any $f \in C([-1,1])$,

$$f(x) = \sum_{n=0}^{\infty} \frac{2n+1}{2} \langle P_n, f \rangle P_n(x) \tag{1.4.6}$$

The expansion (1.4.6) is known as a ***Fourier-Legendre series*** for f.

Fourier-Legendre Series

1.4.5. Example. Using Mathematica, we expand the function

$$f : [-1, 1] \to \mathbb{R}, \qquad f(x) = |x - 1/2|$$

in a Fourier-Legendre series of polynomials. Note that this is an example of a polynomial approximation where a Taylor series would not be applicable, because f is not differentiable at $x = 1/2$.

```
p[x_, m_] := Sum[(2 n + 1)/2 (Integrate[LegendreP[n, y] Abs[y - 1/2], {y, -1, 1}]) LegendreP[n, x], {n, 0, m}];
Simplify[p[x, 4]]
```

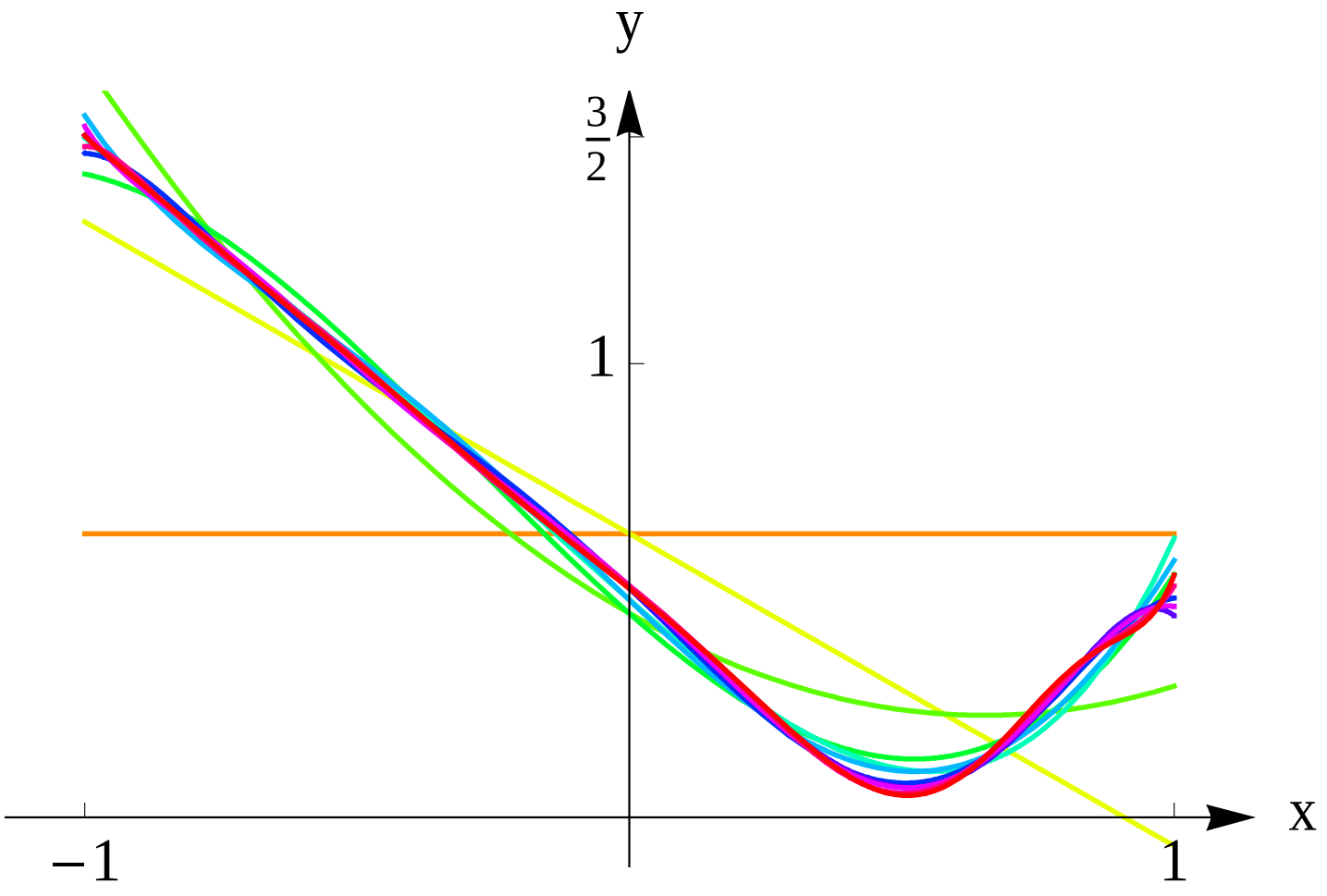
$$\frac{3923 - 8656\,x + 1890\,x^2 + 5040\,x^3 + 2835\,x^4}{8192}$$

```
Simplify[p[x, 8]]
```

$$\frac{1}{33\,554\,432}\left(17\,117\,153 - 35\,623\,456\,x - 17\,089\,380\,x^2 + 8\,981\,280\,x^3 + 102\,972\,870\,x^4 + 44\,108\,064\,x^5 - 104\,108\,004\,x^6 - 35\,212\,320\,x^7 + 34\,459\,425\,x^8\right)$$

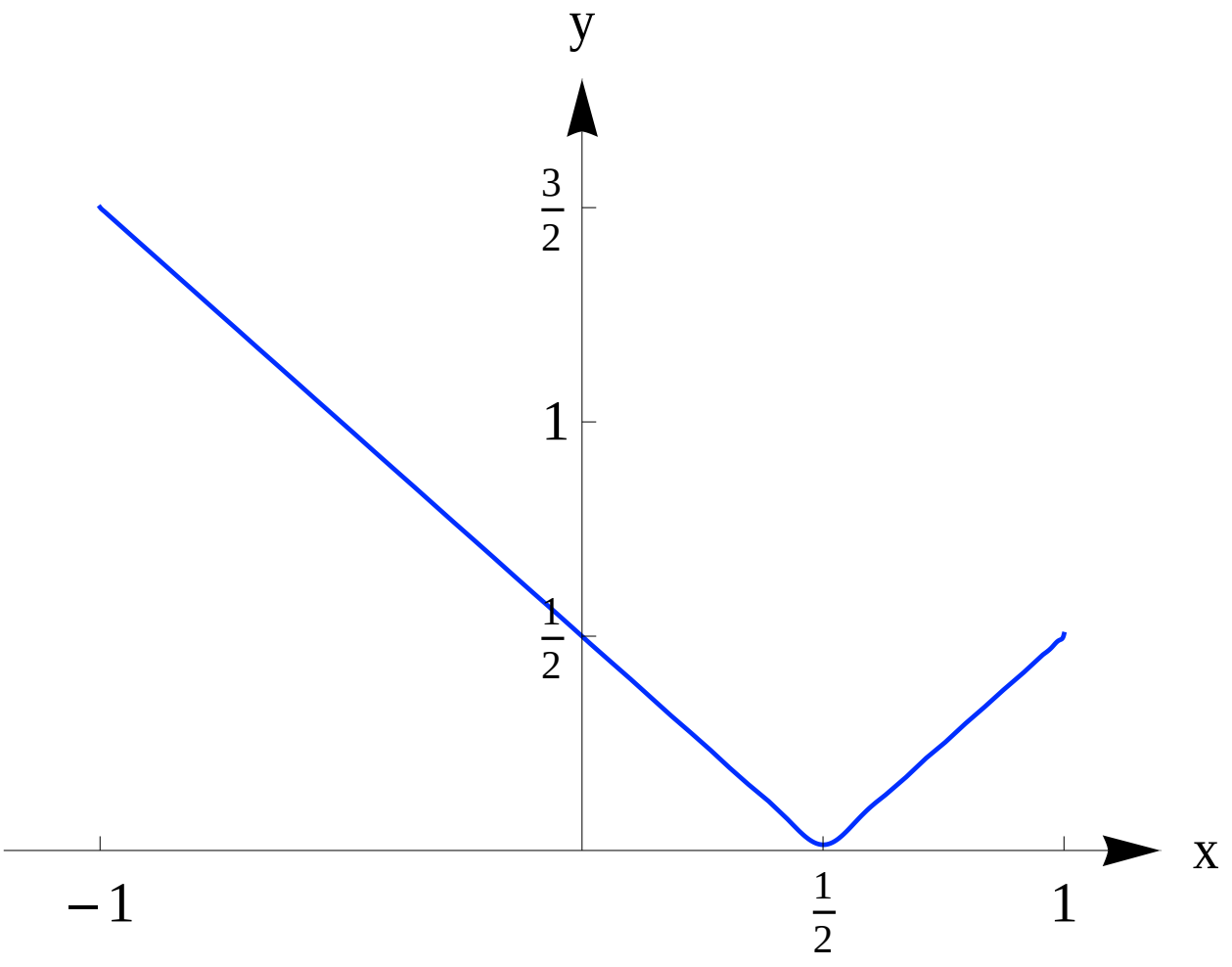
Fourier-Legendre Approximation of $f(x) = |x - 1/2|$

We illustrate the effect of the first 12 terms in the series:



Fourier-Legendre Approximation of $f(x) = |x - 1/2|$

At 40 terms, the approximation is quite good:



Recurrence Relations

Many sequences of orthogonal polynomials can be constructed recursively through so-called ***recurrence relations***. For example, the Legendre polynomials satisfy the relation

$$(2n+1)xP_n(x) = (n+1)P_{n+1}(x) + nP_{n-1}(x). \tag{1.4.7}$$

There exist a large variety of such recurrence relations, some of which we list without proof below:

(i) $P_n(x) = P'_{n+1}(x) - 2xP'_n(x) + P'_{n-1}(x)$,

(ii) $P'_{n+1}(x) - P'_{n-1}(x) = (2n+1)P_n(x)$,

(iii) $xP'_n(x) - P'_{n-1}(x) = nP_n(x)$,

(iv) $P'_n(x) - xP'_{n-1}(x) = nP_{n-1}(x)$,

(v) $(x^2-1)P'_n(x) = nxP_n(x) - nP_{n-1}(x)$.

Recurrence Relations

Proof of (1.4.7).

The proof is based on Rodrigues's formula (1.4.1) and the fact that

$$x\frac{d^n}{dx^n} = \frac{d^n}{dx^n}x - n\frac{d^{n-1}}{dx^{n-1}} \tag{1.4.8}$$

as is easily seen by applying the Leibniz rule of differentiation to $\frac{d^n}{dx^n}(xf(x))$. It follows that

$$\begin{aligned} xP_n(x) &= \frac{1}{2^n n!}\frac{d^n}{dx^n}[x(x^2-1)^n] - \frac{n}{2^n n!}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n \\ &= \frac{1}{2^{n+1}(n+1)!}\frac{d^{n+1}}{dx^{n+1}}(x^2-1)^{n+1} - \frac{n}{2^n n!}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n \\ &= P_{n+1}(x) - \frac{n}{2^n n!}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n \end{aligned} \tag{1.4.9}$$

Recurrence Relations

Proof (continued).

Now

$$xP_n(x) = \frac{x}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n = \frac{x}{2^{n-1}(n-1)!}\frac{d^{n-1}}{dx^{n-1}}[x(x^2-1)^{n-1}]$$

Applying (1.4.8), we obtain

$$\begin{aligned} xP_n(x) &= \frac{1}{2^{n-1}(n-1)!}\frac{d^{n-1}}{dx^{n-1}}[x^2(x^2-1)^{n-1}] \\ &\quad - \frac{n-1}{2^{n-1}(n-1)!}\frac{d^{n-2}}{dx^{n-2}}[x(x^2-1)^{n-1}] \\ &= \frac{1}{2^{n-1}(n-1)!}\frac{d^{n-1}}{dx^{n-1}}[(x^2-1+1)(x^2-1)^{n-1}] \\ &\quad - \frac{n-1}{2^n n!}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n \end{aligned}$$

The Generating Function

Proof (continued).

Expanding and gathering, we have

$$xP_n(x) = P_{n-1} + \frac{n+1}{2^n n!}\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n. \qquad (1.4.10)$$

Taking (1.4.9) divided by n, (1.4.10) divided by $n+1$ and adding the two identities, we obtain the result. □

From the recurrence formula (1.4.7) we obtain the generating function for the Legendre polynomials, which will be the basis for our discussion of potentials..

The Generating Function

1.4.6. Theorem. The Legendre polynomials $P_n(x)$ can be obtained from the ***generating function***

$$\frac{1}{\sqrt{1-2xt+t^2}} = \sum_{n=0}^{\infty} P_n(x)t^n \tag{1.4.11}$$

which for every $x \in [-1,1]$ has radius of convergence 1.

Proof.
The proof is based on the recurrence formula (1.4.7) and is part of the assignments. □

The generating function (1.4.11) occurs naturally in potential problems in physics (which is how it was discovered by Legendre). This includes problems involving the gravitational or electrostatic potentials. We will formulate examples in the electrostatic context, but they can easily be transferred to other setting.

The Electrostatic Potential induced by a Charged Body

A charged body in $\mathbb{R}^3$ induces an electrostatic field $\vec{E}$, which can be described as the gradient of the ***electrostatic potential*** V. Suppose that the charged body can be described through a charge distribution $\varrho\colon \mathbb{R}^3 \to [0, \infty)$, which we assume to be a piecewise continuous function. (This excludes ideal point charges and surface charges.)

In electrostatic cgs units (where $4\pi\varepsilon_0 = 1$), the electrostatic potential induced at a point $p \in \mathbb{R}^3$ is given by

$$V(p) = \iiint_{\mathbb{R}^3} \frac{\varrho(q)}{|p-q|}\, dq. \tag{1.4.12}$$

(This integral can be derived as a so-called ***fundamental solution*** of Poisson's partial differential equation $\Delta V = \varrho$ in $\mathbb{R}^3$. This will be performed in detail in the course Vv557 Methods of Applied Math II.)

Evaluating the integral (1.4.12) is often quite difficult. We now study this in more detail.

Spherical Coordinates

We aim to evaluate (1.4.12) for some fixed $p \in \mathbb{R}^3$. Let us choose an origin and coordinate axes so that p lies on the z-axis at $p = (0, 0, \zeta)$ with $\zeta > 0$. We further introduce spherical coordinates

$$x = r\cos\phi\sin\theta, \qquad y = r\sin\phi\sin\theta, \qquad z = r\cos\theta$$

with $(r, \phi, \theta) \in [0, \infty) \times [0, 2\pi) \times [0, \pi]$. Hence, θ is the angle between a point $q \in \mathbb{R}^3$ and the fixed point p.

For $q = (r\cos\phi\sin\theta, r\sin\phi\sin\theta, r\cos\theta)$ we have

$$\begin{aligned} |p - q| &= \sqrt{(0 - r\cos\phi\sin\theta)^2 + (0 - r\sin\phi\sin\theta)^2 + (\zeta - r\cos\theta)^2} \\ &= \sqrt{r^2 + \zeta^2 - 2\zeta r\cos\theta} \end{aligned} \tag{1.4.13}$$

Hence, in spherical coordinates, (1.4.12) becomes

$$V(0, 0, \zeta) =: v(\zeta) = \int_0^\infty \int_0^{2\pi} \int_0^\pi \frac{\varrho(r, \phi, \theta) r^2 \sin\theta \, d\theta \, d\phi \, dr}{\sqrt{r^2 + \zeta^2 - 2\zeta r\cos\theta}}$$

The Multipole Expansion

Let us assume that the charged body is finite, so $|\varrho(q)| = 0$ if $|q| > a$ for some $a > 0$. If $\zeta > a$, then

$$V(z) = \int_0^a \int_0^{2\pi} \int_0^{\pi} \frac{\varrho(r,\phi,\theta) r^2 \sin\theta \, d\theta \, d\phi \, dr}{\zeta \sqrt{1 + (r/\zeta)^2 - 2(r/\zeta)\cos\theta}}$$

where $r/\zeta < 1$. Using the expansion (1.4.11), we can write

$$\begin{aligned} v(\zeta) &= \frac{1}{\zeta} \int_0^a \int_0^{2\pi} \int_0^{\pi} \varrho(r,\phi,\theta) r^2 \sin\theta \sum_{n=0}^{\infty} P_n(\cos\theta) \left(\frac{r}{\zeta}\right)^n d\theta \, d\phi \, dr \\ &= \sum_{n=0}^{\infty} \frac{q_n}{\zeta^{n+1}} \end{aligned}$$

with

$$q_n = \int_0^a \int_0^{2\pi} \int_0^{\pi} \varrho(r,\phi,\theta) r^{n+2} \sin\theta P_n(\cos\theta) \, d\theta \, d\phi \, dr.$$

The Multipole Expansion

We can write this expansion in coordinate-free form as

$$V(p) = \sum_{n=0}^{\infty} \frac{q_n(\theta)}{|p|^{n+1}}, \tag{1.4.14}$$

where $\theta = \angle(p, q)$ is the angle between p and q and

$$q_n(\theta) = \iiint_{\mathbb{R}^3} |q|^n \varrho(q) P_n(\cos\theta)\, dq.$$

The series (1.4.14) is called the ***multipole expansion*** of the potential. It gives a power series in terms of the distance from the origin. The coefficients q_n, $n \in \mathbb{N}$, have various physical interpretations.

The Monopole Moment

The coefficient q_0 is called the ***monopole moment*** and represents the total charge,

$$q_0(\theta) = \iiint_{\mathbb{R}^3} \varrho(q) P_0(\cos\theta)\, dq = \iiint_{\mathbb{R}^3} \varrho(q)\, dq =: Q. \tag{1.4.15}$$

Hence,

$$V(p) = \frac{Q}{|p|} + \sum_{n=1}^{\infty} \frac{q_n(\theta)}{|p|^{n+1}}.$$

The potential induced by an ideal point charge at the origin is exactly

$$V_{\text{point}}(p) = \frac{Q}{|p|} \tag{1.4.16}$$

so the leading-order (dominating) term in the multipole expansion is the potential of a point charge with the total charge equal to that of the body under consideration.

An Ideal Monopole

1.4.7. Example. Consider a uniformly charged ball of radius $a > 0$, centered at the origin and with charge density $\varrho_0 > 0$. The monopole moment is the total charge,

$$q_0 = \iiint_{\mathbb{R}^3} \varrho(x)\, dx = \frac{4}{3}\pi a^3 \cdot \varrho_0.$$

All other moments vanish, because for $n \geq 1$,

$$\begin{aligned} q_n &= \int_0^a \int_0^{2\pi} \int_0^{\pi} \varrho_0 r^{n+2} \sin\theta P_n(\cos\theta)\, d\theta\, d\phi\, dr \\ &= \frac{2\pi\varrho_0}{n+3} a^{n+3} \int_0^{\pi} \sin\theta P_n(\cos\theta)\, d\theta \\ &= \frac{2\pi\varrho_0}{n+3} a^{n+3} \int_{-1}^{1} 1 \cdot P_n(t)\, dt = \frac{2\pi\varrho_0}{n+3} a^{n+3} \langle P_0, P_n \rangle_{L^2} = 0. \end{aligned}$$

We see that a uniformly charged ball induces exactly the same potential as a point charge.

A Shifted Point Charge

Now suppose that a point charge is shifted from the origin to the position q. Then the potential at a point $p = (0, 0, \zeta)$, $\zeta > |q| > 0$, is given by

$$V_{\text{point}}(p) = \frac{Q}{|p - q|}$$

Using polar coordinates for q we again have (1.4.13), so

$$V_{\text{point}}(p) = \frac{Q}{\sqrt{r^2 + \zeta^2 - 2\zeta r \cos\theta}} = \frac{Q}{\zeta} \sum_{n=0}^{\infty} P_n(\cos\theta) \left(\frac{r}{\zeta}\right)^n$$

Suppose that $|q| < \zeta$. Using again the expansion (1.4.11), we obtain

$$V_{\text{point}}(p) = \frac{Q}{|p|} + \sum_{n=1}^{\infty} P_n(\cos\theta) \frac{Q|q|^n}{|p|^{n+1}},$$

where $\theta = \angle(p, q)$ is the angle between p and q. We now have a much more complicated expansion than (1.4.16). However, note that the term q_0 (the total charge) has stayed the same.

A Shifted Point Charge

1.4.8. Remark. The monopole moment (1.4.15) is independent of the choice of origin (or, equivalently, a position shift of the charged body): if $q' = q + \Delta q$ is the position of the shifted body, then the total charge $q_0 = Q$ does not change.

Now consider two point charges: a positive charge q^+ at position $q \in \mathbb{R}^n$ and an opposite charge q^- at position $-q$. In spherical coordinates,

$$q = \begin{pmatrix} r\cos\phi\sin\theta \\ r\sin\phi\sin\theta \\ r\cos\theta \end{pmatrix}, \quad -q = \begin{pmatrix} -r\cos\phi\sin\theta \\ -r\sin\phi\sin\theta \\ -r\cos\theta \end{pmatrix} = \begin{pmatrix} r\cos\phi\sin(\theta+\pi) \\ r\sin\phi\sin(\theta+\pi) \\ r\cos(\theta+\pi) \end{pmatrix}.$$

We now let

$$r := \frac{d}{2}$$

for $d > 0$.

A Physical Dipole

This arrangement is called a ***physical dipole***. The potential is induced by these charges at $p = (0, 0, \zeta)$ is given by

$$\begin{aligned}
V_{\text{phys. dipole}}(p) &= V_{\text{point};q^+}(p) + V_{\text{point};q^-}(p) \\
&= \sum_{n=0}^{\infty} P_n(\cos\theta)\frac{q^+|q|^n}{|p|^{n+1}} + \sum_{n=0}^{\infty} P_n(-\cos\theta)\frac{q^-|q|^n}{|p|^{n+1}} \\
&= \sum_{n=0}^{\infty} P_n(\cos\theta)\frac{q^+ d^n}{2^n|p|^{n+1}} + \sum_{n=0}^{\infty}(-1)^n P_n(\cos\theta)\frac{q^- d^n}{2^n|p|^{n+1}} \\
&= \frac{q^+ + q^-}{|p|} + \frac{(q^+ - q^-)d\cos\theta}{2|p|^2} \\
&\quad + \sum_{n=2}^{\infty}(q^+ + (-1)^n q^-)P_n(\cos\theta)\frac{d^n}{2^n|p|^{n+1}}
\end{aligned}$$

where we have used that $P_n(-x) = (-1)^n P_n(x)$.

A Mathematical Dipole

Note that if $q^- = -q^+$, then the total charge of the dipole is zero and the dominating term of the potential is

$$\frac{(q^+ - q^-)d\cos\theta}{2|p|^2} = \frac{q^+ d\cos\theta}{|p|^2} = \frac{q^+|p|d\cos\theta}{|p|^3} = \frac{q^+\langle 2q, p\rangle}{|p|^3}.$$

The vector

$$u := q^+ \cdot (2q) \tag{1.4.17}$$

is called the ***dipole moment***. (Here $2q$ is the vector pointing from $-q$ to q.) An ***ideal*** or ***mathematical dipole*** is obtained by letting $d \to 0$ while u remains constant (i.e., $(q^+ - q^-) \to \infty$). This can be made mathematically precise using the theory of distributions, but we omit this here. One then obtains

$$V_{\text{math. dipole}}(p) = \frac{\langle u, p\rangle}{|p|^3}.$$

The Dipole Moment

In the general multipole expansion the second coefficient is

$$q_1(\theta) = \iiint_{\mathbb{R}^3} \varrho(q) \cdot |q| P_1(\cos\theta)\, dq = \iiint_{\mathbb{R}^3} \varrho(q)|q| \cos\angle(p,q)\, dq$$
$$= \frac{1}{|p|} \iiint_{\mathbb{R}^3} \varrho(q)\langle p, q\rangle\, dq = \frac{\langle p, u\rangle}{|p|}$$

where

$$u := \iiint_{\mathbb{R}^3} q\, \varrho(q)\, dq \tag{1.4.18}$$

is called the ***dipole moment***, generalizing (1.4.17). Note that while the monopole moment q_0 is a (constant) scalar, the dipole moment (1.4.18) is a vector. It is independent of p and enters into q_1 by taking the scalar product with $p/|p|$.

Higher-Order Terms and Dependence on Coordinates

While the monopole moment describes the total charge, the dipole moment describes the charge distribution within the body. This is also true of the higher-order terms, such as q_2 (related to the ***quadrupole moment***, a tensor) and q_3 (related to the ***octupole moment***).

The dipole moment is not in general independent of the choice of origin (or, equivalently, a position shift of the charged body): if $q' = q + \Delta q$ is the position of the shifted body and u' denotes the dipole moment of the shifted body, we have

$$u' = \iiint_{\mathbb{R}^3} q\, \varrho(q + \Delta q)\, dq = \iiint_{\mathbb{R}^3} (q - \Delta q)\, \varrho(q)\, dq = u - Q\Delta q.$$

Hence, if the total charge Q vanishes, the dipole moment is independent of the choice of origin. If not, then the origin needs to be explicitly stated for the dipole moment to be defined.

An Ideal Dipole

1.4.9. Example. Consider a ball of unit radius with charge density

$$\varrho(x, y, z) = \frac{z}{\sqrt{x^2 + y^2 + z^2}}.$$

Let $p = (0, 0, \zeta)$, $\zeta > 1$.

$$\begin{aligned} q_n &= \int_0^1 \int_0^{2\pi} \int_0^{\pi} \varrho_0 r^{n+2} \cos\theta \sin\theta P_n(\cos\theta)\, d\theta\, d\phi\, dr \\ &= \frac{2\pi}{n+3} \int_0^{\pi} \cos\theta \sin\theta P_n(\cos\theta)\, d\theta \\ &= \frac{2\pi}{n+3} \int_{-1}^{1} t \cdot P_n(t)\, dt \\ &= \frac{2\pi}{n+3} \langle P_1, P_n \rangle_{L^2} = \begin{cases} \frac{\pi}{3} & n = 1, \\ 0 & n \neq 1 \end{cases} \end{aligned}$$

so on the z-axis the only non-zero term is the dipole term.

On Convergence

In the previous sections, we have imbued the space of continuous functions $C([a, b])$ with the scalar product defined by

$$\langle u, v \rangle := \int_a^b \overline{u(x)} v(x)\, dx. \tag{1.5.1}$$

The induced norm is then

$$\|u\|_2 := \sqrt{\int_a^b |u(x)|^2\, dx} \tag{1.5.2}$$

which is different from the more commonly used

$$\|u\|_\infty := \sup_{x \in [a,b]} |u(x)|.$$

Cauchy Sequences

To understand the implications of the choice of norm, we consider the general setting. A sequence (u_n) in any normed vector space $(V, \|\cdot\|)$ converges to a limit $u \in V$ if and only if

$$\|u_n - u\| \to 0$$

It is then also true (why?) that

$$\underset{\varepsilon>0}{\forall} \underset{N\in\mathbb{N}}{\exists} \underset{n,m>N}{\forall} \quad \|u_n - u_m\| < \varepsilon, \tag{1.5.3}$$

i.e., the terms of the sequence are arbitrarily close to each other if N is sufficiently large. A sequence satisfying (1.5.3) is said to be a ***Cauchy sequence***.

Hence, every convergent sequence is a Cauchy sequence, but is the converse true? Does every Cauchy sequence converge to some limit?

Cauchy Sequences

For illustration, consider a sequence (u_n) of rational numbers, for example

$$u_n = \sum_{k=1}^{n} \frac{1}{k^2}$$

We see that, for $n > m > N$,

$$|u_n - u_m| = \sum_{k=m+1}^{n} \frac{1}{k^2} \leq \sum_{k=m+1}^{n} \frac{1}{(k-1)k} = \frac{1}{m} - \frac{1}{n} < \frac{1}{N}$$

so (u_n) is a Cauchy sequence. But (u_n) converges to an rational number $(\pi^2/6)$. Hence, if we were only considering the set of rational numbers, the sequence (u_n) would not have a limit.

This illustrates that whether or not every Cauchy sequence is a convergent sequence is a property of the normed vector space.

Completeness, Banach and Hilbert Spaces

1.5.1. Definition. A normed vector space $(V, \|\cdot\|)$ is said to be ***complete*** if every Cauchy sequence in V has a limit in V.

(i) A complete normed vector space is called a ***Banach space***.

(ii) An inner product space that is complete with respect to the induced norm is called a ***Hilbert space***.

We will often denote Hilbert spaces by the letter $\mathscr{H}$.

The completeness of a vector space is essential for many basic properties to hold. For example, we can answer the following question: given an orthonormal sequence (e_n) in a Hilbert space, can we choose any numbers λ_n and write out the sum

$$\sum_{n=0}^{\infty} \lambda_n e_n$$

to obtain a meaningful element in $\mathscr{H}$? (In the finite-dimensional case, this is of course trivially the case.)

Convergence and Absolute Convergence of Series

To illustrate the usefulness of the concept of completeness, we give a result concerning series. Given a sequence (a_n) of elements in a Banach space X, we might ask for a condition that ensures that the series

$$\sum_{n=0}^{\infty} a_n$$

converges to some element of X. A useful concept here is that of absolute convergence:

1.5.2. Definition. Let (a_n) be a sequence in a Banach space X. Then we say that the sequence is ***absolutely summable*** or that the series $\sum_{n=0}^{\infty} a_n$ is ***absolutely convergent*** if

$$\sum_{n=0}^{\infty} \|a_n\| < \infty.$$

Convergence and Absolute Convergence of Series

1.5.3. Lemma. Let (a_n) be a sequence in a Banach space X. Then

$$\sum_{n=0}^{\infty} a_n \quad \text{converges} \qquad \text{if} \qquad \sum_{n=0}^{\infty} \|a_n\| < \infty$$

Proof.

Let

$$S_n = \sum_{k=0}^{n} a_k, \qquad s_n = \sum_{k=0}^{n} \|a_k\|.$$

Then,

$$\|S_n - S_m\| = \left\| \sum_{k=m}^{n} a_k \right\| \leq \sum_{k=m}^{n} \|a_k\| = |s_n - s_m|.$$

If (s_n) converges, then (s_n) is Cauchy and so is (S_n). Since X is complete, this implies that (S_n) converges. □

Convergence of Orthonormal Sequences

1.5.4. Theorem. Let $\mathscr{H}$ be a Hilbert space and (e_n) an orthonormal sequence in $\mathscr{H}$.

(i) The series

$$\sum_{n=0}^{\infty} \lambda_n e_n$$

converges to an element $v \in \mathscr{H}$ if and only if

$$\sum_{n=0}^{\infty} |\lambda_n|^2 < \infty.$$

(ii) For any $v \in \mathscr{H}$, the sequence

$$\sum_{n=0}^{\infty} \langle e_n, v \rangle e_n$$

converges.

Convergence of Orthonormal Sequences

For the proof, we use the fact that the real numbers are complete (every real Cauchy sequence converges to some real number). This is usually proven in first-semester calculus.

Proof.

(i) Let

$$S_n = \sum_{k=0}^{n} \lambda_k e_k, \qquad s_n = \sum_{k=0}^{n} |\lambda_k|^2.$$

Then, by Pythagoras's Theorem 1.3.11,

$$\|S_n - S_m\|^2 = \left\| \sum_{k=m}^{n} \lambda_k e_k \right\|^2 = \sum_{k=m}^{n} \|\lambda_k e_k\|^2 = \sum_{k=m}^{n} |\lambda_k|^2 = |s_n - s_m|.$$

Hence (S_n) is Cauchy if and and only if (s_n) is Cauchy, i.e., if and only if (s_n) converges.

Convergence of Orthonormal Sequences

Proof (continued).

(i) Since $\mathscr{H}$ is complete, (S_n) converges if and only if (s_n) converges.

(ii) From the Bessel inequality (1.3.8) we have

$$\sum_{n=0}^{\infty} |\langle e_n, v\rangle|^2 \leq \|v\|^2 < \infty,$$

so applying (i) with $\lambda_n = \langle e_n, v\rangle$ the series

$$\sum_{n=0}^{\infty} \langle e_n, v\rangle e_n$$

converges. □

Criterion for Orthonormal Bases

In Hilbert spaces, we also have a very useful criterion for when an orthonormal system is also a basis.

1.5.5. Theorem. Let $\mathscr{H}$ be a Hilbert space. The span of an orthonormal system $\{e_n\} \subset \mathscr{H}$ is dense in $\mathscr{H}$ if and only if the only vector orthogonal to all e_n is the zero vector, i.e., if and only if

$$\forall_n \, v \perp e_n \qquad \Rightarrow \qquad v = 0$$

for any $v \in \mathscr{H}$.

Proof.

($\Rightarrow$) Suppose that $\operatorname{span}\{e_n\}$ is dense in $\mathscr{H}$. Then $\{e_n\}$ is a basis and

$$v = \sum_{n=1}^{\infty} \langle e_n, v \rangle e_n.$$

Hence, if $\langle v, e_n \rangle = 0$ for all n, then $v = 0$.

Criterion for Orthonormal Bases

Proof (continued).

($\Leftarrow$) By Remark 1.2.12 $\operatorname{span}\{e_n\}$ is dense if and only if

$$\forall_{\varepsilon>0} \; \forall_{w\in\mathscr{H}} \; \exists_{v\in\operatorname{span}\{e_n\}} \quad \|v-w\| < \varepsilon.$$

Suppose that $\operatorname{span}\{e_n\}$ is not dense in $\mathscr{H}$. We will show that then there exists a vector that is orthogonal to all the e_n. First,

$$\exists_{\varepsilon>0} \; \exists_{w\in\mathscr{H}} \; \forall_{v\in\operatorname{span}\{e_n\}} \quad \|v-w\| \geq \varepsilon.$$

Choose such an $\varepsilon > 0$ and a $w \in \mathscr{H}$. This w can not be the zero element (why?), so $w \neq 0$. Now define the sequence (v_n) by

$$v_n := \sum_{i=1}^{n} \langle e_i, w\rangle e_i \in \operatorname{span}\{e_n\}.$$

Criterion for Orthonormal Bases

Proof (continued).

($\Leftarrow$) Then $\|w - v_n\| \geq \varepsilon$ for all n. By Theorem 1.5.4, the sequence (v_n) converges,

$$\left\| w - \sum_{i=1}^{\infty} \langle e_i, w \rangle e_i \right\| \geq \varepsilon.$$

Define $u := w - \sum_{i=1}^{\infty} \langle e_i, w \rangle e_i$. Since $\|u\| \geq \varepsilon$, $u \neq 0$ and

$$\langle e_k, u \rangle = \langle e_k, w \rangle - \langle e_k, w \rangle = 0$$

for all $e_k \in \{e_n\}$. Hence, $\underset{n}{\forall}\, u \perp e_n \not\Rightarrow u = 0$. $\square$

Parseval's Theorem

Another important consequence is Parseval's identity:

1.5.6. Parseval's Theorem. Let $\mathscr{H}$ be a Hilbert space and $\{e_n\}$ an orthonormal sequence in $\mathscr{H}$. Then

$$\|v\|^2 = \sum_{n=0}^{\infty} |\langle e_n, v\rangle|^2 \qquad \text{for all } v \in \mathscr{H} \tag{1.5.4}$$

if and only if span$\{e_n\}$ is dense in $\mathscr{H}$.

For the proof, we refer to [Kreyszig, Theorem 3.6-3].

Many other results we will develop later also depend on the completeness if the inner product space, so it is important to discuss the completeness of the main application so far, the set of continuous functions on an interval.

Completeness of the Space of Continuous Functions

We will show that

- $(C([a,b], \|\cdot\|_\infty)$ is complete but
- $(C([a,b], \|\cdot\|_2)$ is not complete.

Before we prove these statements we remark that they have serious implications:

The norm $\|\cdot\|_2$ is induced by a scalar product and we can use it together with orthonormal bases. However, because the space is not complete, we will have difficulty proving certain results later on.

On the other hand, the space of continuous functions is complete with respect to $\|\cdot\|_\infty$, but we can prove that $\|\cdot\|_\infty$ is not induced by any scalar product, so this norm does not work well with orthonormal bases.

Completeness of $C([a,b])$ with $\|\cdot\|_\infty$

We will first show that $C([a,b])$ is complete in the norm $\|\cdot\|_\infty$. For this, we need a preliminary result, which should be familiar from calculus.

1.6.1. Theorem. Let $[a,b] \subset \mathbb{R}$ be a closed interval. Let (f_n) be a sequence of continuous functions defined on $[a,b]$ such that $f_n(x)$ converges to some $f(x) \in \mathbb{R}$ as $n \to \infty$ for every $x \in [a,b]$. If the sequence (f_n) converges uniformly to the thereby defined function $f : [a,b] \to \mathbb{R}$, then f is continuous.

Proof.
We need to show that f is continuous for all $x \in [a,b]$. We will here deal only with $x \in (a,b)$; the cases $x = a$ and $x = b$ are left to you.

Fix $x \in (a,b)$. We will show that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $|h| < \delta$ implies $|f(x+h) - f(x)| < \varepsilon$ (for h so small that $x + h \in (a,b)$).

Completeness of $C([a,b])$ with $\|\cdot\|_\infty$

Proof (continued).

Fix $\varepsilon > 0$. Then there exists some $N \in \mathbb{N}$ such that

$$\|f_n - f\|_\infty = \sup_{x\in[a,b]} |f_n(x) - f(x)| < \frac{\varepsilon}{3}.$$

for all $n > N$. Since each f_n is continuous on $[a,b]$, there exists some $\delta > 0$ such that $|h| < \delta$ implies

$$|f_n(x) - f_n(x+h)| < \frac{\varepsilon}{3}.$$

Then for $|h| < \delta$ we have

$$\begin{aligned} |f(x+h) - f(x)| &\le |f(x+h) - f_n(x+h)| + |f_n(x+h) - f_n(x)| \\ &\quad + |f_n(x) - f(x)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

□

Completeness of $C([a,b])$ with $\|\cdot\|_\infty$

1.6.2. Theorem. The normed vector space $(C([a,b]), \|\cdot\|_\infty)$ is complete.

Proof.
Let (f_n) be a Cauchy sequence in $C([a,b])$. We will show that $\lim_{n\to\infty} f_n(x)$ exists for every $x \in [a,b]$. First, by definition, for every $\varepsilon > 0$ we have

$$\|f_n - f_m\|_\infty = \sup_{x\in[a,b]} |f_n(x) - f_m(x)| < \varepsilon$$

for n, m sufficiently large. But then, for every fixed $x \in [a,b]$, we have

$$|f_n(x) - f_m(x)| < \varepsilon$$

for n, m sufficiently large. This implies that for every $x \in [a,b]$ the sequence or real numbers $(f_n(x))$ is Cauchy. Since the real numbers are complete, $(f_n(x))$ converges.

Completeness of $C([a,b])$ with $\|\cdot\|_\infty$

Proof (continued).

Hence we can define the limit $f(x) := \lim_{n\to\infty} f_n(x)$ for every $x \in [a,b]$.

Now fix $\varepsilon > 0$ and choose N so that $\|f_n - f_m\| < \varepsilon$ for $n, m > N$. Then for fixed $n > N$ we have

$$\begin{aligned}
\|f - f_n\|_\infty &= \sup_{x\in[a,b]} |f(x) - f_n(x)| \\
&= \sup_{x\in[a,b]} \lim_{m\to\infty} |f_m(x) - f_n(x)| \\
&\leq \sup_{x\in[a,b]} \sup_{m>N} |f_m(x) - f_n(x)| \\
&= \sup_{m>N} \sup_{x\in[a,b]} |f_m(x) - f_n(x)| < \sup_{m\geq N} \varepsilon = \varepsilon.
\end{aligned}$$

Finally, by Theorem 1.6.1, f is continuous, so $f \in C([a,b])$. □

Incompleteness of $C([a,b])$ with $\|\cdot\|_2$

We will first show that $(C([a,b], \|\cdot\|_2)$ is not complete by exhibiting a Cauchy sequence that does not converge to a continuous function. Consider the sequence (u_n) in $C([0,1])$ given by

$$u_n(x) = \begin{cases} \sqrt[4]{n} & 0 \le x \le 1/n, \\ 1/\sqrt[4]{x} & 1/n < x \le 1. \end{cases}$$

For $m > n > N$

$$\|u_n - u_m\|_2^2 = \int_0^1 |u_n(x) - u_m(x)|^2 \, dx \le \int_0^{1/N} \frac{1}{\sqrt{x}} \, dx = \frac{2}{\sqrt{N}}$$

so (u_n) is a Cauchy sequence.

Incompleteness of $C([a,b])$ with $\|\cdot\|_2$

Note that a pointwise limit of the sequence does not exist, since

$$\lim_{n\to\infty} u_n(0) = \lim_{n\to\infty} \sqrt[4]{n} = \infty.$$

By itself, this alone does not preclude a continuous function u from existing such that $u_n \to u$ in the norm $\|\cdot\|_2$.

Suppose a function $u \in C([0,1])$ exists such that $\|u_n - u\|_2 \to 0$ as $n \to \infty$. Then u is bounded, so $u(x) \leq M$ for some $M \in \mathbb{N}$ and all $x \in [0,1]$. Furthermore, $u_n(x) \geq 2M$ for all $x \in [0, 1/(2M)^4]$ if $n \geq (2M)^4$. Hence,

$$\|u_n - u\|_2 \geq \int_0^{1/(2M)^4} |u_n(x) - u(x)|^2 \, dx \geq \frac{1}{16M^2} \not\to 0$$

as $n \to \infty$, giving a contradiction. Hence, the Cauchy sequence (u_n) does not have a limit in $C([0,1])$.

Completion of a Vector Space

If a vector space is not complete, then there are some Cauchy sequences that don't have limits. One then tries to construct the ***completion*** of the vector space, i.e., a slightly larger space in which the original one is embedded and that contains all the missing limits. We illustrate how this is done using the example of the rational numbers.

Given $\mathbb{Q}$, we may consider the space of all sequences in $\mathbb{Q}$ that converge to a limit. Denote this space by $\mathrm{Conv}(\mathbb{Q})$. Each sequence $(a_n) \in \mathrm{Conv}(\mathbb{Q})$ is associated uniquely to a number $a \in \mathbb{Q}$, namely its limit. We can now say that two sequences are equivalent if they have the same limit, i.e.,

$$(a_n) \sim (b_n) \qquad :\Leftrightarrow \qquad \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n. \tag{1.6.1}$$

(This is an ***equivalence relation***.) We then denote the set of all sequences with the same limit as a sequence (a_n) by $[a(a_n)]$. Such a set is called a ***(equivalence) class*** and the set of all classes is denoted $\mathrm{Conv}(\mathbb{Q})/\sim$.

Construction of the Real Numbers

Since each rational number is represented by a class (why?) we have that

$$\mathbb{Q} \simeq \mathrm{Conv}(\mathbb{Q})/\sim .$$

This is just a formal way of saying that the set of rational numbers corresponds to the set of all convergent sequences of rational numbers, if any two sequences with the same limit are considered equivalent. Any rational number corresponds to exactly one class of sequences and vice-versa.

We can now consider a large class of sequences, that of Cauchy sequences of rational numbers, denoted by $\mathrm{Cauchy}(\mathbb{Q})$. Since every convergent sequence is a Cauchy sequence, $\mathrm{Conv}(\mathbb{Q}) \subset \mathrm{Cauchy}(\mathbb{Q})$. Furthermore, we say that two Cauchy sequences are equivalent not if they have the same limit (because they might not converge) but rather if their difference converges to zero:

$$(a_n) \sim (b_n) \qquad :\Leftrightarrow \qquad \lim_{n\to\infty} (a_n - b_n) = 0. \tag{1.6.2}$$

Construction of the Real Numbers

Of course, (1.6.2) is equivalent to (1.6.1) for convergent sequences. We now have the larger set

$$\text{Cauchy}(\mathbb{Q})/\sim \supset \text{Conv}(\mathbb{Q})/\sim \simeq \mathbb{Q}$$

This larger set now incorporates the rational numbers and by its construction every Cauchy sequence (a_n) in the set has a limit, namely precisely the object represented by the class $[(a_n)]$. We denote

$$\mathbb{R} := \text{Cauchy}(\mathbb{Q})/\sim$$

and call this set the ***real numbers***.

1.6.3. Example. Every rational number has a finite decimal representation. We can think of a real number as having an "infinite decimal representation." For example, the sequence

$$3,\ 3.1,\ 3.14,\ 3.141,\ 3.1415,\ 3.14159,\ 3.141592, \ldots$$

may converge to π if the following terms are chosen appropriately.

Construction of the Real Numbers

This "infinite decimal representation" is just the way that real numbers are introduced in middle school. As another example, the sequences

$$(a_n) := (0.4,\ 0.49,\ 0.499,\ 0.4999,\ 0.49999, \ldots)$$

and

$$(b_n) := (0.5,\ 0.5,\ 0.5,\ 0.5,\ 0.5, \ldots)$$

are equivalent in the sense of (1.6.2), since

$$|a_n - b_n| = 10^{-(n+1)} \xrightarrow{n\to\infty} 0.$$

Hence, 0.499999 ... and 0.5 are considered to represent the same real number.

Now the general procedure for the completion of a vector space is exactly the same as for the rational numbers: one takes the set of all Cauchy sequences and defines two Cauchy sequences to be equivalent under the relation (1.6.2). This set is then called the completion of the original space and the original space is embedded in it in a natural way.

The Spaces of p-Integrable Functions

1.6.4. Definition and Theorem. For any $p \geq 1$ the vector space of ***p-integrable functions*** on an interval $[a, b]$ is defined as the closure of $C([a, b])$ with respect to the norm given by

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

This space is denoted by $L^p([a, b])$. Furthermore,

(i) The elements of $L^p([a, b])$ are equivalence classes of functions, where two functions are in the same class if they have the same values ***almost everywhere***.

(ii) The Riemann integral can be extended to the so-called ***Lebesgue integral*** such that the integral of $|u|^p$ exists for all elements of $L^p([a, b])$ (see Example 2.1.12 in the next part). In fact,

$$L^p([a, b]) = \left\{ u \colon [a, b] \to \mathbb{C} \colon \int_a^b |u(x)|^p \, dx < \infty \right\}. \tag{1.6.3}$$

The Spaces of p-Integrable Functions

1.6.5. Remarks.

(i) We will consider two functions identical if they differ only at a finite number of points, e.g., the functions

$$u_1(x) = \begin{cases} 0 & x \le 0, \\ 1 & 0 < x < 1, \\ 0 & x \ge 1, \end{cases} \qquad u_2(x) = \begin{cases} 0 & x < 0, \\ 1 & 0 < x < 1, \\ 0 & x > 1, \\ 1/2 & x = 0 \text{ or } x = 1, \end{cases}$$

are in the same class and considered to be the same function. This is analogous to considering the numbers $0.4999\ldots$ and 0.5 to be the same number.

Actually, two functions are considered identical if they are the same ***almost everywhere***. This means that they differ only on a set of measure zero, for example at countably infinitely many points. However, the technical definition is not important for us at this point.

The Spaces of p-Integrable Functions

(ii) The above convention leads to functions that are not necessarily Riemann integrable. For example, the Riemann integral of the Dirichlet function

$$\chi\colon [0,1] \to \mathbb{R}, \qquad \chi(x) = \begin{cases} 1 & \text{if } x \text{ is rational,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases}$$

does not exist, although the Dirichlet function is in the same class as the zero function. This (very technical) problem can be solved by introducing the Lebesgue integral, which agrees with the Riemann integral where the latter exists and also allows functions such as the Dirichlet function to be integrated.

We will not go into the technical construction of the Lebesgue integral here. The final result is that all functions in the same class have the same integral, so we may just choose an arbitrary representative from each class when calculating an integral.

The Fourier-Legendre Basis

We are particularly interested in $L^2([a,b])$, since it is the only L^p space possessing a norm induced by a scalar product. From now on, we will formulate all of our results in L^2 spaces. As a first step, we will verify that the (normalized) Legendre polynomials give a basis of $L^2([-1,1])$.

1.6.6. Theorem. Let $e_n := P_n/\|P_n\|_2$, where P_n is the nth Legendre polynomial. Then $\mathscr{B} = (e_n)_{n\in\mathbb{N}}$ is an orthonormal basis of $L^2([-1,1])$. This basis is called the ***Fourier-Legendre*** basis of $L^2([-1,1])$.

Proof.
We already know that $\mathscr{B}$ is an orthonormal sequence in $L^2([-1,1])$. We need to show that $\operatorname{span}\mathscr{B}$ is also dense in $L^2([-1.1])$. Now

$$\operatorname{span}\mathscr{B} = \operatorname{span}\underbrace{\{m_0, m_1, m_2, \ldots\}}_{=:\mathscr{M}}$$

where $m_k(x) = x^k$, $k \in \mathbb{N}$.

The Fourier-Legendre Basis

Proof (continued).

The Weierstraß Approximation Theorem 1.2.14 states that $\operatorname{span}\mathscr{M}$ is dense (in the $\|\cdot\|_\infty$ norm) in $C([-1,1])$, so the same is true for $\operatorname{span}\mathscr{B}$.

Hence, for any $\varepsilon > 0$ and any $u \in C([-1,1])$ there exists a polynomial $p \in \operatorname{span}\mathscr{B}$ such that

$$\|u - p\|_\infty = \sup_{x\in[-1,1]} |u(x) - p(x)| < \varepsilon.$$

However, since

$$\begin{aligned}\|u - p\|_2 &= \left(\int_{-1}^{1} |u(x) - p(x)|^2\, dx\right)^{1/2} \leq \sqrt{2} \sup_{x\in[-1,1]} |u(x) - p(x)| \\ &= \sqrt{2}\|u - p\|_\infty\end{aligned}$$

we see that $\operatorname{span}\mathscr{B}$ is also dense in $C([-1,1])$ in the L^2-norm.

The Fourier-Legendre Basis

Proof (continued).

Thus, the closure of $\operatorname{span}\mathscr{B}$ in $L^2([a,b])$ contains at least all continuous functions, i.e.,

$$C([-1,1]) \subset \overline{\operatorname{span}\mathscr{B}}^{\|\cdot\|_2}.$$

Since $L^2([-1,1]) := \overline{C([-1,1])}^{\|\cdot\|_2}$, we can take the closure on both sides of the above subset relation and obtain

$$L^2([-1,1]) \subset \overline{\operatorname{span}\mathscr{B}}^{\|\cdot\|_2}.$$

But since $\overline{\operatorname{span}\mathscr{B}}^{\|\cdot\|_2} \subset L^2([-1,1])$ by definition,

$$\overline{\operatorname{span}\mathscr{B}}^{\|\cdot\|_2} = L^2([-1,1])$$

so $\operatorname{span}\mathscr{B}$ is dense in $L^2([-1,1])$. □

Weighted Square-integrable Functions

In applications, we will often use a slightly generalized version of $L^2([a,b])$, as follows:

1.6.7. Definition. Let $I \subset \mathbb{R}$ be an interval. A continuous function $r\colon I \to [0,\infty)$ such that $r(x) > 0$ almost everywhere is called a ***weight function*** on I.

1.6.8. Definition and Theorem. Let $I \subset \mathbb{R}$ be an interval and $r\colon I \to [0,\infty)$ a weight function on I. Then the set

$$L^2(I; r(x)\,dx) := \left\{ u\colon I \to \mathbb{C} : \int_I |u(x)|^2\, r(x)\,dx < \infty \right\}$$

defines the ***vector space of square-integrable functions f with respect to the weight function r.*** If $r \equiv 1$, we write $L^2(I)$ for short.

Weighted Square-integrable Functions

1.6.9. Definition and Theorem. Let $r\colon I \to [0,\infty)$ be a weight function on I. Then the map $\langle \cdot\,,\,\cdot \rangle_{L^2(I;r(x)\,dx)}\colon L^2(I;r(x)\,dx) \times L^2(I;r(x)\,dx) \to \mathbb{C}$ given by

$$\langle u, v\rangle_{L^2(I;r(x)\,dx)} := \int_I \overline{u(x)}v(x)\,r(x)\,dx, \tag{1.6.4}$$

where $\overline{u(x)}$ denotes the complex conjugate of $u(x)$, defines a scalar product on $L^2(I;r(x)\,dx)$

1.6.10. Remark. We can construct $L^2(I;r(x)\,dx)$ either directly as in Definition 1.6.8 using the concept of the Lebesgue integral or as the completion of $C(I)$, the space of continuous functions on $I \subset \mathbb{R}$, with respect to the norm induced by the scalar product (1.6.4).

Other Orthonormal Systems

There are numerous common orthonormal systems in $L^2(I, r(x)\,dx)$ that occur in various applications. Some examples are given below:

I	$r(x)$	Complete Orthonormal System
$[-\pi, \pi]$	1	Fourier Basis
$[0, 1]$	x	dilated & scaled Bessel functions
$[-1, 1]$	1	normalized Legendre polynomials
$(-1, 1)$	$\frac{1}{\sqrt{1-x^2}}$	normalized Chebychev polynomials
$[0, \infty)$	e^{-x}	normalized Laguerre polynomials
$(-\infty, \infty)$	e^{-x^2}	normalized Hermite polynomials

Hermite and Laguerre polynomials are discussed in [Kreyszig, Section 3.7].

The Real Fourier-Euler Basis of $L^2([-\pi,\pi])$

One of the most important orthonormal bases in $L^2([-\pi,\pi])$ is the ***real Fourier-Euler basis*** given by

$$\mathscr{B}_{\mathscr{F}} = \left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos(nx), \frac{1}{\sqrt{\pi}}\sin(nx) \right\}_{n=1}^{\infty}. \tag{1.7.1}$$

It is easy to check (see assignments) that these functions are actually orthonormal, i.e., for $m, n \in \mathbb{N}$,

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \sin(mx)\sin(nx)\,dx = \begin{cases} 0 & n \neq m, \\ 1 & n = m, \end{cases}$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \cos(mx)\cos(nx)\,dx = \begin{cases} 0 & n \neq m, \\ 1 & n = m \neq 0, \\ 2 & n = m = 0, \end{cases}$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \sin(mx)\cos(nx)\,dx = 0$$

The Real Fourier-Euler Basis of $L^2([-\pi, \pi])$

Although (1.7.1) is easily seen to be an orthonormal system, the proof that $\mathscr{B}_{\mathscr{F}}$ is a basis, i.e., that its span is dense in $L^2([-\pi, \pi])$, is more complicated. We will defer this proof for now.

1.7.1. Theorem. The orthonormal system

$$\left\{ \frac{1}{\sqrt{b-a}}, \sqrt{\frac{2}{b-a}} \cos\left(\frac{2\pi n(x-a)}{b-a}\right), \sqrt{\frac{2}{b-a}} \sin\left(\frac{2\pi n(x-a)}{b-a}\right) \right\}_{n=1}^{\infty}$$

is a basis of $L^2([a, b])$.

The Real Fourier-Euler Basis of $L^2([-\pi,\pi])$

We would now expect that any function $f \in L^2([-\pi,\pi])$ can then be expanded in terms of the basis functions:

$$\begin{aligned} f(x) &= \left\langle \frac{1}{\sqrt{2\pi}}, f \right\rangle_{L^2} \frac{1}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} \left\langle \frac{1}{\sqrt{\pi}} \cos(nx), f \right\rangle_{L^2} \frac{1}{\sqrt{\pi}} \cos(nx) \\ &\quad + \sum_{n=1}^{\infty} \left\langle \frac{1}{\sqrt{\pi}} \sin(nx), f \right\rangle_{L^2} \frac{1}{\sqrt{\pi}} \sin(nx) \\ &= \frac{\langle f, 1 \rangle_{L^2}}{2\pi} + \sum_{n=1}^{\infty} \frac{\langle \cos(nx), f \rangle_{L^2}}{\pi} \cos(nx) + \sum_{n=1}^{\infty} \frac{\langle \sin(nx), f \rangle_{L^2}}{\pi} \sin(nx) \end{aligned} \tag{1.7.2}$$

Such an expansion is called the ***Fourier-Euler series*** of f. However, the actual situation is slightly more complicated, as the following example shows.

Fourier-Euler Series

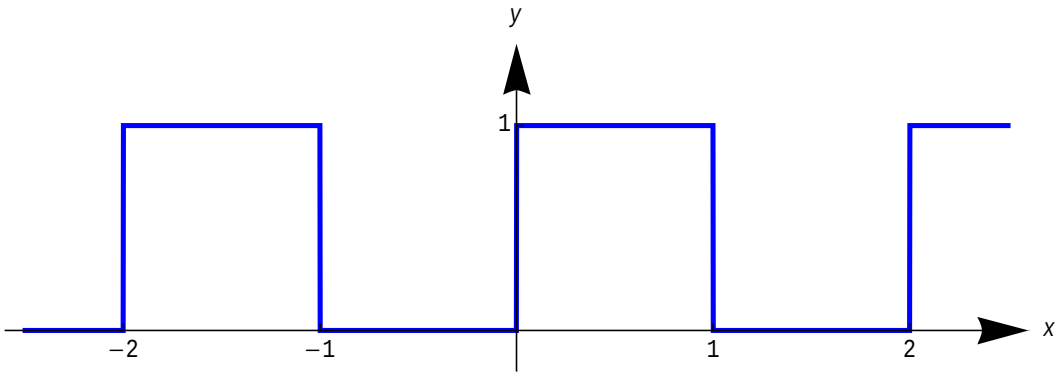
1.7.2. Example. We calculate the Fourier series for the function $f \in L^2([0,2])$ given by

$$f(x) = \begin{cases} 1 & 0 \leq x < 1, \\ 0 & 1 \leq x \leq 2. \end{cases}$$

The representation of f as a Fourier series is

$$f(x) = \frac{\langle f, 1\rangle_{L^2}}{2} + \sum_{n=1}^{\infty} \langle \cos(n\pi x), f\rangle_{L^2} \cos(n\pi x) + \sum_{n=1}^{\infty} \langle \sin(n\pi x), f\rangle_{L^2} \sin(n\pi x)$$

Fourier-Euler Series



$$
\begin{aligned}
\langle f, 1\rangle_{L^2} &= 1, \\
\langle f, \cos(n\pi x)\rangle_{L^2} &= \int_0^1 1 \cdot \cos(n\pi x)\, dx = \frac{1}{n\pi}\sin(n\pi \cdot 1) - \frac{1}{n\pi}\sin(n\pi \cdot 0) \\
&= 0, \\
\langle f, \sin(nx)\rangle_{L^2} &= \int_0^1 1 \cdot \sin(n\pi x)\, dx = \frac{1}{n\pi}\cos(n\pi \cdot 0) - \frac{1}{n\pi}\cos(n\pi \cdot 1) \\
&= \frac{1-(-1)^n}{n\pi}.
\end{aligned}
$$

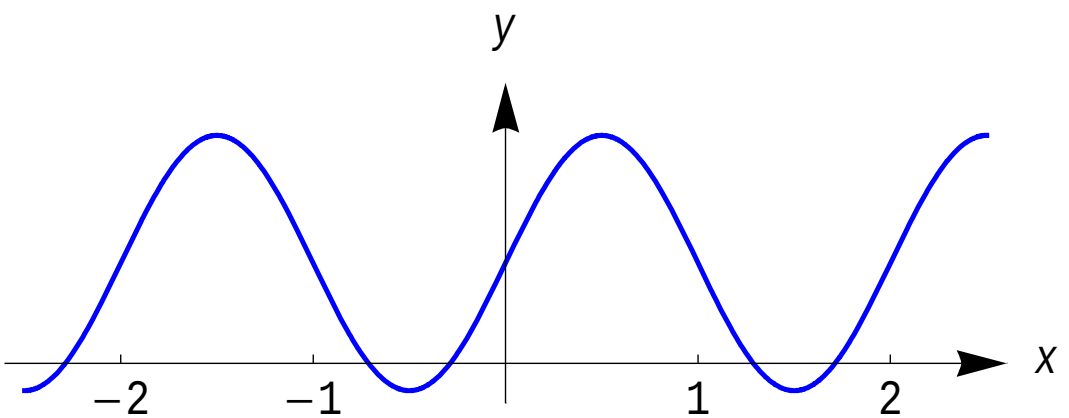
Fourier-Euler Series

Hence, the Fourier series gives

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty} \frac{1-(-1)^n}{n} \sin(n\pi x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty} \frac{\sin((2k+1)\pi x)}{2k+1}$$

For $x = 1/2$ we obtain the well-known formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - + \ldots$$

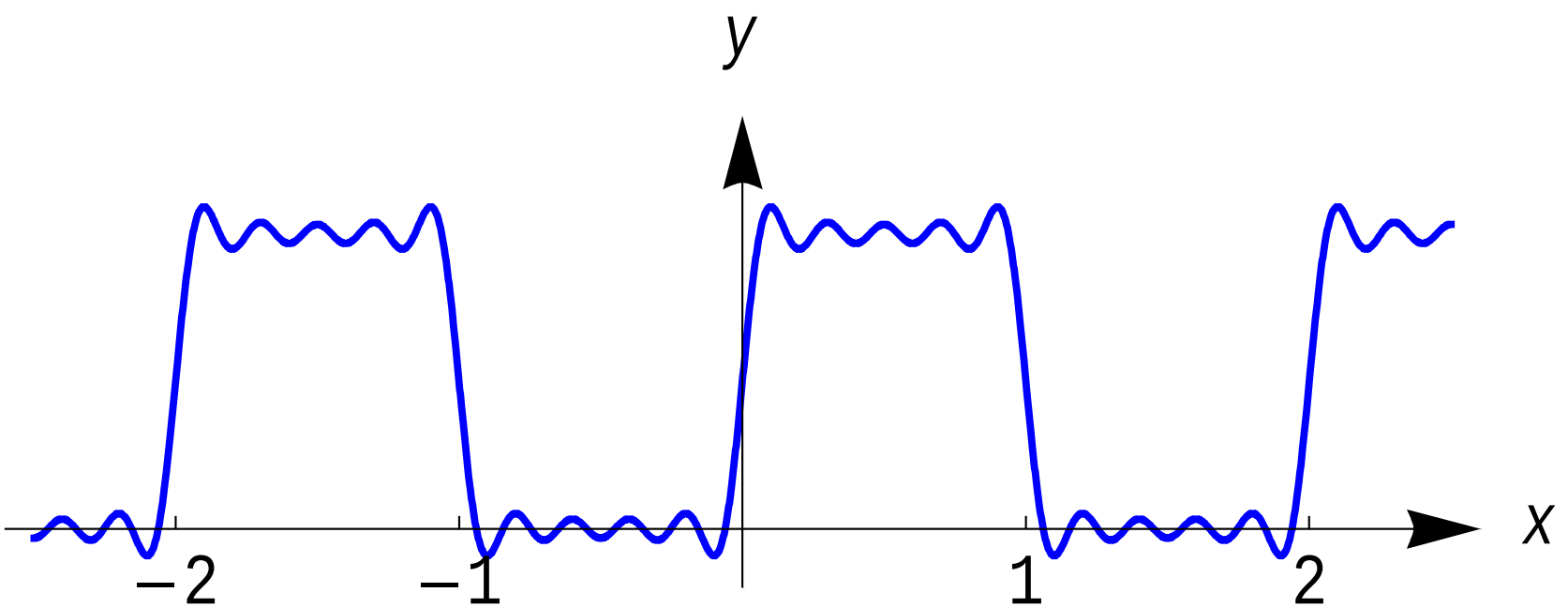


The Fourier expansion with just one term in the series.

Fourier-Euler Series

Hence, the Fourier series gives

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1-(-1)^n}{n}\sin(n\pi x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\sin((2k+1)\pi x)}{2k+1}$$

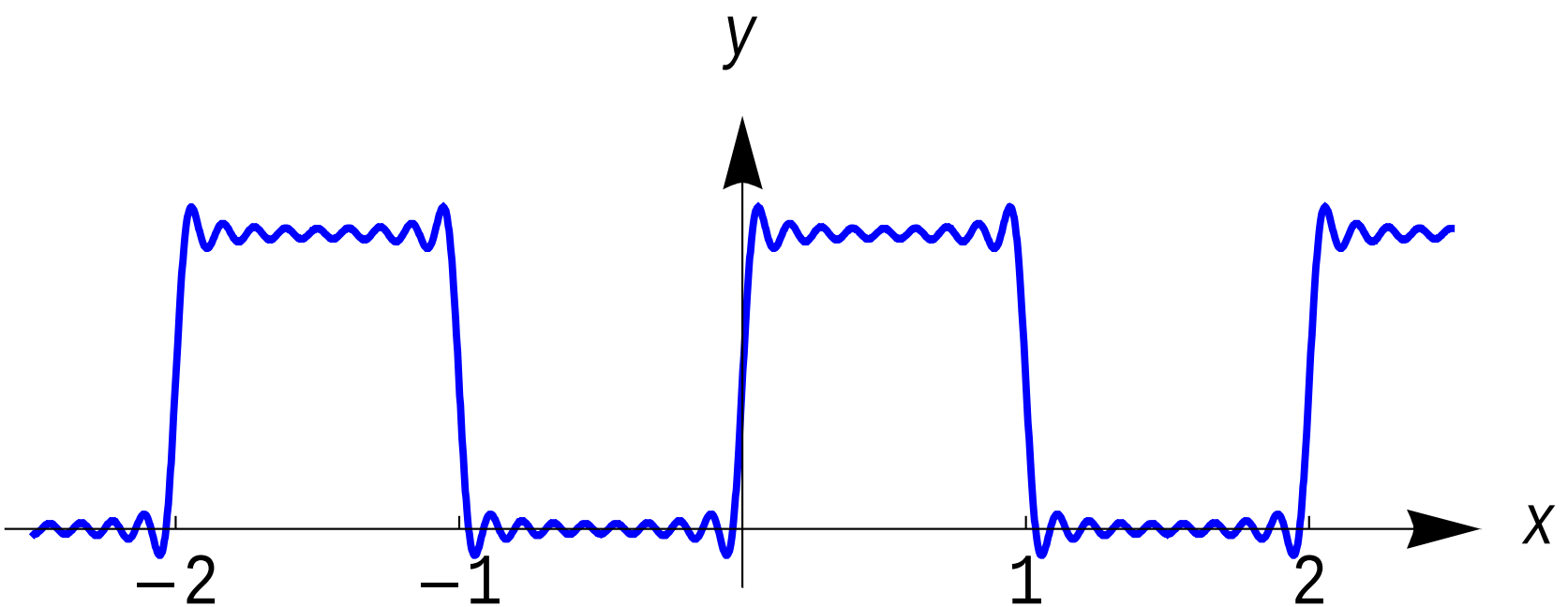


The Fourier expansion with five terms in the series.

Fourier-Euler Series

Hence, the Fourier series gives

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty} \frac{1-(-1)^n}{n} \sin(n\pi x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty} \frac{\sin((2k+1)\pi x)}{2k+1}$$

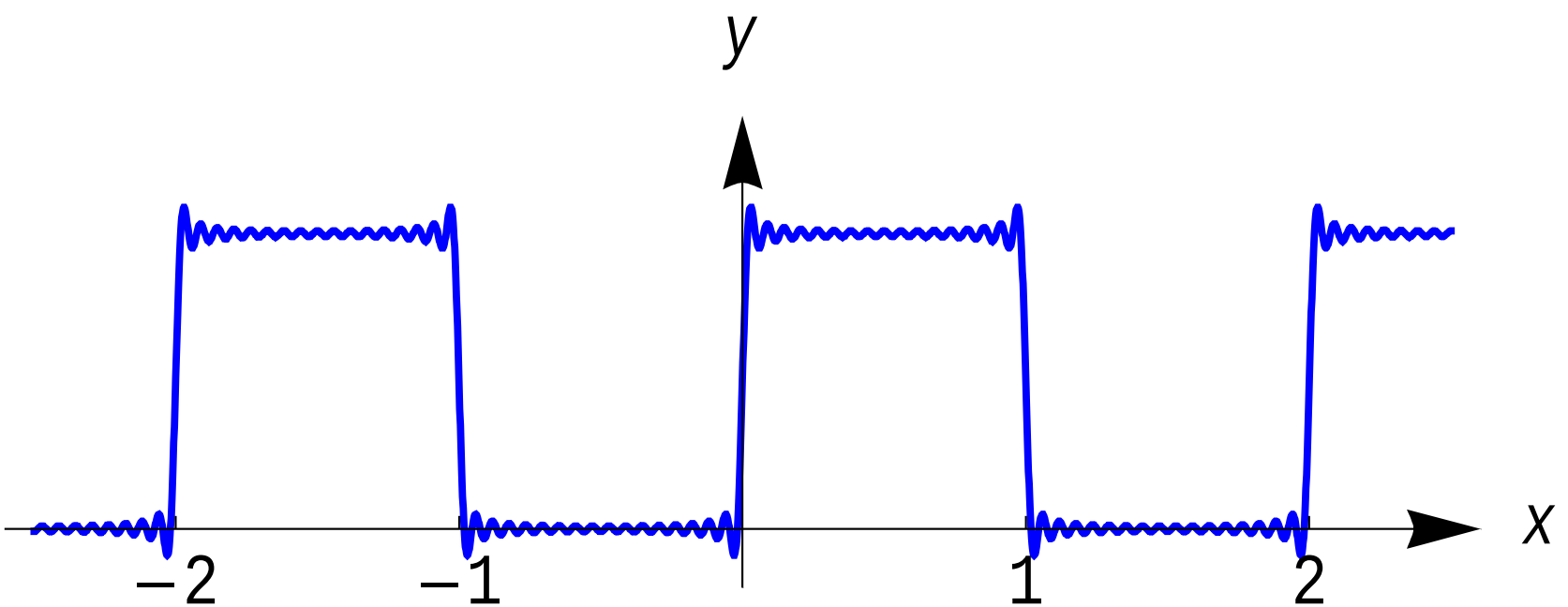


The Fourier expansion with nine terms in the series.

Fourier-Euler Series

Hence, the Fourier series gives

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty} \frac{1-(-1)^n}{n} \sin(n\pi x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty} \frac{\sin((2k+1)\pi x)}{2k+1}$$

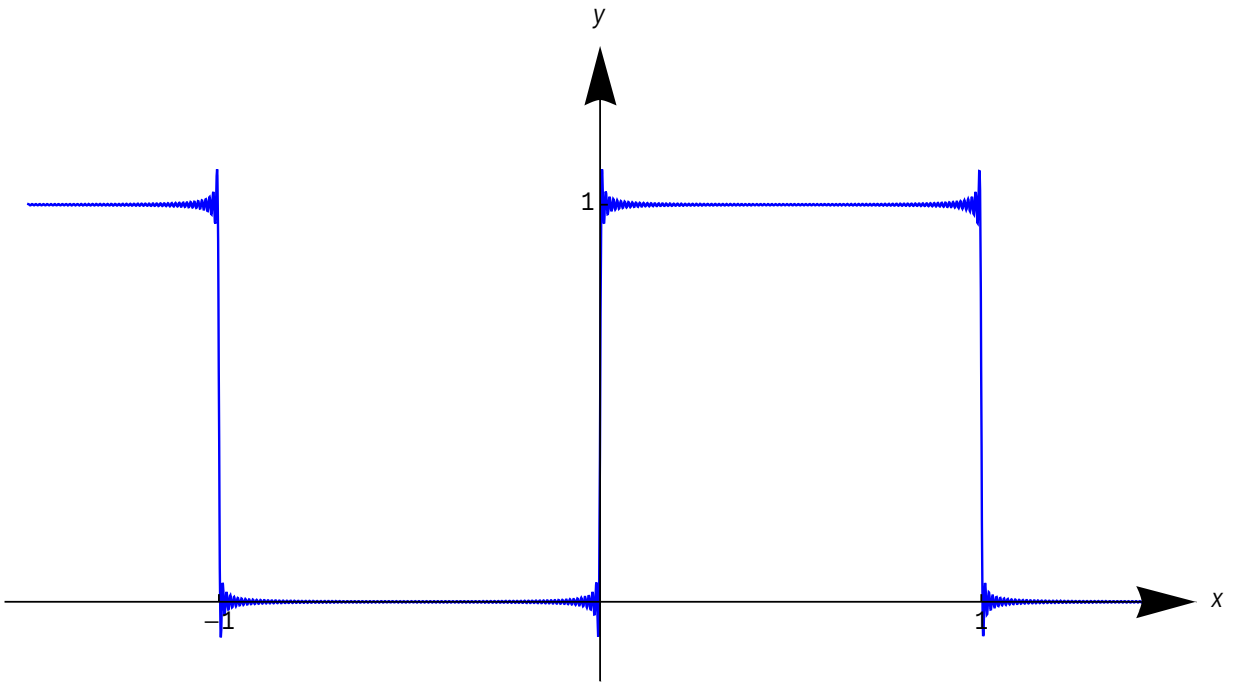


The Fourier expansion with nineteen terms in the series.

Fourier-Euler Series

It follows that

$$f(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{1-(-1)^n}{n}\sin(n\pi x) = \frac{1}{2} + \frac{2}{\pi}\sum_{k=0}^{\infty}\frac{\sin((2k+1)\pi x)}{2k+1}$$



The Fourier expansion with one hundred terms in the series.

Convergence of Fourier Series

It becomes obvious from this example that a Fourier series does not need to converge uniformly to the function; the height of the "peaks" near the jump discontinuities does not decrease. (The occurrence of these peaks is known as the ***Gibbs phenomenon***.)

The reason for this is that convergence is only with respect to the L^2 norm, i.e., for $f \in L^2([-\pi, \pi])$

$$\|S_N - f\|^2_{L^2([-\pi,\pi])} = \int_{-\pi}^{\pi} |S_N(x) - f(x)|^2 \, dx \xrightarrow{N\to\infty} 0, \tag{1.7.3}$$

where

$$S_N(x) = \frac{\langle f, 1\rangle_{L^2}}{2\pi} + \sum_{n=1}^{N} \frac{\langle \cos(nx), f\rangle_{L^2}}{\pi} \cos(nx) + \sum_{n=1}^{N} \frac{\langle \sin(nx), f\rangle_{L^2}}{\pi} \sin(nx).$$

Hence (1.7.3) does not imply pointwise convergence, i.e., that $S_N(x) \to f(x)$ for all $x \in [0, 2]$. In fact, due to the jump discontinuity of f, this is plainly impossible.

Convergence of Fourier Series

We see that care must be taken when writing that "f equals its Fourier series." The precise analysis of the convergence is quite complicated. For example, there exist continuous functions that are nowhere equal to their Fourier series!

The discussion of Fourier series in terms of basis functions gives a good background for understanding which functions can in principle be expanded in terms of trigonometric series. However, for a deep understanding the "abstract generalities" of vector space theory are not sufficient and one needs to do some hard analysis using the specific properties of the sine and cosine functions. This is typical of our current subject: the unifying approach of the abstract theory gives a basic understanding of phenomena, but does not absolve us of concrete, precise calculations when it comes to discussing the more subtle points.

However, in the case of Fourier analysis (which would merit an entire course by itself) we lack the time to go into these details. We merely quote one of the most basic theorems regarding pointwise convergence.

Convergence of Fourier Series

The following result (which we will not prove) clarifies the question of convergence for many applications:

1.7.3. Theorem. Let $f \in L^2([a,b])$ be piecewise continuously differentiable. Then

(i) On any subinterval $[a',b'] \subset [a,b]$ with $a' > a$, $b' < b$ on which f is continuous the Fourier series converges uniformly towards f.

(ii) At any point $x \in [a,b]$, we have the pointwise limit

$$S_N(x) \xrightarrow{N\to\infty} \frac{\lim\limits_{y\nearrow x} f(y) + \lim\limits_{y\searrow x} f(y)}{2}.$$

(This is known as ***Dirichlet's rule***.)

Thus, at jump discontinuities of f the Fourier series converges pointwise towards the "mean value" of f near this point. This is precisely what we have observed in the previous example.

Pure Sine and Cosine Fourier Bases

Other orthonormal bases that are related to the real Fourier basis for $L^2([0, L])$ are the following:

1. The complex Fourier-Euler Basis:

$$\mathscr{B}_1 := \left\{ \frac{1}{\sqrt{L}} e^{2\pi i n x / L} \right\}_{n=-\infty}^{\infty} \tag{1.7.4}$$

2. The Fourier-Cosine Basis:

$$\mathscr{B}_2 := \left\{ \frac{1}{\sqrt{L}}, \sqrt{\frac{2}{L}} \cos\left(\frac{\pi n x}{L}\right) \right\}_{n=1}^{\infty} \tag{1.7.5}$$

3. The Fourier-Sine Basis:

$$\mathscr{B}_3 := \left\{ \sqrt{\frac{2}{L}} \sin\left(\frac{\pi n x}{L}\right) \right\}_{n=1}^{\infty} \tag{1.7.6}$$

Looking back: Finite-Dimensional Vector Spaces

Relationship to Linear Algebra?

Looking back over the previous sections, one may well ask, why certain topics are not discussed in an undergraduate linear algebra course:

- open and closed sets are important in calculus, but are never mentioned in linear algebra
- convergence of sequences, completeness of vector spaces is not a topic of linear algebra
- norms are defined, but the influence of the choice of a norm for a given vector space is never discussed

The reason for these omissions is simple: linear algebra is the study of ***finite-dimensional*** vector spaces, and in such spaces all the above issues vanish. The questions we have dealt with are truly relevant only for infinite-dimensional spaces (although our theorems are of course also valid for finite-dimensional spaces).

We will now discuss this in more detail.

Equivalent Norms

1.8.1. Definition. Let V be a vector space on which we may define two norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Then the two norms are said to be ***equivalent*** if there exist two constants $C_1, C_2 > 0$ such that

$$C_1\|x\|_1 \leq \|x\|_2 \leq C_2\|x\|_1 \qquad \text{for all } x \in V. \tag{1.8.1}$$

1.8.2. Example. In $\mathbb{R}^n$ we have (amongst others) the following two possible choices of norms:

$$\|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, \qquad \|x\|_\infty := \max_{1\leq i\leq n} |x_i|. \tag{1.8.2}$$

It is easily verified that for all $x \in \mathbb{R}^n$,

$$\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2,$$

so the two norms are equivalent.

Equivalent Norms

If two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent, the vector space V endowed with either of these norms, i.e., $(V, \|\cdot\|_1)$ and $(V, \|\cdot\|_2)$ has the same topology. That means, for example, that a sequence (v_n) converges in $(V, \|\cdot\|_1)$ if and only if it converges in $(V, \|\cdot\|_2)$. Similarly, a set $\Omega \subset V$ is open in $(V, \|\cdot\|_1)$ if and only if it is open in $(V, \|\cdot\|_2)$.

Therefore, the following theorem is of fundamental importance:

1.8.3. Theorem. In a finite-dimensional vector space, all norms are equivalent.

A major consequence of Theorem 1.8.3 is that if we have several norms at our disposal in a finite-dimensional space, then we can freely choose a convenient one in order to show openness of sets, convergence of sequences, etc.

The proof of Theorem 1.8.3 requires some preliminary work.

The Theorem of Bolzano-Weierstraß

We recall two basic facts from the theory of sequences of real numbers:

(i) Every bounded and monotonic sequence of real numbers converges.

(ii) Every sequence of real numbers has a monotonic subsequence.

Together, these yield the following fundamental result:

1.8.4. Theorem of Bolzano-Weierstraß. Every bounded sequence of real numbers has a convergent subsequence.

1.8.5. Remark. The Theorem of Bolzano-Weierstraß easily implies that every Cauchy sequence of real numbers converges, because every Cauchy sequence that has a convergent subsequence must itself converge. Since Cauchy sequences are always bounded, every Cauchy sequence in $\mathbb{R}$ converges.

The Theorem of Bolzano-Weierstraß

The basic ingredient in proving that the real numbers (with the usual modulus norm) are complete is the fact that a bounded, monotonic sequence converges. The monotonicity is a specific property of the real numbers, so the proof does not carry over to general vector spaces.

However, we can generalize the Theorem of Bolzano-Weierstraß to $\mathbb{R}^n$.

The Theorem of Bolzano-Weierstraß in $\mathbb{R}^n$

1.8.6. Theorem of Bolzano–Weierstraß in $\mathbb{R}^n$. Let $(x^{(m)})_{m\in\mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^n$, i.e., $x^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$. Suppose that there exists a constant $C > 0$ such that $|x_k^{(m)}| < C$ for all $m \in \mathbb{N}$ and each $k = 1, \dots, n$. Then there exists a subsequence $(x^{(m_j)})_{j\in\mathbb{N}}$ that converges to a vector $y \in \mathbb{R}^n$.

Proof.
Consider the real coordinate sequence $(x_1^{(m)})_{m\in\mathbb{N}}$. By assumption, this sequence is bounded, so by the Theorem of Bolzano-Weierstraß 1.8.4 there exists a convergent subsequence $(x_1^{(m_{j_1})})$ with some limit, say $y_1 \in \mathbb{R}$.

The second coordinate sequence $(x_2^{(m)})$ is also bounded and has a convergent subsequence, but this subsequence does not need to have the same indices as that for $(x_1^{(m)})$.

The Theorem of Bolzano-Weierstraß in $\mathbb{R}^n$

Proof (continued).

We therefore employ a trick: The subsequence $(x_2^{(m_{j_1})})$ that uses the indices from our above subsequence for the first coordinate is of course also bounded and hence has a sub-subsequence $(x_2^{(m_{j_2})})$ that converges, say to $y_2 \in \mathbb{R}$. Taking the corresponding sub-subsequence for the first coordinate, $(x_1^{(m_{j_2})})$ still converges to y_1.

Similarly, a sub-sub-subsequence of the third coordinate will converge to some $y_3 \in \mathbb{R}$ while the corresponding sub-sub-subsequences of the first two coordinates will still converge to y_1 and y_2, respectively. Repeating the procedure n times, the n-fold subsequence $(x_k^{(m_{j_n})})$ converges to some $y_k \in \mathbb{R}$, $k = 1, \ldots, n$. Hence, the subsequence $(x^{(m_{j_n})})$ converges to some $y \in \mathbb{R}^n$. □

A Basic Norm inequality

All our further results are based on the following basic estimate:

1.8.7. Lemma. Let $(V, \|\cdot\|)$ be a finite- or infinite-dimensional normed vector space and $\{v_1, \dots, v_n\}$ an independent set in V. Then there exists a $C > 0$ such that for any $\lambda_1, \dots, \lambda_n \in \mathbb{F}$

$$\|\lambda_1 v_1 + \cdots + \lambda_n v_n\| \geq C\big(|\lambda_1| + \cdots + |\lambda_n|\big). \tag{1.8.3}$$

Proof.

Let $s := |\lambda_1| + \cdots + |\lambda_n|$. If $s = 0$, then all $\lambda_k = 0$ and the inequality (1.8.3) holds trivially for any C, so we can assume $s > 0$. Dividing by s, (1.8.3) becomes

$$\|\mu_1 v_1 + \cdots + \mu_n v_n\| \geq C, \qquad \sum_{k=1}^{n} |\mu_k| = 1, \tag{1.8.4}$$

with $\mu_k = \lambda_k / s$.

A Basic Norm inequality

Proof (continued).

Hence, we need to show

$$\exists_{C>0} \quad \forall_{\substack{\mu_1,\ldots,\mu_n\in\mathbb{F} \\ |\mu_1|+\cdots+|\mu_n|=1}} \quad \|\mu_1 v_1 + \cdots + \mu_n v_n\| \geq C.$$

Suppose that this is false, i.e.,

$$\forall_{C>0} \quad \exists_{\substack{\mu_1,\ldots,\mu_n\in\mathbb{F} \\ |\mu_1|+\cdots+|\mu_n|=1}} \quad \|\mu_1 v_1 + \cdots + \mu_n v_n\| < C.$$

In particular, choosing $C = 1/m$, $m = 1, 2, 3, \ldots$, we can find a sequence of vectors

$$u^{(m)} := \mu_1^{(m)} v_1 + \cdots + \mu_n^{(m)} v_n$$

such that $\|u^{(m)}\| \to 0$ as $m \to \infty$ and $|\mu_1^{(m)}| + \cdots + |\mu_n^{(m)}| = 1$ for all m.

A Basic Norm inequality

Proof (continued).

Hence, for each $k = 1, \dots, n$, $|\mu_k^{(m)}| \leq 1$ and so each coefficient sequence $(\mu_k^{(m)})$ is bounded. Write

$$\mu^{(m)} := (\mu_1^{(m)}, \dots, \mu_n^{(m)})$$

By the Theorem of Bolzano Weierstraß in $\mathbb{R}^n$, there exists a subequence of vectors $(\mu^{(m_j)})_{j\in\mathbb{N}}$ that converges to some $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$. This corresponds to a subsequence $u^{(m_j)}$ of $u^{(m)}$ such that

$$u^{(m_j)} \xrightarrow{j\to\infty} \alpha_1 v_1 + \cdots + \alpha_n v_n =: u \qquad \text{with } |\alpha_1| + \cdots + |\alpha_n| = 1.$$

Since the vectors $v_1, \dots, v_n$ are independent and not all α_k vanish, it follows that $u \neq 0$.

A Basic Norm inequality

Proof (continued).

It is easy to see that $u^{(m_j)} \to u$ as $j \to \infty$ implies

$$\|u^{(m_j)}\| \xrightarrow{j\to\infty} \|u\| \neq 0.$$

But by our construction, $\|u^{(m)}\| \to 0$ as $m \to \infty$, so the subsequence $(\|u^{(m_j)}\|)$ must also converge to zero. This gives a contradiction. □

We can now proceed to prove Theorem 1.8.3.

Equivalence of Norms

Proof of Theorem 1.8.3.

Let V be a finite-dimensional vector space, $\|\cdot\|$ be any norm on V and $\{v_1, \dots, v_n\}$ a basis of V. Let $v \in V$ have the representation $v = \lambda_1 v_1 + \cdots + \lambda_n v_n$ with $\lambda_1, \dots, \lambda_n \in \mathbb{F}$. By the triangle inequality,

$$\|v\| = \|\lambda_1 v_1 + \cdots + \lambda_n v_n\| \leq \sum_{i=1}^{n} |\lambda_i| \|v_i\| \leq C \sum_{i=1}^{n} |\lambda_i|$$

where $C := \max_{1 \leq i \leq n} \|v_i\|$ depends only on the basis and not on v. We hence see that for any norm there are constants $C_1, C_2 > 0$ such that

$$C_1 \sum_{i=1}^{n} |\lambda_i| \leq \|v\| \leq C_2 \sum_{i=1}^{n} |\lambda_i|, \tag{1.8.5}$$

where the first inequality is just (1.8.3). Given two norms $\|\cdot\|_1$ and $\|\cdot\|_2$, it follows from their respective inequalities (1.8.5) that (1.8.1) holds. □

Completeness of Finite-Dimensional Spaces

Another consequence of Lemma 1.8.7 is the following result:

1.8.8. Theorem. Any finite-dimensional normed vector space is complete.

Proof.
Let $(V, \|\cdot\|)$ be a finite-dimensional normed vector space, $\dim V = n$. Let $(v^{(m)})$ be a Cauchy sequence in V and $\{b_1, \dots, b_n\}$ a basis of V. Then we can write

$$v^{(m)} = \sum_{k=1}^{n} \lambda_k^{(m)} b_k$$

and with the estimate (1.8.3) it is easy to see that for each k the coordinate sequence $(\lambda_k^{(m)})$ is also Cauchy. Since the real and complex numbers are complete, these sequences converge, say $\lambda_k^{(m)} \to \lambda_k$ as $n \to \infty$.

Closedness of Finite-Dimensional Subspaces

Proof (continued).
Set

$$v := \sum_{k=1}^{n} \lambda_k b_k.$$

Then it is easy to see that $\|v^{(m)} - v\| \to 0$ as $n \to \infty$, so the Cauchy sequence $(v^{(m)})$ converges. □

1.8.9. Corollary. Any finite-dimensional subspace of a normed vector space is closed.

Proof.
Suppose $(V, \|\cdot\|)$ is a normed vector space and U a finite-dimensional subspace. Suppose that (u_n) is a sequence in U that converges to some $v \in V$. Then (u_n) is a Cauchy sequence in V (and in U). Since U is finite-dimensional, by Theorem 1.8.8 U is complete, so $v \in U$. But this shows that U is closed in V. □

Looking Back

We have seen that the choice of norm is arbitrary in finite-dimensional spaces and that there are no open or dense finite-dimensional subspaces. All such spaces are automatically complete, so the terms Hilbert space and Banach space are not used in linear algebra, as there is no need to distinguish complete spaces.

Moreover, the situation of the Weierstraß Approximation theorem in which the infinite-dimensional space of continuous functions on an interval is the closure of the (infinite-dimensional) subspace of polynomials can not occur in linear algebra. The theory of infinite-dimensional spaces has turned out to be much more complex than just the addition of "infinite bases" and offers many more possibilities useful in applications.

In the following part we will study linear maps on infinite-dimensional spaces. There, too, the theory turns out to be much richer than that of linear maps between finite-dimensional spaces.

First Midterm Exam

The preceding material completes the first third of the course material. It encompasses everything that will be the subject of the First Midterm Exam.

The exam date will be announced on Canvas.

No calculators or other aids will be permitted during the exam.

Part II

Linear Maps

Linear Functionals and Operators

Linear Operators

In this part of the course, we will study linear maps between vector spaces. We first fix some definitions:

2.1.1. Definition. Let U, V be vector spaces over $\mathbb{F}$. Then a map

$$L\colon U \to V$$

satisfying

$$L(\alpha u + \beta u') = \alpha Lu + \beta Lu' \qquad \text{for all } u, u' \in U,\ \alpha, \beta \in \mathbb{F}$$

is called a ***linear operator***, ***linear map*** or ***linear transformation*** from U to V.

(i) If $U = V$ we say that L is a linear operator on V.

(ii) If $V = \mathbb{F}$, L is called a ***linear functional***.

Linear Operators

The ***range*** and ***kernel*** of L are defined by

$$\operatorname{ran} L := \{x \in V : x = Lu \text{ for some } u \in U\},$$
$$\ker L := \{u \in U : Lu = 0\},$$

respectively. (The kernel is also sometimes called the ***null space*** of L.) If $U \subset W$ is a subspace of W, then we say that $U =: \operatorname{dom} L$ is the ***domain*** of L.

2.1.2. Examples.

(i) Let $U = \mathbb{R}^n$. Then $L(x_1, \dots, x_n) := x_1$ is a linear functional on $\mathbb{R}^n$.

(ii) Any $m \times n$ matrix is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix} : \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \to \begin{pmatrix} x_1 + 2x_2 + 3x_3 \\ x_2 + x_3 \end{pmatrix}$$

is a linear operator $\mathbb{R}^3 \to \mathbb{R}^2$.

Linear Operators

(iii) Let $U = \mathscr{P}([0,1]) \subset C([0,1])$ be the space of polynomial functions defined on the interval $[0,1]$ and

$$L = \frac{d}{dx} : \mathscr{P}([0,1]) \to \mathscr{P}([0,1]).$$

Then L is linear and we have

$$\operatorname{dom} L = \operatorname{ran} L = \mathscr{P}([0,1]),$$

$$\ker L = \left\{ p \in \mathscr{P}([0,1]) : \exists_{c \in \mathbb{R}} \forall_{x \in [0,1]} \; p(x) = c \right\}.$$

(iv) Let $U = C([0,1])$ be the space of continuous functions defined on the interval $[0,1]$ and

$$L = \int_0^x : C([0,1]) \to C([0,1]), \qquad f \mapsto \int_0^x f(y)\,dy.$$

Then L is linear and

$$\operatorname{dom} L = C([0,1]), \qquad \operatorname{ran} L = C^1([0,1]), \qquad \ker L = \{0\}.$$

Linear Operators

(v) Let X be a normed vector space. Then the functional

$$L\colon X \to \mathbb{R}, \qquad Lu = \|u\|$$

is not linear (since $L(-u) = Lu \neq -Lu$).

(vi) Let $\mathscr{H}$ be an inner product space and $v \in \mathscr{H}$ be a fixed vector. Then the map

$$L\colon \mathscr{H} \to \mathbb{F}, \qquad Lu = \langle v, u\rangle \tag{2.1.1}$$

is a linear functional on $\mathscr{H}$.

(vii) Let $\mathscr{H}$ be a complex inner product space and $v \in \mathscr{H}$ be a fixed vector. Then the functional

$$L\colon \mathscr{H} \to \mathbb{C}, \qquad Lu = \langle u, v\rangle$$

is not linear (since $L(\alpha u) = \overline{\alpha} Lu \neq \alpha Lu$).

Left- and Right-Shift Operators

2.1.3. Example. We can define the following linear maps on $U = \ell^2$, the space of square-summable complex sequences:

- The ***left-shift operator*** $L\colon \ell^2 \to \ell^2$,

$$L(a_0, a_1, a_2, \ldots) := (a_1, a_2, \ldots),$$

with

$$\operatorname{ran} L = \ell^2, \qquad \ker L = \{(a_n) \in \ell^2 \colon a_n = 0 \text{ for } n > 0\}.$$

- The ***right-shift operator*** $R\colon \ell^2 \to \ell^2$,

$$R(a_0, a_1, a_2, \ldots) := (0, a_0, a_1, a_2, \ldots).$$

with

$$\operatorname{ran} R = \{(a_n) \in \ell^2 \colon a_0 = 0\}, \qquad \ker L = \{0\}.$$

Bounded Linear Operators

2.1.4. Definition. Let X, Y be normed vector spaces, $\Omega \subset X$ and $L\colon \Omega \to Y$ a linear operator. Then L is said to be ***bounded*** if there exists a constant $C > 0$ such that

$$\|Lx\|_Y \leq C \cdot \|x\|_X \qquad \text{for all } x \in \Omega.$$

The smallest such constant is given by

$$\|L\| := \sup_{\substack{x \in X \\ x \neq 0}} \frac{\|Lx\|_Y}{\|x\|_X} \tag{2.1.2}$$

and called the ***operator norm*** (or ***induced norm***) of L.

Bounded Linear Operators

2.1.5. Examples.

(i) The linear functional $L\colon \mathbb{R}^n \to \mathbb{R}$, $L(x_1, \dots, x_n) := x_1$ satisfies

$$|Lx| = |x_1| \leq \|x\|_2$$

and hence is bounded with $\|L\| \leq 1$. Let $x_0 = (1, 0, \dots, 0)$. Then

$$\|L\| = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{|Lx|}{\|x\|} \geq \frac{|Lx_0|}{\|x_0\|_2} = 1,$$

so we see that $\|L\| = 1$.

(ii) The linear functional (2.1.1) is bounded and has norm

$$\|L\| = \|v\|$$

where $\|v\|^2 = \langle v, v \rangle$.

Bounded Linear Operators

(iii) The integral operator in $(C([0,1]), \|\cdot\|_\infty)$

$$L = \int_0^x : C([0,1]) \to C([0,1]), \qquad f \mapsto \int_0^x f(y)\,dy.$$

is bounded, since

$$\|Lf\|_\infty = \sup_{x\in[0,1]} \left| \int_0^x f(y)\,dy \right| \leq = \sup_{x\in[0,1]} \int_0^x |f(y)|\,dy$$

$$= \int_0^1 |f(y)|\,dy \leq 1 \cdot \sup_{x\in[0,1]} |f(x)| = \|f\|_\infty.$$

Hence, $\|L\| \leq 1$. To prove that $\|L\| = 1$, take the function $m_0(x) = 1$. Then

$$\|L\| = \sup_{\substack{f\in C([0,1]) \\ f\neq 0}} \frac{\|Lf\|_\infty}{\|f\|_\infty} \geq \frac{\|Lm_0\|_\infty}{\|m_0\|_\infty} = 1.$$

Bounded Linear Operators

(iv) The operator

$$L = \frac{d}{dx} : \mathscr{P}([0,1]) \to \mathscr{P}([0,1]).$$

is not bounded with respect to the $\|\cdot\|_\infty$ norm: let $m_n(x) = x^n$. Then

$$\|m_n\|_\infty = 1 \qquad \text{but} \qquad \|Lm_n\|_\infty = n,$$

so it is impossible to find a constant C such that $\|Lp\|_\infty \leq C\|p\|_\infty$ for all $p \in \mathscr{P}([0,1])$.

(v) Both the left-shift and the right-shift operator introduced in Example 2.1.3 are bounded and have operator norm

$$\|L\| = \|R\| = 1. \tag{2.1.3}$$

Continuous Linear Operators

The boundedness of linear operators is closely related to their continuity. We first give the formal definition of the latter:

2.1.6. Definition. Let X, Y be Banach spaces, $U \subset X$ a subspace and $L\colon U \to Y$ a linear operator. We say that L is continuous at $u \in U$ if

$$\forall_{\varepsilon>0}\ \exists_{\delta>0}\ \forall_{v\in U}\ \|u - v\|_X < \delta \Rightarrow \|Lu - Lv\|_Y < \varepsilon.$$

We say that L is continuous if L is continuous at every $u \in U$.

2.1.7. Theorem. A linear operator $L\colon U \to Y$ is continuous at $u \in U$ if and only if for any sequence (u_n) in U,

$$u_n \to u \qquad \Rightarrow \qquad Lu_n \to Lu.$$

The proof of the theorem is completely analogous to the proof of the corresponding theorem for real functions and will be omitted.

Bounded Linear Operators

The continuity of linear operators is a crucial property for many calculations; this will become evident in the next section.

It turns out that continuity is precisely equivalent to boundedness:

2.1.8. Theorem. Let X, Y be Banach spaces, $U \subset X$ a subspace and $L\colon U \to Y$ a linear operator. Then the following statements are equivalent:

(i) L is bounded.

(ii) L is continuous.

(iii) L is continuous at 0.

This is the main reason for our interest in bounded linear operators.

Boundedness and Continuity

Proof.

- (i) $\Rightarrow$ (ii) Assume that $L\colon U \to Y$ is linear and bounded. Then we need to show that L is continuous. Let (u_n) be a sequence in U converging to $u \in U$, i.e., $\|u_n - u\|_X \to 0$. Then

$$\|Lu_n - Lu\|_Y = \|L(u_n - u)\|_Y \leq \|L\| \cdot \underbrace{\|u_n - u\|_X}_{\to 0} \to 0.$$

  Thus $u_n \to u$ implies $Lu_n \to Lu$, so L is continuous.
- (ii) $\Rightarrow$ (iii) Trivial.
- (iii) $\Rightarrow$ (i) If L is continuous at 0 we know that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|u\|_X < \delta \qquad \Rightarrow \qquad \|Lu\|_Y < \varepsilon. \tag{2.1.4}$$

  Suppose that L is also not bounded.

Boundedness and Continuity

Proof (continued).

Then for every $c > 0$ there exists a $u \in U$ such that

$$\|Lu\|_Y > c \cdot \|u\|_X. \tag{2.1.5}$$

Now fix $\varepsilon > 0$ and choose $\delta > 0$ so that (2.1.4) holds. Next, set $c = 4\varepsilon/\delta$ and choose a u such that (2.1.5) holds. Set $\widetilde{u} := u \cdot \delta/(2\|u\|_X)$. Then

$$\|\widetilde{u}\|_X = \left\| \frac{u}{\|u\|_X} \cdot \frac{\delta}{2} \right\|_X = \frac{\delta}{2} \frac{\|u\|_X}{\|u\|_X} = \delta/2 < \delta$$

and

$$\|L\widetilde{u}\|_Y = \left\| \frac{1}{\|u\|_X} \cdot \frac{\delta}{2} \cdot Lu \right\|_Y = \frac{\delta}{2} \frac{1}{\|u\|_X} \|Lu\|_Y > \frac{\delta}{2} \frac{1}{\|u\|_X} \frac{4\varepsilon}{\delta} \|u\|_X > 2\varepsilon.$$

But this contradicts (2.1.4). □

The Space of Bounded Linear Operators

2.1.9. Definition and Theorem. Let X, Y be vector spaces and $\Omega \subset X$ a linear subspace. Then the set of all bounded linear operators,

$$\mathscr{L}(\Omega, Y) := \{L \colon \Omega \to Y \colon L \text{ is linear and bounded}\},$$

is a vector space with pointwise addition and scalar multiplication.

If X, Y are Banach spaces, then $(\mathscr{L}(\Omega, Y), \|\cdot\|)$ is also a Banach space with the operator norm (2.1.2).

We omit the proof of the above statements; it is easy to show that $\mathscr{L}(\Omega, Y)$ is a vector space and that the operator norm defines a norm; the proof of completeness of the space is more complicated.

Extension of Bounded Linear Operators

As we have seen in the previous examples, sometimes linear operators can only immediately be defined on a subspace of a vector space that we are interested in. For example, the differentiation operator

$$L = \frac{d}{dx} : \mathscr{P}([0,1]) \to \mathscr{P}([0,1]).$$

can not be defined on the entire space of continuous functions $C([0,1])$. The question we pose now is:

> *Can we define an operator $\overline{L}$ on all of $C([0,1])$ that coincides with L on $\mathscr{P}([0,1])$?*

The operator $\overline{L}$ is called an ***extension*** of L to $C([0,1])$. More generally, we want to extend an operator L from its domain U to the closure $\overline{U}$ of the domain. If the domain is dense, that is the whole space.

Such an extension will exist (even uniquely!) if L is continuous.

Extension of Bounded Linear Operators

2.1.10. B.L.T. Theorem. Let X, Y be Banach spaces and U a subspace of X. Denote by $\overline{U}$ the closure of U. Let $L\colon U \to Y$ be a ***bounded*** linear operator. Then there exists a unique extension $\overline{L}$ of L to a continuous linear map

$$\overline{L}\colon \overline{U} \to Y.$$

Before we prove the B.L.T. Theorem, we note that for unbounded operators this is not possible. In fact, we even have the following theorem:

2.1.11. Theorem. Let X, Y be Banach spaces and U a subspace of X. Let $L\colon U \to Y$ be an unbounded operator. Then there does not exist an extension of L to the entire space U.

Hence, unbounded operators can never be defined on an entire Banach space! In particular, the differentiation operator L can not be extended to $C([0,1])$.

The B.L.T. Theorem

Proof of the B.L.T. Theorem.

The proof proceeds in various steps:

1. We first show that there exists an extension $\overline{L}$ of L to $\overline{U}$. Let $x \in \overline{U}$. Then there exists a sequence (x_n), $x_n \in U$, such that $x_n \to x$. Since (x_n) converges, it is Cauchy. This means that

$$\forall_{\varepsilon>0} \exists_{N\in\mathbb{N}} \forall_{n,m>N} \quad \|x_n - x_m\|_X < \varepsilon.$$

Since $\|Lx_n - Lx_m\|_Y < \|L\| \cdot \|x_n - x_m\|_X$ this implies

$$\forall_{\varepsilon>0} \exists_{N\in\mathbb{N}} \forall_{n,m>N} \quad \|Lx_n - Lx_m\|_Y < \|L\| \cdot \|x_n - x_m\|_X < \varepsilon,$$

so the sequence (Lx_n) is Cauchy in Y. Since Y is complete, the sequence (Lx_n) converges to some $y \in Y$.

The B.L.T. Theorem

Proof of the B.L.T. Theorem (continued).

We hence define

$$\overline{L}x := \lim_{n\to\infty} Lx_n \qquad \text{for } x \in \overline{U}.$$

We need to check that $\overline{L}$ is ***well-defined***. This means that if there are two sequences (x_n) and (x_n') that both converge to x, then we require that

$$\overline{L}x = \lim_{n\to\infty} Lx_n = \lim_{n\to\infty} Lx_n'.$$

We construct a sequence $(x_1, x_1', x_2, x_2', \ldots)$. This sequence will converge to x, so it is Cauchy and hence $(Lx_1, Lx_1', Lx_2, Lx_2', \ldots)$ is Cauchy. Since Y is complete, $(Lx_1, Lx_1', Lx_2, Lx_2', \ldots)$ converges. But then any subsequence also converges to the same limit. Hence

$$\lim_{n\to\infty} Lx_n = \lim_{n\to\infty} Lx_n'.$$

The B.L.T. Theorem

Proof of the B.L.T. Theorem (continued).

2. We next show that $\overline{L}$ is bounded:

$$\begin{aligned}\|\overline{L}x\|_Y &= \|\lim_{n\to\infty} Lx_n\|_Y = \lim_{n\to\infty}\|Lx_n\|_Y \\ &\leq \lim_{n\to\infty}\|L\|\cdot\|x_n\|_X = \|L\|\cdot\|\lim_{n\to\infty} x_n\|_X \\ &= \|L\|\cdot\|x\|_X\end{aligned}$$

where we have used the continuity of $\|\cdot\|_Y$ and $\|\cdot\|_X$. In particular, we see that $\|\overline{L}\| = \|L\|$.

3. We now check that $\overline{L}$ is linear: let $x, y \in \overline{U}$ with $(x_n) \to x$, $(y_n) \to y$. Then

$$\overline{L}(x+y) = \big(\lim_{n\to\infty} L(x_n+y_n)\big) = \lim_{n\to\infty} Lx_n + \lim_{n\to\infty} Ly_n = \overline{L}x + \overline{L}y.$$

The homogeneity is shown similarly.

The B.L.T. Theorem

Proof of the B.L.T. Theorem (continued).

4. Finally we check that $\overline{L}$ is a ***unique*** continuous extension of L. Let $\overline{L}'$ be some other continuous extension of L. Let $x \in \overline{U}$ and $(x_n) \to x$ with $x_n \in U$. Then

$$\overline{L}'x = \overline{L}'\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} \overline{L}'x_n = \lim_{n\to\infty} Lx_n = \overline{L}x. \qquad \square$$

The Lebesgue Integral

2.1.12. Example. By construction (see Definition 1.6.4) the space of square integrable functions $L^2([a,b])$ is the completion of the space of continuous functions $C([a,b])$ with respect to the norm

$$\|u\|_2 := \left(\int_a^b |u(x)|^2\,dx\right)^{1/2}.$$

Here, the (Riemann-)integral is of course defined for all continuous functions and

$$T\colon u \mapsto \int_a^b u(x)\,dx$$

is a bounded linear map on $C([a,b])$ with respect to $\|\cdot\|_2$ (prove this!). The B.L.T. theorem now states that T can be extended to a bounded linear map $\overline{T}$ on $L^2([a,b])$. This map is just the Lebesgue integral alluded to in Remark 1.6.5 ii).

Riesz Representation Theorem

One of the most remarkable results about linear functionals on Hilbert spaces is that they are essentially scalar products. More precisely, any bounded linear functional can be written as a scalar product with a fixed vector:

2.2.1. Riesz Representation Theorem. Let $\mathscr{H}$ be a (possibly infinite-dimensional) Hilbert space and $L\colon \mathscr{H} \to \mathbb{F}$ a bounded linear functional. Then there exists a unique element $v \in \mathscr{H}$ such that

$$Lu = \langle v, u \rangle \qquad \text{for all } u \in \mathscr{H}. \tag{2.2.1}$$

Furthermore, the operator norm of L is equal to the norm of v,

$$\|L\| = \|v\|_{\mathscr{H}}.$$

Riesz Representation Theorem

Proof.

Let L be a given linear functional. If $\ker L = \mathscr{H}$, then $Lu = 0$ for all $u \in \mathscr{H}$ and we can take $v = 0$. Suppose that $\ker L \subsetneq \mathscr{H}$. Then by Theorem 1.5.5 there exists some $v_0 \in (\ker L)^\perp$ different from zero. After multiplying with a suitable constant, we can ensure that $\|v_0\| = 1$ and that $Lv_0 \in \mathbb{R}$. Then for any $u \in \mathscr{H}$,

$$(Lu)v_0 - (Lv_0)u \in \ker L,$$

so $v_0 \perp (Lu)v_0 - (Lv_0)u$. Hence,

$$L(u)\underbrace{\langle v_0, v_0\rangle}_{=1} - (Lv_0)\langle v_0, u\rangle = 0.$$

Since $Lv_0 \in \mathbb{R}$,

$$Lu = \langle (Lv_0)v_0, u\rangle \qquad \text{for any } u \in \mathscr{H},$$

so we simply take $v := (Lv_0)v_0$.

Riesz Representation Theorem

Proof (continued).

We have established the existence of the representation (2.2.1); it remains to show the uniqueness. Suppose that there are two vectors $v, w \in \mathscr{H}$ such that

$$Lu = \langle v, u \rangle = \langle w, u \rangle \qquad \text{for any } u \in \mathscr{H}.$$

Then we have

$$\langle v - w, u \rangle = 0 \qquad \text{for any } u \in \mathscr{H}.$$

Taking $u = v - w$, we see that $\langle v - w, v - w \rangle = \|v - w\|^2 = 0$, so $v = w$.

The proof that $\|L\| = \|v\|_{\mathscr{H}}$ is left to the reader. $\square$

Characterization of Functionals

A linear functional $L\colon \mathbb{R}^n \to \mathbb{R}$ is completely determined by its action on basis vectors: Let $\mathscr{B} = (b_1, \dots, b_n)$ be a basis of $\mathbb{R}^n$ and $x \in \mathbb{R}^n$ given by

$$x = \sum_{i=1}^{n} \lambda_i b_i$$

for some $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Then

$$Lx = L\Big(\sum_{i=1}^{n} \lambda_i b_i\Big) = \sum_{i=1}^{n} \lambda_i L b_i.$$

Hence, if we know the values of $Lb_1, \dots, Lb_n$ the value of Lx can be immediately calculated. Of course, here any finite-dimensional Hilbert space can be substituted for $\mathbb{R}^n$ and this remains true.

Given a basis $\mathscr{B} = (b_1, \dots, b_n)$ in a finite-dimensional space $\mathscr{H}$, any n numbers $v_1, \dots, v_n$ uniquely determine a linear functional L through

$$Lb_i := v_i, \qquad i = 1, \dots, n.$$

Characterization of Functionals

In the infinite-dimensional case the situation is more complicated: suppose $x \in \mathscr{H}$ and $(b_n)_{n\in\mathbb{N}}$ is a basis of $\mathscr{H}$, so $x = \sum_{i=1}^{\infty} \lambda_i b_i$. Then the equality

$$Lx = L\Big(\sum_{i=1}^{\infty} \lambda_i b_i\Big) = L\Big(\lim_{n\to\infty} \sum_{i=1}^{n} \lambda_i b_i\Big) = \lim_{n\to\infty} L\Big(\sum_{i=1}^{n} \lambda_i b_i\Big)$$
$$= \lim_{n\to\infty} \sum_{i=1}^{n} \lambda_i L b_i = \sum_{i=1}^{\infty} \lambda_i L b_i$$

requires the continuity (boundedness) of L. Hence, only bounded linear maps are defined by their action on basis elements.

Dual Basis

We know that if $(e_n)_{n\in\mathbb{N}}$ is an orthonormal basis in an infinite-dimensional Hilbert space $\mathscr{H}$, every $x \in \mathscr{H}$ has a representation

$$x = \sum_{n=0}^{\infty} \langle e_n, x \rangle e_n.$$

If an oblique (non-orthonormal) basis $\mathscr{B} = (b_n)_{n\in\mathbb{N}}$ is given in $\mathscr{H}$, we seek to find an analogous formula for the coefficients λ_n in

$$x = \sum_{n=0}^{\infty} \lambda_n b_n.$$

For any $n \in \mathbb{N}$ define the linear functional $L_n \colon \mathscr{H} \to \mathbb{F}$ by

$$L_n x := \lambda_n \qquad \text{for any } x \in \mathscr{H}, \text{ where } x = \sum_{n=0}^{\infty} \lambda_n b_n.$$

Dual Basis

It is clear that L_n is properly defined for all $x \in \mathscr{H}$ and that L_n is linear. In particular,

$$L_n b_m = \delta_{nm} = \begin{cases} 1 & n = m, \\ 0 & n \neq m. \end{cases}$$

By the Riesz representation theorem, we can find a unique vector $b_n^* \in \mathscr{H}$ such that

$$L_n x = \langle b_n^*, x \rangle.$$

We hence obtain a system of vectors $\mathscr{B}^* := (b_n^*)_{n \in \mathbb{N}}$ which we call the ***dual basis*** to $\mathscr{B}$. Of course, the dual basis can be defined in the same way in the finite-dimensional case.

2.2.2. Remark. If $\mathscr{B}$ is an orthonormal basis, $\mathscr{B}^* = \mathscr{B}$.

Dual Basis

Using the dual basis, we then see that any vector $x \in \mathscr{H}$ can be written as

$$x = \sum_{n=0}^{\infty} \langle b_n^*, x \rangle b_n.$$

2.2.3. Example. Let $\mathscr{H} = \mathbb{R}^2$, $b_1 = \binom{1}{1}$, $b_2 = \binom{1}{2}$, and $\mathscr{B} = (b_1, b_2)$. Then the dual basis is given by

$$b_1^* = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \qquad b_2^* = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Furthermore, if $x = \binom{x_1}{x_2}$, then

$$x = \langle b_1^*, x \rangle b_1 + \langle b_2^*, x \rangle b_2 = (2x_1 - x_2)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + (x_2 - x_1)\begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Matrix Elements

We can generalize the preceding discussion of functionals to operators. Let $\mathscr{H}$ is finite-dimensional and $L\colon \mathscr{H} \to \mathscr{H}$ a linear operator. Then L is determined completely by its action on a basis $\mathscr{B} = (b_1, \dots, b_n)$ as follows: Suppose that $u \in \mathscr{H}$ is given by $u = \sum_{i=1}^{n} \lambda_i b_i$, $\lambda_1, \dots, \lambda_n \in \mathbb{F}$. Then

$$Lu = \sum_{i=1}^{n} \lambda_i L b_i$$

so knowing Lb_i, $i = 1, \dots, n$, allows us to obtain Lu immediately. Since $\mathscr{B}$ is a basis, we can write

$$Lb_j = \sum_{i=1}^{n} \langle b_i^*, Lb_j \rangle b_i$$

where $(b_1^*, \dots, b_n^*)$ is the dual basis to $\mathscr{B}$.

Matrix Elements and Matrices

The n^2 numbers

$$a_{ij} := \langle b_i^*, Lb_j \rangle \in \mathbb{F} \qquad\qquad i, j = 1, \dots, n,$$

determine L completely. These $a_{ij} \in \mathbb{F}$ are called the ***matrix elements*** of L with respect to the basis $\mathscr{B}$. We will usually write the a_{ij} in the form of an array, as

$$\begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} = (a_{ij})_{i,j=1,\dots,n} \tag{2.2.2}$$

The array (2.2.2) is said to be the ***matrix representation of L with respect to the basis $\mathscr{B}$***, or the simply the ***matrix of L*** (with respect to $\mathscr{B}$).

Matrix Elements

We can proceed similarly for bounded operators on infinite-dimensional Hilbert spaces $\mathscr{H}$. For simplicity, assume that $\mathscr{B} = (e_n)_{n\in\mathbb{N}}$ is an orthonormal basis in $\mathscr{H}$. Let L be a bounded linear operator on $\mathscr{H}$ and $u = \sum_{i=0}^{\infty} \lambda_n e_n$. Then, using the continuity of L,

$$Lu = \sum_{i=0}^{\infty} \lambda_n L e_n. \tag{2.2.3}$$

The vectors Le_n can be expressed in terms of the basis $\mathscr{B}$,

$$Le_n = \sum_{m=0}^{\infty} a_{mn} e_m = \sum_{m=0}^{\infty} \langle e_m, Le_n \rangle e_m.$$

and as before we call

$$a_{mn} := \langle e_m, Le_n \rangle, \qquad m, n \in \mathbb{N},$$

the matrix elements of L. Note that the boundedness (continuity) of L is crucial for (2.2.3) to hold and the matrix elements to exist!

Matrix Elements

2.2.4. Example. The space ℓ^2 introduced in Example 2.1.3 has a natural scalar product given by

$$\langle a, b\rangle = \sum_{n=0}^{\infty} \overline{a_n} b_n$$

for sequences $a, b \in l^2$. An orthonormal basis is given by the set $\mathscr{B} = (e_n)_{n\in\mathbb{N}}$, where every e_n is a sequence given by

$$e_n = (\delta_{nm})_{m\in\mathbb{N}} = (0, \dots 0, \underset{\substack{\uparrow \\ n\text{th} \\ \text{entry}}}{1}, 0, \dots), \qquad i = 1, \dots, n,$$

The left-shift operator L acts on a sequence $b = (b_n)_{n\in\mathbb{N}}$ by $Lb = (b_{n+1})_{n\in\mathbb{N}}$. Hence, the matrix elements of L are

$$a_{ij} = \langle e_i, Le_j\rangle = \langle e_i, e_{j+1}\rangle = \delta_{i,j+1}.$$

Matrix Elements and Matrices

Conversely, any array (matrix) of n^2 numbers together with a basis $\mathscr{B}$ defines a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$. Unless stated otherwise, if we are simply given a matrix in $\mathbb{R}^n$ we will assume that $\mathscr{B}$ is the standard basis $(e_1, \dots, e_n)$.

Matrices representing linear maps are studied extensively in linear algebra. We will assume familiarity with basic matrix operations (multiplication, inversion, transposition etc.) and instead investigate another question:

In what sense does an "infinite matrix" define a linear map?

More precisely, let us replace $\mathbb{R}^n$ with ℓ^2 (so an "infinite vector" is just a sequence of numbers and we have an analogous scalar product) and take the orthonormal basis of vectors

$$e_j := (0, \dots, 0, \underset{\substack{\uparrow \\ j}}{1}, 0, \dots) \in \ell^2.$$

Hilbert-Schmidt Operators

If we are now given a set of numbers a_{ij}, $i, j \in \mathbb{N}$, does

$$Le_j := \sum_{i=0}^{\infty} a_{ij} e_i$$

define a bounded linear map $L\colon \ell^2 \to \ell^2$? If $x \in \ell^2$, this would mean that

$$Lx = \left(\sum a_{1j} x_j,\ \sum a_{2j} x_j,\ \sum a_{3j} x_j, \ldots \right) \tag{2.2.4}$$

We will see that a sufficient condition for L to be bounded is that

$$\sum_{i,j=0}^{\infty} |a_{ij}|^2 < \infty. \tag{2.2.5}$$

Operators $L\colon \ell^2 \to \ell^2$ defined by (2.2.4) satisfying (2.2.5) are called ***Hilbert-Schmidt operators*** on ℓ^2.

Adjoint of Bounded Operators

In this section we define two important operators: the adjoint and the inverse.

2.3.1. Definition and Theorem. Let $\mathscr{H}$ be a Hilbert space and $L\colon \mathscr{H} \to \mathscr{H}$ a bounded linear operator. Then the (Hilbert space) ***adjoint*** of L, denoted by L^*, is a map

$$L^*\colon \mathscr{H} \to \mathscr{H}$$

uniquely defined through the relation

$$\langle x, Ly\rangle = \langle L^*x, y\rangle \qquad \text{for all } x, y \in \mathscr{H}. \tag{2.3.1}$$

Furthermore, L^* is bounded with

$$\|L\| = \|L^*\|. \tag{2.3.2}$$

Adjoint of Bounded Operators

Proof.

We need to show that for any bounded operator L the adjoint L^* exists, is unique and has the same operator norm as L.

We may regard $\langle x, L(\,\cdot\,)\rangle$ as a linear functional on $\mathscr{H}$. By the Riesz representation theorem, for any $x \in \mathscr{H}$ we can find a $z_x \in \mathscr{H}$ such that

$$\langle x, Ly\rangle = \langle z_x, y\rangle \qquad \text{for all } y \in \mathscr{H}.$$

The element z_x naturally depends on x, and the dependence is linear: for all $x_1, x_2, x \in \mathscr{H}$ and $\lambda \in \mathbb{F}$,

$$\begin{aligned}\langle z_{x_1+x_2}, y\rangle &= \langle x_1 + x_2, Ly\rangle = \langle x_1, Ly\rangle + \langle x_2, Ly\rangle = \langle z_{x_1}, y\rangle + \langle z_{x_2}, y\rangle \\ &= \langle z_{x_1} + z_{x_2}, y\rangle, \\ \langle z_{\lambda x}, y\rangle &= \langle \lambda x, Ly\rangle = \overline{\lambda}\langle x, Ly\rangle = \overline{\lambda}\langle z_x, y\rangle, = \langle \lambda z_x, y\rangle.\end{aligned}$$

Adjoint of Bounded Operators

Proof (continued).

We hence define

$$L^* x := z_x,$$

giving a well-defined linear map L^* on $\mathscr{H}$. (For every $x \in \mathscr{H}$, $L^* x$ exists and is unique.) This shows the existence of the adjoint L^*.

The uniqueness is easy to prove: suppose some other operator A on $\mathscr{H}$ satisfies $\langle x, Ly \rangle = \langle Ax, y \rangle$ for all $x, y \in \mathscr{H}$. Then

$$\langle (A - L^*)x, y \rangle = 0 \qquad \text{for all } x, y \in \mathscr{H}.$$

This implies $(A - L^*)x = 0$ for all $x \in \mathscr{H}$ and hence $A = L^*$.

Adjoint of Bounded Operators

Proof (continued).

The Cauchy-Schwarz inequality yields

$$\|L^*x\|^2 = \langle L^*x, L^*x\rangle = \langle x, LL^*x\rangle \leq \|x\| \cdot \|LL^*x\| \leq \|x\|\|L\| \cdot \|L^*x\|$$

so we find

$$\|L^*x\| \leq \|L\|\|x\|.$$

Hence L^* is bounded and $\|L^*\| \leq \|L\|$. We note that

$$\langle Ly, x\rangle = \overline{\langle x, Ly\rangle} = \overline{\langle L^*x, y\rangle} = \langle y, L^*x\rangle. \tag{2.3.3}$$

Then, again applying the Cauchy-Schwarz inequality,

$$\|Lx\|^2 = \langle Lx, Lx\rangle = \langle x, L^*Lx\rangle \leq \|x\| \cdot \|L^*Lx\| \leq \|x\|\|L^*\| \cdot \|Lx\|$$

so $\|Lx\| \leq \|L^*\|\|x\|$ and $\|L\| \leq \|L^*\|$. Hence $\|L^*\| = \|L\|$. □

Properties of the Adjoint

2.3.2. Remarks.

(i) From (2.3.3) it follows that $(L^*)^* = L$, i.e., the adjoint of the adjoint is again L, because

$$\langle x, (L^*)^* y \rangle = \langle L^* x, y \rangle = \langle x, Ly \rangle$$

for all $x, y \in \mathscr{H}$.

(ii) (2.3.3) also implies that the matrix elements a_{ij}^* of L^* are given by

$$a_{ij}^* = \langle e_i, L^* e_j \rangle = \langle Le_i, e_j \rangle = \overline{\langle e_j, Le_i \rangle} = \overline{a_{ji}},$$

where a_{ij} are the matrix elements of L.

(iii) The definition of the adjoint of an ***unbounded*** operator is more complicated, since the original operator is not defined on the whole space (see Theorem 2.1.11) and one needs to ensure that (2.3.1) holds on suitable domains. We will not go into details here.

Self-Adjoint Operators

2.3.3. Definition. A bounded linear operator $L\colon \mathscr{H} \to \mathscr{H}$ such that $L = L^*$ is called ***self-adjoint***.

2.3.4. Example. Let L and R denote the left- and right-shift operators of Example 2.1.3. Then

$$\begin{aligned}\langle a, Lb\rangle &= \sum_{n=0}^{\infty} \overline{a_n}(Lb)_n = \sum_{n=0}^{\infty} \overline{a_n} b_{n+1} \\ &= \sum_{m=1}^{\infty} \overline{a_{m-1}} b_m = \sum_{m=0}^{\infty} \overline{(Ra)_m} b_m \\ &= \langle Ra, b\rangle\end{aligned}$$

so $L^* = R$. It follows from Remark 2.3.2 (i) that $R^* = L^{**} = L$.

Furthermore, $(RL)^* = L^*R^* = RL$, so RL is self-adjoint. Of course, $LR = 1$ is also self-adjoint.

Unbounded Operators

The definition of the adjoint of an ***unbounded*** operator is more complicated, since the original operator is not defined on the whole space (see Theorem 2.1.11) and one needs to ensure that (2.3.1) holds on suitable domains. In other words,

$$\operatorname{dom} L \neq \operatorname{dom} L^* \qquad \text{for general unbounded operators.}$$

We will not go into the details of the construction of the adjoint of an unbounded operator. However, we note that if

$$\langle u, Lv \rangle = \langle Lu, v \rangle \qquad \text{for all } u, v \in \operatorname{dom} L$$

then an unbounded operator L is said to be ***symmetric*** (but not self-adjoint). Of course, a bounded, self-adjoint operator will also be symmetric.

Range-Kernel Decomposition

2.3.5. Lemma. Let L be a bounded linear operator defined in a Hilbert space $\mathscr{H}$. Then

$$(\operatorname{ran} L)^{\perp} = \ker L^*.$$

Proof.
Let $x \in (\operatorname{ran} L)^{\perp}$. Then for all $y \in \mathscr{H}$,

$$0 = \langle x, Ly \rangle = \langle L^*x, y \rangle.$$

Since this holds for all $y \in \mathscr{H}$, let $y = L^*x$. Then $\|L^*x\|^2 = 0$, so $L^*x = 0$ and so $x \in \ker L^*$. This shows $(\operatorname{ran} L)^{\perp} \subset \ker L^*$. The proof that $\ker L^* \subset (\operatorname{ran} L)^{\perp}$ is similar. □

We can therefore write $\mathscr{H} = \operatorname{ran} L \oplus \ker L^*$. This is known as the ***range-kernel decomposition*** of $\mathscr{H}$.

Inversion of Linear Operators

The central problem of (linear) operator theory is the finding of solutions of

$$Lu = v, \tag{2.3.4}$$

where $L\colon U \to V$ is a linear map between vector spaces U and V, $v \in V$ is given and $u \in U$ is sought.

- If $U = \mathbb{R}^n$, $V = \mathbb{R}^m$ and L is a matrix, (2.3.4) describes a system of m algebraic equations in n unknowns.
- If L is a (partial or ordinary) differential operator between spaces of functions U and V, (2.3.4) is a (partial or ordinary) differential equation.
- If L is an integral operator between spaces of functions U and V, (2.3.4) is an integral equation.

There are two main concerns in the analysis of (2.3.4):

(i) What is the range of L?

(ii) Is L bijective onto its range?

The Inverse of a Linear Operator

Throughout this section, we assume that X is a Banach space, $U \subset X$ a subspace and that

$$L\colon U \to U' \subset X$$

is a linear operator on X with domain $\operatorname{dom} L = U$ and range $\operatorname{ran} L = U'$.

2.3.6. Definition. An operator $L^{-1}\colon U' \to U$ satisfying

$$L^{-1}(Lu) = u$$

for all $u \in U$ is said to be an ***inverse*** of L.

Note that the inverse of L, if it exists, is unique. The following result, known from linear algebra, is proved in exactly the same way for general linear operators:

2.3.7. Lemma. The inverse of L exists if and only if $\ker L = \{0\}$.

The Inverse of a Linear Operator

Two peculiarities for operators in infinite-dimensional spaces that should be noted:

- If L is bounded, then L^{-1} may be bounded or unbounded.
- if L is invertible with inverse L, it may happen that L^{-1} is not invertible at all.

2.3.8. Example. Consider the operator

$$L\colon \ell^2 \to \ell^2, \qquad (a_n) \mapsto \left(\frac{1}{n+1}a_n\right).$$

Then L is bounded with $\|L\| = 1$ but

$$L^{-1}\colon \operatorname{ran} L \to \ell^2, \qquad (a_n) \mapsto \big((n+1)a_n\big).$$

is unbounded, since $\|Le_n\| = n+1$ while $\|e_n\| = 1$ for any basis sequence e_n.

The Inverse of a Linear Operator

2.3.9. Example. Let L and R be the left- and right-shift operators of Example 2.1.3. Then

$$R^{-1}Ru = LRu = u \qquad \text{for all } u \in \ell^2$$

but, in general,

$$RLu \neq u.$$

In other words, $R^{-1} = L$, but L^{-1} doesn't exist.

Existence of a Bounded Inverse

2.3.10. Definition. We say that L is ***bounded away from zero*** if there exists a $c > 0$ such that

$$\|Lu\| \geq c\|u\| \qquad \text{for all } u \in U. \tag{2.3.5}$$

2.3.11. Theorem. The operator L has a bounded inverse if and only if L is bounded away from zero.

Proof.

If L is bounded away from zero, have $\|Lu\| > 0$ for all $u \neq 0$, so $\ker L = \{0\}$ and L is invertible. If $Lu = v$ we have $u = L^{-1}v$ and hence $\|Lu\| \geq c\|u\|$ implies

$$\|L^{-1}v\| \leq \frac{1}{c}\|v\|, \tag{2.3.6}$$

so L^{-1} is bounded. A similar argument shows that if L^{-1} exists and is bounded, satisfying (2.3.6), then (2.3.5) holds. □

The Range of an Operator

By definition, (2.3.4) has at least one solution u if and only if $v \in \operatorname{ran} L$. Therefore, it is important to characterize the range. However, in many applications it is easier to determine the closure $\overline{\operatorname{ran} L}$ instead of the actual range. Either $\overline{\operatorname{ran} L} = X$, or $\overline{\operatorname{ran} L} \subsetneq X$. In the latter case, there exist elements in the orthogonal complement $\overline{\operatorname{ran} L}^{\perp}$ (which are then also orthogonal to $\operatorname{ran} L$).

Since often both $\operatorname{ran} L$ and X are infinite-dimensional, we may be interested in the dimension of the orthogonal complement of the range, the ***codimension*** of $\operatorname{ran} L$, which we denote by

$$\operatorname{codim} \operatorname{ran} L := \dim(\operatorname{ran} L)^{\perp}$$

The State of an Operator

We are now able to give a basic classification of linear operator L:

I L has a bounded inverse.

II L has an inverse, but L^{-1} is unbounded.

III L has no inverse.

In addition, we differentiate between two cases for the closure of the range:

1 $\overline{\operatorname{ran} L} = X$

2 $\overline{\operatorname{ran} L} \neq X$

We will also sometimes add the subscript c to the arabic numeral to indicate that the range of L is closed and the subscript n to indicate the range is not closed.

2.3.12. Remark. Operators in the state (I,1_c) are often called ***regular operators***.

The State of an Operator

2.3.13. Examples.

(i) The left-shift operator L introduced in Example 2.1.3 has a non-trivial kernel and hence is not invertible. Its range is all of ℓ^2, so L is of type (III,1_c).

(ii) The right-shift operator R has trivial kernel, so R^{-1} exists. The inverse is just the left-shift, so $R^{-1} = L$ is bounded. Furthermore, the range of R is closed (why?) and a strict subset of ℓ^2, so R is of type (I,2_c).

(iii) The operator L of Example 2.3.8 has an unbounded inverse and its range is given by

$$\operatorname{ran} L = \left\{(x_n) \in \ell^2 \colon \sum_{n=0}^{\infty} (n+1)^2 |x_n|^2 < \infty \right\}$$

Since the basis sequences e_n are all in $\operatorname{ran} L$, it follows that $\overline{\operatorname{ran} L} = \ell^2$, so L is of type (II,1_n).

The State of an Operator

(iv) Consider an $n \times n$ matrix A as a map from $\mathbb{R}^n$ to itself (or alternatively on any finite-dimensional Hilbert space). Then

- $\det A \neq 0$ and A is invertible. In that case, the inverse exists and is bounded automatically. Moreover, $\operatorname{ran} A = \mathbb{R}^n$ so A is of type $(\mathrm{I},1_c)$.
- If $\det A = 0$, then A is not invertible and its range is a strict subspace of $\mathbb{R}^n$. Hence, A is of type $(\mathrm{III},2_c)$.

In finite-dimensional vector spaces only these two types of operators occur.

The Resolvent of a Bounded Operator

Throughout this section, we assume T to be a bounded operator on a separable Hilbert space $\mathscr{H}$ with domain dom $T = \mathscr{H}$.

Let $I : \mathscr{H} \to \mathscr{H}$ denote the unit operator. Then for any $\lambda \in \mathbb{C}$, we define

$$T_\lambda := T - \lambda I.$$

The domain of T_λ is of course $\mathscr{H}$, but the range will in general depend on λ. We remark that the adjoint

$$T_\lambda^* := T^* - \overline{\lambda} I.$$

is also defined on $\mathscr{H}$. The inverse

$$R_\lambda(T) := T_\lambda^{-1} = (T - \lambda I)^{-1}$$

is called the ***resolvent*** of T.

The Resolvent Set and the Spectrum

2.4.1. Definition. The ***resolvent set*** $\varrho(T)$ of T is defined as the set of all complex numbers λ for which T_λ has a bounded inverse and ran $T_\lambda = \mathscr{H}$, i.e.,

$$\varrho(T) := \{\lambda \in \mathbb{C} \colon T - \lambda I \text{ is in state } (\mathsf{I},1_c)\}.$$

The ***spectrum*** $\sigma(T)$ of T is defined as the complement of the resolvent set, i.e.,

$$\sigma(T) := \mathbb{C} \setminus \varrho(T).$$

We will not prove the following result:

2.4.2. Proposition. The resolvent set is an open subset of $\mathbb{C}$ and hence the spectrum is closed.

There are several competing approaches to characterizing the spectrum. All of these have advantages and disadvantages; our approach is not the most common, but is very well suited for the present discussion. Its disadvantage lies in not separating the spectrum into disjoint parts.

Division of the Spectrum

2.4.3. Definition. Let $\lambda \in \sigma(T)$.

(i) Suppose that $T - \lambda I$ is in state III ($R_\lambda(T)$ does not exist). Then we say that λ belongs to the ***point spectrum***. By Lemma 2.3.7,

$$Tu = \lambda u$$

has a non-trivial solution, i.e., λ is an eigenvalue of T.

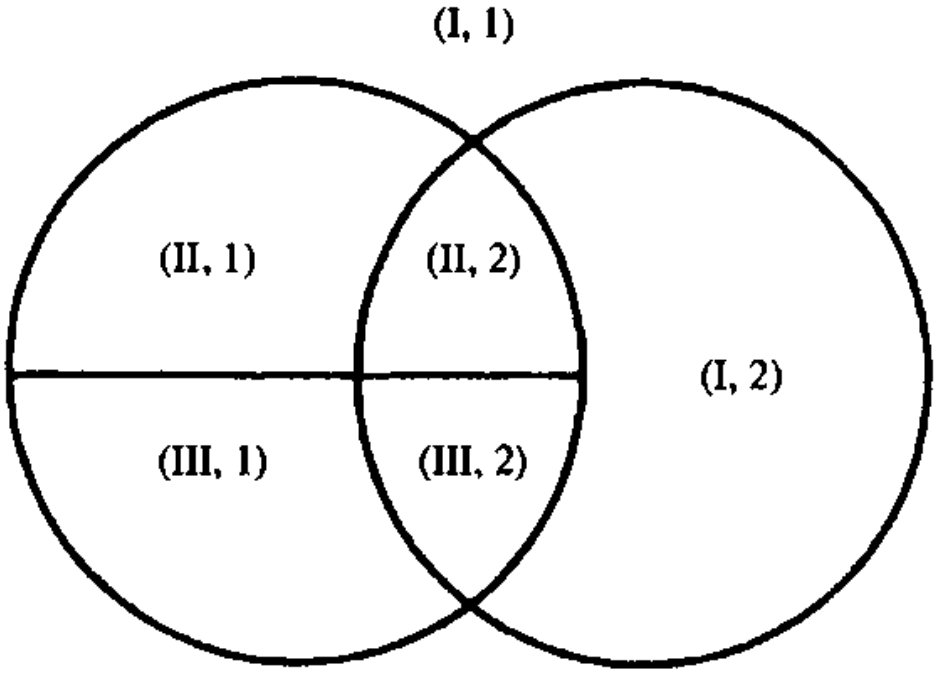
(ii) Suppose that $T - \lambda I$ is in state II ($R_\lambda(T)$ exists but is unbounded). Then we say that λ belongs to the ***continuous spectrum***.

(iii) Suppose that $T - \lambda I$ is in state 2 ($\overline{\operatorname{ran}(T - \lambda I)} \neq \mathscr{H}$). Then we say that λ belongs to the ***compression spectrum***. The range has been compressed and we define the ***deficiency*** of λ as

$$\operatorname{def} \lambda := \operatorname{codim} \operatorname{ran} T - \lambda I.$$

The union of the point spectrum and the continuous spectrum is called the ***approximate point spectrum***.

Division of the Spectrum



The Venn diagram shows the different possible states of an operator. When applied to $T - \lambda I$, the left circle refers to the approximate point spectrum, the right circle refers to the compression spectrum and the region outside the circles represents the resolvent set.

We will often denote the point, continuous and compression spectra of T by

$$\sigma_{\text{point}}(T), \qquad \sigma_{\text{continuous}}(T) \qquad \text{and} \qquad \sigma_{\text{compression}}(T),$$

respectively.

Point and Compression Spectrum

2.4.4. Lemma. A number $\lambda \in \mathbb{C}$ is in the compression spectrum of T if and only if $\overline{\lambda}$ is in the point spectrum of T^*.

Proof.
The proof is based on Lemma 2.3.5:

$$\begin{aligned}
\lambda \in \sigma_{\text{compression}}(T) &\Leftrightarrow \overline{\operatorname{ran}(T-\lambda I)} \subsetneq \mathscr{H} \\
&\Leftrightarrow \exists_{u \neq 0} \; u \in (\operatorname{ran}(T-\lambda I))^{\perp} \\
&\Leftrightarrow \exists_{u \neq 0} \; u \in \ker(T-\lambda I)^* \\
&\Leftrightarrow \exists_{u \neq 0} \; T^* u = \overline{\lambda} u \\
&\Leftrightarrow \overline{\lambda} \in \sigma_{\text{point}}(T^*)
\end{aligned}$$

□

Spectrum of the Left- and Right- Shift Operators

2.4.5. Example. Consider the left- and right-shift operators on ℓ^2 introduced in Example 2.1.3. Since $L^* = R$ and $R^* = L$, it is convenient to discuss both at the same time.

Consider first a complex number λ with $|\lambda| > 1$ and the right-shift operator R. Then, for $a \in \ell^2$,

$$\|Ra - \lambda a\| \geq \big|\|\lambda a\| - \|Ra\|\big| = (|\lambda| - 1)\|a\|,$$

and $R - \lambda I$ is bounded away from zero. By Theorem 2.3.11, $R - \lambda I$ then has a bounded inverse and hence is of type I.

A similar argument for the left-shift operator L shows that $L - \lambda I$ is in state I if $|\lambda| > 1$.

The Resolvent Set

We now show that in fact $R - \lambda I$ and $L - \lambda I$ are in state $(I, 1_c)$ if $|\lambda| > 1$.

We will show that the range of $R - \lambda I$ is ℓ^2. Let $(a_n) \in \ell^2$ and consider a sequence (b_n) such that $(a_n) = (R - \lambda I)(b_n)$, i.e,

$$\begin{aligned}(a_0, a_1, a_2, ...) &= (R - \lambda I)(b_0, b_1, b_2, ...) \\ &= (-\lambda b_0, b_0 - \lambda b_1, b_1 - \lambda b_2, ...)\end{aligned}$$

Hence a pre-image for (a_n) is found recursively from

$$b_0 = -\frac{a_0}{\lambda}, \qquad b_1 = \frac{b_0 - a_1}{\lambda}, \qquad b_2 = \frac{b_1 - a_2}{\lambda}, \dots.$$

An explicit formula is

$$b_n = -\frac{1}{\lambda} \sum_{k=0}^{n} \frac{a_k}{\lambda^{n-k}}.$$

The Resolvent Set

We still need to show that $(b_n) \in \ell^2$. For this, note that

$$(b_n) = -\frac{1}{\lambda}(a_n) * \left(\frac{1}{\lambda^n}\right)$$

where $*$ denotes the convolution of sequences. Then by Young's convolution inequality,

$$\|(b_n)\|_2 \leq \frac{1}{\lambda}\|(a_n)\|_2 \cdot \left\|\left(\frac{1}{\lambda^n}\right)\right\|_1.$$

Since the ℓ^1-norm of $(1/\lambda^n)$ is just a geometric series and $|\lambda| > 1$, we have

$$\|(b_n)\|_2 \leq \frac{\|(a_n)\|_2}{|\lambda| - 1}$$

so that $(b_n) \in \ell^2$. Hence, a pre-image exists for any $a \in \ell^2$ and $\operatorname{ran}(R - \lambda I) = \ell^2$.

A similar discussion shows that $\operatorname{ran}(L - \lambda I) = \ell^2$. It follows that both $R - \lambda I$ and $L - \lambda I$ are in state $(I, 1_c)$.

The Point Spectrum

The eigenvalue equation for the left-shift operator is

$$L(a_0, a_1, a_2, ...) = (a_1, a_2, ...) = \lambda(a_0, a_1, a_2, ...).$$

We obtain $a_1 = \lambda a_0$, $a_2 = \lambda a_1 = \lambda^2 a_0$ etc., so λ is an eigenvalue with eigenvector e_λ if

$$e_\lambda = (1, \lambda, \lambda^2, \lambda^3, ...).$$

Now $e_\lambda \in \ell^2$ if and only if $\sum_{n=0}^{\infty} |\lambda^n|^2 < \infty$ which is the case if and only if $|\lambda| < 1$. We see that the point spectrum of L is given by

$$\sigma_{\text{point}}(L) = \{\lambda \in \mathbb{C} : |\lambda| < 1\}.$$

The Point and the Compression Spectrum

The eigenvalue equation for the right-shift operator is

$$R(a_0, a_1, a_2, ...) = (0, a_0, a_1, a_2, ...) = \lambda(a_0, a_1, a_2, ...).$$

Suppose $\lambda = 0$. Then all $a_k = 0$, so there is no eigenvector. If $\lambda \neq 0$, we obtain $\lambda a_0 = 0$, so $a_0 = 0$, $\lambda a_1 = a_0 = 0$ etc. Hence the right-shift operator does not have any eigenvalues and the point spectrum of R is given by

$$\sigma_{\text{point}}(R) = \emptyset.$$

It follows from Lemma 2.4.4 that

$$\sigma_{\text{compression}}(L) = \emptyset.$$

and

$$\sigma_{\text{compression}}(R) = \{\lambda \in \mathbb{C} : |\lambda| < 1\}.$$

The Continuous Spectrum

Since the spectrum is closed, the circle $|\lambda| = 1$ must be in the spectrum. Since it can not lie in the compression or point spectra, it must lie in the continuous spectrum. It follows that

$$\sigma_{\text{continuous}}(L) = \{\lambda \in \mathbb{C} : |\lambda| = 1\}.$$

The continuous spectrum of R must also contain all λ with $|\lambda| = 1$. However, it might also overlap with the (non-empty) compression spectrum of R. Now, for $|\lambda| < 1$ we have

$$\|Ra - \lambda a\| \geq \big|\|Ra\| - \|\lambda a\|\big| = (1 - |\lambda|)\|a\|,$$

so $R - \lambda I$ is bounded way from zero and hence has a bounded inverse. It follows that λ can not be part of the approximate point spectrum if $|\lambda| < 1$. We deduce

$$\sigma_{\text{continuous}}(R) = \{\lambda \in \mathbb{C} : |\lambda| = 1\}.$$

Approximate Point Spectrum

2.4.6. Lemma. A number $\lambda \in \mathbb{C}$ is in the approximate point spectrum of T if and only if there exists a sequence (u_n) in dom T such that $\|u_n\| = 1$ and $(T - \lambda I)u_n \to 0$.

Proof.

($\Leftarrow$) Suppose that there exists a sequence (u_n) such that $\|u_n\| = 1$ and $(T - \lambda I)u_n \to 0$. Then T_λ can not be bounded away from zero and by Theorem 2.3.11 can not have a bounded inverse. Hence $(T - \lambda I)$ is in state II or III and λ belongs to the approximate point spectrum.

Approximate Point Spectrum

Proof (continued).

($\Rightarrow$) Suppose that $\lambda \in \sigma_{\text{point}}(T)$. Then $(T - \lambda I)u = 0$ for some $u \in \mathscr{H}$, i.e., there exists an eigenvector u to λ. It is sufficient to take the constant sequence given by $u_n = u/\|u\|$.

Suppose that $\lambda \in \sigma_{\text{continuous}}(T)$. Then $R_\lambda(T)$ exists but is unbounded. By Theorem 2.3.11, $(T - \lambda I)$ is not bounded away from zero, so for any $c \in \mathbb{R}$ there exists some $v_c \in \mathscr{H}$ such that

$$\left\| (T - \lambda I)\left(\frac{v_c}{\|v_c\|}\right) \right\| < c.$$

It follows that the sequence of elements $u_n := v_{1/n}/\|v_{1/n}\|$ satisfies the requirements. $\square$

Spectrum of Self-Adjoint Operators

2.4.7. Theorem. Let T be a bounded, self-adjoint operator. Then $\sigma(T) \subset \mathbb{R}$ and

$$\sigma_{\text{compression}}(T) = \sigma_{\text{point}}(T). \tag{2.4.1}$$

Proof.

We will prove that all parts of the spectrum are real by considering the point, continuous and compression spectrum separately.

First note that if T is self-adjoint, then

$$\langle u, Tu\rangle = \langle Tu, u\rangle = \overline{\langle u, Tu\rangle}$$

so $\langle u, Tu\rangle$ is real. Now suppose that $\lambda \in \mathbb{C}$ is in the point spectrum. Then $Tu = \lambda u$ for some $u \neq 0$, so

$$\lambda\|u\|^2 = \lambda\langle u, u\rangle = \langle u, Tu\rangle \in \mathbb{R}$$

Since $\|u\|^2$ is real, $\lambda \in \mathbb{R}$ and $\sigma_{\text{point}}(T) \subset \mathbb{R}$.

Spectrum of Self-Adjoint Operators

Proof (continued).

Now let $\lambda = \xi + i\eta$, $\xi, \eta \in \mathbb{R}$, lie in the continuous spectrum. Suppose that $\eta \neq 0$. Then

$$\|(T - \lambda I)u\|^2 = \|Tu - \xi u\|^2 + \eta^2\|u\|^2 \geq \eta^2\|u\|^2,$$

so T_λ is bounded away from zero and hence has a bounded inverse. But then λ can not lie in the continuous spectrum, so we conclude $\eta = 0$ and $\sigma_{\text{continuous}}(T) \subset \mathbb{R}$.

If λ is in the compression spectrum, then by Lemma 2.4.4 $\overline{\lambda}$ lies in the point spectrum of $T^* = T$. But this implies that $\overline{\lambda} \in \mathbb{R}$ and hence $\lambda \in \mathbb{R}$. The entire spectrum is therefore real.

Equation (2.4.1) then follows from Lemma 2.4.4. □

Eigenpairs and a Bound on the Spectrum

We are interested in finding bounds for the spectrum $\sigma(T) \subset \mathbb{C}$ of a bounded operator T. Suppose that for some $\lambda \in \mathbb{C}$ and some $u \in \mathscr{H}$ we have

$$Tu = \lambda u.$$

We then say that (u, λ) is an ***eigenpair***. Taking the norm of the above expression and also the inner product with u,

$$|\lambda| = \frac{\|Tu\|}{\|u\|}, \qquad \lambda = \frac{\langle u, Tu\rangle}{\|u\|^2}. \tag{2.4.2}$$

If (u, λ) is an eigenpair, by (2.4.2),

$$\|T\| = \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} \frac{\|Tv\|}{\|v\|} \geq \frac{\|Tu\|}{\|u\|} = |\lambda|,$$

so the norm of T is a bound for all eigenvalues. This can be generalized to the complete spectrum.

A Bound on the Spectrum

2.4.8. Proposition. Let T be a bounded linear operator on $\mathscr{H}$. If $\lambda \in \sigma(T)$, then $|\lambda| \leq \|T\|$.

Proof.

We have already proved the theorem if λ is in the point spectrum (i.e., an eigenvalue). If λ is in the compression spectrum, then $\overline{\lambda}$ is an eigenvalue of T^* and

$$|\lambda| = |\overline{\lambda}| \leq \|T^*\| = \|T\|.$$

by Lemma 2.4.4 and (2.3.2). If λ is in the continuous spectrum, by Lemma 2.4.6 there exists a sequence (u_n), $\|u_n\| = 1$, such that $(T - \lambda I)u_n \to 0$. In other words, $w_n := Tu_n - \lambda u_n \to 0$. Then

$$|\lambda| = \|Tu_n - w_n\| \leq \|Tu_n\| + \|w_n\| \leq \|T\| + \|w_n\|.$$

Since $w_n \to 0$, we obtain $|\lambda| \leq \|T\|$. □

A Bound on the Spectrum

2.4.9. Example. For the left- and right-shift operators in ℓ^2 we have seen in Example 2.4.5 that $|\lambda| \leq 1$ for all λ in their spectra.

Together with the fact they are bounded with unit operator norm (see (2.1.3)), Proposition 2.4.8 is confirmed.

In practice, finding the norm of an operator can be quite difficult and it is useful to consider a substitute, the so-called Rayleigh quotient. The Rayleigh quotient also has the advantage of giving lower or upper bounds for eigenvalues if the operators are semi-bounded (see next section).

Eigenpair and Rayleigh Quotient

For any arbitrary non-zero vector $v \in \mathscr{H}$ we define the ***Rayleigh quotient***

$$R(v) := \frac{\langle v, Tv \rangle}{\|v\|^2} \tag{2.4.3}$$

If $v = u$ is an eigenvector with eigenvalue λ, then (2.4.2) gives $R(u) = \lambda$. Furthermore, we define

$$M_T := \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} |R(v)| = \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} \frac{|\langle v, Tv \rangle|}{\|v\|^2} = \sup_{\substack{v \in \mathscr{H} \\ \|v\| = 1}} |\langle v, Tv \rangle|$$

(why is M_T finite?). Then, again by applying (2.4.2), we have

$$M_T \geq |\lambda|.$$

Hence, M_T also gives a bound for the eigenvalues.

Bounded Operators

2.4.10. Remark. We note that by the Cauchy-Schwarz inequality,

$$M_T = \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} \frac{|\langle v, Tv \rangle|}{\|v\|^2} \leq \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} \frac{\|Tv\|}{\|v\|} = \|T\|.$$

2.4.11. Theorem. If T is bounded and self-adjoint, then

$$\|T\| = \sup_{\substack{v \in \mathscr{H} \\ v \neq 0}} \frac{|\langle v, Tv \rangle|}{\|v\|^2} = M_T$$

Before we prove Theorem 2.4.11, we recall the ***parallelogram law*** for a norm induced by an inner product:

$$\|v + w\|^2 + \|v - w\|^2 \leq 2(\|v\|^2 + \|w\|^2), \qquad v, w \in \mathscr{H}.$$

Bounded Operators

Proof of Theorem 2.4.11.

Since $M_T \leq \|T\|$, we just need to verify $M_T \geq \|T\|$. Since T is self-adjoint, $R(v)$ is real for any $v \in \mathscr{H}$ and hence

$$-M_T \leq R(v) \leq M_T \qquad \text{for any } v \in \mathscr{H}.$$

Take any $v, w \in \mathscr{H}$. Then

$$\begin{aligned} \langle v+w, T(v+w)\rangle &\leq M_T \cdot \|v+w\|^2, \\ \langle v-w, T(v-w)\rangle &\geq -M_T \cdot \|v-w\|^2. \end{aligned}$$

From these equations it follows that

$$\begin{aligned} 2(\langle v, Tw\rangle + \langle w, Tv\rangle) &= \langle v+w, T(v+w)\rangle - \langle v-w, T(v-w)\rangle \\ &\leq M_T\big(\|v+w\|^2 + \|v-w\|^2\big) \\ &\leq 2M_T\big(\|v\|^2 + \|w\|^2\big). \end{aligned} \tag{2.4.4}$$

Bounded Operators

Proof of Theorem 2.4.11 (continued).

Suppose that $v \neq 0$ and set

$$w = \frac{\|v\|}{\|Tv\|} Tv.$$

Then (2.4.4) becomes

$$\frac{\|v\|}{\|Tv\|} \big(\langle v, TTv \rangle + \langle Tv, Tv \rangle \big) \leq 2M_T \|v\|^2$$

or, using the self-adjointness of T,

$$\|Tv\| \leq M_T \|v\|.$$

Hence $\|T\| \leq M_T$ and the proof is complete. $\square$

Bounds on the Rayleigh Quotient

2.4.12. Definition. Let T be a symmetric linear operator on a Hilbert space $\mathscr{H}$. Then we define the upper and lower ***Rayleigh bounds***

$$\begin{aligned} L_T &:= \inf_{\substack{v \in \operatorname{dom} T \\ v \neq 0}} R(v) = \inf_{v \in \mathscr{H}} \frac{\langle v, Tv \rangle}{\|v\|^2}, \\ U_T &:= \sup_{\substack{v \in \operatorname{dom} T \\ v \neq 0}} R(v) = \sup_{v \in \mathscr{H}} \frac{\langle v, Tv \rangle}{\|v\|^2}, \end{aligned} \tag{2.4.5}$$

if they exist.

2.4.13. Remark. If T is bounded, both L_T and R_T exist and

$$M_T = \|T\| = \max\{|L_T|, |U_T|\} = \max\{-L_T, U_T\}.$$

Positive Operators

2.4.14. Definition. A (bounded or unbounded) operator T on a Hilbert space $\mathscr{H}$ is said to be ***positive*** if

$$\langle x, Tx\rangle \geq 0 \qquad \text{for all } x \in \operatorname{dom} T.$$

2.4.15. Remark. In a ***complex*** Hilbert space, a bounded positive operator must be self-adjoint, for $\langle x, Tx\rangle \geq 0$ implies $\langle x, Tx\rangle \in \mathbb{R}$ and hence

$$\langle Tx, x\rangle = \overline{\langle x, Tx\rangle} = \langle x, Tx\rangle \qquad \text{for all } x \in \mathscr{H}.$$

Then the polarisation identity implies $\langle Tx, y\rangle = \langle x, Ty\rangle$ for all $x, y \in \mathscr{H}$.

In a real Hilbert space this is not true, since knowing $\langle x, Tx\rangle$ for all $x \in \mathscr{H}$ does not yield $\langle x, Ty\rangle$ for all $x, y \in \mathscr{H}$.

Bounds on the Rayleigh Quotient

2.4.16. Theorem. Let T be a self-adjoint, bounded operator. Then L_T and U_T are in the approximate point spectrum.

Proof.

We show that U_T belongs to the approximate point spectrum. By Lemma 2.4.6, we need to give a sequence (u_n) of unit elements so that $Tu_n - U_T u_n \to 0$. From the definition of U_T there exists a sequence (u_n) of unit elements such that $\langle u_n, Tu_n\rangle \to U_T$.

Suppose that T is positive. Then $U_T = \|T\|$ and

$$\begin{aligned}\|Tu_n - U_T u_n\|^2 &= \|Tu_n\|^2 - 2U_T\langle u_n, Tu_n\rangle + U_T^2\\ &\le \|T\|^2 - 2U_T\langle u_n, Tu_n\rangle + U_T^2\\ &= 2U_T^2 - 2U_T\langle u_n, Tu_n\rangle \xrightarrow{n\to\infty} 0.\end{aligned}$$

Bounds on the Rayleigh Quotient

Proof (continued).

If T is not positive, we can find some $\lambda \in \mathbb{R}$ such that $S = T + \lambda I$ is positive. Then $U_S = U_T + \lambda$ and the sequence (u_n) such that $\langle u_n, Tu_n \rangle \to U_T$ also satisfies $\langle u_n, Su_n \rangle \to U_S$. By our previous calculation, we see that

$$Su_n - (U_T + \lambda)u_n \to 0,$$

i.e., $Tu_n - U_T u_n \to 0$, completing the proof. The argument for L_T is analogous. □

Compact Operators

In the previous section, we have seen that even self-adjoint, bounded operators T are not sufficiently "nice" to ensure, for example, that the bounds on the Rayleigh quotient are eigenvalues. It turns out that properties of this quality are present in an important sub-class of the bounded operators, the compact operators. (These are often denoted with the letter "K", from the german ***kompakt***.)

2.5.1. Definition. Let $\mathscr{H}$ be a Hilbert space and K a linear operator on $\mathscr{H}$. Then K is said to be a ***compact operator*** if for every bounded sequence (u_n) the sequence (Ku_n) has a convergent subsequence.

2.5.2. Remark. A compact operator is bounded: if K is an unbounded operator, the exists a sequence (u_n) of unit elements such that $\|Ku_n\| \to \infty$. We can choose the sequence so that $\|Ku_{n+1}\| > \|Ku_n\|$ and then it is impossible for the sequence to have a convergent subsequence. Hence, unbounded operators can not be compact.

Compact Operators

2.5.3. Examples.

(i) In $\mathbb{R}^n$, any bounded sequence has a convergent subsequence (this is the Theorem of Bolzano-Weierstrass 1.8.6). Since any bounded operator will transform a bounded a sequence into a bounded sequence, the latter of which then also has a convergent subsequence, it follows that every bounded operator on $\mathbb{R}^n$ is compact. Since every linear operator on finite-dimensional spaces is bounded, it actually follows that every linear operator on $\mathbb{R}^n$ is compact.

(ii) The above can be extended to operators whose range is finite-dimensional (these are called ***finite-rank operators***). Every finite-rank operator is compact.

(iii) Hilbert-Schmidt operators are compact, as we shall see.

Limits of Compact Operators

2.5.4. Theorem. Let (K_n) be a sequence of compact operators on $\mathscr{H}$ that converges to an operator T in norm, i.e.,

$$\lim_{n\to\infty} \|K_n - T\| = 0.$$

Then T is compact.

Proof.

Let (u_n) be a bounded sequence. We will show that there is a subsequence (v_n) of (u_n) such that (Tv_n) converges. Since K_1 is compact, there exists a subsequence $(u_n^{(1)})$ such that $(K_1 u_n^{(1)})$ converges. Since K_2 is compact, there exists a subsequence $(u_n^{(2)})$ of $(u_n^{(1)})$ such that $(K_2 u_n^{(2)})$ converges. (Of course, $(K_1 u_n^{(2)})$ still converges as well.) Proceeding iteratively, we can find a subsequence $(u_n^{(m)})$ of $(u_n^{(m-1)})$ such that $(K_m u_n^{(m)})$ converges. We now define the sequence (v_n) by $v_n := u_n^{(n)}$, i.e., we take the nth term of the nth iterative subsequence of (u_n) constructed above.

Limits of Compact Operators

Proof (continued).

Then for any $m \in \mathbb{N}$, $(K_m v_n)$ converges. Furthermore, for any $m, n, k \in \mathbb{N}$,

$$\begin{aligned}\|Tv_n - Tv_m\| &\leq \|Tv_n - K_k v_n\| + \|K_k v_n - K_k v_m\| + \|K_k v_m - Tv_m\| \\ &\leq \|T - K_k\|\big(\|v_m\| + \|v_n\|\big) + \|K_k v_n - K_k v_m\|.\end{aligned}$$

Since (v_n) is a subsequence of (u_n), it is also bounded and we can find a $c > 0$ such that $\|v_m\| + \|v_n\| < c$. By choosing k sufficiently large, we ensure that $\|T - K_k\| < \varepsilon/(2c)$. We then choose m, n large enough so that $\|K_k v_n - K_k v_m\| < \varepsilon/2$. Then

$$\|Tv_n - Tv_m\| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} < \varepsilon$$

for m, n sufficiently large. Hence (Tv_n) is a Cauchy sequence and converges, because $\mathscr{H}$ is complete. □

Hilbert-Schmidt Operators

2.5.5. Example. A ***Hilbert-Schmidt operator*** on $L^2([a,b])$ has the form

$$(Ku)(x) := \int_a^b k(x,y)u(y)\,dy,$$

where the kernel k satisfies $\int_a^b \int_a^b |k(x,y)|^2\,dx\,dy =: M^2 < \infty$. We will show that such an operator is always compact.

For simplicity, we consider $[a,b] = [0,1]$. Then

$$\int_0^1 \int_0^1 |k(x,y)|^2\,dx\,dy < \infty$$

means that $k \in L^2([0,1]\times[0,1])$, a generalization of the L^2 space (1.6.3) to functions of two variables.

Hilbert-Schmidt Operators

It can be shown that the functions

$$\left\{e^{2\pi i(mx+ny)}\right\}_{m,n\in\mathbb{Z}}$$

are an orthonormal basis of this space with respect to the scalar product

$$\langle u, v\rangle := \int_0^1 \int_0^1 \overline{u(x,y)}v(x,y)\,dx\,dy.$$

We can hence expand $k(x,y)$ into a "two-dimensional" Fourier series, writing

$$k(x,y) = \sum_{m,n\in\mathbb{Z}} c_{mn}e^{2\pi i(mx+ny)}$$

where $c_{mn} = \langle e^{2\pi i(mx+ny)}, k\rangle$.

Hilbert-Schmidt Operators

Define

$$(K_N u)(x) := \int_a^b k_N(x,y)u(y)\,dy,$$

where

$$k_N(x,y) = \sum_{m,n=-N}^{N} c_{mn} e^{2\pi i(mx+ny)}$$

is the Fourier approximation to $k(x,y)$. Then K_N is a finite-rank operator and therefore compact. Furthermore,

$$\begin{aligned} |Ku(x) - K_N u(x)|^2 &= \left| \int_0^1 \big(k(x,y) - k_N(x,y)\big) u(y)\,dy \right|^2 \\ &\leq \int_0^1 \big|k(x,y) - k_N(x,y)\big|^2\,dy \cdot \int_0^1 |u(y)|^2\,dy \end{aligned}$$

Hilbert-Schmidt Operators

so that

$$\begin{aligned}\|(K - K_N)u\|_{L^2}^2 &= \int_0^1 |Ku(x) - K_N u(x)|^2 \, dx \\ &\leq \int_0^1 \int_0^1 \left|k(x,y) - k_N(x,y)\right|^2 dy \, dx \cdot \|u\|_{L^2}^2\end{aligned}$$

and hence

$$\|K - K_N\| \leq \left(\int_0^1 \int_0^1 \left|k(x,y) - k_N(x,y)\right|^2 dy \, dx \right)^{1/2}.$$

Since k_N is just the partial sum of the series expansion of k,

$$\|K - K_N\| \xrightarrow{N\to\infty} 0.$$

Hence, K is the limit of finite-rank operators and therefore compact.

Inverse of Compact Operators

Compact operators are "convergence-inducing" in that they transform a sequence that is merely bounded into a sequence with a convergent subsequence. Of course, the inverse of a compact operator then has the opposite effect and is thus not a very well-behaved object.

2.5.6. Theorem. If K is a compact operator on $\mathscr{H}$ and (e_n) an infinite orthonormal sequence in $\mathscr{H}$, then $\lim_{n\to\infty} Ke_n = 0$.

In particular, if K is invertible, then K^{-1} is unbounded.

Proof.

Let (e_n) be an orthonormal sequence and suppose that (Ke_n) does not converge to zero. Then there exists a subsequence (f_n) of (e_n) and some $\varepsilon > 0$ such that $\|Kf_n\| > \varepsilon$ for all n. Since an orthonormal sequence is bounded and K is compact, we can find a subsequence of (g_n) of (f_n) such that (Kg_n) converges to some $u \in \mathscr{H}$. Since $\|Kg_n\| > \varepsilon$ for all n, it follows that $\|u\| \geq \varepsilon$.

Inverse of Compact Operators

Proof (continued).

By the continuity of the inner product,

$$\langle Kg_n, u\rangle \to \langle u, u\rangle = \|u\|^2 \geq \varepsilon^2.$$

However, by the Riemann-Lebesgue Lemma 1.3.23,

$$\langle Kg_n, u\rangle = \langle g_n, K^* u\rangle \xrightarrow{n\to\infty} 0$$

since (g_n) is a subsequence of an orthonormal system. This gives a contradiction and establishes the first part of the theorem.

Furthermore, since $Ke_n \to 0$ for an orthonormal system, we see that K is not bounded away from zero, so if K is invertible, the inverse can not be bounded. □

Inverse of Compact Operators

A typical example is the following:

2.5.7. Example. Let K be the operator on $L^2([0,1])$ defined by

$$(Ku)(x) := \int_0^x u(y)\,dy.$$

This is a Hilbert-Schmidt operator with kernel $k(x,y) = H(x-y)$, where H is the Heaviside function

$$H(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Hence, K is compact. The inverse of K is given by

$$K^{-1} = \frac{d}{dx},$$

which is an unbounded operator.

PDEs and Separation of Variables

We now introduce an important application of our study of general linear operators. The classical ***heat equation*** in n dimensions is

$$\varrho(x)c(x)\frac{\partial u(x,t)}{\partial t} = \operatorname{div}(k(x)\operatorname{grad} u(x,t)) + q(x,t), \tag{2.6.1}$$

where

- u is the temperature at position $x \in \Omega \subset \mathbb{R}^n$ and time $t \in \mathbb{R}$,
- ϱ is the density of the material,
- c is the specific heat capacity,
- k is the heat conduction coefficient and
- q represents the heat source density.

(If k is a constant function of x, the term $\operatorname{div}(k \operatorname{grad} u)$ reduces to $k\Delta u$.)

One usually specifies boundary conditions for u on $\partial\Omega$ and an initial condition

$$u(x,0) = f(x). \tag{2.6.2}$$

PDEs and Separation of Variables

Let us simplify the notation and consider an equation of the form

$$\frac{\partial u}{\partial t} + Lu = q(x,t),$$

where L is a linear differential operator with respect to the x coordinates. Furthermore, we consider the case where q vanishes identically, so the equation is homogeneous,

$$\frac{\partial u}{\partial t} + Lu = 0. \tag{2.6.3}$$

We make "separation of variables" ansatz by setting

$$u(x,t) = X(x)T(t)$$

for unknown functions X, T. Inserting into (2.6.3) yields

$$-\frac{T'}{T} = \frac{LX}{X}.$$

PDEs and Separation of Variables

Since the left-hand side is independent of t and the right-hand side is independent of x, both sides must be constant, say equal to $\lambda \in \mathbb{C}$. Hence we need to solve the equations

$$T' = -\lambda T, \qquad LX = \lambda X. \tag{2.6.4}$$

In addition, there are boundary conditions for X on Ω.

Essentially, we need to solve the eigenvalue problem for the linear differential operator L. The separation of variables approach will yield a solution only if

- L has an eigenvalue.

It turns out that in many cases

- L has a countably infinite number of eigenvalues λ_n, $n \in \mathbb{N}$, and eigenfunctions X_n.

PDEs and Separation of Variables

The general solution of (2.6.3) is then

$$u(x,t) = \sum_{n=0}^{\infty} u_n e^{-\lambda_n t} X_n(x), \qquad u_n \in \mathbb{C},$$

providing that the series converges; this relies on the fact that

- $\lambda_n \to +\infty$ as $n \to \infty$.

In order to satisfy the initial condition (2.6.2), we require

$$u(x,0) = \sum_{n=0}^{\infty} u_n X_n(x) = f(x)$$

for suitable functions f. This is possible for any $f \in L^2$ if

- The set of eigenfunctions $\{X_n\}$ is a basis of a suitable L^2 space.

Differential and Compact Operators

Our goal in this section is to lay the groundwork for analyzing differential operators such as the L of (2.6.3). In particular, we would like to prove the various properties of the eigenvalues and -functions of L that have been mentioned on the preceding slides. The main result that makes this possible will be the discovery that

- The resolvent of a differential operator such as L is in many cases a compact operator and
- The eigenvalues and -functions of the resolvent are closely related to the eigenvalues and -functions of L.

For this reason, we will first study the spectral theory of compact operators in more detail.

Existence of Eigenvalues

An important question for linear operators is the existence of eigenvalues; in finite-dimensional, complex Hilbert spaces (essentially, in $\mathbb{C}^n$), the existence of eigenvalues is guaranteed by the existence of (complex) zeroes of the characteristic polynomial.

In the infinite-dimensional case, the situation is more complicated: a bounded linear operator (such as the right-shift operator on ℓ^2 in Example 2.4.5) does not need to have any eigenvalues at all. However, self-adjoint, compact operators always have eigenvalues, as we now show.

2.6.1. Theorem. Let K be a self-adjoint, compact operator on a Hilbert space $\mathscr{H}$. Suppose $\lambda \neq 0$ is in the approximate point spectrum of K. Then λ is an eigenvalue of K.

Existence of Eigenvalues

Proof.

Let $\lambda \in \sigma_{\text{point}}(L) \cup \sigma_{\text{continuous}}(L)$. Then there exists a sequence (u_n) of unit elements such that $Ku_n - \lambda u_n \to 0$. Since K is compact, a subsequence (v_n) of (u_n) will have the property that Kv_n converges. Then (λv_n) converges since

$$\lambda v_n = \underbrace{Kv_n - \lambda v_n}_{\to 0} + \underbrace{Kv_n}_{\text{converges}}.$$

Since $\lambda \neq 0$, this implies that (v_n) converges to some unit element v. Since K is continuous, v satisfies

$$Kv = K\left(\lim_{n\to\infty} v_n\right) = \lim_{n\to\infty} Kv_n = \lim_{n\to\infty} \lambda v_n = \lambda v,$$

so λ is an eigenvalue with eigenvector v. $\square$

Existence of Eigenvalues

Hence, any compact, self-adjoint operator on a Hilbert space has eigenvalues:

2.6.2. Corollary. Let K be a compact, self-adjoint operator on a Hilbert space $\mathscr{H}$.

(i) The Rayleigh bounds U_K and L_K are in the approximate point spectrum by Theorem 2.4.16. If either is non-zero, it is an eigenvalue.

(ii) If K is not the zero operator, then $\|K\|$ or $-\|K\|$ is an eigenvalue. If $Ku = 0$ for all $u \in \mathscr{H}$, then $\lambda = 0$ is an eigenvalue.

The Spectrum of Compact Operators

We now know a certain amount about the spectrum of self-adjoint, compact operators:

(i) The point spectrum is non-empty (there exists an eigenvalue).

(ii) The compression spectrum coincides with the point spectrum (by self-adjointness; cf. Theorem 2.4.7).

(iii) $\lambda = 0$ is in the spectrum (because $K_\lambda = K - 0 \cdot I = K$ can not have a bounded inverse).

Moreover, Theorem 2.6.1 immediately implies the followinG::

2.6.3. Fredholm Alternative. Let K be a compact, self-adjoint operator on a Hilbert space $\mathscr{H}$. Let $\lambda \in \mathbb{C}$, $\lambda \neq 0$. Then

- either $\lambda \in \varrho(K)$
- or $\lambda \in \sigma_{\text{point}}(K)$.

The Fredholm Alternative

2.6.4. Remark. The Fredholm alternative is also true for non-self-adjoint compact operators, but the proof is more complicated. For our purposes, this simplified version is sufficient.

2.6.5. Remark. The Fredholm alternative can be rephrased as follows:

- Either the equation $(K - \lambda)u = v$ has a unique solution u for any given v
- or the equation $(K - \lambda)u = 0$ has a non-trivial solution $u \neq 0$.

This is similar to the situation for matrices, where

- either the equation $Ax = y$ has a unique solution for any given $y \in \mathbb{R}^n$ (if $\det A \neq 0$).
- or the equation $Ax = 0$ has a non-trivial solution (if $\det A = 0$).

The Spectral Theorem for Compact Operators

2.6.6. Spectral Theorem. Let K be a compact, self-adjoint operator on a (separable) Hilbert space $\mathscr{H}$. Then there exists an orthonormal basis (e_n) of $\mathscr{H}$ and numbers $\lambda_n \in \mathbb{R}$ such that $Ke_n = \lambda_n e_n$.

If $\mathscr{H}$ is infinite-dimensional, then the eigenvalues λ_n can be arranged in a monotonically decreasing sequence with $|\lambda_n| \searrow 0$.

Proof.
In the trivial case $K = 0$, we have the eigenvalue $\lambda = 0$ only and we can take any orthonormal basis of $\mathscr{H}$.

If $K \neq 0$, we have an eigenvalue $\lambda^{(1)} = \pm\|K\|$. Let M_1 be the space spanned by the eigenvectors for this $\lambda^{(1)}$ (called the eigenspace for $\lambda^{(1)}$). The eigenspace M_1 must be finite-dimensional (why?). We choose an orthonormal basis of M_1. If $\overline{M_1} = M_1 = \mathscr{H}$, we are finished.

The Spectral Theorem for Compact Operators

Proof (continued).

If $M_1 \neq \mathscr{H}$, the orthogonal complement $M_1^\perp$ contains a non-zero element. Furthermore, since $Ku \in M_1$ for all $u \in M_1$, we also have $Kv \in M_1^\perp$ for all $v \in M_1^\perp$ (why?). It follows that

$$K_1 := K|_{M_1^\perp} : M_1^\perp \to M_1^\perp$$

is a well-defined operator that remains self-adjoint and compact with $\|K_1\| \leq \|K\|$. If $\|K_1\| = 0$, we take an arbitrary orthonormal basis in $M_1^\perp$ for the eigenvalue $\lambda = 0$ and we are finished.

If $K_1 \neq 0$, there exists an eigenvalue $\lambda^{(2)} = \pm\|K_1\|$ and we can repeat the above argument, finding a space M_2 spanned by the eigenvectors of K to the eigenvalue $\lambda^{(2)}$. We then consider the orthogonal complement of M_2 within $M_1^\perp$ etc., etc.

The Spectral Theorem for Compact Operators

Proof (continued).

We hence obtain a sequence of distinct eigenvalues with $|\lambda^{(1)}| > |\lambda^{(2)}| > \dots$. If $\mathscr{H}$ is finite-dimensional, the iterative procedure terminates when $M_n^{\perp} = \{0\}$ for some $n \in \mathbb{N}$. The union of the orthonormal bases in $M_1, \dots, M_n$ then gives an orthonormal basis for $\mathscr{H}$ with the required properties.

If $\mathscr{H}$ is infinite-dimensional, we obtain an infinite sequence of finite-dimensional eigenspaces M_n and a decreasing sequence of eigenvalues $|\lambda^{(1)}| > |\lambda^{(2)}| > \dots$ and orthonormal eigenvectors. Let us denote the sequence of eigenvectors by (e_k) and the corresponding eigenvalues by λ_k, where we adjust the previous notation to allow $\lambda_k = \lambda_j$ for $j \neq k$ if a given eigenvalue has more than one independent eigenvector.

The Spectral Theorem for Compact Operators

Proof (continued).

We claim that the sequence (λ_k) converges to zero, which we show as follows: It is sufficient to establish that $|\lambda_n| \to 0$ as $n \to \infty$. The sequence $(|\lambda_n|)$ is decreasing and bounded below, so it converges. Suppose that $\lim|\lambda_n| = \Lambda$. Then K applied to the sequence of (orthonormal) eigenvectors gives

$$\|Ke_n - Ke_m\|^2 = \|\lambda_n e_n - \lambda_m e_m\|^2 = |\lambda_n|^2 + |\lambda_m|^2 > 2\Lambda.$$

If $\Lambda > 0$, then it is impossible for the sequence (Ke_n) to contain a subsequence that is Cauchy (i.e., converges). But this contradicts the compactness of K. Hence, $\Lambda = 0$.

The Spectral Theorem for Compact Operators

Proof (continued).

Finally, we show that the sequence of eigenvectors (e_n) is indeed a basis of $\mathscr{H}$. Let $M = \operatorname{span}\{e_n\}$ be the span of all eigenvectors of K. Then $Ku \in M$ if $u \in M$ and $Kv \in M^\perp$ if $v \in M^\perp$. Now for every $n \in \mathbb{N}$, we have $M_n \subset M$, so $M^\perp \subset M_n^\perp$ (why?). It follows that

$$\|K|_{M^\perp}\| \leq \|K|_{M_n^\perp}\| = \|K_n\| = |\lambda^{(n+1)}| \xrightarrow{n\to\infty} 0.$$

This implies that $K|_{M^\perp} = 0$, i.e., $M^\perp = \ker K$. If $M^\perp = \{0\}$, we are finished. If the kernel of K is non-trivial, $\lambda = 0$ is an eigenvalue and an orthonormal basis of $M^\perp$ consists of eigenvectors to this eigenvalue.

This completes the proof of the theorem. □

Principal Axis Transformation for Symmetric Matrices

As an example, consider a square $n \times n$ matrix A with real coefficients. Then $A\colon \mathbb{R}^n \to \mathbb{R}^n$ is a compact operator and A is self-adjoint if $A = A^T$.

The spectral theorem then states that there exist n orthonormal eigenvectors $v_1, \dots, v_n$ with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$.

These eigenvectors are an orthonormal basis of $\mathbb{R}^n$. If $U = (v_1, \dots, v_n)$ is the $n \times n$ matrix whose columns are these eigenvectors, then

$$UAU^{-1} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

is "the matrix representation of A in the basis of eigenvectors". (Here the right-hand side denotes a matrix which is zero everywhere except on the diagonal, where it has the entries $\lambda_1, \dots, \lambda_n$.) This is just the principal axis transformation familiar from linear algebra.

Spectral Decomposition of Compact Operators

2.6.7. Corollary. Let K be a compact, self-adjoint operator on $\mathscr{H}$ and (e_n) an orthonormal basis of eigenvectors of K. Then

$$Ku = \sum_n \lambda_n \langle e_n, u \rangle e_n \qquad \text{for all } u \in \mathscr{H}. \tag{2.6.5}$$

This follows simply from $u = \sum_n \langle e_n, u \rangle e_n$ and the continuity of K.

2.6.8. Example. If $A \in \mathrm{Mat}(n \times n; \mathbb{C})$ is self-adjoint with eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and orthonormal eigenvectors $e_1, \dots, e_n$, then

$$Ax = U^T \operatorname{diag}(\lambda_1, \dots, \lambda_n) U x.$$

where $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$ is the $n \times n$ matrix with the eigenvalues on the diagonal and all other entries vanishing.

Second Midterm Exam

The preceding material completes the second third of the course material. It encompasses everything that will be the subject of the Second Midterm Exam.

The exact exam date will be announced on SAKAI.

No calculators or other aids will be permitted during the exam.

Part III

Applications of Operator Theory

Sturm-Liouville Boundary Value Problems

The Rayleigh-Ritz Method

Positive Operators and the Polar Decomposition

The Singular Value Decomposition for Compact Operators and Matrices

Sturm-Liouville Eigenvalue Problems

Let us now return to the previous discussion of the heat equation (2.6.3). In order to solve

$$\frac{\partial u}{\partial t} + Lu = 0$$

an ansatz of the form $u(x,t) = X(x)T(t)$ yields an eigenvalue problem (2.6.4) for X.

In the case of a single space dimension, L is an ordinary differential operator and the boundary conditions are imposed on an interval $I \subset \mathbb{R}$. Such problems are called ***Sturm-Liouville problems***.

Sturm-Liouville Eigenvalue Problems

We will consider the operator

$$L := -\frac{1}{r(x)}\left(\frac{d}{dx}\left(p(x)\frac{d}{dx}\right) + q(x)\right) \tag{3.1.1}$$

which encompasses the operator

$$Lu = -\frac{1}{\varrho(x)c(x)}\left(\operatorname{div}(k(x)\operatorname{grad} u(x,t)) + q(x)\right)$$

of (2.6.1) in one space dimension.

Generality of the Sturm-Liouville Operator

3.1.1. Remark. Formally, any second-order operator can be written in the form (3.1.1): let

$$L = a_2(x)\frac{d^2}{dx^2} + a_1(x)\frac{d}{dx} + a_0(x)$$

is given. Set

$$p(x) = e^{\int \frac{a_1}{a_2}}, \qquad r(x) = -\frac{p(x)}{a_2(x)}, \qquad q(x) = -a_0(x)r(x). \qquad (3.1.2)$$

Then L is given by (3.1.1) with p, q, r as in (3.1.2).

3.1.2. Example. Let $L = x^2\frac{d^2}{dx^2} + \frac{d}{dx} + x^3$. With

$$p(x) = e^{\int \frac{a_1}{a_2}} = e^{-\frac{1}{x}}, \qquad r(x) = -\frac{1}{x^2}e^{-\frac{1}{x}}, \qquad q(x) = xe^{-\frac{1}{x}},$$

we can write

$$L = x^2 e^{\frac{1}{x}}\left(\frac{d}{dx}\left(e^{-\frac{1}{x}}\frac{d}{dx}\right) + xe^{-\frac{1}{x}}\right).$$

Regular Sturm-Liouville Problems (ODE Point of View)

We suppose that $I = (a, b)$ is a bounded interval, $p, p', q, r \in C([a, b])$ and $p(x), r(x) > 0$ for all $x \in [a, b]$. Then the equation

$$\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + (q(x) + \lambda r(x))\, u = 0, \qquad x \in (a, b), \tag{3.1.3a}$$

is said to be a ***regular Sturm-Liouville equation***. We impose boundary conditions

$$\begin{aligned} B_a u &:= \alpha_1 u(a) + \beta_1 u'(a) = 0, \\ B_b u &:= \alpha_2 u(b) + \beta_2 u'(b) = 0, \end{aligned} \tag{3.1.3b}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$ and $|\alpha_1| + |\beta_1| \neq 0$, $|\alpha_2| + |\beta_2| \neq 0$. We sometimes refer to B_a and B_b as ***boundary operators***.

The problem (3.1.3) is said to be a ***regular Sturm-Liouville problem***.

Sturm-Liouville Problems (Operator Point of View)

A regular Sturm-Liouville BVP may be regarded as an eigenvalue problem for the Sturm-Liouville operator L on $L^2([a,b]; r(x)\,dx)$ with domain

$$U := \left\{ u \in C^2([a,b]) \colon B_a u = B_b u = 0 \right\}, \tag{3.1.4}$$

where

$$L^2([a,b]; r(x)\,dx) = \left\{ u \colon [a,b] \to \mathbb{R} \colon \int_a^b |u(x)|^2\, r(x)\,dx < \infty \right\}$$

is the space of weighted square-integrable functions. Eigenvalues λ and eigenfunctions $u_\lambda \in U$ will constitute solutions of the Sturm-Liouville problem.

We will combine the differential equations with the operator point of view to analyze the Sturm-Liouville problem.

The Wronskian

For $u, v \in C^2(a,b) \cap C([a,b])$ we define the Wronskian

$$W(u(x), v(x)) := \det \begin{pmatrix} u(x) & v(x) \\ p(x)u'(x) & p(x)v'(x) \end{pmatrix} = p(x)\big(u(x)v'(x) - v(x)u'(x)\big).$$

3.1.3. Lemma. Let $\lambda \in \mathbb{C}$ and $u, v \in C^2(a,b) \cap C([a,b])$ be any two solutions of the Sturm-Liouville equation (3.1.3a).

(i) The Wronskian $W(u(x), v(x))$ vanishes if and only if u and v are dependent.

(ii) The Wronskian $W(u(x), v(x))$ is constant.

The Wronskian

Proof.

(i) Since $p(x) > 0$ for all $x \in [a, b]$, the Wronskian vanishes if and only if $u(x) = \alpha v(x)$ and $u'(x) = \alpha v'(x)$ for some $\alpha \in \mathbb{R}$. Hence u and v are multiples of each other and therefore linearly dependent.

(ii) For $u, v \in C^2(a, b) \cap C([a, b])$ we have

$$\begin{aligned} r(uLv - vLu) &= -u(pv')' - uqv + v(pu')' + vqu \\ &= \big(p(vu' - uv')\big)' \\ \Rightarrow uLv - vLu &= \frac{\big(p(vu' - uv')\big)'}{r} = \frac{1}{r} W(u, v)' \end{aligned} \tag{3.1.5}$$

The equation (3.1.5) is called the ***Lagrange identity*** for L. If u, v satisfy $Lu = \lambda u$ and $Lv = \lambda v$, the left-hand side vanishes and the Wronskian is constant. □

Symmetry

3.1.4. Lemma. The regular Sturm-Liouville operator L is ***symmetric***, i.e.,

$$\langle u, Lv\rangle_{L^2([a,b];r(x)\,dx)} = \langle Lu, v\rangle_{L^2([a,b];r(x)\,dx)}.$$

for all $u, v \in U$.

Proof.

Integrating (3.1.5) over (a, b) we obtain ***Green's formula*** for L,

$$\int_a^b \big(u(x)Lv(x) - v(x)Lu(x)\big)r(x)\,dx = [p(vu' - uv')]_a^b. \tag{3.1.6}$$

Then

$$\begin{aligned}\langle u, Lv\rangle - \langle Lu, v\rangle &= \int_a^b \big(u(x)Lv(x)r(x) - v(x)Lu(x)r(x)\big)\,dx \\ &= [p(vu' - uv')]_a^b. \end{aligned} \tag{3.1.7}$$

Symmetry

Proof (continued).

For $u, v \in U$ we know that the boundary conditions (3.1.3b) hold. Thus, assuming that $\beta_1, \beta_2 \neq 0$ we have

$$u'(a) = -\frac{\alpha_1}{\beta_1}u(a), \qquad u'(b) = -\frac{\alpha_2}{\beta_2}u(b),$$
$$v'(a) = -\frac{\alpha_1}{\beta_1}v(a), \qquad v'(b) = -\frac{\alpha_2}{\beta_2}v(b).$$

Hence

$$\begin{aligned} v'(a)u(a) - u'(a)v(a) &= -\frac{\alpha_1}{\beta_1}(v(a)u(a) - u(a)v(a)) = 0, \\ v'(b)u(b) - u'(b)v(b) &= -\frac{\alpha_2}{\beta_2}(v(b)u(b) - u(b)v(b)) = 0. \end{aligned} \tag{3.1.8}$$

The same result is true if $\beta_1 = 0$ or $\beta_2 = 0$. It follows that the right-hand side of (3.1.7) vanishes. □

Simple Eigenvalues

3.1.5. Lemma. Let $\lambda \in \mathbb{R}$ be an eigenvalue of L. Then λ is ***simple***, i.e., there can not exist two independent eigenfunctions $u, v \in U$ with $Lu = \lambda u$ and $Lv = \lambda v$.

Proof.

Suppose that $u, v \in U$ satisfy $Lu = \lambda u$, $Lv = \lambda v$. By Lemma 3.1.3 it is sufficient to check that the Wronskian vanishes at a single point. Since $B_a u = 0$ and $B_a v = 0$, we have

$$W(u(a), v(a)) = p(a)\big(u(a)v'(a) - v(a)u'(a)\big) = 0$$

by (3.1.8). □

The Resolvent of the Sturm-Liouville Operator

The inhomogeneous Sturm-Liouville equation has the form

$$Lu - \lambda u = v \tag{3.1.9}$$

for some $v \in L^2([a, b]; r(x)\,dx)$. If λ is not an eigenvalue (so $L - \lambda I$ is invertible), (3.1.9) can be "resolved" by setting

$$u = (L - \lambda I)^{-1} v$$

where $(L - \lambda I)^{-1} = R_\lambda(L)$ is the resolvent of L.

(In fact, this exact problem was one of the main motivations for the development Hilbert space theory in the early 20^{th} century.)

The resolvent $R_\lambda(L)$ can be constructed explicitly using methods from the theory of differential equations. We now summarise the construction.

Construction of the Resolvent

Suppose that u_1 and u_2 satisfy the equations

$$\begin{aligned} (L-\lambda I)u_1 &= 0, & B_a u_1 &= 0, & u_1 &\neq 0, \\ (L-\lambda I)u_2 &= 0, & B_b u_2 &= 0, & u_2 &\neq 0. \end{aligned}$$

Then the solution of $(L-\lambda I)u = v$ is given by

$$u(x) = (L-\lambda I)^{-1} v(x) = \int_a^b g(x,y)v(y)\,r(y)\,dy, \tag{3.1.10}$$

where

$$g(x,y) := \begin{cases} \dfrac{u_1(x)u_2(y)}{W(u_1,u_2)} & \text{if } y \leq x, \\ \dfrac{u_1(y)u_2(x)}{W(u_1,u_2)} & \text{if } y > x. \end{cases}$$

Construction of the Resolvent

Since $g(x, y)$ is a continuous function, the resolvent (3.1.10) is a Hilbert-Schmidt operator and therefore compact. Furthermore, since $g(x, y) = g(y, x)$, the resolvent is self-adjoint.

We can now apply the theory of compact operators to the resolvent. The following theorem relates the properties of the resolvent to the properties of L.

3.1.6. Theorem. Let L be a linear operator on a Hilbert space $\mathscr{H}$. Let $\mu \in \mathbb{R}$ be such that μ is ***not*** an eigenvalue of L and let $u \in \mathscr{H}$. Then

- $\mu + \dfrac{1}{\lambda}$ is an eigenvalue of L with eigenfunction u and $\lambda \neq 0$

if and only if

- λ is an eigenvalue of $(L - \mu I)^{-1}$ with the same eigenfunction u.

The Sturm-Liouville Eigenvalue Problem

3.1.7. Theorem. Let L be a linear operator on a Hilbert space $\mathscr{H}$. Let $\mu \in \mathbb{R}$ be such that μ is ***not*** an eigenvalue of L and let $u \in \mathscr{H}$. Then

- $\mu + \dfrac{1}{\lambda}$ is an eigenvalue of L with eigenfunction u and $\lambda \neq 0$

if and only if

- λ is an eigenvalue of $(L - \mu I)^{-1}$ with the same eigenfunction u.

Proof.

($\Rightarrow$) Suppose that $\lambda \neq 0$ and $Lu = \left(\mu + \frac{1}{\lambda}\right) u$. Then

$$u = \lambda(L - \mu I)u$$

and, applying $(L - \mu I)^{-1}$ to both sides,

$$(L - \mu I)^{-1} u = \lambda u.$$

The Sturm-Liouville Eigenvalue Problem

Proof.

($\Leftarrow$) Conversely, suppose that

$$(L - \mu I)^{-1} u = \lambda u$$

for $u \in \operatorname{ran}(L - \mu I)$. Then, with $u = (L - \mu I)v$,

$$v = \lambda (L - \mu I) v$$

and

$$Lv = \left(\mu + \frac{1}{\lambda}\right) v.$$

Applying $L - \mu I$ to both sides, we note $(L - \mu I)Lv = L(L - \mu I)v = Lu$ and so

$$Lu = \left(\mu + \frac{1}{\lambda}\right) u.$$

□

The Sturm-Liouville Operator is Bounded Below

3.1.8. Definition. Let L be a symmetric operator with dense domain on a Hilbert space $\mathscr{H}$.

We say that L is ***bounded below*** if there exists a constant $c \in \mathbb{R}$ such that

$$\langle u, Lu \rangle \geq c\|u\|^2 \qquad \text{for all } u \in \operatorname{dom} L. \tag{3.1.11}$$

We say that L is ***bounded above*** if $-L$ is bounded below.

3.1.9. Theorem. The Sturm-Liouville operator L with $\operatorname{dom} L = U$ is bounded below.

The proof, which involves some rather fine analysis, is part of this week's homework.

It follows that the lower Rayleigh bound (see (2.4.5)) of L is finite and that there is a lower bound on the eigenvalues of L.

In particular, there exists a number μ which is ***not*** an eigenvalue of L.

Spectral Theorem for the Sturm-Liouville Operator

We obtain the spectral theorem for regular Sturm-Liouville operators:

3.1.10. Spectral Theorem. Let L be a regular Sturm-Liouville operator (3.1.1) on $L^2([a,b]; r(x)\,dx)$ with domain (3.1.4). Then

$$\sigma_{\text{point}}(L) \neq \emptyset$$

The eigenvalues of L are simple and form a countable, increasing sequence (λ_n) with $\lim_{n\to\infty} \lambda_n = \infty$. The corresponding normed eigenvectors are an orthonormal basis of $L^2([a,b]; r(x)\,dx)$.

3.1.11. Remark. The spectral theorem ensures the conditions discussed in Slide 293 for the separation-of-variables method to succeed.

Spectral Theorem for the Sturm-Liouville Operator

Proof.

Since L is bounded below, we can find $\mu \in \mathbb{R}$ such that μ is not an eigenvalue of L. The resolvent $R_\mu(L)$, given by (3.1.10), is compact and self-adjoint so that by the Spectral Theorem for compact operators 2.6.6 there exists a sequence of eigenvalues (λ_n) of $R_\mu(L)$ such that $\lambda_n \to 0$ as $n \to \infty$.

By Theorem 3.1.7, $(\mu + 1/\lambda_n)$ is then the sequence of eigenvalues for L (there are no other eigenvalues) with the same eigenfunctions. Furthermore,

$$\mu + \frac{1}{\lambda_n} \xrightarrow{n\to\infty} +\infty$$

(Since L is bounded below, the convergence can not be to $-\infty$.) The eigenfunctions of $R_\mu(L)$ are a basis of $L^2([a, b]; r(x)\,dx)$. Since they coincide with the eigenfunctions of L, this proves the last assertion of the theorem. □

Spectral Theorem for the Sturm-Liouville Operator

3.1.12. Example. The Sturm-Liouville operator $L = -\frac{d^2}{dx^2}$ on $L^2([0,\pi])$ with domain

$$U = \{u \in C^2([0,\pi]) \colon u(0) = u(\pi) = 0\}$$

has eigenvalues $\lambda_n = n^2$, $n \in \mathbb{N} \setminus \{0\}$, and (normed) eigenfunctions

$$e_n(x) = \frac{2}{\sqrt{\pi}} \sin(nx), \qquad n \in \mathbb{N}.$$

By the Spectral Theorem, the sequence (e_n) of eigenfunctions is an orthonormal basis of $L^2([0,\pi])$.

This establishes that the Fourier-sine orthonormal system of functions (1.7.6) is actually a basis.

Estimating Eigenvalues

Throughout this section we denote by K an operator that is ***compact***, ***self-adjoint*** and ***positive*** (see Definition 2.4.14) on a separable Hilbert space $\mathscr{H}$. We have seen that the spectrum of K consists of a sequence of (possibly repeated) eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots \geq 0$$

with $\lambda_n \to 0$ as $n \to \infty$. The number 0 is either an eigenvalue or in the continuous spectrum of K. We denote by (e_n) the orthonormal basis of $\mathscr{H}$ associated to K and write

$$M_0 := \{0\}, \qquad M_n := \operatorname{span}\{e_1, \ldots, e_n\}, \quad n \geq 1.$$

For simplicity, we assume that the sequence of eigenvalues is infinite; if there are only m non-zero eigenvalues, we set $\lambda_n = 0$ and assume that $e_n \in \ker K$ for $n > m$.

An Expression for the Eigenvalues

3.2.1. Theorem. The $(n+1)$st eigenvalue satisfies

$$\lambda_{n+1} = \max_{u \in M_n^\perp} R(u),$$

where $R(u)$ is the Rayleigh quotient (2.4.3).

Proof.
By (2.6.5) we have $\langle u, Ku\rangle = \sum_{k=1}^{\infty} \lambda_k |\langle e_k, u\rangle|^2$ for $u \in \mathscr{H}$. If $u \in M_n^\perp$, $\langle e_k, u\rangle = 0$ for $k \le n$ and

$$\langle u, Ku\rangle = \sum_{k=n+1}^{\infty} \lambda_k |\langle e_k, u\rangle|^2 \le \lambda_{n+1} \sum_{k=n+1}^{\infty} |\langle e_k, u\rangle|^2 = \lambda_{n+1} \|u\|^2.$$

Hence, $R(u) \le \lambda_{n+1}$. Furthermore, $R(e_{n+1}) = \lambda_{n+1}$, so the theorem is proven. □

The Weyl-Courant Minimax Theorem

Theorem 3.2.1 gives a basic expression for the eigenvalues of K, but it is of limited usefulness, as it requires knowledge of the eigenfunctions. A more practical approach is to replace M_n with a subspace E_n spanned by "wrong" functions that may not be eigenfunctions. The following theorem states that λ_{n+1} can be found by minimizing the result over all possible E_n:

3.2.2. Weyl-Courant Minimax Theorem. Let E_n be any n-dimensional subspace of $\mathscr{H}$ and set

$$\nu(E_n) := \max_{u \in E_n^\perp} R(u).$$

Then

$$\lambda_{n+1} = \min_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \nu(E_n) = \min_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \max_{u \in E_n^\perp} R(u).$$

The Weyl-Courant Minimax Theorem

Proof.

From Theorem 3.2.1 we have $\lambda_{n+1} = \nu(M_n)$, so

$$\min_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \nu(E_n) \leq \lambda_{n+1}.$$

We now show the reverse inequality as follows: for each choice of E_n we find an element $w \in E_n^{\perp}$ such that $R(w) \geq \lambda_{n+1}$. Then $\nu(E_n) \geq \lambda_{n+1}$ for all E_n and the theorem is proven.

Let E_n be given with basis $\{v_1, \ldots, v_n\}$. Then we can find numbers $c_1, \ldots, c_{n+1}$ such that

$$w = c_1 e_1 + \cdots + c_n e_n + c_{n+1} e_{n+1}$$

is non-zero and $\langle w, v_k \rangle = 0$ for all $k = 1, \ldots, n$.

The Weyl-Courant Minimax Theorem

Proof (continued).

(This is possible because the $c_1, \ldots, c_{n+1}$ are determined by a homogeneous system of equations with $n+1$ unknowns and n equations - there is always a non-trivial solution.) Then

$$R(w) = \frac{\langle Kw, w\rangle}{\|w\|^2} = \frac{\sum_{k=1}^{n+1} \lambda_k |c_k|^2}{\sum_{k=1}^{n+1} |c_k|^2} \geq \lambda_{n+1}.$$

□

The Weyl-Courant Minimax Theorem

The Weyl-Courant Theorem provides a method of estimating the eigenvalue λ_{n+1} by calculating $\nu(E_n)$ for a special case V_n of E_n. Then the eigenvalue λ_n will not be greater that the approximation:

$$\lambda_{n+1} = \min_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \nu(E_n) \leq \nu(V_n)$$

However, if $\mathscr{H}$ is infinite-dimensional, then so is $V_n^{\perp}$ and calculating $\nu(V_n) = \max_{u \in V_n^{\perp}} R(u)$ can be quite difficult.

The Weyl-Courant principle is instead quite useful for proving certain properties of operators.

The Rayleigh-Ritz Method

Another approach, based on the original Theorem 3.2.1, is used by the ***Rayleigh-Ritz method***. Suppose we are interested in the first (highest) eigenvalue λ_1. Then, by Theorem 3.2.1,

$$\lambda_1 = \max_{u \in \mathscr{H}} R(u).$$

If we restrict the maximum to only those u from a subspace $V_n = \text{span}\{v_1, \ldots, v_n\}$ we have

$$\lambda_1 \geq \max_{v \in V_k} R(v) = \max_{c_1, \ldots, c_k \in \mathbb{F}} R(c_1 v_1 + \cdots + c_n v_n). \tag{3.2.1}$$

The elements $v_1, \ldots, v_n$ are called ***trial vectors*** (or ***trial functions*** if $\mathscr{H}$ is a space of functions) and are selected in a "suitable" way. The goal is, of course, for the maximum in (3.2.1) to be as close to λ_1 as possible.

The Rayleigh-Ritz Method

Suppose trial vectors $v_1, \ldots, v_n$ are given. then

$$R(c_1 v_1 + \cdots + c_n v_n) = \frac{\langle c_1 v_1 + \cdots + c_n v_n, K(c_1 v_1 + \cdots + c_n v_n)\rangle}{\|c_1 v_1 + \cdots + c_n v_n\|^2}$$
$$= \frac{\sum_{i,j=1}^{n} \overline{c_i} c_j k_{ij}}{\sum_{i,j=1}^{n} \overline{c_i} c_j \alpha_{ij}} \qquad (3.2.2)$$

where

$$\alpha_{ij} = \langle v_i, v_j\rangle = \overline{\alpha}_{ji} \qquad \text{and} \qquad k_{ij} = \langle v_i, Kv_j\rangle = \overline{k}_{ji}$$

are known and can be calculated in advance. It is obviously a good idea numerically to choose the trial vectors to be orthonormal (or normalized and "nearly" orthogonal).

This works well for estimating the first eigenvalue. However, to apply this method for the second and further eigenvalues requires some more discussion.

The Galerkin Equation

We are effectively trying to find approximate eigenvectors for K in the space $V_n = \text{span}\{v_1, \dots, v_n\}$ and corrsponding approximations to eigenvalues. In other words, we would like to find approximate solutions to

$$Ku = \lambda u \tag{3.2.3}$$

by taking $u \in V_n$. However, $u \in V_n$ does not necessarily imply $Ku \in V_n$, which makes (3.2.3) impossible to solve exactly.

Define the orthogonal projection $P : \mathscr{H} \to V_n$. Then we can instead consider the eigenvalue problem

$$PKv = \Lambda v, \qquad v \in V_n, \tag{3.2.4}$$

which makes sense. Note that if K is compact, symmetric and positive, then so is PK (why?). Hence, $R(c_1 v_1 + \cdots + c_n v_n)$ is just the Rayleigh quotient for PK and maximizing it finds the largest eigenvalue Λ_1.

The Galerkin Equation

Given that $\{v_1, \dots, v_n\}$ is a basis of V_n, we can write out (3.2.4) in coordinate form by noting that it holds if and only if

$$\langle PKv - \Lambda v, v_j \rangle = 0, \qquad j = 1, \dots, k.$$

For $v = c_1 v_1 + \cdots + c_n v_n$ this reduces to the equations

$$\sum_{i=1}^{n} \langle Kv_i, v_j \rangle c_i = \Lambda \sum_{i=1}^{n} \langle v_i, v_j \rangle, \qquad j = 1, \dots, n. \tag{3.2.5}$$

this is just the equation found when maximizing (3.2.2).

The equation (3.2.4) as well as its coordinate form (3.2.5) are called the ***Galerkin equation***.

The Eigenvalues of the Galerkin Equation

Since PK is a symmetric and positive operator on the finite-dimensional space V_n, there are exactly n eigenvalues

$$\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_n \geq 0.$$

We note that, since we are in a finite-dimensional space,

$$\Lambda_1 = \max_{u \in V_n} R(u) = \max_{u \in V_n} \frac{\langle PKu, u \rangle}{\|u\|^2}, \tag{3.2.6}$$

$$\Lambda_n = \min_{u \in V_n} R(u) = \min_{u \in V_n} \frac{\langle PKu, u \rangle}{\|u\|^2}. \tag{3.2.7}$$

We would like to establish a relationship between these eigenvalues and the first k eigenvalues of K.

Poincaré's Theorem

3.2.3. Poincaré's Theorem.

$$\Lambda_1 \leq \lambda_1, \qquad \dots, \qquad \Lambda_n \leq \lambda_n$$

The proof is very similar to that of the Weyl-Courant Theorem 3.2.2.

Proof.
We already know that $\Lambda_1 \leq \lambda_1$. Now let $k = 2, \dots, n$. Then by Theorem 3.2.1

$$\lambda_k = \max_{u \in M_{k-1}^{\perp}} R(u).$$

We choose a vector $w \in M_{k-1}^{\perp}$ such that $w \neq 0$ and

$$w = d_1 w_1 + \cdots + d_k w_k,$$

where $w_1, \dots, w_k$ are the first k eigenvectors of PK.

Poincaré's Theorem

Proof (continued).

Then $R(w) \leq \lambda_k$ and

$$R(w) = \frac{\langle PKw, w \rangle}{\|w\|^2} = \frac{\sum_{i=1}^{k} \Lambda_i |d_i|^2}{\sum_{i=1}^{k} |d_i|^2} \geq \Lambda_k.$$

This shows $\Lambda_k \leq \lambda_k$. □

The Poincaré Maximin Theorem

3.2.4. Poincaré Maximin Theorem. Let E_n be any n-dimensional subspace of $\mathscr{H}$ and set

$$\mu(E_n) := \min_{u \in E_n} R(u).$$

Then

$$\lambda_n = \max_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \mu(E_n) = \max_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \min_{u \in E_n} R(u).$$

Proof.

From (3.2.7) we have $\mu(E_n) = \Lambda_n$ and by Poincaré's Theorem 3.2.3 we have $\Lambda_n \leq \lambda_n$. Therefore,

$$\max_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \mu(E_n) \leq \lambda_n.$$

However, $\mu(M_n) = \lambda_n$ (why?), so $\max_{\substack{E_n \subset \mathscr{H} \\ \dim E_n = n}} \mu(E_n) = \lambda_n$. □

The Poincaré Maximin Theorem

3.2.5. Remarks.

(i) All of the previous results work for operators that are negative instead of positive (T is negative if $-T$ is positive) if the words "maximum" and "minimum" are interchanged and all the inequalities are reversed.

If an operator is neither negative nor positive, then the original results work for the positive end of the spectrum and the modified results work for the negative end. Generally speaking, if a symmetric operator is simply bounded below or bounded above and the spectrum at the bounded end consists of eigenvalues only, then these eigenvalues can be estimated by using the above results.

(ii) Usually, the approximation Λ_1 to λ_1 is better than that of Λ_2 to λ_2 and so on.

(iii) By increasing k, the approximation improves. In theory, the eigenvalues Λ_i will converge to λ_i as $k \to \infty$, since PK converges to K in norm as $k \to \infty$.

Application to Sturm-Liouville Operators

Recall that regular Sturm-Liouville operators are symmetric and bounded below. Their spectrum consists of an increasing sequence of eigenvalues

$$-\infty < \lambda_1 \leq \lambda_2 \leq \dots$$

so (as per Remark 3.2.5 i)) we can apply the Rayleigh-Ritz procedure. For example, to find an estimate and lower bound for the first (and lowest) eigenvalue, we use trial vectors $v_1, \dots, v_n$ spanning a subspace V_n. Then

$$\lambda_1 \leq \min_{v \in V_n} R(v) = \min_{c_1,\dots,c_n \in \mathbb{F}} R(c_1 v_1 + \dots + c_n v_n). \tag{3.2.8}$$

Of course, now we have to ensure that $V_n \subset \operatorname{dom} L$. This is not an issue for compact operators but becomes relevant for the (unbounded) Sturm-Liouville case.

A Sturm-Liouville Problem

3.2.6. Example. Let us consider the Sturm-Liouville problem

$$Lu = -u'' = \lambda u \quad \text{on } (0,1), \qquad u(0) = 0, \qquad u(1) = 0. \tag{3.2.9}$$

Using Mathematica 10.3, we can find an exact solution:

```
DSolve[{D[y[x], {x, 2}] + λ y[x] == 0, y[0] == 0, y[1] == 0}, y[x], x]
```

$$\left\{\left\{y[x] \to \begin{cases} C[1]\,\text{Sin}\left[x\sqrt{\lambda}\right] & \dot{n} \in \text{Integers} \;\&\&\; \dot{n} \geq 1 \;\&\&\; \lambda == \dot{n}^2\,\pi^2 \\ 0 & \text{True} \end{cases}\right\}\right\}$$

We hence have normed eigenfunctions

```
ψexact[x_, n_] := Assuming[n ∈ Integers, Normalize[Sin[n π x], √(∫_0^1 Abs[#]^2 𝕕x) &]];
ψexact[x, n]
```

$$\sqrt{2}\,\text{Sin}[n\,\pi\,x]$$

Polynomial Approximation to the Eigenfunctions

We define the differential operator L, the Rayleigh quotient R and our trial functions, which will be polynomials of degree n:

```
L := -D[#, {x, 2}] &;
```

$$\texttt{R} := \frac{\int_0^1 \# \; \texttt{L}[\#] \; \mathbb{d}\texttt{x}}{\int_0^1 \#^2 \; \mathbb{d}\texttt{x}} \;\&;$$

$$\texttt{p}[x\_, m\_] := a_0 + \sum_{k=1}^{m} a_k \, x^k$$

We now determine a polynomial of order $n = 2$ that satisfies the boundary conditions (lies in $U = \operatorname{dom} L = \left\{u \in C^2([a,b]) \colon u(0) = u(1) = 0\right\}$):

```
m = 2;
coeff = Solve[{Evaluate[p[0, m]] == 0, p[1, m] == 0}, Table[a_k, {k, 0, m}]]
```

Solve::svars : Equations may not give solutions for all "solve" variables. ≫

$$\{\{a_0 \to 0,\ a_2 \to -a_1\}\}$$

Estimating the Lowest Eigenvalue

This yields a single normalized trial function:

```
v[x_] := p[x, m] /. coeff[[1]]; v[x]
```

$$x\,a_1 - x^2\,a_1$$

```
ψest[x_] := Normalize[Evaluate[v[x] /. a1 → 1], Sqrt[∫_0^1 Abs[#]^2 dx] &];
ψest[x]
```

$$\sqrt{30}\,\left(x - x^2\right)$$

The Rayleigh quotient is

```
R[ψest[x]]
```

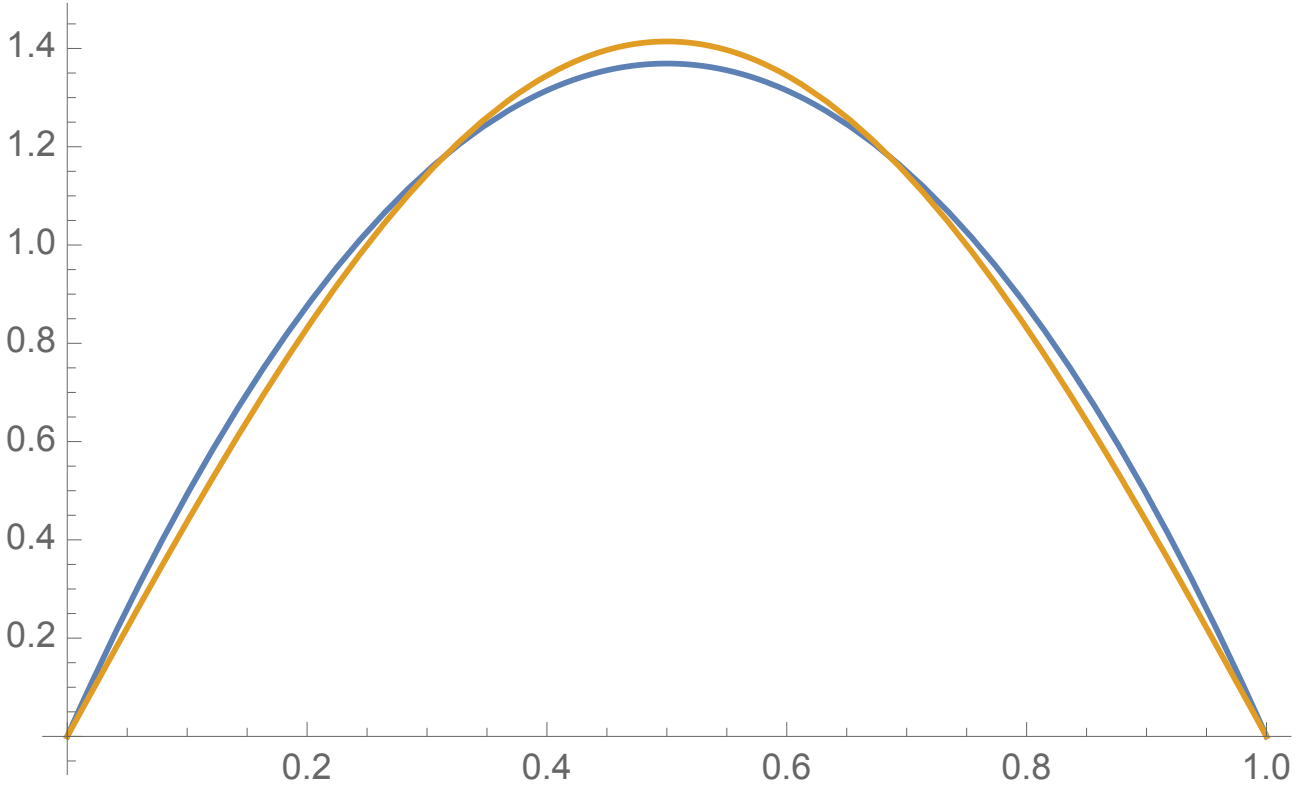
$$10$$

which is a good approximation to the true value $\pi^2 \approx 9.869604$. Since there is only a single trial function, no maximization needs to take place.

Estimating the Lowest Eigenfunction

The approximate eigenfunction is fairly close to the true eigenfunction:

```
Plot[Evaluate[{ψest[x], ψexact[x, 1]}], {x, 0, 1}]
```



Improving the Estimates for the Lowest Eigenvalue

We can take a fourth-order polynomial to obtain an improved estimate:

```
m = 4;
coeff = Solve[{Evaluate[p[0, m]] == 0, p[1, m] == 0}, Table[a_k, {k, 0, m}]];
v[x_] := p[x, m] /. coeff[[1]];
v[x]
```

Solve::svars : Equations may not give solutions for all "solve" variables. ≫

$$x\,a_1 + x^2\,a_2 + x^4\,(-a_1 - a_2 - a_3) + x^3\,a_3$$

The Rayleigh quotient now has three parameters:

```
R[v[x]]
```

$$\frac{\frac{9\,a_1^2}{7} + \frac{48\,a_1\,a_2}{35} + \frac{44\,a_2^2}{105} + \frac{4\,a_1\,a_3}{7} + \frac{13\,a_2\,a_3}{35} + \frac{3\,a_3^2}{35}}{\frac{a_1^2}{9} + \frac{13\,a_1\,a_2}{126} + \frac{8\,a_2^2}{315} + \frac{7\,a_1\,a_3}{180} + \frac{5\,a_2\,a_3}{252} + \frac{a_3^2}{252}}$$

Improving the Estimates for the Lowest Eigenvalue

We find the minimum of the Rayleigh quotient:

```
min = FindMinimum[R[v[x]], Table[a_k, {k, 0, m}]]
```

$$\{9.86975,\ \{a_0 \to 1.,\ a_1 \to 2.39859,\ a_2 \to 0.319349,\ a_3 \to -5.43587,\ a_4 \to 1.\}\}$$

Note that this minimum is a very good approximation to $\pi^2 \approx 9.869604$. The estimated eigenfunction is :

```
ψ_est[x_] := Normalize[Evaluate[v[x] /. min[[2]]], Sqrt[∫_0^1 Abs[#]^2 dx] &];
```

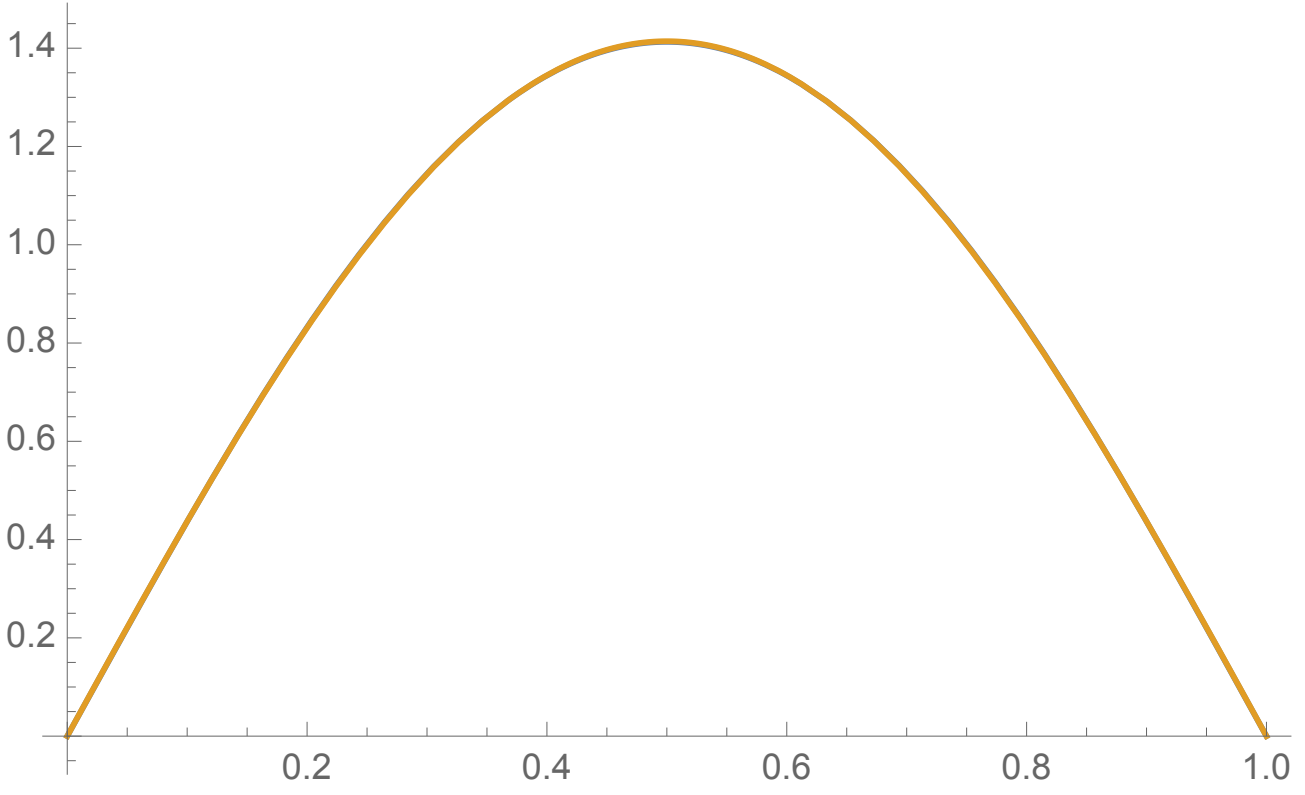
```
ψ_est[x]
```

$$1.83608\left(2.39859\,x + 0.319349\,x^2 - 5.43587\,x^3 + 2.71794\,x^4\right)$$

Estimating the Lowest Eigenfunction

There is no immediately visible difference between the approximate and the true eigenfunction:

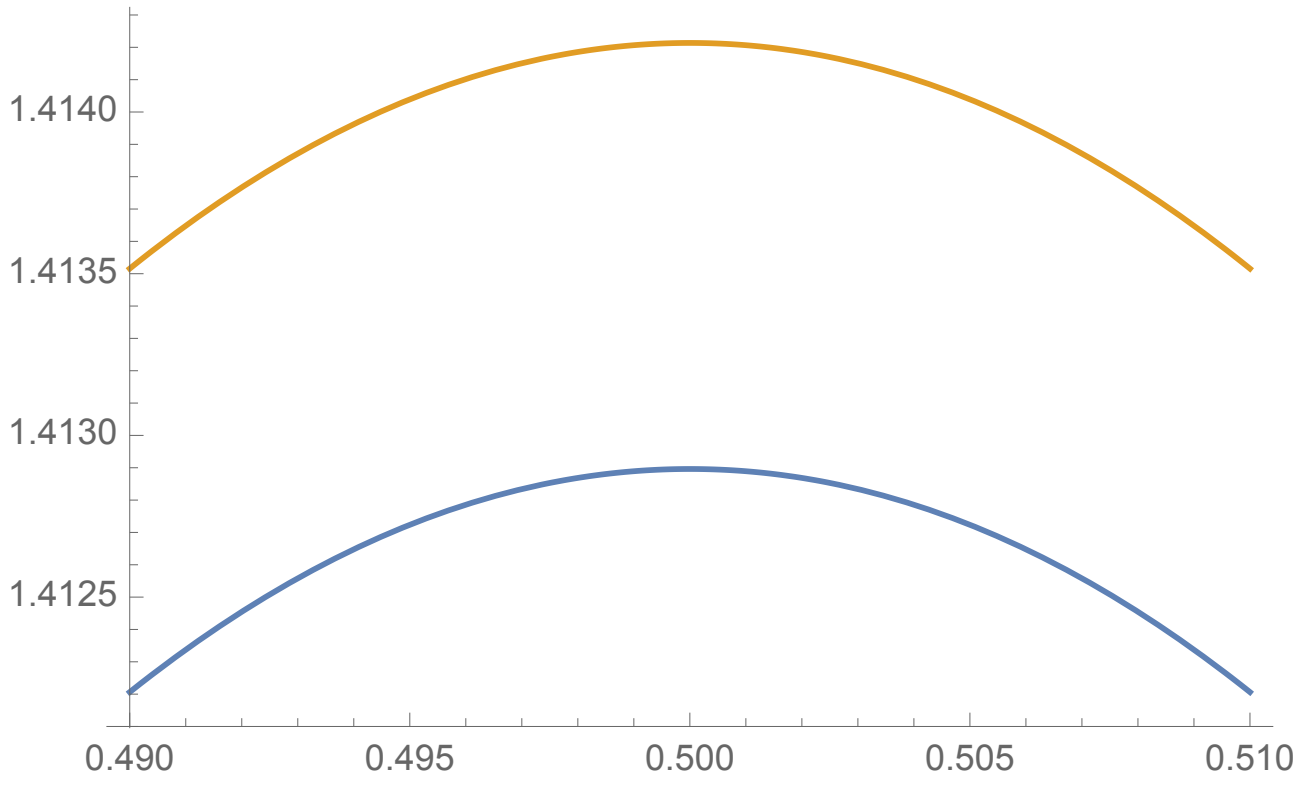
```
Plot[Evaluate[{ψest[x], ψexact[x, 1]}], {x, 0, 1}]
```



Estimating the Lowest Eigenfunction

Magnification shows the actual difference:

```
Plot[Evaluate[{ψest[x], ψexact[x, 1]}], {x, 0.49, 0.51}]
```



Finding Estimates for the Next Eigenvalues

In the calculation so far, we have effectively found Λ_1, once for a space of trial functions $U \cap \mathscr{P}_2$ (the domain of L intersected with the polynomials of degree not larger than 2) and once for $U \cap \mathscr{P}_4$. To find an estimate for the second eigenvalue λ_2, we need to find Λ_2 for some suitable space.

The space $U \cap \mathscr{P}_2$ is one-dimensional, so the operator only has the eigenvalue Λ_1 there and we can not use it to find an approximation to λ_2. However, $U \cap \mathscr{P}_4$ is three-dimensional and we can find two more eigenvalues Λ_2 and Λ_3 with their corresponding eigenfunctions.

We find these eigenvalues and -functions by restricting to the orthogonal complement of the previously determined eigenfunctions for Λ_1 (and Λ_2).

Finding Estimates for the Second Eigenvalue

```
n = 4;
coeff = Solve[{Evaluate[p[0, n]] == 0, p[1, n] == 0, ∫_0^1 p[x, n] ψtrial[x, 1] dx == 0},
  Table[ak, {k, 0, n}]]
```

Solve::svars : Equations may not give solutions for all "solve" variables. ≫

$$\{\{a_0 \to 0.,\ a_3 \to 0. - 6.27867\, a_1 - 2.75955\, a_2,\ a_4 \to 0. + 5.27867\, a_1 + 1.75955\, a_2\}\}$$

```
ψtrial[x_] := p[x, n] /. coeff
R[ψtrial[x]]
```

$$\left\{\frac{1.07691\, a_1^2 + 0.432691\, a_1\, a_2 + 0.0467966\, a_2^2}{0.0233759\, a_1^2 + 0.00879238\, a_1\, a_2 + 0.000862572\, a_2^2}\right\}$$

```
min = FindMinimum[R[ψtrial[x]], {a1, a2}]
```

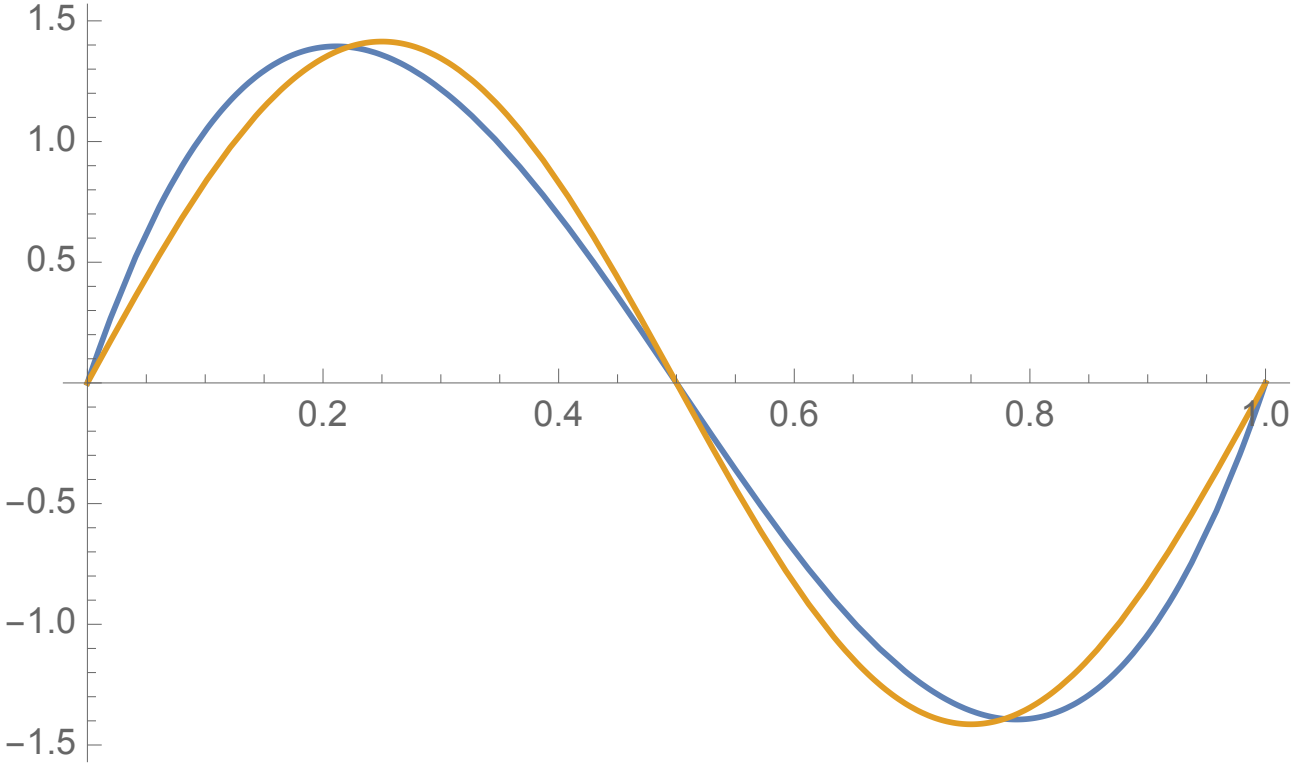
$$\{42.,\ \{a_1 \to 2.23743,\ a_2 \to -6.71231\}\}$$

```
ψtrial[x_, 2] := Normalize[0.` + 2.2374334666115634` x - 6.712314638301565` x^2
    + 4.474886320874187` x^3 - 5.1491841865924926`*^-6 x^4, Sqrt[∫_0^1 #^2 dx] &]
```

Estimating the Second Eigenfunction

Note that the upper bound of 42 is not very close to the true eigenvalue $4\pi^2 \approx 39.48$. Also, the approximation to the second eigenfunction is not as good as the approximation of the first eigenfunction:

```
Plot[Evaluate[{ψtrial[x, 2], ψexact[x, 2]}], {x, 0, 1}]
```



Finding Estimates for the Third Eigenvalue

```
n = 4;
coeff = Solve[{Evaluate[p[0, n]] == 0, p[1, n] == 0, ∫_0^1 p[x, n] ψ_trial[x, 1] dx == 0,
    ∫_0^1 p[x, n] ψ_trial[x, 2] dx == 0}, Table[a_k, {k, 0, n}]]
```

Solve::svars : Equations may not give solutions for all "solve" variables. ≫

$$\{\{a_0 \to 0.,\ a_2 \to 0. - 5.63314\, a_1,\ a_3 \to 0. + 9.26627\, a_1,\ a_4 \to 0. - 4.63314\, a_1\}\}$$

```
ψ_trial[x_] := p[x, n] /. coeff
R[ψ_trial[x]]
```

$$\left\{\frac{0. + 0.124457\, a_1^2}{0. + 0.00121861\, a_1^2}\right\}$$

```
min = FindMinimum[R[ψ_trial[x]], {a_1, a_2}]
```

$$\{102.13,\ \{a_1 \to 1.,\ a_2 \to 1.\}\}$$

```
f[x_] := p[x, n] /. coeff /. min[[2]]
f[x]
```

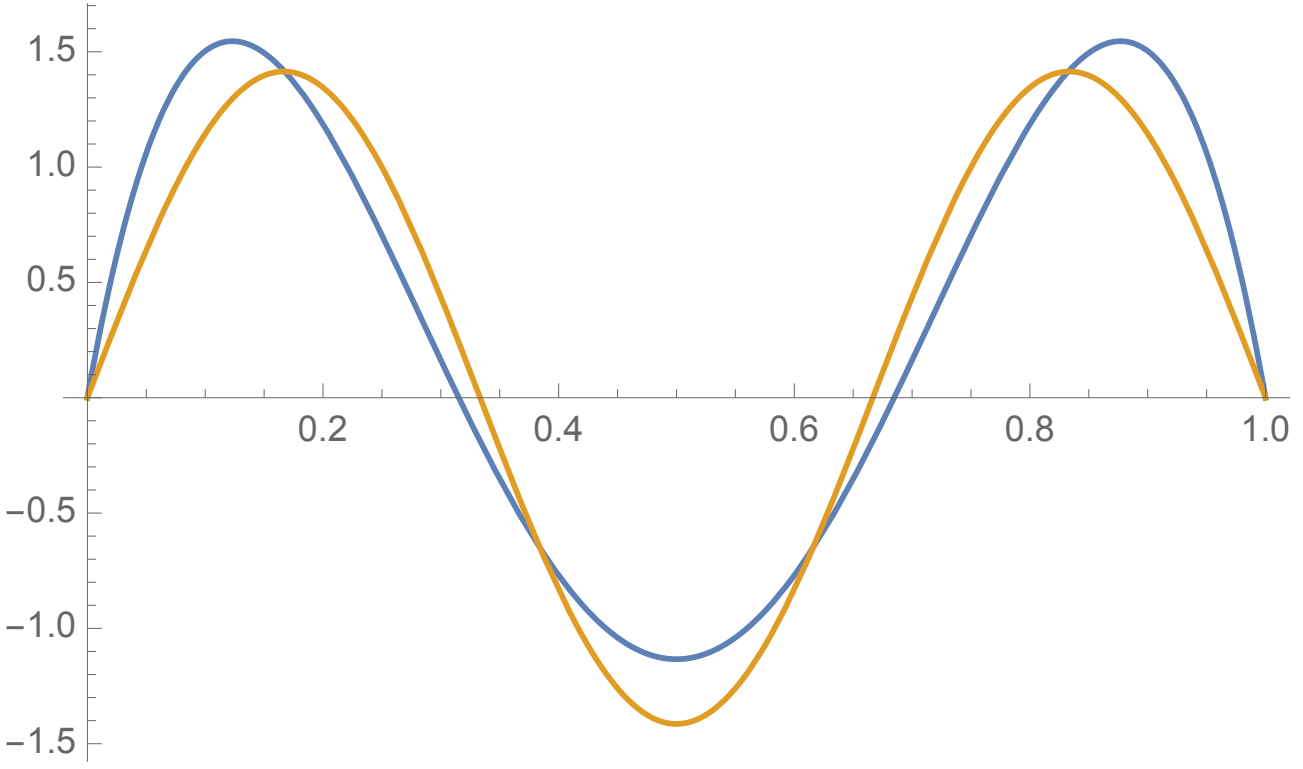
$$\left\{0. + 1.\, x - 5.63314\, x^2 + 9.26627\, x^3 - 4.63314\, x^4\right\}$$

```
ψ_trial[x_, 3] := Normalize[0.` + 1.` x - 5.63313597332083` x^2 + 9.26627194658505` x^3
    - 4.63313597326422` x^4, Sqrt[∫_0^1 #^2 dx] &]
```

Estimating the Third Eigenfunction

The bound of 102.13 on $9\pi^2 \approx 88.83$ is again not very good. The approximation to the third eigenfunction is qualitatively correct but quantitatively poor:

```
Plot[Evaluate[{ψtrial[x, 3], ψexact[x, 3]}], {x, 0, 1}]
```



Polar Representation of Complex Numbers and Operators

The goal of this section is to find a ***polar representation*** of bounded linear operators that is analogous to the polar representation of complex numbers.

A complex number $z \in \mathbb{C}$ mar be expressed as $z = |z|e^{i \arg(z)}$. Now fix $z \in \mathbb{C}$ and consider the linear map

$$T_z : \mathbb{C} \to \mathbb{C}, \qquad w \mapsto zw = e^{i \arg(z)}|z|w$$

Then we can write T_z as a composition of two maps:

- multiplication with $|z| > 0$;
- multiplication with $e^{i \arg(z)}$.

The first is a positive operator in the sense that

$$\langle w, |z|w \rangle = |z|\overline{w}w = |z| \cdot |w|^2 > 0 \qquad \text{if } w \neq 0.$$

Polar Representation of Complex Numbers and Operators

The second is an ***isometry***, meaning that

$$|e^{i\arg(z)}w| = |e^{i\arg(z)}| \cdot |w| = |w|,$$

i.e., the length of w remains unchanged. In this section we will analogously define and prove the decomposition

$$A = U|A|$$

for any bounded linear operator A, where $|A|$ is a positive operator and U is a (partial) isometry.

Positive Operators

Recall that a bounded linear operator B on a Hilbert space $\mathscr{H}$ is said to be ***positive*** if

$$\langle x, Bx\rangle \geq 0 \qquad \text{for all } x \in \mathscr{H}$$

In this case, we write $B \geq 0$. We write $A \geq B$ if $A - B \geq 0$.

3.3.1. Remark. If A is a bounded linear operator, then A^*A is self-adjoint and positive, since

$$(A^*A)^* = A^*(A^*)^* = A^*A$$

and

$$\langle x, A^*Ax\rangle = \langle Ax, Ax\rangle = \|Ax\|^2 \geq 0.$$

Commutation of Operators

We will now work to define the ***modulus*** of an operator in analogy to the complex modulus. To do this, we need to define the square root of positive operators. In preparation, we formalize a preliminary concept that is essential to the calculus of operators:

3.3.2. Definition. Let A and B be two linear operators on a vector space V. The ***commutator*** of A and B is defined by

$$[A, B] := AB - BA$$

with domain $\mathrm{dom}(AB) \cap \mathrm{dom}(BA)$. The operators A and B are said to ***commute*** if

$$AB = BA,$$

i.e., if $A(Bv) = B(Av)$ for all $v \in \mathrm{dom}(AB) \cap \mathrm{dom}(BA)$.

The Binomial Series

In the proof of the Weierstraß Approximation Theorem 1.2.14, we have already encountered the binomial series (1.2.4),

$$\sqrt{1-z} = 1 - \sum_{n=1}^{\infty} \frac{1}{2^{2n-1}} \binom{2n-2}{n-1} \frac{z^n}{n}. \tag{3.3.1}$$

We have also proved in (1.2.5) that

$$\sum_{n=1}^{\infty} \frac{1}{2^{2n-1} n} \binom{2n-2}{n-1} \leq 1 \tag{3.3.2}$$

and, in particular, the series (3.3.1) converges absolutely when $|z| \leq 1$.

The Square Root Lemma

3.3.3. Square Root Lemma. Let A be a linear, bounded, self-adjoint and positive operator on $\mathscr{H}$. Then there exists a unique linear, self-adjoint operator B such that

$$B \geq 0 \qquad \text{and} \qquad B^2 = A.$$

Furthermore, B commutes with any bounded operator which commutes with A. We write $B =: \sqrt{A}$.

The Square Root Lemma

Proof.

We may suppose that $\|A\| \leq 1$ (why is this sufficient?). Then

$$0 \leq \langle v, Av \rangle \leq \|v\|^2 \qquad \text{for any } v \in \mathscr{H}.$$

This estimate together with

$$\langle v, (I - A)v \rangle = \|v\|^2 - \langle v, Av \rangle$$

implies that

$$0 \leq \langle v, (I - A)v \rangle \leq \|v\|^2.$$

Then by Theorem 2.4.11,

$$0 \leq \|I - A\| = \sup_{v \in \mathscr{H}} \frac{|\langle v, (I - A)v \rangle|}{\|v\|^2} \leq 1 \tag{3.3.3}$$

The Square Root Lemma

Proof (continued).

Therefore, the series of numbers

$$1 - \sum_{n=1}^{\infty} \frac{1}{n2^{2n-1}} \binom{2n-2}{n-1} \|I - A\|^n$$

converges. From this we see that the series of operators in $\mathscr{L}(\mathscr{H}, \mathscr{H})$

$$I - \sum_{n=1}^{\infty} \frac{1}{n2^{2n-1}} \binom{2n-2}{n-1} (I - A)^n$$

converges absolutely and by Lemma 1.5.3 then converges to an operator $B \in \mathscr{L}(\mathscr{H}, \mathscr{H})$.

The Square Root Lemma

Proof (continued).

Since

$$\left(1-\sum_{n=1}^{\infty}\frac{1}{2^{2n-1}}\binom{2n-2}{n-1}\frac{(1-z)^n}{n}\right)^2=z$$

for any $z\in\mathbb{C}$ with $|z|\leq 1$ and the series converges absolutely when $1-z$ is replaced by the operator $I-A$ and z is replaced by A, this shows that $B^2=A$.

Next, $B=B^*$ since $(I-A)^*=I-A$ and

$$B^*=I-\sum_{n=1}^{\infty}\frac{1}{n2^{2n-1}}\binom{2n-2}{n-1}[(I-A)^n]^*.$$

The Square Root Lemma

Proof (continued).

We next prove that $B \geq 0$. From (3.3.3),

$$0 \leq \langle v, (I-A)^n v \rangle \leq \|(I-A)^n\| \cdot \|v\|^2 \leq \|I-A\|^n \cdot \|v\|^2 \leq \|v\|^2.$$

and so, for $v \in \mathscr{H}$

$$\begin{aligned}
\langle v, Bv \rangle &= \langle v, v \rangle - \sum_{n=1}^{\infty} \frac{1}{n 2^{2n-1}} \binom{2n-2}{n-1} \langle v, (I-A)^n v \rangle \\
&\geq \|v\|^2 \left[1 - \sum_{n=1}^{\infty} \frac{1}{n 2^{2n-1}} \binom{2n-2}{n-1} \right] \\
&\geq 0
\end{aligned}$$

where we have used (3.3.2).

The Square Root Lemma

Proof (continued).

Since the series for B converges absolutely, we can take any operator C such that $AC = CA$ and find

$$\begin{aligned} CB &= C\left(I - \sum_{n=1}^{\infty} \frac{1}{n2^{2n-1}} \binom{2n-2}{n-1} (I-A)^n\right) \\ &= C - \sum_{n=1}^{\infty} \frac{1}{n2^{2n-1}} \binom{2n-2}{n-1} C(I-A)^n \\ &= C - \sum_{n=1}^{\infty} \frac{1}{n2^{2n-1}} \binom{2n-2}{n-1} (I-A)^n C \\ &= BC. \end{aligned}$$

This proves that B commutes with any operator that commutes with A.

The Square Root Lemma

Proof (continued).

Finally, we prove the uniqueness of B: Suppose there exists some other self-adjoint operator B' such that $B' \geq 0$ and $B'^2 = A$. Then

$$B'A = B'^3 = AB'$$

and so B' commutes with B. This implies

$$\begin{aligned}(B - B')B(B - B') + (B - B')B'(B - B') &= (B^2 - B'^2)(B - B') \\ &= (A - A)(B - B') \\ &= 0.\end{aligned}$$

Since both $(B - B')B(B - B')$ and $(B - B')B'(B - B')$ are positive (why?) this implies that both operators vanish (why?). Then

$$(B - B')^3 = (B - B')B(B - B') - (B - B')B'(B - B') = 0.$$

The Square Root Lemma

Proof (continued).

This implies that

$$\|(B - B')^4\| \leq \|B - B'\| \cdot \|(B - B')^3\| = 0. \tag{3.3.4}$$

Using Theorem 2.4.11 and noting that $B - B'$ is self-adjoint, we see that

$$\begin{aligned}\|(B - B')^2\| &= \sup_{v \in \mathscr{H}} \frac{|\langle v, (B - B')^2 v\rangle|}{\|v\|^2} = \sup_{v \in \mathscr{H}} \frac{\|(B - B')v\|^2}{\|v\|^2} \\ &= \|B - B'\|^2.\end{aligned}$$

The same argument, applied once more, shows that

$$\|(B - B')^4\| = \|B - B'\|^4.$$

With (3.3.4) this implies $\|B - B'\| = 0$ and we conclude $B = B'$. □

The Modulus of an Operator

Hence, for any bounded operator A on $\mathscr{H}$ the modulus

$$|A| := \sqrt{A^*A}$$

is a well-defined, bounded, positive and self-adjoint linear operator on $\mathscr{H}$.

Note that

- $|\lambda A| = |\lambda| \cdot |A|$ for $\lambda \in \mathbb{C}$,
- but in general $|A| \neq |A^*|$,
- in general, $|A + B| \not\leq |A| + |B|$.

However, if $A = A^*$, then

$$|A|^2 = A^*A = A^2.$$

The Modulus of a Compact Operator

If K is a compact operator, the modulus $|K|$ can be calculated using the spectral representation (2.6.5) for K^*K. This is based on the fact that if K is compact, then so is K^*K, since K^* will be a bounded linear operator and the composition of a bounded with a compact operator is compact. Since K^*K is self-adjoint and positive, we can apply the spectral theorem to obtain:

3.3.4. Lemma. Let K be a compact operator on a Hilbert space $\mathscr{H}$ and denote by $\lambda_n > 0$ the eigenvalues and by v_n the eigenvectors of K^*K. Then $|K|$ is compact and

$$|K| = \sum_{n \in I} \sqrt{\lambda_n} \langle v_n, \cdot \rangle v_n. \tag{3.3.5}$$

3.3.5. Corollary. The representation (3.3.5) implies that the eigenvalues and -vectors of $|K|$ are given by $\sqrt{\lambda_n}$ and v_n, respectively.

The Modulus of a Compact Operator

Proof.

We first show (3.3.5). Let

$$T := \sum_{n \in I} \sqrt{\lambda_n} \langle v_n, \cdot \rangle v_n.$$

It is easy to verify that

- $T^2 = K^* K$,
- $T = T^*$,
- $T \geq 0$,

so T is the unique square root of K^*K, i.e., $T = |K|$. Furthermore, $|K|$ is compact, since it is the norm limit of the finite-rank operators

$$K_N := \sum_{n \leq N} \sqrt{\lambda_n} \langle v_n, \cdot \rangle v_n.$$

with $\|K - K_N\| \leq \sqrt{\lambda_{N+1}}$ and $\sqrt{\lambda_N} \to 0$ as $N \to \infty$. $\square$

Polar Decomposition

In analogy to the polar representation of complex numbers,

$$z = |z| \cdot e^{i \arg z}, \qquad (e^{i \arg z})^{-1} = \overline{e^{i \arg z}},$$

we would like to write

$$A = U|A| \tag{3.3.6}$$

for a suitable operator U. However, while

$$e^{i \arg z} \cdot \overline{e^{i \arg z}} = 1,$$

we may not be able to achieve

$$U^*U = UU^* = I$$

since U or U^* may have a non-trivial kernel.

Partial Isometries

3.3.6. Example. Let $R\colon \ell^2 \to \ell^2$ be the right-shift operator. Then $R^* = L$ (the left-shift operator) and $R^*R = I$, so $|R| = I$. This means that we would have to take $U = R$ in (3.3.6). But then

$$R^*R = I, \qquad RR^* = I - \langle e_1, \cdot \rangle e_1$$

where $e_1 = (1, 0, 0, \ldots)$.

3.3.7. Definition. An operator U on $\mathscr{H}$ is said to be an ***isometry*** if $\|Ux\| = \|x\|$ for all $x \in \mathscr{H}$.

The operator U is said to be a ***partial isometry*** if it is an isometry when restricted to the closed subspace $(\ker U)^\perp$.

The Polar Decomposition

The following theorem then allows us to define the polar decomposition for bounded linear operators:

3.3.8. Theorem. Let A be a bounded linear operator on a Hilbert space $\mathscr{H}$. Then there exists a partial isometry U such that

$$A = U|A| \tag{3.3.7}$$

The partial isometry is uniquely determined by requiring $\ker U = \ker A$. Moreover, $\operatorname{ran} U = \overline{\operatorname{ran} A}$.

Proof.

In order to achieve (3.3.7), we need to define a partial isometry

$$U\colon\ \operatorname{ran}|A| \to \operatorname{ran} A$$

which is most obviously done by setting

$$Uw = U(|A|v) = Av \qquad \text{for any } w = |A|v \in \operatorname{ran}|A|. \tag{3.3.8}$$

The Polar Decomposition

Proof (continued).

However, it could be that $w = |A|v_1 = |A|v_2$ with $v_1 \neq v_2$. Then it is not clear if U is well-defined, because the action of U might depend on which v is used. We note that

$$\||A|v\|^2 = \langle |A|v, |A|v\rangle = \langle v, |A|^2 v\rangle = \langle v, A^*Av\rangle = \|Av\|^2 \qquad (3.3.9)$$

and hence

$$\||A|v_1 - |A|v_2\|^2 = \||A|(v_1 - v_2)\|^2 = \|A(v_1 - v_2)\|^2 = \|Av_1 - Av_2\|^2$$

so $Av_1 = Av_2$ if and only if $|A|v_1 = |A|v_2$. This shows that U is well-defined.

From (3.3.9) we also see that $\|Uw\| = \|w\|$ for all $w \in \text{ran}|A|$.

The Polar Decomposition

Proof (continued).

Our goal is now to extend U (currently defined only on $\operatorname{ran}|A|$) to an operator on all of $\mathscr{H}$ such that

$$\operatorname{ran} U = \overline{\operatorname{ran} A} \qquad \text{and} \qquad \ker u = \ker A.$$

First, we use the B.L.T. Theorem 2.1.10, to extend U to a map $\overline{\operatorname{ran}|A|}$ to $\overline{\operatorname{ran} A}$. (Explain why then $\operatorname{ran} U = \overline{\operatorname{ran} A}$.).

Then, we simply define $Ux = 0$ for all $x \in (\operatorname{ran}|A|)^{\perp}$. Since

$$\mathscr{H} = \operatorname{ran}|A| \oplus (\operatorname{ran}|A|)^{\perp},$$

this defines U on $\mathscr{H}$.

The Polar Decomposition

Proof (continued).

We next prove that $\ker U = \ker A$. By our construction, $\ker U \supset (\text{ran}|A|)^{\perp}$. But does the kernel contain any other elements?

If $w = |A|v \in \text{ran}|A|$, then

$$Uw = 0 \quad \Leftrightarrow \quad U(|A|v) = 0 \quad \Leftrightarrow \quad Av = 0 \quad \Leftrightarrow \quad |A|v = w = 0$$

where we have used (3.3.9). Thus, there are no other elements in the kernel and by Lemma 2.3.5 and (3.3.9),

$$\ker U = (\text{ran}|A|)^{\perp} = \ker|A| = \ker A.$$

It follows that U has the desired properties. The proof of uniqueness is left to the reader. □

Sturm-Liouville Boundary Value Problems

The Rayleigh-Ritz Method

Positive Operators and the Polar Decomposition

The Singular Value Decomposition for Compact Operators and Matrices

Singular Value Decomposition

For compact, self-adjoint operators K on a Hilbert space $\mathscr{H}$, the Spectral theorem 2.6.6 allowed us to obtain the representation (2.6.5) in terms of their real eigenvalues and orthonormal eigenvectors,

$$Ku = \sum_n \lambda_n \langle e_n, u \rangle e_n \qquad \text{for all } u \in \mathscr{H}.$$

(This representation is equivalent to the diagonalization of square, self-adjoint matrices.)

In this section, we will obtain a similar representation that

- is valid even for compact operators that are not self-adjoint;
- has the same form, but the eigenvalues λ_n are replaced with strictly positive numbers σ_n.

This representation will also be useful for obtaining the polar decomposition of a compact operator.

Singular Value Decomposition

3.4.1. Theorem. Let K be a compact operator on a Hilbert space $\mathscr{H}$. Then there exist families of orthonormal vectors $\{v_n\}_{n\in I}$ and $\{u_n\}_{n\in I}$, $I \subset \mathbb{N}$, and strictly positive real numbers $\{\sigma_n\}_{n\in I}$, such that

$$K = \sum_{n\in I} \sigma_n \langle v_n, \cdot \rangle u_n.$$

The numbers σ_n are called the ***singular values*** of K, the vectors v_n the ***right-singular vectors*** and the vectors u_n the ***left-singular vectors***.

Singular Value Decomposition

Proof.

Since K is compact, so is K^*K (why?). Thus, K^*K is compact, self-adjoint and positive by Remark 3.3.1. By the Spectral Theorem for compact operators, there exists an orthonormal system (not necessarily a basis) of eigenvectors $\{v_n\}_{n\in I}$ such that

$$K^*Kv_n = \lambda_n v_n \qquad \text{with } \lambda_n \neq 0,\ n \in I,$$

and

$$K^*Kx = 0 \qquad \text{for } x \in (\text{span}\{v_n\})^{\perp}.$$

Since $K^*K \geq 0$, all $\lambda_n > 0$.

Singular Value Decomposition

Proof (continued).

Define

$$\sigma_n := \sqrt{\lambda_n}$$

and set

$$u_n := \frac{1}{\sigma_n} K v_n.$$

Then

$$\langle u_i, u_j \rangle = \frac{1}{\sigma_i \sigma_j} \langle K v_i, K v_j \rangle = \frac{1}{\sigma_i \sigma_j} \langle v_i, K^* K v_j \rangle = \frac{\sigma_j}{\sigma_i} \langle v_i, v_j \rangle = \delta_{ij}$$

so $\{u_n\}_{n \in I}$ is an orthonormal system.

Singular Value Decomposition

Proof (continued).

We next show that $\ker K^*K = \ker K$. Note that if $K^*Kx = 0$, then

$$\|Kx\|^2 = \langle Kx, Kx\rangle = \langle K^*Kx, x\rangle = 0,$$

so $x \in \ker K$. This gives $\ker K^*K \subset \ker K$. Since $\ker K \subset \ker K^*K$, the two kernels are equal.

We may now write

$$\begin{aligned}\mathscr{H} &= \operatorname{span}\{v_n\} \oplus (\operatorname{span}\{v_n\})^\perp \\ &= \operatorname{span}\{v_n\} \oplus \ker K^*K \\ &= \operatorname{span}\{v_n\} \oplus \ker K.\end{aligned}$$

Hence, each $x \in \mathscr{H}$ may be expressed in the form

$$x = \sum_{n\in I}\langle v_n, x\rangle v_n + w, \qquad \text{where } w \in \ker K.$$

Singular Value Decomposition

Proof (continued).
We then have

$$Kx = \sum_{n \in I} \langle v_n, x \rangle K v_n + 0 = \sum_{n \in I} \sigma_n \langle v_n, x \rangle u_n,$$

which is the desired representation. □

3.4.2. Remark. The singular values are just the square roots of the eigenvalues of K^*K. From Lemma 3.3.4 we know that the singular values are equal to the eigenvalues of $\sqrt{K^*K} = |K|$. This gives a connection to the polar decomposition, as we will see.

Relationship to the Spectral Representation

While the singular value decomposition is useful as an alternative to the spectral decomposition for non-selfadjoint, compact operators, it is worth mentioning the following relationship between the two decompositions.

3.4.3. Lemma. Let K be a self-adjoint, compact operator with spectral representation

$$K = \sum_{n \in I} \lambda_n \langle e_n, \cdot \rangle e_n.$$

Then the corresponding singular value decomposition of K is

$$K = \sum_{n \in I} |\lambda_n| \langle e_n, \cdot \rangle u_n \qquad u_n = \frac{\lambda_n}{|\lambda_n|} e_n. \tag{3.4.1}$$

Relationship to the Spectral Decomposition

Proof.

We have

$$K^*K = K^2 = \sum_{n \in I} \lambda_n^2 \langle e_n, \cdot \rangle e_n$$

This implies that the eigenvectors of K^*K are just the eigenvectors e_n of K and the eigenvalues of K^*K are given by λ_n^2. Then $\sigma_n = |\lambda_n|$ and $u_n = \frac{1}{\sigma_n} K e_n = \frac{\lambda_n}{|\lambda_n|} e_n$, yielding the representation (3.4.1). □

It is worth noting the following result that we used in the proof:

3.4.4. Remark. Let K be a self-adjoint, compact operator. Then K^2 has the same eigenvectors as K and the eigenvalues of K^2 are precisely the squares of the eigenvalues of K.

Relationship to the Polar Decomposition

Let K be a compact operator. Then from Lemma 3.3.4 we know that

$$|K| = \sum_{n \in I} \sigma_n \langle v_n, \cdot \rangle v_n \tag{3.4.2}$$

where the v_n are the eigenvectors of K^*K and σ_n the singular values of K.

Define further $u_n := \frac{1}{\sigma_n} K v_n$ and set

$$U = \sum_{n \in I} \langle v_n, \cdot \rangle u_n. \tag{3.4.3}$$

Then for $x \in \operatorname{span}\{v_n\}$ we have

$$\|Ux\|^2 = \sum_{n \in I} |\langle v_n, x \rangle|^2 = \|x\|^2$$

and

$$\ker U = \operatorname{span}\{v_n\}^\perp = \ker K.$$

Hence, U is a partial isometry with the same kernel as K.

Relationship to the Polar Decomposition

Therefore, given σ_n, u_n, v_n for $n \in I$ in the Singular Value Decomposition (Theorem 3.4.1), we can construct the polar decomposition

$$K = U|K|$$

by defining U and $|K|$ by (3.4.3) and (3.4.2), respectively.

Generalizing the Singular Value Decomposition

In principle, the singular value decomposition can be defined for linear operators

$$K : \mathscr{H}_1 \to \mathscr{H}_2,$$

where $\mathscr{H}_1$ and $\mathscr{H}_2$ are distinct Hilbert spaces. A suitably-defined adjoint would then be a map $K^* : \mathscr{H}_2 \to \mathscr{H}_1$ and we would have

$$K^*K : \mathscr{H}_1 \to \mathscr{H}_1,$$

allowing us to apply the spectral representation of K^*K as before.

Instead of developing this general theory (including a suitable generalization of the adjoint) we will discuss only the case of matrices

$$A : \mathbb{F}^n \to \mathbb{F}^m, \qquad A \in \mathrm{Mat}(m \times n; \mathbb{F}).$$

where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Singular Value Decomposition for Matrices

For $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ we simply take $A^* = \overline{A^T}$. Then

$$A^*A \in \mathrm{Mat}(n \times n; \mathbb{F}) \qquad \text{and} \qquad AA^* \in \mathrm{Mat}(m \times m; \mathbb{F})$$

It is clear that both A^*A and AA^* are square, self-adjoint and positive and we can calculate their square roots, giving $|A|$ and $|A^*|$. (Note that $|A|$ does not have the same size as A!)

3.4.5. Lemma. Let $\lambda > 0$ be an eigenvalue of A^*A with eigenvector $v \in \mathbb{R}^n$. Then λ is also an eigenvalue of AA^* with eigenvector $Av \in \mathbb{R}^m$.

Proof.
Suppose that $A^*Av = \lambda v \neq 0$. Then $Av \neq 0$ and

$$(AA^*)Av = A(A^*Av) = A\lambda v = \lambda \cdot Av. \qquad \square$$

Singular Values and Left- and Right-Singular Vectors

3.4.6. Corollary. If $A \in \mathrm{Mat}(m \times n; \mathbb{F})$, then A^*A and AA^* have at most $\min(m, n)$ non-zero eigenvalues. These eigenvalues must be strictly positive.

We can now make essentially the same definitions as before, with a few additional comments.

We assume that $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ and that $\lambda_1, \ldots, \lambda_r > 0$, $r \leq \min(m, n)$, are the strictly positive eigenvalues of A^*A.

- The numbers $\lambda_1, \ldots, \lambda_r$ are also the non-zero eigenvalues of AA^*. The singular values of A and A^* are both given by $\sigma_i := \sqrt{\lambda_i}$, $i = 1, \ldots, r$.
- The orthonormal eigenvectors for the λ_i are the right-singular vectors $v_i \in \mathbb{R}^n$.
- The left-singular vectors are $u_i = \frac{1}{\sigma_i} A v_i$, where $u_i \in \mathbb{R}^m$.

As before, the sets $\{v_i\}_{i=1}^r$ and $\{u_i\}_{i=1}^r$ are orthonormal systems in their respective spaces.

Singular Vectors

3.4.7. Remark. It is not difficult to see that $\sigma > 0$ is a singular value and two normalized vectors $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ are right- and left-singular vectors for $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ if and only if

$$Av = \sigma u, \qquad A^* u = \sigma v$$

This gives a clear sense of the way in which singular values are generalizations of eigenvalues.

The matrices

$$U_r := (u_1, \dots, u_r) \in \mathrm{Mat}(m \times r; \mathbb{F}),$$
$$V_r := (v_1, \dots, v_r) \in \mathrm{Mat}(n \times r; \mathbb{F}),$$

are in general partial isometries, as can be easily checked. If $r < \min(m, n)$, they will both have a non-trivial kernel.

Orthogonality of Singular Vectors

Since $Av_i = \sigma_i u_i$ and the $u_1, \ldots, u_r$ are orthonormal,

$$\begin{aligned} U_r^* A V_r &= \begin{pmatrix} u_1^* \\ \vdots \\ u_r^* \end{pmatrix} A(v_1, \ldots, v_r) = \begin{pmatrix} u_1^* \\ \vdots \\ u_r^* \end{pmatrix} (Av_1, \ldots, Av_r) \\ &= \begin{pmatrix} \langle u_1, Av_1 \rangle & \ldots & \langle u_1, Av_r \rangle \\ \vdots & & \vdots \\ \langle u_r, Av_1 \rangle & \ldots & \langle u_r, Av_r \rangle \end{pmatrix} \\ &= \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \\ &=: \Sigma_r. \end{aligned}$$

Using the usual matrix algebra, we can rewrite this in the form

$$A = U_r \Sigma_r V_r^*$$

The Compact Singular Value Decomposition

We have hence proved the following result:

3.4.8. Compact Singular Value Decomposition. Let $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ and let $\sigma_1, \dots, \sigma_r > 0$, $r \leq \min(m, n)$, be the singular values of A. Then there exist partial isometries U_r, V_r such that

$$A = U_r \Sigma_r V_r^*$$

where $\Sigma_r := \mathrm{diag}(\sigma_1, \dots, \sigma_r)$.

3.4.9. Remark. One usually orders the left- and right-singular vectors in such a way that the singular values of A are decreasing, i.e.,

$$\Sigma_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. In this way, the matrix Σ_r is determined uniquely (but U_r and V_r are of course not unique).

The Compact Singular Value Decomposition

3.4.10. Example. Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Then

$$A^*A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix} \qquad \text{and} \qquad AA^* = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The eigenvalues of A^*A are $\lambda_1 = 3$, $\lambda_2 = 1$ and $\lambda_3 = 0$. The eigenvalues λ_1 and λ_2 are also the eigenvalues of AA^*. The non-zero singular values of A are

$$\sigma_1 = \sqrt{3}, \qquad \sigma_2 = 1.$$

The Compact Singular Value Decomposition

The right-singular vectors of A are the normed eigenvectors of A^*A:

$$v_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \qquad v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$$

The left-singular vectors are

$$u_1 = \frac{1}{\sqrt{3}} A v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad u_2 = A v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Hence,

$$U_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad V_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & -\sqrt{3} \\ 1 & \sqrt{3} \\ 2 & 0 \end{pmatrix}$$

The Compact Singular Value Decomposition

We can verify directly that

$$U_2^T A V_2 = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{pmatrix} = \Sigma_2$$

and

$$U_2 \Sigma_2 V_2^T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = A.$$

Note that the matrices U and V are not uniquely determined, because we could have changed the sign of the right-singular vectors.

3.4.11. Example. We calculate the singular value decomposition of

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & -1 \end{pmatrix}$$

For reference, we note that the eigenvalues of A are $-\sqrt{2}, \sqrt{2}$ and 1.

The Singular Values

We have

$$A^*A = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The eigenvalues of A^*A are

$$\lambda_1 = 2 + \sqrt{2}, \qquad \lambda_2 = 2, \qquad \lambda_3 = 2 - \sqrt{2}$$

so the singular values of A are

$$\sigma_1 = \sqrt{2 + \sqrt{2}}, \qquad \sigma_2 = \sqrt{2}, \qquad \sigma_3 = \sqrt{2 - \sqrt{2}}.$$

Hence,

$$\Sigma_3 = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3) = \begin{pmatrix} \sqrt{2 + \sqrt{2}} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2 - \sqrt{2}} \end{pmatrix}$$

The Right-Singular Vectors

Orthonormalized eigenvectors for the λ_i, i.e., the right-singular vectors, are

$$v_1 = \frac{1}{\sqrt{4-2\sqrt{2}}} \begin{pmatrix} 1 \\ -1+\sqrt{2} \\ 0 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

$$v_3 = \frac{1}{\sqrt{4-2\sqrt{2}}} \begin{pmatrix} 1-\sqrt{2} \\ 1 \\ 0 \end{pmatrix}.$$

Hence, we have

$$V = \begin{pmatrix} \frac{1}{\sqrt{4-2\sqrt{2}}} & 0 & \frac{1-\sqrt{2}}{\sqrt{4-2\sqrt{2}}} \\ \frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}} & 0 & \frac{1}{\sqrt{4-2\sqrt{2}}} \\ 0 & 1 & 0 \end{pmatrix}$$

The Left-Singular Vectors

We can either find the left-singular vectors directly as $u_i = \frac{1}{\sigma_i} A v_i$, $i = 1, 2, 3$, or, since Σ_3 is invertible, through $\Sigma_3 = U^* A V$:

$$U = AV\Sigma_3^{-1} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{\sqrt{2}} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix}$$

The orthonormal columns of U are the left-singular eigenvectors u_1, u_2, u_3.

The Full Singular Value Decomposition

We can extend the compact singular value decomposition so that the decomposition involves quadratic matrices U and V, which are then full (not just partial) isometries.

Let $v_1, \dots, v_r$, $r < n$, be the orthonormal right-singular vectors for all non-zero singular values of A. We complement these with arbitrary orthonormal vectors $v_{r+1}, \dots, v_n$ such that $v_1, \dots, v_n$ gives an orthonormal basis of $\mathbb{R}^n$.

We similarly add vectors $u_{r+1}, \dots, u_n$ to the left-singular vectors so that $u_1, \dots, u_n$ is an orthonormal basis. Then

$$u_i^* A v_j = \langle u_i, A v_j \rangle = 0 \qquad \text{if } i > r \text{ or } j > r.$$

We see this as follows: if $j > r$, then $A^* A v_j = 0$. This implies $\langle A v_j, A v_j \rangle = \langle v_j, A^* A v_j \rangle = 0$ and hence $A v_j = 0$. If $i > r$ and $j \leq r$, then $A v_j = \sigma_j u_j$ and the expression vanishes because $\langle u_i, u_j \rangle = 0$.

The Full Singular Value Decomposition

3.4.12. (Full) Singular Value Decomposition. For $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ there exist isometries $U \in \mathrm{Mat}(m \times m; \mathbb{F})$, $V \in \mathrm{Mat}(n \times n; \mathbb{F})$ such that

$$A = U \Sigma V^*,$$

where $\Sigma \in \mathrm{Mat}(m \times n; \mathbb{F})$ is a not-necessarily-square matrix whose diagonal lists the non-zero singular values of A and is zero elsewhere.

3.4.13. Example. Consider again the matrix of Example 3.4.10,

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

We had found

$$U_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad V_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & -\sqrt{3} \\ 1 & \sqrt{3} \\ 2 & 0 \end{pmatrix}$$

The Full Singular Value Decomposition

The matrix U_2 is already unitary, and we add an orthonormal vector to right-singular vectors to obtain

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad V = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 & -\sqrt{3} & \sqrt{2} \\ 1 & \sqrt{3} & \sqrt{2} \\ 2 & 0 & -\sqrt{2} \end{pmatrix}$$

Then

$$U^T A V = \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

3.4.14. Example. Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 2 & 1 \end{pmatrix}$$

The Full Singular Value Decomposition

Here

$$A^*A = \begin{pmatrix} 2 & 1 & 3 & 2 \\ 1 & 2 & 3 & 1 \\ 3 & 3 & 6 & 3 \\ 2 & 1 & 3 & 2 \end{pmatrix}, \qquad AA^* = \begin{pmatrix} 3 & 1 & 4 \\ 1 & 2 & 3 \\ 4 & 3 & 7 \end{pmatrix}$$

and the eigenvalues of A^*A are $\lambda_1 = 6 + \sqrt{21}$, $\lambda_2 = 6 - \sqrt{21}$, $\lambda_3 = \lambda_4 = 0$. Omitting the (very messy) calculations, the compact singular value decomposition yields isometric U_2, V_2 such that

$$U_2^* A V_2 = \begin{pmatrix} \sqrt{6+\sqrt{21}} & 0 \\ 0 & \sqrt{6-\sqrt{21}} \end{pmatrix},$$

while the full singular value decomposition gives unitary U, V such that

$$U^* A V = \begin{pmatrix} \sqrt{6+\sqrt{21}} & 0 & 0 & 0 \\ 0 & \sqrt{6-\sqrt{21}} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The Truncated Singular Value Decomposition

One of the major applications of the singular value decomposition is to approximate a given matrix A by another matrix $\widetilde{A}$ obtained by reconstruction from a reduced number of singular values of A:

3.4.15. Truncated Singular Value Decomposition. For $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ have non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, $r \leq \min(m, n)$. Let $t \leq r$,

$$\Sigma_t := \mathrm{diag}(\sigma_1, \ldots, \sigma_t)$$

and U_t and V_t the corresponding matrices of left- and right-singular vectors. Then

$$\widetilde{A} = U_t \Sigma_t V_t^*$$

is called the truncated singular value decomposition of A.

The Truncated Singular Value Decomposition

3.4.16. Example. Consider once more the matrix of Example 3.4.11,

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 0 \\ -1 & 0 & -1 \end{pmatrix}$$

The singular values of A were

$$\sigma_1 = \sqrt{2+\sqrt{2}}, \qquad \sigma_2 = \sqrt{2}, \qquad \sigma_3 = \sqrt{2-\sqrt{2}}.$$

We calculate the truncated SVD using only the two largest singular values, so

$$\Sigma_2 := \begin{pmatrix} \sqrt{2+\sqrt{2}} & 0 \\ 0 & \sqrt{2} \end{pmatrix}.$$

The Truncated Singular Value Decomposition

The matrices of left- and right-singular vectors are

$$U_2 = \begin{pmatrix} \frac{1}{2} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{2} & -\frac{1}{\sqrt{2}} \end{pmatrix}, \qquad V_2 = \begin{pmatrix} \frac{1}{\sqrt{4-2\sqrt{2}}} & 0 \\ \frac{\sqrt{2}-1}{\sqrt{4-2\sqrt{2}}} & 0 \\ 0 & 1 \end{pmatrix}.$$

Then

$$U_2^T A V_2 = \Sigma_2$$

and

$$\widetilde{A} = U_2 \Sigma_2 V_2^T = \frac{\sqrt{3+2\sqrt{2}}}{2} \begin{pmatrix} 1/\sqrt{2} & 1-1/\sqrt{2} & -\sqrt{3-2\sqrt{2}} \\ 1 & \sqrt{2}-1 & 0 \\ -1/\sqrt{2} & 1/\sqrt{2}-1 & -\sqrt{3-2\sqrt{2}} \end{pmatrix}.$$

Low-Rank Approximation

The truncated SVD can be used to approximate a given matrix through a lower-rank matrix; in fact, it is the best such approximation with respect to the so-called ***trace norm***, defined as

$$\|A\|_{\mathrm{tr}} := \sqrt{\mathrm{tr}\, A^* A} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

for $A \in \mathrm{Mat}(m \times n; \mathbb{F})$.

3.4.17. Eckart-Young Theorem. Let $A \in \mathrm{Mat}(m \times n; \mathbb{F})$ have the non-zero singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. Then the matrix M of rank t that minimizes $\|A - M\|_{\mathrm{tr}}$ is given by the truncated singular value decomposition using the t largest singular values of A.

Low-Rank Approximation

We recall that

$$\operatorname{tr}(a_{ij})_{i,j=1}^{n} = \sum_{i=1}^{n} a_{ii}.$$

and that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any square matrices A and B.

Proof.

Using the full singular value decomposition $A = U\Sigma V^*$, we have

$$\begin{aligned}
\|A - M\|_{\text{tr}}^2 &= \|U\Sigma V^* - M\|_{\text{tr}}^2 = \operatorname{tr}\big((U\Sigma V^* - M)^*(U\Sigma V^* - M)\big) \\
&= \operatorname{tr}\big((V\Sigma U^* - M^*)(U\Sigma V^* - M)\big) \\
&= \operatorname{tr}\big(VV^*(V\Sigma U^* - M^*)(U\Sigma V^* - M)\big) \\
&= \operatorname{tr}\big((\Sigma U^* - V^*M^*)UU^*(U\Sigma - MV)\big) \\
&= \operatorname{tr}\big((\Sigma - V^*M^*U)(\Sigma - U^*MV)\big) \\
&= \|\Sigma - U^*MV\|_{\text{tr}}^2
\end{aligned}$$

Since Σ is diagonal, the trace norm is minimized if $S := U^*MV$ is diagonal.

Low-Rank Approximation

Proof (continued).

Then

$$\|\Sigma - S\|_{\text{tr}}^2 = \sum_{i=1}^{n} (\sigma_i - s_{ii})^2.$$

Since S is to have rank r, only r diagonal entries of S can be non-zero. These non-zero entries s_{ii} should equal σ_i to minimize the sum. Then

$$\|\Sigma - S\|_{\text{tr}}^2 = \sum_{s_{ii}=0} \sigma_i^2.$$

This is minimized if the sum is over the $n - r$ smallest singular values, i.e., if the diagonal elements of S are the r largest singular values. It follows that $S = \Sigma_r$ and

$$M = U\Sigma_r V^*.$$

□

Image Compression

Reference Example 3.4.11 and the following application are taken from S. Beaver, ***The Singular Value Decomposition and a Democratic Method of Orthogonalization***, http://www.wou.edu/~beavers/Talks/TalksPage.html

The singular value decomposition can be used for image compression: A 256 grayscale image of size 320×200 pixel may be represented as a data matrix $A \in \mathrm{Mat}(320 \times 200; \mathbb{R})$ with entries between 0 and 1 corresponding to the grayscale. Each entry in the matrix takes up 1 byte (8 bits; a number between 0 and 255) of storage space, so the total amount of storage space needed for the image is $320 \cdot 200 = 64000$ bytes.

We perform a singular value decomposition on A, obtaining $A = U\Sigma V^*$. Replacing Σ with the truncated SVD Σ_r, we obtain

$$A_r = U\Sigma_r V^*$$

as the best rank-r approximation of A.

Image Compression

To store A_r, we need $320 \cdot r$ bytes to store the r vectors $\sigma_1 v_1, \ldots \sigma_r v_r$ and $200 \cdot r$ bytes to store the vectors $u_1, \ldots, u_r$. For $r = 20$ this is $520 \cdot 20 = 10400$ bytes, less than $1/6$ the original storage space.

We demonstrate the compression using Mathematica. The image below is a 320×200, 256 grayscale bitmap image:

```
clown = Import["clownpic1.tif"]
```

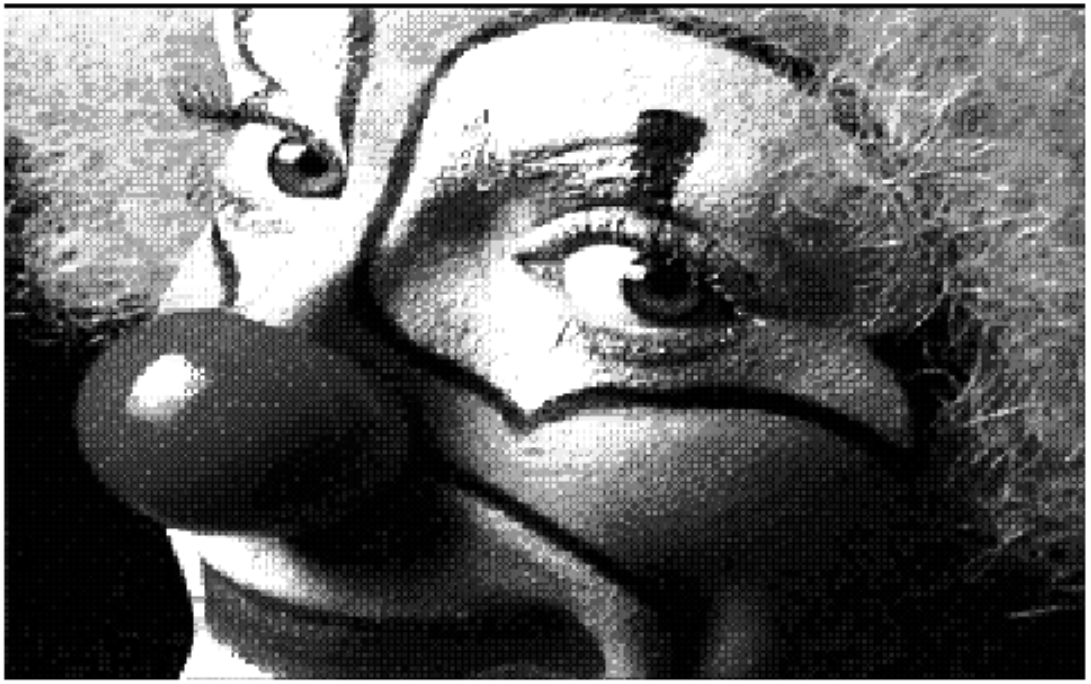


Image Compression

We verify the size of the image, the number of channels (byte/pixel) and obtain the singular value decomposition; we use lower case letters (u, σ, v) for (U, Σ, V). Here $\sigma \in \mathrm{Mat}(320 \times 200; \mathbb{R})$.

```
ImageDimensions[clown]
```

```
{320, 200}
```

```
ImageChannels[clown]
```

```
1
```

```
data := ImageData[clown]
```

```
{u, σ, v} = SingularValueDecomposition[data]
```

```
Length[σ]
```

```
200
```

```
Length[Transpose[σ]]
```

```
320
```

Image Compression

The original image is regained from the singular value decomposition:

```
Image[u.σ.Transpose[v]]
```

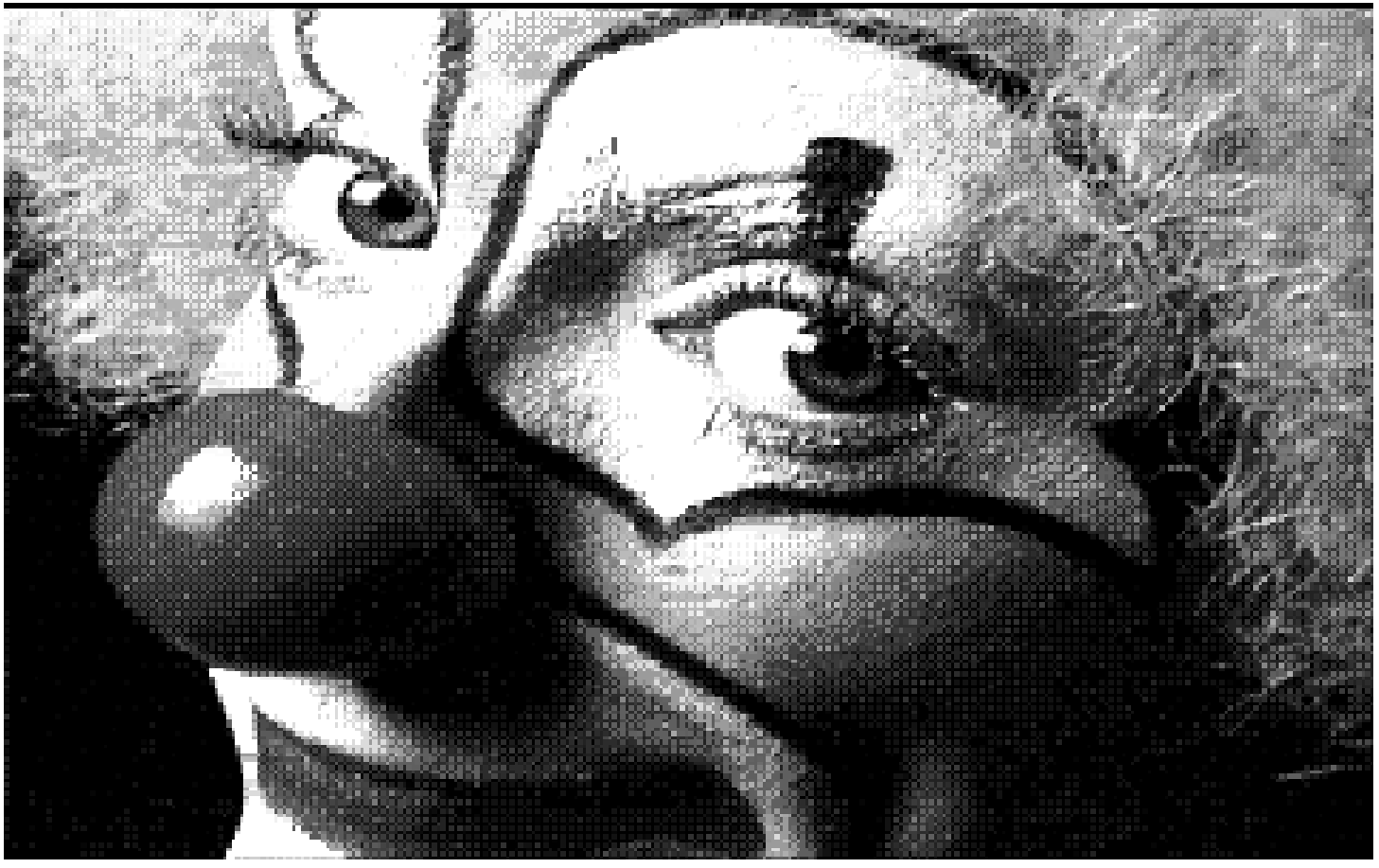


Image Compression

```
s[i_, j_] := If[i == j && i ≤ 20, σ[[i]][[j]], 0];
S := Array[s, {200, 320}];
Image[u.S.Transpose[v]]
```

Image Compression

On the previous slide, we have first constructed the matrix Σ_r, $r = 20$ (here denoted by S) and then displayed the compressed image. The image quality is of course worse than that of the original, but given a compression by more than 80% it is quite satisfactory.

The next slide shows the image for $r = 60$, which corresponds to 31200 bytes (compression by 50%). The image quality is quite good.

Image Compression

```
s[i_, j_] := If[i == j && i ≤ 60 , σ[[i]][[j]], 0];
S := Array[s, {200, 320}];
Image[u.S.Transpose[v]]
```