Linear Methods

Markus Grasmair

Trondheim, Norway

Lecture notes for the course TMA4145 - Linear Methods, Autumn 2023, NTNU

Contents

CHAPTER 1

Linear Algebra

1. Vector Spaces and Linear Transformations

1.1. Basic Definition. We start with a formal definition of vector spaces. Most if not all of this section has already been discussed in previous mathematics classes (e.g. Mathematics 3 or Linear Algebra with Applications), though possibly in less detail and possibly in a slightly less abstract way.

DEFINITION 1.1.1. A *vector space* over the real numbers $\mathbb{R}$ (the complex numbers $\mathbb{C}$) is a set V together with the operations $+\colon V \times V \to V$ and $\cdot\colon \mathbb{R} \times V \to V$ ($\cdot\colon \mathbb{C} \times V \to V$), such that the following properties hold:

(1) Commutativity: For all u, $v \in V$ we have $u + v = v + u$.
(2) Associativity: For all u, v, $w \in V$ we have $(u + v) + w = u + (v + w)$, and for all $v \in V$ and all λ, $\mu \in \mathbb{R}$ (λ, $\mu \in \mathbb{C}$) we have $\lambda(\mu v) = (\lambda\mu)v$.
(3) Additive identity: There exists an element $0 \in V$ such that $v + 0 = v$ for all $v \in V$.
(4) Additive inverse: For every element $u \in V$ there exists an element $v \in V$ such that $u + v = 0$.
(5) Multiplicative identity: For all $v \in V$ we have that $1v = v$.
(6) Distributivity: For all u, $v \in V$ and all λ, $\mu \in \mathbb{R}$ (λ, $\mu \in \mathbb{C}$) we have that $\lambda(u + v) = \lambda u + \lambda v$, and $(\lambda + \mu)v = \lambda v + \mu v$.

DEFINITION 1.1.2. An element of a vector space is called a *vector*.

REMARK 1.1.3. For simplicity, we will in the following denote by $\mathbb{K}$ either the set $\mathbb{R}$ of real numbers or the set $\mathbb{C}$ of complex numbers in order to simplify the notation. There will be some situations later on, when we will have to distinguish between real and complex vector spaces, but many of the simpler results are the same in both cases.

The letter $\mathbb{K}$ here stands for *field*, which in Norwegian is *kropp* (and in German *Körper*). This is the most general algebraic structure for which it makes sense to define vector spaces. For more information, please see the course TMA4150 – Algebra.

EXAMPLE 1.1.4. The following are some standard examples of vector spaces that will be used throughout the course. It is left to the reader as an exercise[1] to verify that all the requirements for a vector space (that is, Items (1)–(6) in Definition 1.1.1) are indeed satisfied in each example.

(1) For every $n \in \mathbb{N}$, the set $\mathbb{K}^n$ of n-tuples of real (complex) numbers is a real (complex) vector space. Here we use the componentwise addition and scalar multiplication, that is, for $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n) \in \mathbb{K}^n$ and $\lambda \in \mathbb{K}$ we define

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$$

[1] These "exercises for the reader" can actually provide you with suitable training for the exam! They are *not exclusively* here, because I was too lazy to formulate a complete proof.

and

$$\lambda \cdot (x_1, x_2, \dots, x_n) = (\lambda x_1, \lambda x_2, \dots, \lambda x_n).$$

(2) For all n, $m \in \mathbb{N}$, the sets $\mathbf{Mat}_{n,m}(\mathbb{K}) = \mathbb{K}^{m \times n}$ and $\mathbf{Mat}_n(\mathbb{K}) = \mathbb{K}^{n \times n}$ of $(n \times m)$-dimensional and $(n \times n)$-dimensional matrices, respectively, are vector spaces.

(3) The set $\mathcal{P}(\mathbb{K})$ of all polynomials (of arbitrary degree) with coefficients in $\mathbb{K}$ is a vector space.

(4) Assume that S is an arbitrary set and denote by

$$\mathbb{K}^S := \{f : S \to \mathbb{K}\}$$

the set of $\mathbb{K}$-valued functions on S. On $\mathbb{K}^S$ we define addition and scalar multiplication of functions by

$$(f+g)(s) := f(s) + g(s) \qquad \text{and} \qquad (\lambda f)(s) := \lambda f(s)$$

for f, $g \in \mathbb{K}^S$, $\lambda \in \mathbb{K}$, and $s \in S$. With these operations, the set $\mathbb{K}^S$ becomes a vector space.

Remark 1.1.5. The usage of $+$ and $\cdot$ for the addition and scalar multiplication in Definition 1.1.1 is intended to be suggestive and follows the standard notation of these operations. However, it somewhat hides how restrictive the assumptions actually are.

A more formal (but much less intuitive) way would be to define a vector space as a set V together with two functions $f : V \times V \to V$ and $g : \mathbb{R} \times V \to V$ that satisfy Items (1)–(6). With this notation, commutativity and parts of associativity and distributivity, for instance, would read as

$$\begin{aligned}
&\text{Commutativity:} && f(u,v) = f(v,u),\\
&\text{Associativity:} && f\big(f(u,v),w\big) = f\big(u,f(v,w)\big),\\
&\text{Distributivity:} && g(\lambda+\mu,v) = f\big(g(\lambda,v),g(\mu,v)\big),
\end{aligned}$$

respectively.

1.2. General Properties. In the following, we will prove some general results concerning vector spaces that follow directly from the definition of a vector space. Amongst others, we will show that the additive identity and the additive inverse are unique, although this seems not to be required in the definition. This is obviously true in all the examples provided above. What these results, however, show, is that it is impossible to construct a vector space with two different 0-elements. If one would need such an object for whatever theoretical or practical reason, one would need to search for it outside the realm of vector spaces.

This section is mainly intended to demonstrate by means of simple examples how elementary algebraic proofs look like. Feel free to skip over this part, if you are already familiar with this type of proofs.

Lemma 1.1.6. *Assume that V is a vector space. Then the additive identity $0 \in V$ is unique. That is, there exists no vector $v \in V$ with $v \neq 0$ such that $u + v = u$ for all $u \in V$.*

Proof. Assume that $v \in V$ satisfies $u + v = u$ for all $u \in V$. With $u = 0$ we have in particular that

$$0 = 0 + v \underbrace{=}_{(1)} v + 0 \underbrace{=}_{(3)} v,$$

which shows that $v = 0$. □

Lemma 1.1.7. *Assume that V is a vector space and that $u \in V$. Then there exists a unique vector $v \in V$ such that $u + v = 0$.*

PROOF. Since V is a vector space, we already know that such a vector v exists. Thus it remains to show that v is unique.

Assume therefore that v_1, $v_2 \in V$ satisfy $u + v_1 = 0$ and $u + v_2 = 0$. Then

$$v_1 \underbrace{=}_{(3)} v_1+0 = v_1+(u+v_2) \underbrace{=}_{(2)} (v_1+u)+v_2 \underbrace{=}_{(1)} (u+v_1)+v_2 = 0+v_2 \underbrace{=}_{(1)} v_2+0 = v_2,$$

that is, $v_1 = v_2$. This proves the uniqueness of the additive inverse. □

DEFINITION 1.1.8. Let V be a vector space and $v \in V$. By $-v \in V$ we denote the additive inverse of v, that is, the unique vector satisfying $v + (-v) = 0$. Moreover, for $u \in V$, we generally abbreviate $u - v := u + (-v)$.

LEMMA 1.1.9. *Let V be a vector space and let $v \in V$. Then $0 \cdot v = 0$.*

PROOF. We have that

$$v + 0 \cdot v \underbrace{=}_{(5)} 1 \cdot v + 0 \cdot v \underbrace{=}_{(6)} (1 + 0) \cdot v = 1 \cdot v \underbrace{=}_{(5)} v,$$

which shows that $0 \cdot v$ is an additive identity in V. Because of the uniqueness of the additive identity (see Lemma 1.1.6), it follows that $0 \cdot v = 0$. □

LEMMA 1.1.10. *Let V be a vector space and $v \in V$. Then $-v = (-1) \cdot v$.*

PROOF. We have that

$$v + (-1) \cdot v \underbrace{=}_{(5)} 1 \cdot v + (-1) \cdot v \underbrace{=}_{(6)} (1 - 1) \cdot v = 0 \cdot v \underbrace{=}_{\text{Lemma 1.1.9}} 0.$$

This shows that $(-1) \cdot v$ is an additive inverse for v. Now the identity $(-1) \cdot v = -v$ follows from the uniqueness of the additive inverse. □

1.3. Subspaces. We will now introduce the notion of subspaces of vector spaces.

DEFINITION 1.1.11. Assume that V is a vector space and $U \subset V$ a subset of V. We say that U is a subspace of V, if U is in itself a vector space (with the same addition and scalar multiplication as on U, and in particular over the same field $\mathbb{K}$).

LEMMA 1.1.12. *Assume that V is a vector space and $U \subset V$. Then U is a subspace of V, if and only if the following conditions hold:*

(1) (Additive identity:) $0 \in U$.
(2) (Closedness under addition:) If u, $v \in U$, then also $u + v \in U$.
(3) (Closedness under scalar multiplication:) If $u \in U$ and $\lambda \in \mathbb{R}$, then also $\lambda u \in U$.

PROOF. Items (2) and (3) guarantee that the addition and scalar multiplication indeed map $U \times U$ and $\mathbb{K} \times U$ onto U (and not arbitrarily into V). Moreover, Item (1) guarantees that U contains an additive identity, and Item (3) guarantees that U contains additive inverses (namely the element $(-1) \cdot u$ for $u \in U$). The other properties of a vector space are satisfied, because they already hold on the larger space V. □

EXAMPLE 1.1.13. Consider the following examples of vector spaces and subsets, which in some cases are subspaces and in others are not:[2]

(1) For every vector space V, the sets $\{0\}$ and V are subspaces of V.
(2) The set $U := \big\{(x_1, x_2) \in \mathbb{R}^2 : x_2 = -x_1\big\}$ is a subspace of $\mathbb{R}^2$.

[2] As an exercise, you can verify in each of the examples whether this is actually true.

(3) Let $V = \mathbb{K}^n$ and $U := \mathbb{K}^{n-1} \times \{0\} = \big\{(x_1, \dots, x_{n-1}, x_n) \in \mathbb{K}^n : x_n = 0\big\}$. Then U is a subspace of V.
(4) For every $n \in \mathbb{N}$, the sets $\mathbf{Sym}_n(\mathbb{K}) := \big\{A \in \mathbf{Mat}_n(\mathbb{K}) : A^T = A\big\}$ of symmetric $n \times n$-matrices and $\mathbf{Skew}_n(\mathbb{K}) := \big\{A \in \mathbf{Mat}_n(\mathbb{K}) : A^T = -A\big\}$ of skewsymmetric $n \times n$-matrices are sub-spaces of $\mathbf{Mat}_n(\mathbb{K})$.
(5) For every $n \in \mathbb{N}$, the set $\mathcal{P}_n(\mathbb{K})$ of polynomials of degree at most n is a subspace of the vector space $\mathcal{P}(\mathbb{K})$ of all polynomials.
(6) The set $C(\mathbb{K})$ of continuous $\mathbb{K}$-valued functions $f\colon \mathbb{R} \to \mathbb{K}$ is a sub-space of the space $\mathbb{R}^{\mathbb{K}}$ of all $\mathbb{K}$-valued functions on $\mathbb{R}$.
(7) Denote by $V := \big\{f \in C(\mathbb{K}) : f(7) = 0\big\}$. Then V is a subspace of $C(\mathbb{K})$.
(8) Let $V = \mathbb{K}^n$ and $U := \mathbb{K}^{n-1} \times \{1\} = \big\{(x_1, \dots, x_{n-1}, x_n) \in \mathbb{K}^n : x_n = 1\big\}$. Then U is **not** a subspace of V.
(9) Let $U := \big\{(x_1, x_2) \in \mathbb{K}^2 : x_1 = 0 \text{ or } x_2 = 0\big\}$. Then U is **not** a subspace of $\mathbb{K}^2$.
(10) Let $n \in \mathbb{N}$, $n \geq 2$, and $U := \big\{A \in \mathbf{Mat}_n(\mathbb{K}) : \det A = 0\big\}$ the set of singular matrices. Then U is **not** a subspace of $\mathbf{Mat}_n(\mathbb{K})$.

1.4. Linear Transformations.

DEFINITION 1.1.14. Assume that U, V are vector spaces and that $T\colon U \to V$ is some mapping. We say that T is *linear* (or a *linear transformation*) if the following conditions hold:

- For all u, $v \in U$ we have $T(u + v) = T(u) + T(v)$.
- For all $u \in U$ and $\lambda \in \mathbb{K}$ we have $T(\lambda u) = \lambda\, T(u)$.

REMARK 1.1.15. For linear transformations it is common to drop the parantheses around the argument. That is, if $T\colon U \to V$ is linear, one usually writes Tu instead of $T(u)$.

EXAMPLE 1.1.16. Consider the following examples of linear mappings:
(1) The mapping $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA := A^T$ is linear.
(2) Let $P \in \mathbf{Mat}_n(\mathbb{K})$ be a fixed matrix. Then the mapping $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA := PAP$ is linear.
(3) The mapping $D\colon \mathcal{P}(\mathbb{K}) \to \mathcal{P}(\mathbb{K})$, $p \mapsto Dp := p'$ is linear.

If $T\colon U \to V$ is a linear transformation, we have two important associated subspaces of U and V:

DEFINITION 1.1.17. Let U, V be vector spaces and $T\colon U \to V$ be a linear transformation.

- The *kernel* of T is defined as
$$\ker(T) := \big\{u \in U : Tu = 0\big\} \subset U.$$
- The *range* of T is defined as
$$\operatorname{ran}(T) := \big\{v \in V : \text{ there exists } u \in U \text{ with } Tu = v\big\} \subset V.$$

Put differently, the kernel of T is the set of all solutions of the equation $Tu = v$; the range of T is the set of all right hand sides v for which the equation $Tu = v$ has a solution.

LEMMA 1.1.18. *The range and the kernel of a linear transformation are subspaces of the respective vector spaces.*

PROOF. Exercise! □

EXAMPLE 1.1.19. Consider the mapping $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA := A^T + A$.

- The kernel $\ker(T)$ is the set of all matrices A that satisfy $A^T + A = 0$, that is, the set of all skew-symmetric matrices.
- The range $\operatorname{ran}(T)$ is the set of all matrices B that can be written in the form $B = A + A^T$ for some matrix A. It is easy to see that this is the case if and only if B is symmetric.

Thus we have that

$$\ker(T) = \mathbf{Skew}_n(\mathbb{K}) \qquad \text{and} \qquad \operatorname{ran}(T) = \mathbf{Sym}_n(\mathbb{K}).$$

2. Linear Independence and Bases

2.1. Linear Span.

Definition 1.2.1. Assume that V is a vector space and that $S := \{v_i : i \in I\} \subset V$ is a (possibly infinite) subset of V. We define the *linear span* $\operatorname{span}(S) \subset V$ of S as the set of all (finite) linear combinations of elements in S.

That is, a vector $u \in V$ is an element of $\operatorname{span}(S)$, if and only if there exist $N \in \mathbb{N}$, vectors $v_{i_1}, \ldots, v_{i_N} \in S$, and scalars $\lambda_1, \ldots, \lambda_N \in \mathbb{K}$ such that

$$u = \sum_{j=1}^{N} \lambda_j v_{i_j}.$$

Remark 1.2.2. It can be useful to also talk about the span of the empty set $\emptyset$, mainly in order to avoid having to deal with annoying special cases in various theoretical results. For that, we simply define

$$\operatorname{span}(\emptyset) := \{0\} \subset V.$$

Proposition 1.2.3. *Assume that V is a vector space and $S \subset V$ is a subset. Then $\operatorname{span}(S)$ is the smallest subspace of V containing S. That is, $\operatorname{span}(S)$ is a subspace of V, and if $U \subset V$ is any other subspace of V with $S \subset U$, then also $\operatorname{span}(S) \subset U$.*

Proof. We show first that $\operatorname{span}(S)$ is actually a subspace of V. To that end, we note first that, trivially, $0 \in V$. Next assume that $u, w \in \operatorname{span}(S)$. Then we can write

$$u = \sum_{j=1}^{N} \lambda_j v_{i_j} \qquad \text{and} \qquad w = \sum_{j=N+1}^{N+M} \lambda_j v_{i_j}$$

for some $N, M \in \mathbb{N}$, vectors $v_{i_1}, \ldots, v_{i_{N+M}} \in S$, and scalars $\lambda_1, \ldots, \lambda_{N+M} \in \mathbb{K}$. As a consequence,

$$u + w = \sum_{j=1}^{N+M} \lambda_j v_{i_j} \in \operatorname{span}(S).$$

Similarly, if $u \in \operatorname{span}(S)$ and $\alpha \in \mathbb{K}$, we can write

$$u = \sum_{j=1}^{N} \lambda_j v_{i_j}$$

for some $N \in \mathbb{N}$, vectors $v_{i_1}, \ldots, v_{i_N} \in S$, and scalars $\lambda_1, \ldots, \lambda_N \in \mathbb{K}$. Then

$$\lambda u = \alpha \sum_{j=1}^{N} \lambda_j v_{i_j} = \sum_{j=1}^{N} \alpha \lambda_j v_{i_j},$$

which shows that also $\lambda u \in \operatorname{span}(S)$. According to Lemma 1.1.12, this shows that $\operatorname{span}(S)$ is a subspace of V.

Now assume that $U \subset V$ is any subspace of V containing S. Let moreover $u \in \operatorname{span}(S)$. We have to show that also $u \in U$.

Because $u \in \operatorname{span}(S)$, we can write

$$u = \sum_{j=1}^{N} \lambda_j v_{i_j}$$

for some $N \in \mathbb{N}$, vectors $v_{i_1}, \ldots, v_{i_N} \in S$, and scalars $\lambda_1, \ldots, \lambda_N \in \mathbb{K}$. By assumption, $v_{i_j} \in U$ for every j, and therefore also $\lambda_j v_{i_j} \in U$, as U is a subspace of V. Next, we have that

$$\lambda_1 v_{i_1} + \lambda_2 v_{i_2} + \ldots + \lambda_N v_{i_N} \in U,$$

because each term is contained in U and, again, U is a subspace of V. This shows that $u \in U$, which concludes the proof. □

Example 1.2.4. Consider the space $\mathcal{P}(\mathbb{K})$ of all polynomials with coefficients in $\mathbb{K}$. Let moreover $M := \{1, x, x^2, \ldots\}$ be the set of all monomials. Then $\operatorname{span}(M) = \mathcal{P}(\mathbb{K})$, as we can write every polynomial as a linear combination of finitely many polynomials.

2.2. Linear Independence and Bases.

Definition 1.2.5. Assume that V is a vector space and $S = \{v_i : i \in I\}$ is a subset of V. We say that S is *linearly independent*, if the following holds:

For all $N \in \mathbb{N}$, vectors $v_{i_1}, \ldots, v_{i_N} \in S$, and scalars $\lambda_1, \ldots, \lambda_N \in \mathbb{K}$, if

$$\sum_{j=1}^{N} \lambda_j v_{i_j} = 0,$$

then

$$\lambda_j = 0 \text{ for all } j = 1, \ldots, N.$$

If S is not linearly independent, we call S *linearly dependent*.

Proposition 1.2.6. *Assume that V is a vector space and $S = \{v_i : i \in I\}$ is a subset of V. Then S is linearly independent, if and only if every vector $u \in \operatorname{span}(S)$ can be written in a unique way as a finite linear combination of elements of S.*

Proof. To do! □

Example 1.2.7. Consider the space $\mathcal{P}(\mathbb{K})$ of all polynomials with coefficients in $\mathbb{K}$. Let moreover $M := \{1, x, x^2, \ldots\}$ be the set of all monomials. Then M is linearly independent.

Definition 1.2.8. Let V be a vector space and $B \subset V$ a subset. We say that B is a *basis* of V, if B is linearly independent and $\operatorname{span}(B) = V$.

Remark 1.2.9. In view of Proposition 1.2.6, we can equivalently say that B is a basis of V, if *every* vector $v \in V$ can be written in a *unique* way as a *finite* linear combination of elements of B.

Example 1.2.10. The set M of monomials is a basis of the vector space of all polynomials.

Theorem 1.2.11. *Assume that V is a vector space and $S \subset V$ is some linearly independent subset. Then there exists a basis B of V with $S \subset B$.*

In particular, every vector space has a basis.

Proof. This is a surprisingly tricky result and relies heavily on a completely non-intuitive axiom of set theory. If time permits, I will add a proof (together with the required prerequisites); if eventually added, this proof will of course **not** be part of the pensum. □

The existence of a basis of every vector space is, for pure mathematicians, an extremely interesting result. For practical applications, however, it turns out to be largely useless in that generality, as one can actually show that there cannot be a universally applicable method for actually constructing a basis. Because of that, we will turn instead to finite dimensional vector spaces, where the situation appears to be much simpler. This, however, requires us first to define the notion of the dimension of a vector space.

DEFINITION 1.2.12. A vector space V is called *finite dimensional*, if there exists a finite set S with $\operatorname{span}(S) = V$. Else, V is called *infinite dimensional.*

THEOREM 1.2.13. *Assume that V is finite dimensional. Then all bases of V are finite and contain the same number of elements, called the* dimension *of V.*

PROOF. Again, I will add a proof if time permits. In contrast to the proof of the existence of bases for arbitrary vector spaces, the proof of this result is conceptually not that difficult, only rather annoyingly technical. □

3. Direct Sums and Invariant Subspaces

The goal, we are currently working towards, is the decomposition of linear transformations of finite dimensional vector spaces: Given a linear transformation $T\colon V \to V$, we want to write T in an "as simple as possible form." In order to do so, we will first have to introduce the necessary notation to be able to define what we mean by a decomposition.

DEFINITION 1.3.1. Assume that V is a vector space and that $S_1, \ldots, S_N$ are subsets of V. We define

$$S_1+\ldots+S_n := \bigl\{v \in V : \text{there exist } u_i \in S_i,\ i = 1, \ldots, N, \text{ with } v = u_1+\ldots+u_N\bigr\}.$$

We will mostly consider sums of subspaces of a vector space. In this case, it is easy to show that this is again a subspace.

LEMMA 1.3.2. *Assume that V is a vector space and that $U_1, \ldots, U_N$ are subspaces of V. Then $U_1 + \ldots U_N$ is again a subspace of V.*

PROOF. Exercise! □

EXAMPLE 1.3.3. We consider the vector space $\mathbf{Mat}_n(\mathbb{K})$ of $n \times n$-dimensional matrices over $\mathbb{K}$. We denote by $\mathbf{TriU}^*_n(\mathbb{K}) \subset \mathbf{Mat}_n(\mathbb{K})$ the subspace of strictly upper triangular matrices, that is, matrices A of the form

$$A = \begin{pmatrix} 0 & * & \ldots & * \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & * \\ 0 & \ldots & \ldots & 0 \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

The sum

$$W := \mathbf{TriU}^*_n(\mathbb{K}) + \mathbf{Skew}_n(\mathbb{K})$$

is the subspace of all matrices that can be written as a sum of a strict upper triangular matrix and a skew-symmetric matrix. It is easy to see that W consists of all matrices A where all the diagonal entries are equal to 0.

Now denote by $\mathbf{Diag}_n(\mathbb{K}) \subset \mathbf{Mat}_n(\mathbb{K})$ the subspace of all diagonal matrices. Then we have that

$$\mathbf{Mat}_n(\mathbb{K}) = \mathbf{TriU}^*_n(\mathbb{K}) + \mathbf{Skew}_n(\mathbb{K}) + \mathbf{Diag}_n(\mathbb{K}),$$

that is, we can write every matrix as a sum of a strict upper triangular matrix, a skew-symmetric matrix, and a diagonal matrix.

The sum of subspaces U_i, $i = 1, \ldots, N$, thus consists of all vectors w that can in some way be written as a sum $w = u_1 + \ldots + u_N$, where each of the vectors u_i is an element of the corresponding subspace U_i. These vectors u_i, however, need in general not be unique. That is, there might be different ways of decomposing the vector w into a sum of vectors in U_i.

Definition 1.3.4. Let V be a vector space and let $U_1, \ldots, U_N$ be subspaces of V. Denote moreover $W := U_1 + \ldots + U_N$. We say that W is the *direct sum* of the subspaces $U_1, \ldots, U_N$, denoted

$$W = U_1 \oplus \ldots \oplus U_N,$$

if every vector $w \in W$ can be uniquely written in the form $w = u_1 + \ldots + u_N$ with $u_i \in U_i$ for $i = 1, \ldots, N$.

More explicitly, we have that $W = U_1 \oplus \ldots \oplus U_N$, if we can conclude from the fact that we can write

$$\begin{aligned} w &= u_1 + \ldots + u_N \qquad \text{with } u_i \in U_i \text{ for all } i = 1, \ldots, N, \\ w &= v_1 + \ldots + v_N \qquad \text{with } v_i \in U_i \text{ for all } i = 1, \ldots, N, \end{aligned}$$

that necessarily

$$u_i = v_i \text{ for all } i = 1, \ldots, N.$$

Example 1.3.5. We have that

$$\mathbf{Mat}_n(\mathbb{K}) = \mathbf{TriU}_n^*(\mathbb{K}) \oplus \mathbf{Skew}_n(\mathbb{K}) \oplus \mathbf{Diag}_n(\mathbb{K}),$$

since for every matrix $A \in \mathbf{Mat}_n(\mathbb{K})$ there is one and only one way how we can write A as a sum of a strict upper triangular matrix, a skew-symmetric matrix, and a diagonal matrix: The diagonal entries of A need to be taken care of by the diagonal matrix; the strict lower triangular part needs to go into the skew-symmetric matrix; what remains of A after subtracting the diagonal and the skew-symmetric part is strictly upper triangular.

In the case of a sum of two sets, there is a simpler way for checking whether a sum is direct or not, as the following result shows.

Lemma 1.3.6. *Let V be a vector space and let U_1, $U_2 \subset V$ be subspaces of V. The sum $U_1 + U_2$ is direct, if and only if $U_1 \cap U_2 = \{0\}$.*

Proof. Assume first that $U_1 \cap U_2 = \{0\}$, and assume that $w \in U_1 + U_2$ can be written as

$$w = u_1 + u_2 = v_1 + v_2$$

with u_1, $v_1 \in U_1$, and u_2, $v_2 \in U_2$. Since U_1, U_2 are subspaces of V, it follows that $u_1 - v_1 \in U_1$ and $u_2 - v_2 \in U_2$, and we also have that $-(u_1 - v_1) \in U_1$ and $-(u_2 - v_2) \in U_2$. However, we also have that

$$0 = w - w = (u_1 - v_1) + (u_2 - v_2),$$

and therefore

$$u_1 - v_1 = -(u_2 - v_2).$$

This shows that $u_1 - v_1 = -(u_2 - v_2) \in U_1 \cap U_2 = \{0\}$, and thus $u_1 - v_1 = -(u_2 - v_2) = 0$, or $u_1 = v_1$ and $u_2 = v_2$. Thus the sum $U_1 + U_2$ is direct.

Now assume that the sum $U_1 + U_2$ is direct, and let $v \in U_1 \cap U_2$. Then we can write

$$v = v + 0 = 0 + v$$

in two ways as sum of one vector in U_1 and one vector in U_2. Since the sum is direct, this is only possible for $v = 0$, which shows that $U_1 \cap U_2 = \{0\}$. $\square$

REMARK 1.3.7. The result from Lemma 1.3.6 does not hold in the same/similar form for sums of more than two subspaces. That is, if we have three subspaces U_1, U_2, U_3 and we have that all pairwise intersections are trivial, that is, $U_1 \cap U_2 = U_1 \cap U_3 = U_2 \cap U_3 = \{0\}$, then we cannot conclude that the sum $U_1 + U_2 + U_3$ is direct. As a simple counterexample, consider the spaces $U_1 = \operatorname{span}\{(1,0)\}$, $U_2 = \operatorname{span}\{(0,1)\}$, $U_3 = \operatorname{span}\{(1,1)\}$ of $\mathbb{K}^2$. Their pairwise intersection is trivial, but their sum is not direct: The vector $(1,1)$, for instance, is an element of U_3, but it can also be written in the form $(1,1) = (1,0) + (0,1)$ as a sum of vectors in U_1 and U_2.

PROPOSITION 1.3.8. *Assume that V is a finite dimensional vector space and that $U_1, \ldots, U_N$ are subspaces of V. Then*

$$\dim(U_1 + \ldots + U_N) \leq \dim(U_1) + \ldots \dim(U_N),$$

and we have equality if and only if the sum is direct.

PROOF. Exercise! □

LEMMA 1.3.9. *Assume that V is a vector space and $U \subset V$ is a subspace. Then there exists a subset $W \subset V$ such that $V = U \oplus W$.*

PROOF. Let $\mathcal{B} = \{u_i : i \in I\}$ be a basis of U. In particular, $\mathcal{B}$ is a linearly independent subset of V and thus, by Theorem 1.2.11, can be extended to a basis of the whole space V. That is, there exist a set $\mathcal{C} = \{w_j : j \in J\} \subset V$ with $\mathcal{C} \cap \mathcal{B} = \emptyset$ such that $\mathcal{B} \cup \mathcal{C}$ forms a basis of V.

Define now $W := \operatorname{span}(\mathcal{C})$. Then

$$U + W = \operatorname{span}(\mathcal{B}) + \operatorname{span}(\mathcal{C}) \underbrace{=}_{\text{Exercise!}} \operatorname{span}(\mathcal{B} \cup \mathcal{C}) = V,$$

since $\mathcal{B} \cup \mathcal{C}$ is a basis of V.

It remains to show that the sum is direct. To that end, assume that $v \in U \cap V$. Since $\mathcal{B} \cup \mathcal{C}$ is a basis of V, there exists a unique way of writing v in the form

$$v = \sum_{j=1}^{N} \lambda_j u_{i_j} + \sum_{k=1}^{M} \mu_k v_{i_k}.$$

However, because $v \in V$, it follows that the coefficients μ_k have to be equal to 0 for all k. Similarly, because $v \in U$, it follows that the coefficients λ_j have all to be equal to 0. Thus $v = 0$, which shows that $U \cap W = \{0\}$. This shows that the sum is direct, which in turn completes the proof. □

It is now time to remind ourselves of our goal: We want to write a linear transformation $T\colon V \to V$ in an as simple as possible way. For that, we will decompose the vector space V into a direct sum of subspaces, on which T behaves "nicely." The next step is thus to explain what we mean by a nice behaviour of T on a subspace.

DEFINITION 1.3.10. Assume that V is a vector space and that $T\colon V \to V$ is a linear transformation. Assume moreover that $U \subset V$ is a subspace. We say that U is a *T-invariant* subspace of V, if $T(U) \subset U$, that is, for all $u \in U$ we have that $Tu \in U$.

EXERCISE 1.3.1. Show that $\ker(T)$ and $\operatorname{ran}(T)$ are always T-invariant subspaces.

EXAMPLE 1.3.11. Consider the transformation $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA = A^T$. Then the sets $\mathbf{Sym}_n(\mathbb{K})$ and $\mathbf{Skew}_n(\mathbb{K})$ are T-invariant subspaces: If $A \in \mathbf{Sym}_n(\mathbb{K})$, then $TA = A^T = A \in \mathbf{Sym}_n(\mathbb{K})$ as well, and similarly, if $A \in \mathbf{Skew}_n(\mathbb{K})$, then $TA = A^T = -A \in \mathbf{Skew}_n(\mathbb{K})$ as well.

EXAMPLE 1.3.12. Denote by $C^\infty(\mathbb{R})$ the space of all arbitrarily differentiable real-valued functions on $\mathbb{R}$, and let $\mathcal{P}$ be the space of all real polynomials. Define moreover $T\colon C^\infty(\mathbb{R})$, $f \mapsto Tf := f'$. Then $\mathcal{P}$ is a T-invariant subspace of $C^\infty(\mathbb{R})$: If p is a polynomial, then $Tp = p'$ is a polynomial as well. That is, if $p \in \mathcal{P}$, then also $Tp \in \mathcal{P}$, which is precisely the definition of T-invariance.

Until now, it is most probably not immediately obvious, why we should consider direct sums and invariant subspaces at all, if our goal is a simple description of a linear transformation. In the next section, we will therefore discuss matrix forms of linear transformations and in particular see how a direct sums and T-invariant subspaces translate into matrix form.

4. Direct Sums and Invariant Subspaces in Matrix Form

Assume that U, V are finite dimensional vector spaces (over the same field $\mathbb{K}$) with $n = \dim(U)$ and $m = \dim(V)$. Let moreover $\mathcal{B} = \{u_1, \ldots, u_n\} \subset U$ and $\mathcal{C} = \{v_1, \ldots, v_m\} \subset V$ be bases of U and V, respectively.

Since $\mathcal{B}$ and $\mathcal{C}$ are bases of the respective spaces, we can write each vector $u \in U$ uniquely as a linear combination

$$u = \sum_{j=1}^{n} \alpha_j u_j \qquad \text{with } \alpha_j \in \mathbb{K} \text{ for } j = 1, \ldots, n,$$

and similarly each vector $v \in V$ as a linear combination

$$v = \sum_{i=1}^{m} \beta_i v_i \qquad \text{with } \beta_i \in \mathbb{K} \text{ for } i = 1, \ldots, m.$$

The vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{K}^n$ is called the coordinate vector for u with respect to the basis $\mathcal{B}$, and the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T \in \mathbb{K}^m$ the coordinate vector for v with respect to $\mathcal{C}$. Commonly one says that one "identifies u and v with their coordinates $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$." Be mindful, though, that this identification only works (or is well-defined) *once we have fixed the bases on the spaces*; if we change the bases, the coordinate representations of the same vectors will usually change as well.

Assume now that $T\colon U \to V$ is a linear transformation and that $u \in U$ and $v \in V$ satisfy the relation

$$Tu = v.$$

In the following, we will discuss how this relation between the vectors translates into a relation between their coordinate representations $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

To start with, we consider the image of each of the basis vectors u_j, $j = 1, \ldots, m$, separately. Since $\mathcal{C}$ is a basis of V and $Tu_j \in V$, there exist, for each j, unique coefficients a_{ij}, $i = 1, \ldots, n$, such that

$$Tu_j = \sum_{i=1}^{n} a_{ij} v_i. \tag{1}$$

Now consider the vector

$$u = \sum_{j=1}^{m} \alpha_j u_j$$

with coordinates $(\alpha_1, \ldots, \alpha_m)$. The linearity of T implies that

$$Tu = \sum_{j=1}^{m} \alpha_j \, Tu_j = \sum_{j=1}^{m} \alpha_j \Bigl(\sum_{i=1}^{n} a_{ij} v_i\Bigr) = \sum_{j=1}^{m} \sum_{i=1}^{n} \alpha_j a_{ij} v_i = \sum_{i=1}^{n} \Bigl(\sum_{j=1}^{m} a_{ij} \alpha_j\Bigr) v_i. \tag{2}$$

Now assume that $Tu = v$ has the coordinate representation

$$Tu = v = \sum_{i=1}^{n} \beta_i v_i. \tag{3}$$

Since $\{v_1, \dots, v_n\}$ is a basis of V, the coefficients in the representation are unique. By comparing the right hand sides of (2) and (3), we obtain that

$$\sum_{j=1}^{m} a_{ij}\alpha_j = \beta_i \qquad \text{for } i = 1, \dots, n. \tag{4}$$

Now collect the coefficients a_{ij} in an $(m \times n)$-dimensional matrix A. Recalling the definition of matrix-vector multiplication, we then see that the system of equations (4) translates into the matrix-vector system

$$A\boldsymbol{\alpha} = \boldsymbol{\beta}.$$

In other words, the matrix $A \in \mathbb{K}^{m\times n}$ with entries a_{ij} defined by the relation (1) represents the linear transformation T in the bases $\mathcal{B}$ and $\mathcal{C}$.

Example 1.4.1. Consider the mapping $T\colon \mathcal{P}_3 \to \mathcal{P}_2$, $p \mapsto Tp := p'$ (that is, differentiation of 3rd degree polynomials). In $\mathcal{P}_3$ we may choose the monomial basis $\mathcal{B} = \{1, x, x^2, x^3\}$, and in $\mathcal{P}_2$ the essentially same basis $\mathcal{C} = \{1, x, x^2\}$ (though we are missing the x^3 term, as we are only considering 2nd degree polynomials there).

In order to find the matrix representation A of T in the monomial basis, we apply the operation T to the basis elements in $\mathcal{B}$ and write the result in terms of the elements of $\mathcal{C}$. We have

$$\begin{aligned} T1 &= 0 = 0\cdot 1 + 0\cdot x + 0\cdot x^2, \\ Tx &= 1 = 1\cdot 1 + 0\cdot x + 0\cdot x^2, \\ Tx^2 &= 2x = 0\cdot 1 + 2\cdot x + 0\cdot x^2, \\ Tx^3 &= 3x^2 = 0\cdot 1 + 0\cdot x + 3\cdot x^2, \end{aligned}$$

The column entries of the matrix A can then be directly read off from the coefficients on the right hand side. For instance, the third column corresponds to the result of Tx^2, where we have the coefficients $(0, 2, 0)$. In total, we obtain the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}.$$

Now we consider again the general case of a linear transformation $T\colon U \to V$, but we assume that we have a decomposition of the sub-spaces U and V as direct sums, and are given bases of the corresponding subspaces. That is, we assume that $U = U_1 \oplus U_2$ and $V = V_1 \oplus V_2$, and that we have bases $\mathcal{B}_1 = \{u_1, \dots, u_k\}$ of U_1, $\mathcal{B}_2 = \{u_{k+1}, \dots, u_m\}$ of U_2, $\mathcal{C}_1 = \{v_1, \dots, v_\ell\}$ of V_1, and $\mathcal{C}_2 = \{v_{\ell+1}, \dots, v_n\}$ of V_2. Let now $u \in U$ and $v \in V$ such that $Tu = v$ holds. Then we can decompose u uniquely as

$$u = u^{(1)} + u^{(2)} \qquad \text{with } u^{(1)} \in U_1 \text{ and } u^{(2)} \in U_2,$$

and we have a similar decomposition

$$v = v^{(1)} + v^{(2)} \qquad \text{with } v^{(1)} \in V_1 \text{ and } v^{(2)} \in V_2.$$

Our next goal is to investigate how this decomposition translates into the matrix of the transformation. For that, we note that we can compute a matrix representation of the transformation T in the same way as above after having concatenated

$\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ into bases of the whole spaces U and V. Now we can split up the resulting matrix in to sub-blocks

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

corresponding to the decomposition of the spaces U and V. That is, $A_{11} \in \mathbb{K}^{\ell \times k}$, $A_{12} \in \mathbb{K}^{\ell \times (m-k)}$, $A_{21} \in \mathbb{K}^{(n-\ell) \times k}$, and $A_{22} \in \mathbb{K}^{(n-\ell) \times (m-k)}$.

If we now denote by $\boldsymbol{\alpha}^{(j)}$, $j = 1, 2$, the coefficients of $u^{(j)}$ with respect to the basis $\mathcal{B}_j$, and by $\boldsymbol{\beta}^{(i)}$, $i = 1, 2$, the coefficients of $v^{(i)}$ with respect to $\mathcal{C}_i$, then we obtain the relation

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{pmatrix} \qquad \text{or} \qquad \begin{aligned} A_{11}\boldsymbol{\alpha}^{(1)} + A_{12}\boldsymbol{\alpha}^{(2)} &= \boldsymbol{\beta}^{(1)}, \\ A_{21}\boldsymbol{\alpha}^{(1)} + A_{22}\boldsymbol{\alpha}^{(2)} &= \boldsymbol{\beta}^{(2)}. \end{aligned}$$

That is, a decomposition of the spaces U and V into direct sums translates into a block decomposition of the matrix A. Moreover, the block A_{11} describes the part of the transformation that maps U_1 into V_1, A_{12} describes the part mapping U_2 into V_1, and so on.

Now consider in particular the case of an operator $T\colon V \to V$, where we have a decomposition $V = V_1 \oplus V_2$, and consider again a corresponding matrix in block form as described above. In addition, assume that V_1 is a T-invariant subspace, that is, $Tv \in V_1$ whenever $v \in V_1$. Now denote again the coordinates of v and Tv by $(\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)})$ and $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$, respectively. Then the condition $v \in V_1$ means that $\boldsymbol{\alpha}^{(2)} = 0$ while $\boldsymbol{\alpha}^{(1)}$ can be arbitrary, whereas the condition $Tv \in V_1$ means that $\boldsymbol{\beta}^{(2)} = 0$ while again $\boldsymbol{\beta}^{(1)}$ can be arbitrary. Now recall the relation $A_{21}\boldsymbol{\alpha}^{(1)} + A_{22}\boldsymbol{\alpha}^{(2)} = \boldsymbol{\beta}^{(2)}$ that we have derived above. The T-invariance of V_1 then means that

$$A_{21}\boldsymbol{\alpha}^{(1)} = 0 \qquad \text{for all } \boldsymbol{\alpha}^{(1)} \in \mathbb{K}^{(n-k) \times k},$$

which is only possible if $A_{21} = 0$ and thus

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}.$$

Conversely, if $A_{21} = 0$, then V_1 will be a T-invariant subspace of V.

Thus we get the following result:

Lemma 1.4.2. *Assume that $T\colon V \to V$ is a linear transformation and that we have a decomposition $V = V_1 \oplus V_2$. Then V_1 is T-invariant if and only if the corresponding matrix A is block-upper-triangular.*

Even more, if in addition the space V_2 is T-invariant, then a similar argumentation as above shows that the sub-matrix A_{12} has to be equal to 0, and thus A has to have a block-diagonal form.

Proposition 1.4.3. *Assume that $T\colon V \to V$ is a linear transformation and that we have a decomposition $V = V_1 \oplus \ldots V_N$ such that every subspace V_j is T-invariant. Choose a basis of each of the subspaces V_j and denote by A the corresponding matrix representation of T. Then A is block diagonal, that is, A has the form*

$$A = \begin{pmatrix} A_{11} & 0 & \ldots & 0 \\ 0 & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & A_{NN} \end{pmatrix},$$

with $A_{jj} \in \mathbb{K}^{\dim(V_j) \times \dim(V_j)}$ for $j = 1, \ldots, N$.

5. Eigenvalues and Eigenspaces

DEFINITION 1.5.1. Assume that V is a vector space and $T\colon V \to V$ is a linear transformation. Then $\lambda \in \mathbb{K}$ is an *eigenvalue* of T, if there exists $v \in V$ with $v \neq 0$ such that

$$Tv = \lambda v.$$

Such a vector v is called an *eigenvector* of T (for the eigenvalue λ).

The *eigenspace* $E(\lambda, T)$ of T for the eigenvalue λ is defined as

$$E(\lambda, T) := \ker(T - \lambda I) = \bigl\{v \in V : Tv = \lambda v\bigr\}.$$

EXAMPLE 1.5.2. Consider the transformation $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA := A^T$. If $A \in \mathbf{Sym}_n(\mathbb{K})$ is symmetric, then $TA = A^T = A$, that is, 1 is an eigenvalue of T, and every symmetric matrix with exception of the 0-matrix is a corresponding eigenvector.

Similarly, if $A \in \mathbf{Skew}_n(\mathbb{K})$ is skew-symmetric, then $TA = A^T = -A$, and thus -1 is an eigenvalue as well, and the corresponding eigenvectors are all non-zero skew-symmetric matrices.

LEMMA 1.5.3. *Assume that V is a vector space and $T\colon V \to V$ is linear. Assume moreover that $U \subset V$ is a one-dimensional subspace of V. Then U is T-invariant, if and only if there exists an eigenvector v of T with $U = \operatorname{span}(\{v\})$.*

PROOF. Assume first that U is T-invariant. Since U is one-dimensional, we can write $U = \operatorname{span}(\{v\})$ for some $v \neq 0$ (with $\{v\}$ being a basis of U). Then $Tv \in U = \operatorname{span}(\{v\})$, which implies that there exists $c \in \mathbb{K}$ with $Tv = cv$. Thus v is an eigenvector of T (for the eigenvalue c).

Conversely, assume that v is an eigenvector of T for the eigenvalue $\lambda \in \mathbb{K}$ and that $U = \operatorname{span}(\{v\})$. Let $u \in U$. Then there exists $c \in \mathbb{K}$ with $u = cv$. Since v is an eigenvector of T we have that

$$Tu = T(cv) = c\,Tv = c\lambda\, v \in \operatorname{span}(\{v\}) = U.$$

Thus U is T-invariant. □

For the following characterisation of eigenvalues, we have to recall some basic notation of linear algebra:

Assume that U, V are vector spaces, and that $T\colon U \to V$ is a linear transformation.

- The transformation T is *surjective* if $\operatorname{ran}(T) = V$.

  Equivalently, T is surjective, if the equation $Tu = v$ has for every $v \in V$ *at least* one solution $u \in U$.
- The transformation T is *injective*, if for all $v \in V$, the equation $Tu = v$ has *at most* one solution.

  This is (for linear transformations) equivalent to the equality $\ker(T) = \{0\}$.
- The transformation T is *bijective*, if it is injective and surjective. That is, for all $v \in V$ the equation $Tu = v$ has *precisely* one solution u.

  In this case, we can define an *inverse transformation* $T^{-1}\colon V \to U$, which maps $v \in V$ to the unique solution u of the equation $Tu = v$. One can show[3] that the inverse T^{-1} is linear, if T is linear.

LEMMA 1.5.4. *Assume that U, V are vector spaces, U is finite dimensional, and that $T\colon U \to V$ is bijective. Assume moreover that $\mathcal{B} = \{u_1, \dots, u_n\}$ is a*

[3]Try to do this!

basis of U. Then $T\mathcal{B} = \{Tu_1, \dots, Tu_n\}$ is a basis of V. In particular, V is finite dimensional and $\dim(U) = \dim(V)$.

Proof. Assume that $v \in V$. Since T is surjective, we have that $V = \operatorname{ran}(T)$, and thus there exists $u \in U$ with $v = Tu$. Since $\mathcal{B}$ is a basis of U we can write $u = c_1u_1 + \dots + c_nu_n$ for some $c_n \in \mathbb{K}$. As a consequence,

$$v = Tu = c_1Tu_1 + \dots + c_nTu_n.$$

This shows that $V = \operatorname{span}\{Tu_1, \dots, Tu_n\}$.

It remains to show that the set $\{Tu_1, \dots, Tu_n\}$ is linearly independent. Assume to that end that

$$0 = c_1Tu_1 + \dots + c_nTu_n.$$

We can rewrite this as

$$0 = T\big(c_1u_1 + \dots + c_nu_n\big),$$

that is,

$$c_1u_1 + \dots + c_nu_n \in \ker(T).$$

Since T is injective, we have that $\ker(T) = \{0\}$, and thus

$$c_1u_1 + \dots + c_nu_n = 0.$$

Now the linear independence of the set $\{u_1, \dots, u_n\}$ implies that $c_j = 0$ for all j, which in turn proves the linear independence of the set $\{Tu_1, \dots, Tu_n\}$. □

Theorem 1.5.5. *Assume that U, V are vector spaces, U is finite dimensional, and that $T\colon U \to V$ is linear. Then*

$$\dim U = \dim(\ker T) + \dim(\operatorname{ran} T).$$

Proof. By Lemma 1.3.9, there exists a subspace $W \subset U$ such that $U = (\ker T) \oplus W$. Now consider the transformation $S\colon W \to \operatorname{ran} T$, $w \mapsto Sw := Tw$. That is, S is the restriction of T to W, seen as a mapping from W to $\operatorname{ran} T$. We will show that S is a bijection.

We note first that $\ker S = W \cap \ker T = \{0\}$, as the sum of W and $\ker T$ is direct. Thus S is injective. In order to show that S is surjective, assume that $v \in \operatorname{ran} T$. Then there exists $u \in U$ with $v = Tu$. Now decompose $u = u_0 + w$ with $u_0 \in \ker T$ and $w \in W$. Then $v = Tu = Tu_0 + Tw = 0 + Sw = Sw$. Thus $v \in \operatorname{ran} S$, which shows that S is surjective as well.

As a consequence, we have that $\dim(W) = \dim(\operatorname{ran} T)$. Next, since the sum of $\ker T$ and W is direct, we have that $\dim U = \dim(\ker T) + \dim(W)$, and thus

$$\dim U = \dim(\ker T) + \dim(W) = \dim(\ker T) + \dim(\operatorname{ran} T).$$

□

Theorem 1.5.6. *Assume that V is a finite dimensional vector space and that $T\colon V \to V$ is linear. The following are equivalent:*

(1) T is bijective.
(2) T is injective.
(3) T is surjective.

Proof. The implications (1) $\implies$ (2) and (1) $\implies$ (3) hold trivially.

We now show the implication (2) $\implies$ (3). Since T is injective and thus $\ker T = \{0\}$, we have

$$\dim(V) = \dim(\ker T) + \dim(\operatorname{ran} T) = 0 + \dim(\operatorname{ran} T) = \dim(\operatorname{ran} T).$$

Thus $\operatorname{ran} T$ is a subspace of V of the same dimension as V. This already implies that $V = \operatorname{ran} T$ and thus T is surjective.

Finally we show the implication (3) $\implies$ (1). Here it remains to show that T is injective. Since T is surjective, we have that $V = \operatorname{ran} T$ and in particular $\dim V = \dim(\operatorname{ran} T)$. Thus

$$\dim(V) = \dim(\ker T) + \dim(\operatorname{ran} T) = \dim(\ker T) + \dim(V),$$

which implies that $\dim(\ker T)$. This is only possible if $\ker T = \{0\}$, that is, T is injective. $\square$

DEFINITION 1.5.7. Let V be a vector space. We denote by $\mathrm{Id}_V \colon V \to V$ the identity operator on V given by $\mathrm{Id}_V\, v = v$ for all $v \in V$. If the space V is clear from the context, we omit the subscript $_V$ and write Id instead.

THEOREM 1.5.8. *Assume that V is finite dimensional, $T \colon V \to V$ is a linear transformation, and that $\lambda \in \mathbb{K}$.*

The following are equivalent:

(1) λ is an eigenvalue of T.
(2) $T - \lambda\,\mathrm{Id}$ is not injective.
(3) $T - \lambda\,\mathrm{Id}$ is not surjective.
(4) $T - \lambda\,\mathrm{Id}$ is not bijective.

PROOF. The equivalences (2) $\iff$ (3) $\iff$ (4) follow from Theorem 1.5.6.

Moreover, we have that λ is an eigenvalue of T, if and only if there exists $v \neq 0$ with $Tv = \lambda v$. This is in turn equivalent to stating that there exists $v \neq 0$ such that $(T - \lambda\,\mathrm{Id})v = 0$, which in turn is equivalent to the non-injectivity of $T - \lambda\,\mathrm{Id}$. This proves the equivalence (1) $\iff$ (2). $\square$

6. Existence of Eigenvalues

DEFINITION 1.6.1. Assume that U, V, W are sets and that $T \colon U \to V$ and $S \colon V \to W$ are mappings. Then the composition of S and T is the mapping $S \circ T \colon U \to W$ defined as

$$S \circ T(u) := S\big(T(u)\big) \qquad \text{for all } u \in U.$$

LEMMA 1.6.2. *Assume that U, V, W are vector spaces and that $T \colon U \to V$ and $S \colon V \to W$ are linear transformations. Then the composition $S \circ T$ is linear as well.*

PROOF. Let u, $v \in U$, and λ, $\mu \in \mathbb{K}$. Then

$$\begin{aligned} S \circ T(\lambda u + \mu v) = S\big(T(\lambda u + \mu v)\big) &\underbrace{=}_{T \text{ linear}} S\big(\lambda\, Tu + \mu\, Tv\big) \\ &\underbrace{=}_{S \text{ linear}} \lambda\, S(Tu) + \mu\, S(Tv) = \lambda\, S \circ T(u) + \mu\, S \circ T(v). \end{aligned}$$

$\square$

If we have in particular the situation that $T \colon V \to V$ is a linear transformation of the space V, we can define the composition of T with itself. In order to simplify notation, we abbreviate $T^2 := T \circ T$, and similarly $T^3 = T \circ T \circ T$, and so on. This further allows us to evaluate polynomials in linear transformations:

DEFINITION 1.6.3. Assume that V is a vector space, $T \colon V \to V$ is a linear transformation, and p is a polynomial with coefficients in $\mathbb{K}$, say

$$p(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n.$$

We define the mapping $p(T) \colon V \to V$ by

$$p(T)u := c_0 u + c_1 Tu + c_2 T^2 u + \ldots c_n T^n u \qquad \text{for all } u \in V.$$

PROPOSITION 1.6.4. *Assume that V is a vector space, $T: V \to V$ is linear, and p, q are polynomials with coefficients in $\mathbb{K}$. Denote by $r(x) := p(x)q(x)$ the product of p and q. Then*

$$r(T) = p(T) \circ q(T) = q(T) \circ p(T).$$

PROOF. Exercise. □

THEOREM 1.6.5. *Assume that V is a finite dimensional complex vector space, $V \neq \{0\}$, and that $T: V \to V$ is a linear transformation. Then T has at least one eigenvalue $\lambda \in \mathbb{C}$.*

PROOF. Denote by $n := \dim(V)$ the dimension of V. Let $v \in V$ be arbitrary with $v \neq 0$ and consider the vectors

$$v, Tv, T^2v, \ldots, T^nv.$$

These are $n+1$ vectors in an n-dimensional spaces, which implies that they cannot be linearly independent. That is, there exist numbers $c_0, \ldots, c_n \in \mathbb{K}$, not all of them equal to 0, such that

$$c_0v + c_1Tv + c_2T^2v + \ldots + c_nT^nv = 0. \tag{5}$$

Define now the polynomial

$$p(x) := c_0 + c_1x + \ldots + c_nx^n.$$

Then we can write (5) as

$$p(T)v = 0.$$

Now let m be the largest index for which $c_m \neq 0$ (it is possible that $c_n = 0$, so it can happen that $m < n$). Then we can factorise the polynomial p as

$$p(x) = c_m(x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_m)$$

with $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$ being the complex zeroes of p (possible with higher multiplicities). Thus, after dividing by $c_m \neq 0$, the equation (5) can be further written as

$$(T - \lambda_1 \operatorname{Id}) \circ (T - \lambda_2 \operatorname{Id}) \circ \ldots \circ (T - \lambda_m \operatorname{Id})v = 0.$$

Since $v \neq 0$ by assumption, this implies that the composition $(t - \lambda_1 \operatorname{Id}) \circ \ldots \circ (T - \lambda_m \operatorname{Id})$ cannot be injective (as v in its kernel). This, however, is only possible if at least one of the factors $T - \lambda_j \operatorname{Id}$ is not injective. This, in turn, is equivalent to λ_j being an eigenvalue of T. □

REMARK 1.6.6. Note that the proof breaks down if we consider real vector spaces, as in this case the factorisation of the polynomial is no longer possible (and polynomials need not have real zeroes). Indeed, there exist linear transformations in real vector spaces that do not have (real!) eigenvalues, the simplest example being that of a rotation in $\mathbb{R}^2$ (by an angle that is not an integer multiple of π).

7. Triangularisation

LEMMA 1.7.1. *Assume that the matrix*

$$B = \begin{pmatrix} \beta_1 & * & \ldots & * \\ 0 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \ldots & 0 & \beta_n \end{pmatrix} \in \mathbf{Mat}_n(\mathbb{K})$$

is upper triangular. Then the matrix B is invertible, if and only if all diagonal entries $\beta_1, \ldots, \beta_n$ are different from zero.

Idea of proof. The matrix B is invertible, if and only if its columns are linearly independent. If all diagonal entries are different from zero, this is clearly the case (if you want, you can prove this by induction over n).

Conversely, if one of the diagonal entries is equal to zero, say $\beta_j = 0$, then the first j columns cannot be linearly independent, as only the first $j-1$ entries of each of these columns are possibly different from zero. □

Proposition 1.7.2. *Assume that V is a finite dimensional vector space and that $T\colon V \to V$ is linear. Assume moreover that V has a basis with respect to which that matrix representation A of T is upper triangular. Then the eigenvalues of T are precisely the distinct diagonal entries of A.*

Proof. The matrix A is upper triangular, and thus it has the form

$$A = \begin{pmatrix} \alpha_1 & * & \dots & * \\ 0 & \alpha_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \alpha_n \end{pmatrix}$$

for some $\alpha_j \in \mathbb{K}$, $j = 1, \dots, n$. Now note that the matrix representation of $T - \lambda \operatorname{Id}$ in the same basis is the matrix

$$A - \lambda \operatorname{Id} = \begin{pmatrix} \alpha_1 - \lambda & * & \dots & * \\ 0 & \alpha_2 - \lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \dots & 0 & \alpha_n - \lambda \end{pmatrix}.$$

By Lemma 1.7.1, this matrix is invertible if and only if its diagonal entries are non-zero.

Now recall that λ is an eigenvalue of T, if and only if the transformation $T - \lambda \operatorname{Id}$ is not bijective. This is the case, if and only if its matrix representation (in any basis) is not invertible, which, as we have just seen, is the case if and only if $\lambda = \alpha_j$ for some j, which proves the assertion. □

Theorem 1.7.3. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. Then there exists a basis of V with respect to which the matrix representation of T is upper triangular.*

Proof. We prove this result by induction over $n := \dim V$. For $n = 1$, this result is trivial, as every (1×1)-matrix is automatically upper triangular (and even diagonal!).

Assume therefore that we have proven the claim for every vector space U with $\dim U < n$ and every mapping $S\colon U \to U$.

Let λ be an eigenvalue of V (the existence of which follows from Theorem 1.6.5). If $T = \lambda \operatorname{Id}$, then the result holds trivially, as the matrix of T with respect to *any* basis is simply the matrix $\lambda \operatorname{Id}$. Thus we may assume without loss of generality that $T \neq \lambda \operatorname{Id}$ Define now $U := \operatorname{ran}(T - \lambda \operatorname{Id})$. Since λ is an eigenvalue of T, it follows that $T - \lambda \operatorname{Id}$ is not surjective, which implies that $\dim(U) < \dim(V)$. Since $T - \lambda \operatorname{Id} \neq 0$, we also obtain that $\dim(U) \geq 1$.

We now want to use our induction assumption on the space U. For that, we show first that U is T-invariant: Indeed, assume that $u \in U$. Then there exists some $v \in V$ with $(T - \lambda \operatorname{Id})v = u$. Thus

$$Tu = T(T - \lambda \operatorname{Id})v = (T - \lambda \operatorname{Id})Tv \in \operatorname{ran}(T - \lambda \operatorname{Id}) = U.$$

Here the second equality follows from the computation rules for polynomials in Proposition 1.6.4.

Define now the transformation $S\colon U \to U$, $u \mapsto Su := Tu$. That is, the mapping S is the restriction $T|_U$ of T to U, but seen as a mapping from U to itself (which is possible because U is T-invariant). Thus there exists a basis $\{u_1, \ldots, u_m\}$ of U for which the matrix representation of S is upper triangular.

Now let $W \subset V$ be such that $V = U \oplus W$, and choose any basis $\{w_1, \ldots, w_{n-m}\}$ of W. We now consider the matrix representation A of T with respect to the basis $\{u_1, \ldots, u_m\} \cup \{w_1, \ldots, w_{n-m}\}$. Following our considerations from Section 4, the fact that U is T-invariant implies that A has a block upper triangular structure

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}.$$

Moreover, the matrix A_{11} is precisely the matrix representation of S with respect to the basis $\{u_1, \ldots, u_m\}$ and thus upper triangular. Finally, we have for all w_j that

$$Tw_j = (Tw_j - \lambda w_j) + \lambda w_j = (T - \lambda \operatorname{Id})w_j + \lambda w_j.$$

Since $(T - \lambda \operatorname{Id})w_j \in \operatorname{ran}(T - \lambda \operatorname{Id}) = U$, it follows that the matrix A_{22} (which represents the part of T that maps W into W) is $A_{22} = \lambda \operatorname{Id}$. Thus the matrix A as a whole is upper triangular. □

8. Diagonalisation

DEFINITION 1.8.1. Let V be a finite dimensional vector space and let $T\colon V \to V$ be linear. We say that T is *diagonalisable* if there exists a basis of V such that the corresponding matrix representation of T is diagonal.

In the following, we will derive different equivalent conditions for the diagonalisability of a linear operator. For that, we show first that eigenvectors for different eigenvalues are necessarily linearly independent.

THEOREM 1.8.2. *Assume that V is a vector space and $T\colon V \to V$ is linear. Let moreover $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ be* distinct *eigenvalues (that is, $\lambda_i \neq \lambda_j$ for $i \neq j$), and let $v_1, \ldots, v_m \in V$ be corresponding eigenvectors. Then the set $\{v_1, \ldots, v_m\}$ is linearly independent.*

PROOF. We apply induction over m. For $m = 1$, the result is trivial, as every eigenvector is non-zero.

Assume now that we have shown the result for $m - 1$. That is, the set $\{v_1, \ldots, v_{m-1}\}$ is linearly independent.

Let now $c_1, \ldots, c_m \in \mathbb{K}$ be such that

$$0 = c_1 v_1 + \ldots + c_m v_m.$$

We have to show that $c_j = 0$ for all $j = 1, \ldots, m$.

First we consider the case where $c_m = 0$. In this case, we have that

$$0 = c_1 v_1 + \ldots + c_{m-1} v_{m-1}.$$

Since by our induction assumption, the set $\{v_1, \ldots, v_{m-1}\}$ is linearly independent, it follows that the coefficients $c_1, \ldots, c_{m-1}$ are all equal to zero as well, which proves the claim in this case.

Now assume that $c_m \neq 0$. Defining $d_j := -c_j/c_m$ for $j = 1, \ldots, m-1$, we then obtain the equation

$$v_m = d_1 v_1 + \ldots + d_{m-1} v_{m-1}. \tag{6}$$

If we apply the transformation T to (6) we then obtain that

$$\begin{aligned}\lambda_m v_m = Tv_m &= T\big(d_1 v_1 + \ldots + d_{m-1} v_{m-1}\big)\\ &= d_1 T v_1 + \ldots + d_{m-1} T v_{m-1} = d_1 \lambda_1 v_1 + \ldots + d_{m-1}\lambda_{m-1} v_{m-1}.\end{aligned}$$

If we now multiply (6) by λ_m and subtract this last equation, we obtain that

$$0 = d_1(\lambda_m - \lambda_1)v_1 + \ldots + d_{m-1}(\lambda_m - \lambda_{m-1})v_{m-1}.$$

Now the assumed linear independence of the set $\{v_1, \ldots, v_{m-1}\}$ implies that all the coefficients $d_j(\lambda_m - \lambda_j)$, $j = 1, \ldots, m-1$, are equal to zero. Since $\lambda_m \neq \lambda_j$ for $j = 1, \ldots, m-1$, this implies that also $d_j \neq 0$. Finally, this shows that $c_j = -c_m d_j \neq 0$, which concludes the proof. □

As a consequence of this result, we obtain that the sum of eigenspaces is direct.

LEMMA 1.8.3. *Assume that $T\colon V \to V$ is linear and that $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ are distinct eigenvalues of T. Then the sum of the eigenspaces $E(\lambda_1, T), \ldots, E(\lambda_m, T)$ is direct.*

PROOF. Let $v \in E(\lambda_1, T) + \ldots + E(\lambda_m, T)$. Assume that we can write v in two ways as

$$\begin{aligned}v &= u_1 + \ldots + u_m,\\ v &= v_1 + \ldots + v_m,\end{aligned}$$

such that u_j, $v_j \in E(\lambda_j, T)$ for all $j = 1, \ldots, m$. We have to show that in this case $u_j = v_j$ for all j.

By subtracting the two representations of v from each other, we see that

$$0 = (u_1 - v_1) + \ldots + (u_m - v_m) \tag{7}$$

with $(u_j - v_j) \in E(\lambda_j, T)$ for all $j = 1, \ldots, m$. Now recall that we shown in Theorem 1.8.2 that eigenvectors for distinct eigenvalues are linearly independent. Thus (7) can only hold if the vectors $(u_j - v_j)$ are *not* eigenvectors of T, which is only possible if they are equal to zero. □

THEOREM 1.8.4. *Assume that V is a finite dimensional vector space and that $T\colon V \to V$ is linear. Denote by $\lambda_1, \ldots, \lambda_m \in \mathbb{K}$ the distinct eigenvalues of T. The following are equivalent:*

(1) T is diagonalisable.
(2) V has a basis of eigenvectors of T.
(3) We can decompose $V = U_1 \oplus U_2 \oplus \ldots \oplus U_n$ such that each space U_j is a one-dimensional T-invariant subspace of V.
(4) $V = E(\lambda_1, T) \oplus \ldots \oplus E(\lambda_m, T)$.
(5) $\dim V = \dim\big(E(\lambda_1, T)\big) + \ldots + \dim\big(E(\lambda_m, T)\big)$.

PROOF. It is easy to see that the matrix A with respect to a basis $v_1, \ldots, v_n$ of V is diagonal, if and only if there exist $\mu_i \in \mathbb{K}$, $i = 1, \ldots, n$, (the diagonal entries of A) such that $Tv_i = \mu_i v_i$ for all i. This shows the equivalence (1) $\iff$ (2). Also, the equivalence (1) $\iff$ (3) follows from the considerations in Section 4.

Next we show the equivalence (2) $\iff$ (4): If V has a basis of eigenvectors, then necessarily V is the sum of the eigenspaces of T. Moreover, this sum is direct by Lemma 1.8.3. Conversely, if (4) holds, we can choose a basis of each eigenspace, and concatenate the different bases to a basis of V.

The implication (4) $\implies$ (5) follows from Proposition 1.3.8. Finally, assume that (5) holds. By Lemma 1.8.3, the sum of the eigenspaces is direct, and thus, by Proposition 1.3.8,

$$\dim\big(E(\lambda_1, T) + \ldots + E(\lambda_m, T)\big) = \dim\big(E(\lambda_1, T)\big) + \ldots + \dim\big(E(\lambda_m, T)\big) = \dim V.$$

Since we already know that the sum is direct, it follows that

$$V = E(\lambda_1, T) \oplus \ldots \oplus E(\lambda_m, T).$$

□

In particular, we obtain the following result concerning diagonalisability of operators:

COROLLARY 1.8.5. *Assume that* $\dim V = n$ *and that the linear transformation* $T\colon V \to V$ *has* n *distinct eigenvalues. Then* T *is diagonalisable.*

9. Generalised Eigenspaces

In the previous section, we have introduce the notion of diagonalisability of a linear transformation. Moreover, we have seen that a transformation is diagonalisable if and only if it has a basis of eigenvectors. Unfortunately, it turns out that this is not the case for all transformations. As an example, consider the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Since this matrix is upper triangular, we can read off the eigenvalues from its diagonal, which is 0. Thus the only eigenvalue of A is 0. However, the matrix A cannot be diagonalisable: If it were, it could be transformed by a basis change into a diagonal matrix where all the diagonal entries are equal to 0; in other words, the matrix would have be 0.

With the same argumentation we obtain that strictly upper triangular matrices cannot be diagonalisable, unless they are already equal to 0. Moreover, the same holds if we add the same constant λ to the diagonal: A matrix of the form

$$A = \begin{pmatrix} \lambda & * & \ldots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \ldots & 0 & \lambda \end{pmatrix} \tag{8}$$

can only be diagonalisable, if all the off-diagonal elements are equal to 0 and $A = \lambda\,\mathrm{Id}$.

In this section, we will show that these is in a sense the most general form of a non-diagonalisable matrix. More precisely, we will show that every linear operator T on a finite dimensional, complex vector space has a basis such that the corresponding matrix of T has a block-diagonal form where every block is of the form (8). In order to arrive at this representation, we will have to introduce some notation, though, and then do some hard work.

Nilpotent Transformations.

DEFINITION 1.9.1. Assume that V is a vector space and $T\colon V \to V$ is a linear transformation. We say that T is *nilpotent*, if there exists $j \in \mathbb{N}$ such that $T^j = 0$.

It is easy to show that nilpotent transformations can have no other eigenvalue than 0:

LEMMA 1.9.2. *Assume that* V *is a vector space and* $T\colon V \to V$ *is a nilpotent linear transformation. Then the only eigenvalue of* T *is* 0.

PROOF. Assume that $\lambda \in \mathbb{K}$ is an eigenvalue of T, and let $v \in V$, $v \neq 0$, be a corresponding eigenvector. Then $Tv = \lambda v$. Moreover, since T is nilpotent, there exists $j \in \mathbb{N}$ such that $T^j = 0$, and in particular $T^j v = 0$. Thus we have that

$$0 = T^j v = T^{j-1} T v = T^{j-1} \lambda v = \lambda T^{j-1} v = \ldots = \lambda^j v.$$

Since $v \neq 0$, this implies that $\lambda = 0$. □

LEMMA 1.9.3. *Assume that the matrix $A \in \mathbf{Mat}_n(\mathbb{K})$ is* strictly *upper triangular, that is, its entries a_{ij} satisfy $a_{ij} = 0$ for all $i \geq j$. Then $A^n = 0$.*

PROOF. This is a simple calculation. □

PROPOSITION 1.9.4. *Assume that V is a complex finite dimensional vector space and that $T : V \to V$ is nilpotent. Then*

$$T^{\dim(V)} = 0.$$

PROOF. By Theorem 1.7.3, we can find a basis of V such that the matrix A of T is upper triangular. Next we know from Lemma 1.9.2 that the only eigenvalue of T is 0. Thus all the diagonal elements of A (which are the eigenvalues of T) are 0, which means that the matrix A is actually strictly upper triangular. Now the claim follows from Lemma 1.9.3. □

Now consider the situation where the operator T has a single eigenvalue λ. After choosing a suitable basis, we can then write the matrix A of T in the form given in (8). This, however, implies that the matrix $A - \lambda \operatorname{Id}$ is nilpotent and thus $(A - \lambda \operatorname{Id})^n = 0$. Now note that the matrix $(A - \lambda \operatorname{Id})^n$ is precisely the matrix of the operator $(T - \lambda \operatorname{Id})^n$ in the same basis. Thus the operator $T - \lambda \operatorname{Id}$ is nilpotent and every element v in the vector space V satisfies $(T - \lambda \operatorname{Id})^n v = 0$. In particular, there exists a basis of V consisting of vectors that satisfy $(T - \lambda \operatorname{Id})^n v = 0$.

DEFINITION 1.9.5. Let V be a vector space, let $T : V \to V$ be linear and $\lambda \in \mathbb{K}$. A vector $v \in V$, $v \neq 0$, is called a *generalised eigenvector* of T for the eigenvalue λ, if there exists $j \in \mathbb{N}$ such that

$$(T - \lambda \operatorname{Id})^j v = 0.$$

The linear span of all generalised eigenvectors of T for the eigenvalue λ is called the *generalised eigenspace* of T for λ, and denoted by $G(\lambda, T)$.

REMARK 1.9.6. It is easy to see that the generalised eigenspace for the eigenvalue λ consists of all the generalised eigenvectors for λ together with the vector 0.

REMARK 1.9.7. If $v \neq 0$ satisfies the equation $(T - \lambda \operatorname{Id})^j v = 0$, then λ necessarily has to be an eigenvalue of T: Let k be the largest non-negative integer such that $w := (T - \lambda \operatorname{Id})^k v \neq 0$. Then

$$(T - \lambda \operatorname{Id}) w = (T - \lambda \operatorname{Id})^{k+1} v = 0,$$

and thus $Tw = \lambda w$, which in particular shows that λ is an eigenvalue (and w a corresponding eigenvector).

Preparatory Technical Results.

We will in the following show that every vector space has a basis consisting of generalised eigenvectors. For that, we will need several additional technical results.

LEMMA 1.9.8. *Let V be a vector space, let $T : V \to V$ be linear, and let p be a polynomial. Then the subspaces $\ker(p(T))$ and $\operatorname{ran}(p(T))$ are T-invariant.*

PROOF. Assume that $u \in \ker(p(T))$. Then $p(T) = 0$ and thus $p(T)(Tu) = T(p(T)u) = T(0) = 0$, which shows that also $Tu \in \ker(p(T))$. Thus $\ker(p(T))$ is T-invariant.

Now assume that $u \in \operatorname{ran}(p(T))$, say $u = p(T)v$ for some $v \in V$. Then $Tu = T(p(T)v) = p(T)(Tv) \in \operatorname{ran}(p(T))$, which shows that $\operatorname{ran}(p(T))$ is T-invariant as well. □

LEMMA 1.9.9. *Assume that V is a vector space and that $T\colon V \to V$ is linear. Then*

$$\{0\} \subset \ker T \subset \ker T^2 \subset \ker T^3 \subset \dots$$

PROOF. Let $k \in \mathbb{N}$ and let $u \in \ker T^k$. Then $T^k u = 0$ and thus $T^{k+1}u = T(T^k u) = T(0) = 0$ as well, which shows that $u \in \ker T^{k+1}$. This proves that $\ker T^k \subset \ker T^{k+1}$ for all k. $\square$

LEMMA 1.9.10. *Assume that V is a vector space and that $T\colon V \to V$ is linear. Assume moreover that for some $k \in \mathbb{N}$ we have that $\ker T^k = \ker T^{k+1}$. Then $\ker T^{k+m} = \ker T^k$ for all $m \geq 0$.*

PROOF. We prove this result by induction over m. For $m = 0$, this result is trivial.

Now assume that $\ker T^{k+m} = \ker T^k$. We have to show that also $\ker T^{k+m+1} = \ker T^k$. By Lemma 1.9.9 we know that $\ker T^k \subset \ker T^{k+m+1}$. Thus we only have to show the inclusion $\ker T^{k+m+1} \subset \ker T^k$. Assume therefore that $u \in \ker T^{k+m+1}$. Then $0 = T^{k+m+1}u = T^{k+1}(T^m u)$, and thus $T^m u \in \ker T^{k+1}$. Since $\ker T^{k+1} = \ker T^k$, we then obtain that $T^m u \in \ker T^k$, which in turn shows that $T^{k+m}u = T^k(T^m u) = 0$. Thus $u \in \ker T^{k+m}$, which shows that $\ker T^{k+m+1} \subset \ker T^{k+m}$. $\square$

LEMMA 1.9.11. *Assume that V is a finite dimensional vector space and that $T\colon V \to V$ is linear. Then*

$$\ker T^{\dim V} = \ker T^{\dim V+1} = \ker T^{\dim V+2} = \dots$$

PROOF. By Lemma 1.9.10 it is sufficient to show that the equality $\ker T^{\dim V} = \ker T^{\dim V+1}$ holds. Assume thus that this is not the case, that is, that $\ker T^{\dim V} \subsetneq \ker T^{\dim V+1}$. Then Lemma 1.9.10 implies that $\ker T^k \subsetneq \ker T^{k+1}$ for all $k \leq \dim V$. Thus we have that $\dim(\ker T^{k+1}) \geq \dim(\ker T^k) + 1$ for all $k \leq \dim V$. This, however, implies that $\dim(\ker T^{\dim V+1}) \geq \dim V + 1$, which is impossible. $\square$

PROPOSITION 1.9.12. *Assume that V is a finite dimensional vector space and $T\colon V \to V$ is linear. Then*

$$V = (\ker T^{\dim V}) \oplus (\operatorname{ran} T^{\dim V}).$$

PROOF. We show first that $(\ker T^{\dim V}) \cap (\operatorname{ran} T^{\dim V}) = \{0\}$. Assume to that end that $v \in (\ker T^{\dim V}) \cap (\operatorname{ran} T^{\dim V})$. Then we can write $v = T^{\dim V}u$ for some $u \in V$. Since $v \in \ker T^{\dim V}$, it then follows that $T^{2\dim V}u = T^{\dim V}(T^{\dim V}u) = T^{\dim V}v = 0$, that is, $u \in \ker T^{2\dim V}$. Now we know from Lemma 1.9.11 that $\ker T^{2\dim V} = \ker T^{\dim V}$, and thus we actually have that $u \in \ker T^{\dim V}$. This, however, implies that $v = T^{\dim V}u = 0$.

We now have shown that the sum of the spaces $\ker T^{\dim V}$ and $\operatorname{ran} T^{\dim V}$ is direct and thus

$$\dim(\ker T^{\dim V}) + \dim(\operatorname{ran} T^{\dim V}) = \dim\bigl((\ker T^{\dim V}) \oplus (\operatorname{ran} T^{\dim V})\bigr).$$

Next we know from Theorem 1.5.5 that

$$\dim V = \dim(\ker T^{\dim V}) + \dim(\operatorname{ran} T^{\dim V}).$$

Thus $(\ker T^{\dim V}) \oplus (\operatorname{ran} T^{\dim V})$ is a subspace of V of the same dimension as V. This immediately implies that it is the whole space, which finishes the proof. $\square$

LEMMA 1.9.13. *Let V be a complex finite dimensional vector space, $T\colon V \to V$ linear, and $\lambda \in \mathbb{K}$ an eigenvalue of T. Then*

$$G(\lambda, T) = \ker\bigl((T - \lambda \operatorname{Id})^{\dim V}\bigr).$$

PROOF. Assume that $v \in G(\lambda, T)$. Then there exists $j \in \mathbb{N}$ such that $(T - \lambda I)^j v = 0$, that is, $v \in \ker\big((T - \lambda \operatorname{Id})^j\big)$. From Lemma 1.9.11 we know that we may assume without loss of generality that $j \le \dim V$. Then, however, we can use Lemma 1.9.9 and obtain that $v \in \ker\big((T - \lambda \operatorname{Id})^{\dim V}\big)$, which proves the assertion. $\square$

Linear Independence of Generalised Eigenvectors.

THEOREM 1.9.14. *Let V be a vector space and let $T\colon V \to V$ be linear. Assume moreover that $\lambda_1, \dots, \lambda_m$ are distinct eigenvalues of T and that $v_1, \dots, v_m$ are corresponding generalised eigenvectors. Then the set $\{v_1, \dots, v_m\}$ is linearly independent.*

PROOF. Assume that $c_1, \dots, c_m \in \mathbb{K}$ are such that

$$0 = c_1 v_1 + \ldots + c_m v_m. \tag{9}$$

For each $j = 1, \dots, m$, denote by k_j the largest non-negative integer such that $(T - \lambda_j \operatorname{Id})^{k_j} v_j \ne 0$. That is,

$$(T - \lambda_j \operatorname{Id})^{k_j} v_j \ne 0 \qquad \text{but} \qquad (T - \lambda_j \operatorname{Id})^{k_j+1} v_j = 0.$$

Such an integer exists for each j, as the vectors v_j themselves are non-zero, but $(T - \lambda_j \operatorname{Id})^k v_j = 0$ for some sufficiently large k.

Now denote

$$w := (T - \lambda_1)^{k_1} v_1.$$

Then $w \ne 0$, but

$$(T - \lambda_1) w = (T - \lambda_1)^{k_1+1} v_1 = 0,$$

which shows that w is an eigenvector of T for the eigenvalue λ_1.

In particular, this implies that

$$(T - \lambda_j \operatorname{Id}) w = Tw - \lambda_j w = \lambda_1 w - \lambda_j w = (\lambda_1 - \lambda_j) w,$$

and thus also

$$(T - \lambda_j \operatorname{Id})^k w = (\lambda_1 - \lambda_j)^k w.$$

Now consider the operator

$$S = (T - \lambda_1 \operatorname{Id})^{k_1} \circ (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ (T - \lambda_3 \operatorname{Id})^{k_3+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1}.$$

Note that S is a polynomial of T, and thus we can rearrange the factors in any way we want in order to evaluate S. In particular, we have

$$\begin{aligned} S v_1 &= (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1} \circ (T - \lambda_1 \operatorname{Id})^{k_1} v \\ &= (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1} w \\ &= (\lambda_1 - \lambda_2)^{k_2+1} \cdots (\lambda_1 - \lambda_m)^{k_m+1} w. \end{aligned}$$

On the other hand, we obtain for $j = 2, \dots, m$ that

$$\begin{aligned} S v_j = (T - \lambda_1 \operatorname{Id})^{k_1} \circ (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ \cdots \circ (T - \lambda_{j-1} \operatorname{Id})^{k_{j-1}+1} \circ \\ \circ (T - \lambda_{j+1} \operatorname{Id})^{k_{j+1}+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1} \circ (T - \lambda_j \operatorname{Id})^{k_j+1} v_j = 0. \end{aligned}$$

Now apply the operator S to both sides of the equation (9). Then we obtain that

$$\begin{aligned} 0 = S(0) = c_1 S(v_1) + c_2 S(v_2) + \ldots + c_m S(v_m) \\ = c_1 (\lambda_1 - \lambda_2)^{k_2+1} \cdots (\lambda_1 - \lambda_m)^{k_m+1} w. \end{aligned}$$

Since $w \ne 0$ and $\lambda_1 \ne \lambda_j$ for $j = 2, \dots, m$, this is only possible if $c_1 = 0$.

With a similar argumentation, we now obtain that also $c_2 = c_3 = \ldots = c_m = 0$, which proves the linear independence of the set $\{v_1, \dots, v_m\}$. $\square$

Decomposition into Generalised Eigenspaces.

THEOREM 1.9.15. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. Denote by $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$ the distinct eigenvalues of T. Then*

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T) \tag{10}$$

and each subspace $G(\lambda_j, T)$ is T-invariant.

In particular, there exists a basis of V such that the corresponding matrix A of T has the blockdiagonal form

$$A = \begin{pmatrix} A_{11} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & A_{mm} \end{pmatrix},$$

and each non-zero block A_{jj} is upper triangular with diagonal elements λ_j, that is,

$$A_{jj} = \begin{pmatrix} \lambda_j & * & \ldots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \ldots & 0 & \lambda_j \end{pmatrix}.$$

PROOF. Since $G(\lambda_j, T) = \ker\bigl((T - \lambda_j \operatorname{Id})^{\dim V}\bigr)$, the generalised eigenspaces are the kernels of polynomials of T and thus T-invariant. Moreover, once we have established that V is the direct sum of the generalised eigenspaces, we can find a basis of each of spaces $G(\lambda_j, V)$ for which the restriction of T to $G(\lambda_j, V)$ is upper triangular. After concatenating these bases to a basis of the whole space V, we then obtain from our previous considerations that the matrix A of T has the above given form.

Thus it remains to establish the decomposition (10), which we will do by induction over the dimension $n = \dim V$ of the space V.

For $\dim V = 1$, the result trivially holds.

Assume now that we have shown (10) for every finite dimensional complex vector space U of dimension $\dim U < n$, and every linear transformation $S\colon U \to U$.

Let $\lambda_1 \in \mathbb{C}$ be an eigenvalue of T, the existence of which follows from Theorem 1.6.5. Define moreover

$$U := \operatorname{ran}\bigl((T - \lambda_1 \operatorname{Id})^n\bigr).$$

Then we can write

$$V = \ker\bigl((T - \lambda_1 \operatorname{Id})^n\bigr) \oplus \operatorname{ran}\bigl((T - \lambda_1 \operatorname{Id})^n\bigr) = G(\lambda_1, T) \oplus U.$$

If $U = \{0\}$, we have the desired decomposition and are done. Assume therefore that this is not the case.

Since U is the range of a polynomial of T it is T-invariant. Thus we can define the operator $S\colon U \to U$, $u \mapsto Su := Tu$. That is, S is the restriction of T to U, interpreted as a mapping from U to itself. We now note that the eigenvalues of S are a subset of $\{\lambda_2, \ldots, \lambda_m\}$. Indeed, if $u \in U \subset V$ is an eigenvector of S for an eigenvalue $\lambda \in \mathbb{K}$, then we have that $Tu = Su = \lambda u$, and thus u is already an eigenvector of T for the same eigenvalue. However, if $Su = \lambda_1 u$, then

$$\begin{aligned} u \in \ker(S - \lambda_1 \operatorname{Id}) &= \ker(T - \lambda_1 \operatorname{Id}) \cap U \\ &\subset \ker(T - \lambda_1 \operatorname{Id})^n \cap U = G(\lambda_1, T) \cap U = \{0\}. \end{aligned}$$

Thus λ_1 cannot be an eigenvalue of S.

Since $\dim U = \dim V - \dim(G(\lambda_1, T)) < \dim V = n$, we can apply the induction hypothesis to the transformation $S\colon U \to U$. Thus we can write

$$U = G(\lambda_2, S) \oplus \ldots \oplus G(\lambda_m, S), \tag{11}$$

and consequently

$$V = G(\lambda_1, T) \oplus G(\lambda_2, S) \oplus \ldots \oplus G(\lambda_m, S).$$

It is therefore sufficient to show that $G(\lambda_j, S) = G(\lambda_j, T)$ for all $j = 2, \ldots, m$.

We note first that

$$G(\lambda_j, S) = \bigl\{u \in U : (T - \lambda_j \operatorname{Id})^{\dim U} = 0\bigr\},$$
$$G(\lambda_j, T) = \bigl\{v \in V : (T - \lambda_j \operatorname{Id})^{\dim V} = 0\bigr\}.$$

Since $U \subset U$ and $\dim U < \dim V$, we immediately obtain that $G(\lambda_j, S) \subset G(\lambda_j, T)$ for all $j = 2, \ldots, m$.

For the converse inclusion assume that $v \in G(\lambda_j, T)$. Because of the decomposition $V = G(\lambda_1, T) \oplus U$ we can write

$$v = v_1 + u$$

with $v_1 \in G(\lambda_1, T)$ and $u \in U$. Next, the decomposition (10) implies that

$$u = u_2 + \ldots + u_m$$

with $u_j \in G(\lambda_j, S)$ for $j = 2, \ldots, m$. Thus

$$v = v_1 + u_2 + \ldots + u_m,$$

which we can rewrite as

$$0 = v_1 + u_2 + \ldots + u_{j-1} + (u_j - v) + u_{j+1} + \ldots + u_m.$$

Now note that $v_1 \in G(\lambda_1, T)$, $u_\ell \in G(\lambda_\ell, S) \subset G(\lambda_\ell, T)$ for $\ell = 2, \ldots, m$, $\ell \neq j$, and also $(u_j - v) \in G(\lambda_j, T) + G(\lambda_j, S) = G(\lambda_j, T)$.

Thus we have written 0 as a sum of elements of the generalised eigenspaces of T, which each consist of the generalised eigenvectors of T for the corresponding eigenvalue together with the vector 0. Since generalised eigenvectors of T for different eigenvalues are linearly independent, they cannot occur on the right hand side of this sum. Thus we have to conclude that all the terms on the right hand side are equal to 0, and in particular that $(u_j - v) = 0$. This shows that $v = u_j \in G(\lambda_j, S)$. Since $v \in G(\lambda_j, T)$ was arbitrary, this proves that $G(\lambda_j, T) \subset G(\lambda_j, S)$, which concludes the proof. □

10. Characteristic and Minimal Polynomial

Definition 1.10.1 (Multiplicities of eigenvalues). Assume that V is a finite dimensional vector space, that $T\colon V \to V$ is linear, and that $\lambda \in \mathbb{K}$ is an eigenvalue of T.

- The *algebraic multiplicity* of λ is defined as $\dim(G(\lambda, T))$.
- The *geometric multiplicity* of λ is defined as $\dim(E(\lambda, T))$.

Remark 1.10.2. Assume that V is a finite dimensional complex vector space and $T\colon V \to V$ is linear. We have just shown that V can be decomposed into the direct sum of the generalised eigenspaces of T. In particular, this shows that the sum of the algebraic multiplicities of all the eigenvalues of T is equal to the dimension of V.

In general, though, the geometric multiplicities may be smaller than the algebraic multiplicities, and their sum may be smaller than the dimension of V. More precisely, we obtain that the geometric multiplicities of all the eigenvalues sum up to the dimension of V, if and only if T is diagonalisable.

REMARK 1.10.3. Assume that the operator $T\colon V \to V$ has for some basis an upper triangular matrix with not necessarily distinct diagonal entries $\mu_1, \mu_2, \ldots, \mu_n$ (with $n = \dim V$). Then the characteristic polynomial of T is

$$\chi(z) = (z - \mu_1)(z - \mu_2) \cdots (z - \mu_n).$$

Note here, though, that the same linear factor may repeat several times.

DEFINITION 1.10.4. Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. Denote be $\lambda_1, \ldots, \lambda_m$ the distinct eigenvalues of T, and let $d_1, \ldots, d_m$ be their respective algebraic multiplicities. By

$$\chi(z) := (z - \lambda_1)^{d_1} \cdots (z - \lambda_m)^{d_m}$$

we denote the *characteristic polynomial* of T.

REMARK 1.10.5. This definition of the characteristic polynomial turns out to be the same as the one using determinants, which you have seen in previous maths courses. That is, one can show that

$$\chi(z) = \det(z \operatorname{Id} - T).$$

THEOREM 1.10.6 (Cayley–Hamilton). *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear, and denote by χ the characteristic polynomial of T. Then*

$$\chi(T) = 0.$$

PROOF. Write

$$\chi(z) := (z - \lambda_1)^{d_1} \cdots (z - \lambda_m)^{d_m},$$

where $\lambda_1, \ldots, \lambda_m$ are the distinct eigenvalues of T and $d_1, \ldots, d_m$ the corresponding algebraic multiplicities.

We show first that the restriction of $\chi(T)$ to each of the generalised eigenspaces of T is equal to 0. Since $G(\lambda_j, T)$ is T-invariant, we can interpret the restriction of T to $G(\lambda_j, T)$ as a transformation of the space $G(\lambda_j, T)$. Now we have defined the generalised eigenspace as $G(\lambda_j, T) = \ker(T - \lambda_j \operatorname{Id})^{\dim V}$, and thus $(T - \lambda_j \operatorname{Id})^{\dim V} v = 0$ for every $v \in G(\lambda_j, T)$. In other words, the restriction of $T - \lambda_j \operatorname{Id}$ to $G(\lambda_j, T)$ is nilpotent. This, however, implies that we already have that $(T - \lambda_j \operatorname{Id})^{\dim(G(\lambda_j, T))} v = 0$ for all $v \in G(\lambda_j, T)$. Since $\dim G(\lambda_j, T) = d_j$, this shows that $(T - \lambda_j \operatorname{Id})^{d_j} v = 0$ for all $v \in G(\lambda_j, T)$. Now the definition of the characteristic polynomial (and the fact that we can freely switch the order of the linear factors) shows that $\chi(T)v = 0$ for all $v \in G(\lambda_j, T)$ and all $j = 1, \ldots, m$.

Now let $v \in V$ be arbitrary. Since V is the direct sum of its generalised eigenspaces, we can write

$$v = v_1 + \ldots + v_m \qquad \text{with } v_j \in G(\lambda_j, T) \text{ for } j = 1, \ldots, m.$$

Thus

$$\chi(T)v = \chi(T)v_1 + \ldots + \chi(T)v_m = 0.$$

Since this holds for all $v \in V$, it follows that $\chi(T) = 0$. □

DEFINITION 1.10.7. A polynomial p is called *monic*, if its leading order coefficient is equal to 1. That is, a monic polynomial is of the form

$$p(z) = z^m + c_{m-1} z^{m-1} + c_{m-2} z^{m-2} + \ldots + c_1 z + c_0.$$

LEMMA 1.10.8. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. There exists a unique monic polynomial p of minimal degree such that $p(T) = 0$.*

PROOF. From Theorem 1.10.6 we know that there exists a polynomial such that $p(T) = 0$ (namely the characteristic polynomial). Thus there also exists a monic polynomial of minimal degree with this property. It remains to show that this is unique.

Assume therefore to the contrary that p and q are monic polynomials of minimal degree such that $p(T) = q(T) = 0$, and that $p \neq q$. In particular, $\deg p = \deg q$, and thus we can Write

$$\begin{aligned} p(z) &= z^m + a_{m-1}z^{m-1} + \ldots + a_1 z + a_0, \\ q(z) &= z^m + b_{m-1}z^{m-1} + \ldots + b_1 z + b_0, \end{aligned}$$

for some coefficients a_j, $b_j \in \mathbb{K}$. Let k be the largest index such that $a_k \neq b_k$ (such an index exists, as $p \neq q$). Now define the polynomial $r(z) = (p(z)-q(z))/(a_k-b_k)$. Then $r(T) = p(T) - q(T) = 0$ and r is a monic polynomial of degree $k < m$, which is a contradiction to the definitions of p and q. □

DEFINITION 1.10.9. Assume that V is a finite dimensional complex vector space and that $T: V \to V$ is linear. The unique monic polynomial p of minimal degree satisfying $p(T) = 0$ is called the *minimal polynomial* of T.

EXAMPLE 1.10.10. Assume that $T: \mathbb{C}^3 \to \mathbb{C}^3$ is the linear transformation given by the matrix

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

with respect to the standard basis on $\mathbb{C}^3$. Since A is triangular, we can immediately read of the characteristic polynomial of T and obtain

$$\chi(T) = (z-2)^3.$$

It turns out, however, that we already have that $(A - 2\,\mathrm{Id})^2 = 0$. Thus in this case the minimal polynomial is different from the characteristic polynomial. In fact, one can show that the minimal polynomial of T is $p(z) = (z-2)^2$.

Next we establish some results about the relation between the minimal polynomial and the characteristic polynomial.

PROPOSITION 1.10.11. *Assume that V is a finite dimensional complex vector space and that $T: V \to V$ is linear, and denote by $p(z)$ the minimal polynomial of T. Assume moreover that $q(z)$ is another polynomial satisfying $q(T) = 0$. Then q is a polynomial multiple of p. That is, there exists a polynomial s such that $q(z) = s(z)p(z)$.*

PROOF. By performing polynomial division with remainder, we can write

$$q(z) = s(z)p(z) + r(z),$$

where r and s are polynomials, and $\deg r < \deg p$. Now, since $q(T) = 0$ and $p(T) = 0$, we obtain that also $r(T) = 0$. However, p is the minimal polynomial of T and $\deg r < \deg p$, which is only possible if $r(z) = 0$ is the 0-polynomial. Thus we obtain the desired result. □

PROPOSITION 1.10.12. *Assume that V is a finite dimensional complex vector space and that $T: V \to V$ is linear. Then the zeroes of the minimal polynomial p of T are precisely the eigenvalues of T. Moreover, p is of the form*

$$p(z) = (z-\lambda_1)^{s_1}(z-\lambda_2)^{s_2}\cdots(z-\lambda_m)^{s_m}$$

with $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of T, and $1 \leq s_j \leq d_j$ for all j, with d_j being the algebraic multiplicity of the eigenvalue λ_j.

Proof. We know already that the characteristic polynomial is a polynomial multiple of p, which shows that p is of the form

$$p(z) = (z - \lambda_1)^{s_1}(z - \lambda_2)^{s_2} \cdots (z - \lambda_m)^{s_m}$$

with $0 \le s_j \le d_j$. It remains to show that $s_j \ge 1$ for all j. Assume therefore to the contrary that $s_j = 0$ for some j, that is, that the corresponding linear factor does not appear in the minimal polynomial. Let now v_j be an eigenvector for the eigenvalue λ_j. Then a simple calculation (similar to those we needed for the linear independence of generalised eigenspaces) shows that

$$p(T)v_j = (\lambda_j - \lambda_1)^{s_1} \cdots (\lambda_j - \lambda_{j-1})^{s_{j-1}}(\lambda_j - \lambda_{j+1})^{s_{j+1}} \cdots (\lambda_j - \lambda_m)^{s_m} v_j.$$

Since $v_j \neq 0$ and $\lambda_k - \lambda_j \neq 0$ for $k \neq j$, this shows that $p(T)v_j \neq 0$, which contradicts the assumption that $p(T) = 0$. Therefore $s_j \ge 1$ for all j, which proves the claim. □

11. Jordan's Normal Form

We will now discuss a further refinement of the decomposition shown in Theorem 1.9.15, which describes the, in a sense, simplest possibility of representing an arbitrary linear transformation of a finite dimensional complex vector space.

Theorem 1.11.1. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. Then there exists a basis of V such that the matrix representation A of T has the blockdiagonal form*

$$A = \begin{pmatrix} A_{11} & 0 & \dots & 0 \\ 0 & A_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & A_{mm} \end{pmatrix},$$

where each block A_{mm} itself has a block-diagonal form

$$A_{jj} = \begin{pmatrix} J_j^{(1)} & 0 & \dots & 0 \\ 0 & J_j^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_j^{k_j} \end{pmatrix},$$

and the sub-blocks are upper triangular matrices of the form

$$J_j^{(\ell)} = \begin{pmatrix} \lambda_j & 1 & 0 & \dots & 0 \\ 0 & \lambda_j & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \lambda_j \end{pmatrix}$$

(or possibly $J_j^{(\ell)} = (\lambda_j)$ for sub-blocks of size 1), where $\lambda_1, \dots, \lambda_m$ are the distinct eigenvalues of T. Such a basis is called a Jordan basis *of T, and the sub-blocks $J_j^{(\ell)}$ are called* Jordan blocks *of T.*

Moreover the following properties are satisfied:

- *The blocks A_{jj} are $(d_j \times d_j)$-matrices, where d_j is the algebraic multiplicity of the eigenvalue λ_j.*
- *The number k_j of different blocks for the eigenvalue λ_j is the geometric multiplicity of λ_j.*

- *The minimal polynomial p of T has the form*
$$p(z) = (z-\lambda_1)^{s_1}(z-\lambda_s)^{s_2}\cdots(z-\lambda_m)^{s_m},$$
where s_j is the size of the largest Jordan block $J_j^{(\ell)}$ for the eigenvalue λ_j.
- *Apart from the order, the Jordan blocks are unique.*

EXAMPLE 1.11.2. Assume that the operator $T\colon \mathbb{C}^4 \to \mathbb{C}^4$ has the only eigenvalue $\lambda = 2$. Then we have the following possibilities for the Jordan normal form of T:

(1) A decomposition in a single Jordan block of size 4:
$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$
Here the geometric multiplicity of the eigenvalue 2 is 1, and the minimal polynomial of T is $p(z) = (z-2)^4$.

(2) A decomposition in a Jordan block of size 3 and a block of size 1:
$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$
Here the geometric multiplicity of the eigenvalue 2 is 2, and the minimal polynomial of T is $p(z) = (z-2)^3$.

(3) A decomposition in two Jordan blocks of size 2:
$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$
Here the geometric multiplicity of the eigenvalue 2 is 2, and the minimal polynomial of T is $p(z) = (z-2)^2$.

(4) A decomposition in one Jordan blocks of size 2 and two blocks of size 1:
$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$
Here the geometric multiplicity of the eigenvalue 2 is 3, and the minimal polynomial of T is $p(z) = (z-2)^2$.

(5) A decomposition in four Jordan blocks of size 1:
$$A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$
Here the geometric multiplicity of the eigenvalue 2 is 4, and the minimal polynomial of T is $p(z) = (z-2)$.

The algebraic multiplicity of 2 is in all cases equal to 4, and the characteristic polynomial is always $\chi(z) = (z-2)^4$.

CHAPTER 2

Inner Product Spaces and Singular Value Decomposition

1. Inner Products

Recall that the Euclidean (or standard) inner product on $\mathbb{R}^n$ is given as

$$\langle u, v\rangle = \sum_{i=1}^{n} u_i v_i,$$

and that the corresponding Euclidean norm is

$$\|u\| := \Big(\sum_{i=1}^{n} u_i^2\Big)^{1/2}.$$

The intuition behind these notion is that $\|u\|$ would be the length of the vector $u \in \mathbb{R}^n$, whereas the inner product $\langle u, v\rangle$ is related to the angle between the vectors u and v. More precisely, we have

$$\langle u, v\rangle = \|u\|\,\|v\|\,\cos(\alpha),$$

where α is the angle between the vectors u and v.

In the following, we will develop a generalisation of this notion of inner products for arbitrary real or complex vector spaces.

Definition 2.1.1. Let V be a vector space. An *inner product* on V is a mapping $\langle\cdot,\cdot\rangle \colon V \to \mathbb{K}$ with the following properties:

(1) Linearity in the first component:

$$\langle \lambda u + \mu v, w\rangle = \lambda\langle u, w\rangle + \mu\langle v, w\rangle$$

for all u, v, $w \in V$ and λ, $\mu \in \mathbb{K}$.

(2) Conjugate symmetry:

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

for all u, $v \in V$.

(3) Positive definiteness:

$$\langle v, v\rangle \in \mathbb{R}_{\geq 0}$$

for all $v \in V$. Moreover $\langle v, v\rangle = 0$ if and only if $v = 0$.

A vector space V with an inner product is called an *inner product space.*[1]

Example 2.1.2. Consider the following examples of inner product spaces:

(1) The space $\mathbb{C}^n$ is an inner product space with the inner product

$$\langle u, v\rangle := \sum_{i=1}^{n} u_i \overline{v_i},$$

which we can also write as

$$\langle u, v\rangle = v^H u,$$

[1]Sometimes one also uses the name *pre-Hilbert space* for an inner product space.

where
$$v^H := (\overline{v})^T$$
is the Hermitian conjugate (or: "conjugate transpose") of v.

Here the positive definiteness follows from the fact that
$$\langle v, v\rangle = \sum_{i=1}^n v_i\overline{v_i} = \sum_{i=1}^n |v_i|^2,$$
which is non-negative (and real) for all $v \in \mathbb{C}^n$, and equal to zero if and only if $v = 0$.

(2) We can make the space $C([0,1],\mathbb{C})$ of complex-valued continuous functions on $[0,1]$ to an inner product space by defining
$$\langle u, v\rangle := \int_0^1 u(x)\overline{v(x)}\,dx$$
for u, $v \in C([0,1],\mathbb{C})$.

We obtain here the positive definiteness by writing
$$\langle u, u\rangle = \int_0^1 |u(x)|^2\,dx,$$
which obviously is real and non-negative. Moreover, since the integrand $|u(x)|^2$ is a continuous and non-negative function, the integral is equal to zero if and only if $|u(x)| = 0$ for all x, which is the case if and only if $u = 0$.

LEMMA 2.1.3. *Assume that V is an inner product space with inner product $\langle\cdot,\cdot\rangle$. Then we have for all u, v, $w \in V$ and λ, $\mu \in \mathbb{K}$ that*
$$\langle u, \lambda v + \mu w\rangle = \overline{\lambda}\langle u, v\rangle + \overline{\mu}\langle u, w\rangle.$$
That is, $\langle\cdot,\cdot\rangle$ is conjugate linear *in the second component.*

PROOF. We can write
$$\langle u, \lambda v + \mu w\rangle = \overline{\langle \lambda v + \mu w, u\rangle} = \overline{\lambda\langle v, u\rangle + \mu\langle w, u\rangle}$$
$$= \overline{\lambda}\overline{\langle v, u\rangle} + \overline{\mu}\overline{\langle w, u\rangle} = \overline{\lambda}\langle u, v\rangle + \overline{\mu}\langle u, w\rangle.$$
□

REMARK 2.1.4. It is also somewhat common, especially in physics, to define inner products slightly differently by requiring linearity in the second component and conjugate linearity in the first component. Then one would for instance write the inner product on $\mathbb{C}^n$ as $\langle u, v\rangle = u^H v$. From a theoretical point of view, this makes no real difference, but it can be pretty annoying in practice, as all concrete formulas look slightly different.

REMARK 2.1.5. The linearity of an inner product in the first component together with the conjugate linearity in the second component is sometimes called *sesquilinearity* from the Latin word *sesqui* meaning one-and-a-half.

DEFINITION 2.1.6. Let V be an inner product space with inner product $\langle\cdot,\cdot\rangle$. The associated *norm* on V is the mapping $\|\cdot\|\colon V \to \mathbb{R}_{\geq 0}$,
$$\|v\| := \big(\langle v, v\rangle\big)^{1/2}.$$

LEMMA 2.1.7. *The norm on an inner product space V has the following properties:*

(1) Positivity: $\|v\| \geq 0$ for all $v \in V$, and $\|v\| = 0$ if and only if $v = 0$.
(2) Positive homogeneity: $\|\lambda v\| = |\lambda|\|v\|$ for all $v \in V$ and $\lambda \in \mathbb{K}$.

PROOF. The positivity follows immediately from the positive definiteness of the inner product. For the positive homogeneity we note that

$$\|\lambda v\|^2 = \langle \lambda v, \lambda v\rangle = \lambda\overline{\lambda}\langle v, v\rangle = |\lambda|^2\|v\|.$$

$\square$

DEFINITION 2.1.8. Let V be an inner product space with inner product $\langle\cdot,\cdot\rangle$, and let u, $v \in V$. We say that u and v are *orthogonal*, if $\langle u, v\rangle = 0$.

THEOREM 2.1.9 (Pythagoras). *Let V be an inner product space, and let u, $v \in V$ be orthogonal. Then*

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

PROOF. We have that

$$\begin{aligned}\|u + v\|^2 = \langle u + v, u + v\rangle = \langle u, u\rangle + \langle u, v\rangle + \langle v, u\rangle + \langle v, v\rangle \\ = \langle u, u\rangle + \langle v, v\rangle = \|u\|^2 + \|v\|^2.\end{aligned}$$

$\square$

REMARK 2.1.10. In an inner product space we can also write

$$\|u + v\|^2 = \langle u, u\rangle + \langle u, v\rangle + \langle v, u\rangle + \langle v, v\rangle = \|u\|^2 + \|v\|^2 + 2\Re(\langle u, v\rangle),$$

where $\Re(z)$ denotes the real part of $z \in \mathbb{C}$. Thus Pythagoras' theorem actually holds more generally for vectors u, v for which the inner product $\langle u, v\rangle$ is purely imaginary.

THEOREM 2.1.11 (Cauchy–Schwarz–Bunyakovsky). *Assume that V is an inner product space and that u, $v \in V$. Then*

$$|\langle u, v\rangle| \le \|u\|\|v\|,$$

and equality holds if and only if u and v are linearly dependent.

PROOF. If $v = 0$, then the result holds (with equality, and u and v are linearly dependent). We may therefore assume without loss of generality that $v \neq 0$.

Define now

$$w := u - \frac{\langle u, v\rangle}{\|v\|^2}v.$$

Then

$$\langle w, v\rangle = \langle u, v\rangle - \frac{\langle u, v\rangle}{\|v\|^2}\langle v, v\rangle = 0,$$

and thus

$$\|u\|^2 = \left\|w + \frac{\langle u, v\rangle}{\|v\|^2}v\right\|^2 = \|w\|^2 + \frac{|\langle u, v\rangle|^2}{\|v\|^4}\|v\|^2 = \|w\|^2 + \frac{|\langle u, v\rangle|^2}{\|v\|^2}.$$

Thus

$$\|u\|^2\|v\|^2 = \|v\|^2\|w\|^2 + |\langle u, v\rangle|^2 \le |\langle u, v\rangle|^2,$$

and we have equality if and only if $w = 0$. This, however, implies that u and v are linearly independent.

Finally, assume that u and v are linearly independent (and still $v \neq 0$), say $u = \lambda v$ for some $\lambda \in \mathbb{K}$. then

$$|\langle u, v\rangle| = |\langle \lambda v, v\rangle| = |\lambda||\langle v, v\rangle| = |\lambda|\|v\|^2 = \|\lambda v\|\|v\| = \|u\|\|v\|.$$

$\square$

THEOREM 2.1.12 (Triangle inequality). *Assume that V is an inner product space and u, $v \in V$. Then*

$$\|u+v\| \le \|u\| + \|v\|,$$

and we have equality, if and only if either $v = 0$ or $u = \lambda v$ for some $\lambda \in \mathbb{R}_{\ge 0}$.

PROOF. We can write

$$\begin{aligned}\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2\Re(\langle u, v\rangle) &\le \|u\|^2 + \|v\|^2 + 2|\langle u, v\rangle| \\ &\le \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| = (\|u\| + \|v\|)^2.\end{aligned}$$

This shows that the inequality $\|u+v\| \le \|u\| + \|v\|$ holds.

Moreover, we have an equality, if and only if we have an equality in all the steps of our estimate above. The third step was the Cauchy–Schwarz–Bunyakovsky inequality, where we have an equality if and only if u and v are linearly independent, that is, if either $v = 0$ or $u = \lambda v$ for some $\lambda \in \mathbb{K}$. For the second step, we have an equality if and only if $\Re(\langle u, v\rangle) = |\langle u, v\rangle|$. If $v = 0$, this trivially holds; if $u = \lambda v$ for some $\lambda \in \mathbb{K}$, we need the additional condition that $\lambda \in \mathbb{R}_{\ge 0}$, which proves the assertion. □

2. Bases of Inner Product Spaces

DEFINITION 2.2.1. Let V be an inner product space. A set $\{v_1, \ldots, v_m\} \subset V$ of vectors is called *orthonormal*, if $\|v_i\| = 1$ for all $i = 1, \ldots, m$, and $\langle v_i, v_j\rangle = 0$ for all i, $j = 1, \ldots, m$ with $i \ne j$.

LEMMA 2.2.2. *Assume that V is an inner product space and that $\{v_1, \ldots, v_m\} \subset V$ is orthonormal. Then $\{v_1, \ldots, v_m\}$ is linearly independent.*

PROOF. Assume that $c_1, \ldots, c_m \in \mathbb{K}$ are such that

$$0 = c_1 v_1 + \ldots + c_m v_m.$$

A repeated application of Pythagoras' theorem implies that

$$\begin{aligned}0 = \|c_1 v_1 + \ldots + c_m v_m\|^2 &= \|c_1 v_1\|^2 + \ldots + \|c_m v_m\|^2 \\ &= |c_1|^2 \|v_1\|^2 + \ldots + |c_m|^2 \|v_m\|^2 = |c_1|^2 + \ldots + |c_m|^2.\end{aligned}$$

Thus $c_1 = c_2 = \ldots = c_m = 0$, which proves the linear independence of the set $\{v_1, \ldots, v_m\}$. □

DEFINITION 2.2.3. Let V be an inner product space. An *orthonormal basis* of V is an orthonormal set that is a basis of V.

THEOREM 2.2.4 (Gram–Schmidt Orthogonalisation). *Assume that V is a vector space and that $S = \{v_1, v_2, \ldots\} \subset V$ is a linearly independent subset of V. Then we can construct a set $E = \{e_1, e_2, \ldots\}$ as follows:*

- *Start with setting*
$$e_1 = \frac{v_1}{\|v_1\|}.$$
- *Assume we have already constructed orthonormal vectors $e_1, \ldots, e_m$. We define first*
$$f_{m+1} = v_{m+1} - \langle v_{m+1}, e_1\rangle e_1 - \langle v_{m+1}, e_2\rangle e_2 - \ldots - \langle v_{m+1}, e_m\rangle e_m$$
and then set
$$e_{m+1} = \frac{f_{m+1}}{\|f_{m+1}\|}.$$

Then the set $\{e_1, e_2, \ldots\}$ is a well-defined orthonormal set. Moreover, we have that $\operatorname{span}\{v_1, v_2, \ldots, v_m\} = \operatorname{span}\{e_1, e_2, \ldots, e_m\}$ *for all m.*

PROOF. We show by induction over m that the set $\{e_1, e_2, \ldots, e_m\}$ is well-defined and orthonormal, and satisfies $\operatorname{span}\{v_1, \ldots, v_m\} = \operatorname{span}\{e_1, \ldots, e_m\}$.

For $m = 1$, this is obvious (as $v_1 \neq 0$).

Now assume that we have shown the result for m. Then we have for all $1 \leq j \leq m$ that

$$\begin{aligned}\langle f_{m+1}, e_j\rangle = \langle v_{m+1}, e_j\rangle - \langle v_{m+1}, e_1\rangle\langle e_1, e_j\rangle - \ldots - \langle v_{m+1}, e_m\rangle\langle e_m, e_j\rangle \\ = \langle v_{m+1}, e_j\rangle - \langle v_j, e_j\rangle\langle e_j, e_j\rangle = 0.\end{aligned}$$

Here we have used that $\langle e_i, e_j\rangle = 0$ for $i \neq j$ and $\langle e_j, e_j\rangle = 1$.

In order to show that e_{m+1} is well-defined, we need to show that $f_{m+1} \neq 0$. Assume therefore to the contrary that $f_{m+1} = 0$. Then we can write

$$\begin{aligned}v_{m+1} = \langle v_{m+1}, e_1\rangle e_1 + \langle v_{m+1}, e_2\rangle e_2 + \ldots + \langle v_{m+1}, e_m\rangle e_m \\ \in \operatorname{span}\{e_1, \ldots, e_m\} = \operatorname{span}\{v_1, \ldots, v_m\},\end{aligned}$$

which contradicts the linear independence of the set $\{v_1, \ldots, v_m, v_{m+1}\}$. Thus $f_{m+1} \neq 0$, and e_{m+1} is well-defined and satisfies $\|e_{m+1}\| = 1$. Therefore the set $\{e_1, \ldots, e_m, e_{m+1}\}$ is orthonormal.

Finally, by construction every vector e_j can be written as a linear combination of the vectors $v_1, \ldots, v_j$, and thus $\operatorname{span}\{e_1, \ldots, e_m, e_{m+1}\} \subset \operatorname{span}\{v_1, \ldots, v_m, v_{m+1}\}$. Since those spaces have the same dimension (namely $m + 1$) they have to be equal. □

COROLLARY 2.2.5. *Every finite dimensional inner product space has an orthonormal basis.*

PROOF. Start with any basis of the vector space, and apply Gram–Schmidt orthogonalisation. □

LEMMA 2.2.6. *Assume that V is a finite dimensional inner product space with orthonormal basis $\{v_1, \ldots, v_n\}$. Assume that u, $w \in V$ have the coordinates $x = (x_1, \ldots, x_n)^T \in \mathbb{K}^n$ and $y = (y_1, \ldots, y_n)^T \in \mathbb{K}^n$, respectively, that is,*

$$u = x_1 v_1 + \ldots + x_n v_n, \qquad w = y_1 v_1 + \ldots + y_n v_n.$$

Then

$$\langle u, w\rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^H x.$$

PROOF. We have

$$\langle u, w\rangle = \Big\langle \sum_{i=1}^{n} x_i v_i, \sum_{j=1}^{n} y_j v_j \Big\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i \overline{y_j}\langle v_i, v_j\rangle = \sum_{i=1}^{n} x_i \overline{y_i}.$$

□

DEFINITION 2.2.7. Assume that V is a finite dimensional inner product space and that $\{v_1, \ldots, v_n\}$ is a basis of V. The *Gram matrix* for this basis is the matrix $G \in \mathbf{Mat}_n(\mathbb{K})$ with entries $G_{ij} = \langle v_j, v_i\rangle$. That is,

$$G := \begin{pmatrix} \langle v_1, v_1\rangle & \langle v_2, v_1\rangle & \ldots & \langle v_n, v_1\rangle \\ \langle v_1, v_2\rangle & \langle v_2, v_2\rangle & \ldots & \langle v_n, v_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_1, v_n\rangle & \langle v_2, v_n\rangle & \ldots & \langle v_n, v_n\rangle \end{pmatrix}.$$

REMARK 2.2.8. We see immediately that the Gram matrix for a basis is the identity if and only if the basis is orthonormal.

PROPOSITION 2.2.9. *Assume that V is a finite dimensional inner product space, that $\{v_1,\dots,v_n\}$ is a basis of V, and that G is the corresponding Gram matrix. Assume that u, $w \in V$ have the coordinates $x = (x_1,\dots,x_n)^T \in \mathbb{K}^n$ and $y = (y_1,\dots,y_n)^T \in \mathbb{K}^n$, respectively. Then*

$$\langle u, w\rangle = y^H G x.$$

PROOF. We have

$$\langle u, w\rangle = \sum_{i=1}^n \sum_{j=1}^n x_i \overline{y_j} \langle v_i, v_j\rangle = \sum_{j=1}^n \sum_{i=1}^n \overline{y}_j G_{ji} x_i = y^H G x.$$

□

DEFINITION 2.2.10. A matrix $A \in \mathbf{Mat}_n(\mathbb{K})$ is called *Hermitian*, if $A^H = A$, that is, if $a_{ij} = \overline{a_{ji}}$ for all i, $j = 1,\dots,n$.

DEFINITION 2.2.11. A matrix $A \in \mathbf{Mat}_n(\mathbb{K})$ is called *positive definite*, if

$$x^H A x \in \mathbb{R}_{>0}$$

for all $x \in \mathbb{K}^n$ with $x \neq 0$.

LEMMA 2.2.12. *Assume that V is a finite dimensional inner product space and that $\{v_1,\dots,v_n\}$ is a basis of V, and denote by G the corresponding Gram matrix. Then G is Hermitian and positive definite.*

PROOF. Since

$$G_{ij} = \langle v_j, v_i\rangle = \overline{\langle v_i, v_j\rangle} = \overline{G_{ji}},$$

the matrix G is Hermitian.

Now let $x \in \mathbb{K}^n$ with $x \neq 0$, and let $v = \sum_{i=1}^n x_i v_i$ be the vector with coordinates $(x_1,\dots,x_n)^T$. Then $v \neq 0$ and thus

$$x^H G x = \langle v, v\rangle = \|v\|^2 > 0,$$

which proves that G is positive definite. □

REMARK 2.2.13. Conversely, assume that $\{v_1,\dots,v_n\}$ is a basis of V and $G \in \mathbf{Mat}_n(\mathbb{K})$ is positive definite and Hermitian. Then we can define an inner product on V by setting

$$\langle u, v\rangle_G := y^H G x,$$

if x, $y \in \mathbb{K}^n$ are the coordinates of the vectors u, $v \in V$. (It is straight-forward to show that this is indeed an inner product.) Thus we have a one-to-one correspondence between positive definite Hermitian matrices and inner products.

PROPOSITION 2.2.14. *Assume that V is a finite dimensional inner product space and that $\{v_1,\dots,v_n\}$ is a basis of V, denote by G the corresponding Gram matrix and by G^{-1} its inverse, and let $v \in V$. Then the coordinates $x_1,\dots,x_n$ are given as*

$$x_i = \sum_{j=1}^n G_{ij}^{-1} \langle v, v_j\rangle.$$

PROOF. Exercise! □

COROLLARY 2.2.15. *Assume that V is a finite dimensional inner product space and that $\{e_1,\dots,e_n\}$ is an orthonormal basis of V. Then*

$$v = \sum_{i=1}^n \langle v, e_i\rangle e_i$$

for all $v \in V$. In particular, the coordinate vector of v is $(\langle v, e_1\rangle,\dots,\langle v, e_n\rangle)$.

3. Orthogonal Complements and Projections

Definition 2.3.1. Assume that V is an inner product space and $U \subset V$ is a subset. We denote by

$$U^\perp := \bigl\{v \in V : \langle u, v\rangle = 0 \text{ for all } u \in U\bigr\}$$

the *orthogonal complement* of U.

Lemma 2.3.2. *Let V be an inner product space and $U \subset V$ a subset. Then $U^\perp$ is a subspace of V and $U \cap U^\perp = \{0\}$.*

Proof. We note first that $0 \in U^\perp$, since $\langle u, 0\rangle = 0$ for all $u \in V$. Now assume that v, $w \in U^\perp$. Then we have for all $u \in U$ that

$$\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle = 0 + 0 = 0,$$

and thus $v + w \in U^\perp$. Similarly, if $v \in U^\perp$ and $\lambda \in \mathbb{K}$, then

$$\langle u, \lambda v\rangle = \overline{\lambda}\langle u, v\rangle = 0$$

for all $u \in U$ and therefore $\lambda v \in U^\perp$. This shows that $U^\perp$ is a subspace.

Finally assume that $u \in U \cap U^\perp$. Then we have in particular that $\langle u, u\rangle = 0$, and thus $u = 0$. □

Lemma 2.3.3. *Let V be an inner product space and U, W be subsets of V satisfying $U \subset W$. Then $W^\perp \subset U^\perp$.*

Proof. Assume that $v \in W^\perp$. Then $\langle w, v\rangle = 0$ for all $w \in W$. Since $U \subset W$, this implies that, in particular, $\langle u, v\rangle = 0$ for all $u \in U$, which in turn shows that $v \in U^\perp$. □

Proposition 2.3.4. *Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Then*

$$V = U \oplus U^\perp.$$

Proof. Denote $n = \dim V$ and $m = \dim U$. Let $\mathcal{B} = \{u_1, \ldots, u_m\}$ be an orthonormal basis of U. Using the Gram–Schmidt algorithm, we can extend this to an orthonormal basis of V. That is, we can find a set $\mathcal{C} = \{v_1, \ldots, v_{n-m}\}$ such that $\mathcal{B} \cup \mathcal{C}$ is an orthonormal basis of V. We claim that $U^\perp = \operatorname{span}\mathcal{C}$.

For this, we note first that $\mathcal{C} \subset U^\perp$ by construction (as every vector v_j is orthogonal to all the basis elements of U). Since $U^\perp$ is a subspace, this shows that $\operatorname{span}\mathcal{C} \subset U^\perp$. In order to show the converse inclusion, we recall that $U^\perp \cap U = \{0\}$, which implies that the sum of U and $U^\perp$ is direct. In particular,

$$\dim U^\perp = \dim(U \oplus U^\perp) - \dim U \le n - m = \dim(\operatorname{span}\mathcal{C}).$$

Since $\operatorname{span}\mathcal{C} \subset U^\perp$, this implies that $U^\perp = \operatorname{span}\mathcal{C}$.

As a consequence, we have that

$$V = \operatorname{span}\mathcal{B} \oplus \operatorname{span}\mathcal{C} = U \oplus U^\perp,$$

which concludes the proof. □

Proposition 2.3.5. *Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Then $(U^\perp)^\perp = U$.*

Proof. Assume that $u \in U$. Then we have for every $v \in U^\perp$ that $\langle u, v\rangle = 0$. This, however, implies that $u \in (U^\perp)^\perp$. This shows that $U \subset (U^\perp)^\perp$

Now assume that $v \in (U^\perp)^\perp$. Since $V = U \oplus U^\perp$, we can write v in the form $v = u + w$ with $u \in U$ and $w \in U^\perp$. We have to show that $w = 0$. Since $u \in U \subset (U^\perp)^\perp$, it follows that $w = v - u \in (U^\perp)^\perp$. On the other hand, we know that $w \in U^\perp$. Since $U^\perp \cap (U^\perp)^\perp = \{0\}$, this implies that $w = 0$. Thus $v = u \in U$, which shows that $(U^\perp)^\perp \subset U$. □

DEFINITION 2.3.6. Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Since $V = U \oplus U^\perp$, we can decompose each vector $v \in V$ *uniquely* as $v = u + w$ with $u \in U$ and $w \in U^\perp$. The mapping $\pi_U : V \to V$ maps v to the corresponding vector $u \in U$ is called the *orthogonal projection onto* U.

THEOREM 2.3.7. *Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Then the projection π_U onto U has the following properties:*

(1) *$\pi_U : V \to V$ is a linear transformation.*
(2) *$\pi_U u = u$ for all $u \in U$.*
(3) *$\pi_U w = 0$ for all $w \in U^\perp$.*
(4) *$\operatorname{ran} \pi_U = U$ and $\ker \pi_U = U^\perp$.*
(5) *$v - \pi_U v \in U^\perp$ for all $v \in V$.*
(6) *$\pi_U^2 = \pi_U$.*
(7) *$\|v\|^2 = \|\pi_U v\|^2 + \|v - \pi_U v\|^2$.*
(8) *$\|\pi_U v\| \le \|v\|$ for all $v \in V$, and we have equality if and only if $v \in U$.*
(9) *$\pi_U v$ is the unique solution of the optimisation problem*

$$\min_{u \in U} \|v - u\|^2.$$

(10) *If $\{e_1, \ldots, e_m\}$ is an orthonormal basis of U, then*

$$\pi_U v = \langle v, e_1 \rangle e_1 + \ldots + \langle v, e_m \rangle e_m.$$

PROOF. Let $\{e_1, \ldots, e_m\}$ be an orthonormal basis of U, and let $\{e_{m+1}, \ldots, e_n\}$ be an orthonormal basis of $U^\perp$, such that $\{e_1, \ldots, e_n\}$ is an orthonormal basis of V. Then we can write

$$v = \sum_{i=1}^n \langle v, e_i \rangle e_i = \Big(\sum_{i=1}^m \langle v, e_i \rangle e_i \Big) + \Big(\sum_{i=m+1}^n \langle v, e_i \rangle e_i \Big) \in U \oplus U^\perp.$$

From the definition of $\pi_U v$, it now follows that

$$\pi_U v = \sum_{i=1}^m \langle v, e_i \rangle e_i \qquad \text{and} \qquad v - \pi_U v = \sum_{i=m+1}^n \langle v, e_i \rangle e_i,$$

which shows (10), and also proves (1). Next, (2) and (3) follow directly from the definition of the orthogonal projection. By definition, we have that $\operatorname{ran} \pi_U \subset U$. Since $\pi_U u = u$ for every $u \in U$, we then obtain that, actually, $\operatorname{ran} \pi_U = U$. Because of the decomposition $V = U \oplus U^\perp$ and the definition of π_U we obtain next that $\ker \pi_U = U^\perp$, which proves (4). The item (5) follows from the definition of the orthogonal projection. Since $\pi_U v \in U$ for all $v \in V$, we obtain from (1) that $\pi_U(\pi_U v) = \pi_U v$ for all $v \in V$, which shows (6).

Since $\langle \pi_U v, v - \pi_U v \rangle = 0$, we can apply Pythagoras' Theorem and obtain that

$$\|v\|^2 = \|\pi_U v\|^2 + \|v - \pi_U v\|^2,$$

which proves (7). In addition, we obtain that $\|\pi_U v\|^2 \le \|v\|^2$ with equality if and only if $\|v - \pi_U v\| = 0$, that is, $v = \pi_U v$, which in turn is the case if and only if $v \in U$. This proves (8).

Finally, assume that $v \in V$. Then we have for all $u \in U$ that

$$\|v - u\|^2 = \|v - \pi_U v + \pi_U v - u\|^2 = \|v - \pi_U v\|^2 + \|\pi_U v - u\|^2,$$

since $v - \pi_U v \in U^\perp$ and $\pi_U v - u \in U$. Thus the minimisation problem $\min_{u \in U} \|v - u\|^2$ is equivalent to the problem

$$\min_{u \in U} \|\pi_U v - u\|^2,$$

which has the obvious unique solution $u = \pi_U v$. This proves (9), which concludes the proof. □

4. Singular Value Decomposition

THEOREM 2.4.1 (Singular Value Decomposition (SVD)). *Assume that U and V are finite dimensional inner product spaces with inner products $\langle\cdot,\cdot\rangle_U$ and $\langle\cdot,\cdot\rangle_V$, respectively, and that $T\colon U \to V$ is linear. Denote by $p := \operatorname{rank} T = \dim(\operatorname{ran} T)$ the rank of T. Then there exist orthonormal sets $\{u_1,\ldots,u_p\} \subset U$ and $\{v_1,\ldots,v_p\} \subset V$, and real numbers $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_p > 0$ such that*

$$Tu = \sum_{i=1}^{p} \sigma\langle u, u_i\rangle v_i$$

for all $u \in U$.

The numbers σ_j are called the singular values *of T, and the systems $\{u_1,\ldots,u_p\}$ and $\{v_1,\ldots,v_p\}$* singular systems *of T.*

PROOF. We use induction over $n = \dim U$. For $n = 0$, the claim trivially holds (with $T\colon \{0\} \to V$ being the zero operator, $p = 0$, empty orthonormal systems and no singular values).

Assume therefore that $n \ge 1$ and that the claim holds for all linear transformations between inner product spaces, where the dimension of the domain is strictly smaller than n. If $Tu = 0$ for all $u \in U$, then there is nothing to show (we can again choose $p = 0$, empty orthonormal systems, and no singular values). Assume therefore without loss of generality that there exists some $u \in U$ with $Tu \neq 0$.

Denote now by $\|\cdot\|_U$ and $\|\cdot\|_V$ the norms on U and V, respectively, and consider the optimisation problem

$$\max_{\|u\|_U=1} \|Tu\|_V. \tag{12}$$

The function $u \mapsto \|Tu\|_V$ is continuous, and the set $\bigl\{u \in U : \|u\|_U = 1\bigr\}$ is closed and bounded, and thus this optimisation problem admits a solution.[2]

Now choose a solution u_1 of (12), define $\sigma_1 = \|Tu_1\|$ and $v_1 = Tu_1/\sigma_1$.[3] Then we have in particular that

$$\|Tu\|_V = \|u\|_U \Bigl\| T\Bigl(\frac{u}{\|u\|_U}\Bigr)\Bigr\|_V \le \sigma_1 \|u\|_U \tag{13}$$

for all $u \in U$ with $u \neq 0$, and the same inequality $\|Tu\|_V \le \sigma_1\|u\|_U$ obviously holds for $u = 0$.

Define now $U_1 = \operatorname{span}\{u_1\}$ and $U_2 = U_1^\perp$, and define similarly $V_1 = \operatorname{span}\{v_1\}$ and $V_2 = V_1^\perp$. We will show in the following that $Tw \in V_2$ for all $w \in U_2$.

Assume therefore that $w \in U_2$, and let $\alpha \in \mathbb{K}$. Then

$$T(\alpha u_1 + w) = \alpha T u_1 + Tw = \alpha\sigma_1 v_1 + Tw.$$

[2]Here we have actually used quite a number of results from calculus in several variables, which may not seem immediately obvious.

[3]Note here that the solution u_1 can never be unique, as $-u_1$ is another obvious solution.

As a consequence, we have that

$$\begin{aligned}
\|T(\alpha u_1 + w)\|_V^2 &= \|\alpha\sigma_1 v_1 + Tw\|_V^2 \\
&= \|\alpha\sigma_1 v_1 + \pi_{V_1}(Tw)\|_V^2 + \|Tw - \pi_{V_1} w\|_V^2 \\
&= |\alpha\sigma_1 + \langle Tw, v_1\rangle_V|^2 + \|Tw - \pi_{V_1} w\|_V^2 \\
&= \sigma_1^2|\alpha|^2 + 2\sigma_1|\alpha||\langle Tw, v_1\rangle_V| + |\langle Tw, v_1\rangle_V|^2 + \|Tw - \pi_{V_1} w\|_V^2 \\
&= \sigma_1^2|\alpha|^2 + 2\sigma_1|\alpha||\langle Tw, v_1\rangle_V| + \|\pi_{V_1} Tw\|_V^2 + \|Tw - \pi_{V_1} w\|_V^2 \\
&= \sigma_1^2|\alpha|^2 + 2\sigma_1|\alpha||\langle Tw, v_1\rangle_V| + \|Tw\|_V^2.
\end{aligned}$$

On the other hand, we have by (13) that

$$\|T(\alpha u_1 + w)\|_V^2 \le \sigma_1^2\|\alpha u_1 + w\|_U^2 = \sigma_1^2|\alpha|^2 + \sigma_1^2\|w\|_U^2.$$

Combining these estimates, we obtain that

$$\sigma_1^2|\alpha|^2 + 2\sigma_1|\alpha||\langle Tw, v_1\rangle_V| + \|Tw\|_V^2 \le \sigma_1^2|\alpha|^2 + \sigma_1^2\|w\|_U^2,$$

or

$$2\sigma_1|\alpha||\langle Tw, v_1\rangle_V| \le \sigma_1^2\|w\|_U^2 - \|Tw\|_V^2.$$

Since $\alpha \in \mathbb{K}$ was arbitrary and $\sigma_1 > 0$, this is only possible if $\langle Tw, v_1\rangle_V = 0$, that is, if $Tw \in V_1^\perp = V_2$.

We have thus shown that $Tw \in V_2$ whenever $w \in U_2$. Consider therefore the mapping $S\colon U_2 \to V_2$, $w \mapsto Sw := Tw$. Since $\dim U_2 < \dim U$, we can apply the induction hypothesis. We can easily see that $\operatorname{rank} S = p - 1$ (one dimension of the range of T being the space V_1). Thus there exist orthonormal systems $\{u_2, \ldots, u_p\}$ and $\{v_2, \ldots, v_p\}$ in U_2 and V_2, respectively, and positive numbers $\sigma_2 \ge \ldots \ge \sigma_p > 0$ such that

$$Sw = \sum_{i=2}^{p} \sigma_i \langle w, u_i\rangle v_i.$$

Since U_2 and U_1 are orthogonal, it follows that $\{u_1, u_2, \ldots, u_p\}$ is an orthonormal system in U, and similarly $\{v_1, v_2, \ldots, v_p\}$ is an orthonormal system in V. Moreover, we have that

$$Tu = T(\pi_{U_1} u) + T(\pi_{U_2} u) = T(\langle u, u_1\rangle u_1) + T(\pi_{U_2} u) = \sigma_1\langle u, u_1\rangle v_1 + \sum_{i=2}^{p}\langle u, u_i\rangle v_i$$

for all $u \in U$. Finally, we note that $v_2 = \sigma_2 u_2$ and thus, by (13),

$$\sigma_2 = \sigma_2\|v_2\| = \|u_2\| \le \sigma_1\|u_2\| = \sigma_1.$$

This shows that $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_p > 0$, which concludes the proof. □

Remark 2.4.2. One can show that the singular values of a mapping are unique. In addition, one obtains a limited uniqueness result for the singular vectors: If the singular values satisfy $\sigma_{j-1} > \sigma_j = \sigma_{j+1} = \ldots = \sigma_{j+k} > \sigma_{j+k+1} \ldots$, then the linear spans $\operatorname{span}\{u_j, \ldots, u_{j+k}\}$ and $\operatorname{span}\{v_j, \ldots, v_{j+k}\}$ are

Lemma 2.4.3. *Assume that the mapping $T\colon U \to V$ has the singular value decomposition*

$$Tu = \sum_{i=1}^{p} \sigma_i\langle u, u_i\rangle v_i.$$

Then

$$\operatorname{ran} T = \operatorname{span}\{v_1, \ldots, v_p\} \qquad \textit{and} \qquad \ker T = \operatorname{span}\{u_1, \ldots, u_p\}^\perp.$$

Proof. This is immediately obvious from the decomposition. □

We will come back soon to the singular value decomposition in order to discuss important properties, but also slightly different representations. For that, we will need some further notation, though.

5. Unitary Transformations and Adjoints

In real vector spaces, we have the notion of orthogonal transformations, which are linear mappings that preserve inner products between vectors. In complex vector spaces, we have the same idea, but we call these transformation unitary instead:

DEFINITION 2.5.1. Let U and V be inner product spaces, and let $T\colon U \to V$ be linear. Then T is called unitary, if

$$\langle Tu, Tv\rangle_V = \langle u, v\rangle_U$$

for all u, $v \in U$.

LEMMA 2.5.2. *Let U and V be inner product spaces, and let $T\colon U \to V$ be unitary. Then T is injective.*

PROOF. Assume that $u \in U$ is such that $Tu = 0$. Then

$$0 = \|Tu\|_V^2 = \langle Tu, Tu\rangle_V = \langle u, u\rangle_U = \|u\|_U^2$$

and thus $u = 0$. □

LEMMA 2.5.3. *Assume that U and V are finite dimensional inner product spaces with bases $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_m\}$, respectively, and denote by G_U and G_V the corresponding Gram matrices. Let moreover $T\colon U \to V$ be a linear transformation with matrix representation $A \in \mathbb{K}^{m\times n}$. Then T is unitary, if and only if*

$$G_U = A^H G_V A.$$

PROOF. Let u, $v \in U$ be arbitrary with coordinate vectors x, $y \in \mathbb{K}^n$. Then

$$\langle u, v\rangle_U = y^H G_U x,$$

and

$$\langle Tu, Tv\rangle_V = (Ax)^H G_V Ay = x^H (A^H G_V A)y.$$

Thus T is unitary, if and only if

$$y^H G_U x = x^H (A^H G_V A)y$$

for all x, $y \in \mathbb{K}^n$, which is the case if and only if $G_U = A^H G_V A$. □

REMARK 2.5.4. If $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_m\}$ are orthonormal bases of U and V, then the corresponding Gram matrices are just the identity matrices of the corresponding dimensions. In this case, the transformation is unitary, if and only if

$$A^H A = \mathrm{Id}_n,$$

that is, *matrix A is unitary.*

DEFINITION 2.5.5 (Adjoint of a linear transformation). Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Assume moreover that T has the singular value decomposition

$$Tu = \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle v_i$$

for all $u \in U$. The *adjoint transformation* $T^*\colon V \to U$ is defined as

$$T^* v = \sum_{i=1}^{p} \sigma_i \langle v, v_i\rangle u_i$$

for all $v \in V$.

THEOREM 2.5.6. *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. The adjoint $T^*\colon V \to U$ is the unique linear transformation satisfying*

$$\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U \qquad \text{for all } u \in U \text{ and } v \in V. \tag{14}$$

PROOF. The linearity of T^* is obvious from its definition.

We next show that (14) holds. Let therefore $u \in U$ and $v \in V$. Then

$$\langle Tu, v\rangle_V = \Big\langle \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle_U v_i, v\Big\rangle_V = \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle_U \langle v_i, v\rangle_V.$$

On the other hand, we have that

$$\begin{aligned}\langle u, T^*v\rangle_V = \Big\langle u, \sum_{i=1}^{p} \sigma_i \langle v, v_i\rangle_V u_i\Big\rangle_U \\ = \sum_{i=1}^{p} \sigma_i \overline{\langle v, v_i\rangle_V} \langle u, u_i\rangle_U = \sum_{i=1}^{p} \sigma_i \langle v_i, v\rangle_V \langle u, u_i\rangle_U.\end{aligned}$$

The resulting expressions are the same in the two cases, which shows that (14) holds.

It remains to show that T^* is the unique linear transformation with this property. Assume therefore that $S\colon V \to U$ is another linear transformation such that $\langle Tu, v\rangle_V = \langle u, Sv\rangle_U$ for all $u \in U$ and $v \in V$. Then

$$\langle u, T^*v\rangle_U = \langle Tu, v\rangle_V = \langle u, Sv\rangle_U$$

for all $u \in U$ and $v \in V$, and thus

$$\langle u, (T^* - S)v\rangle_U = 0$$

for all $u \in U$ and $v \in V$. That is, $(T^* - S)v \in U^\perp = \{0\}$ for all $v \in V$, which is the same as saying that $T^*v = Sv$ for all $v \in V$, that is, $T^* = S$. □

LEMMA 2.5.7. *Assume that U and V are inner product spaces and that $T\colon U \to V$ is linear. Then*

$$T = (T^*)^*.$$

PROOF. Exercise!

(Use Theorem 2.5.6!) □

LEMMA 2.5.8. *Assume that U, V, W are inner product spaces and that $T\colon U \to V$ and $S\colon V \to W$ are linear. Then*

$$(S \circ T)^* = T^* \circ S^*.$$

PROOF. Exercise!

(Use Theorem 2.5.6!) □

LEMMA 2.5.9. *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Then*

$$\ker T = (\operatorname{ran} T^*)^\perp \qquad \text{and} \qquad \operatorname{ran} T = (\ker T^*)^\perp.$$

PROOF. We have that $u \in \ker T$ if and only if $Tu = 0$. This is equivalent to u satisfying $\langle Tu, v\rangle_V = 0$ for all $v \in V$. Using Theorem 2.5.6, this in turn is equivalent to u satisfying $\langle u, T^*v\rangle_V = 0$ for all $v \in V$, which is equivalent to the inclusion $u \in (\operatorname{ran} T^*)^\perp$. This shows that $\ker T = (\operatorname{ran} T^*)^\perp$.

By applying this result to T^* and recalling that $(T^*)^* = T$, we obtain that $\ker T^* = (\operatorname{ran} T)^\perp$. Taking orthogonal complements on both sides, we finally obtain that $\operatorname{ran} T = (\ker T^*)^\perp$ as claimed. □

Adjoints of unitary transformations.

Using the notion of the adjoint of a mapping, we can find a different characterisation of unitary transformations:

PROPOSITION 2.5.10. *Assume that U and V are inner product spaces and that $T\colon U \to V$ is a linear transformation. Then T is unitary, if and only if*

$$T^*T = \mathrm{Id}_U\,.$$

PROOF. Assume first that $T^*T = \mathrm{Id}_U$. Then we have for all u, $v \in U$ that

$$\langle Tu, Tv\rangle_V = \langle u, T^*Tv\rangle_U = \langle u, v\rangle_U,$$

which shows that T is unitary.

Conversely, assume that T is unitary. Then

$$\langle u, T^*Tv\rangle_U = \langle Tu, Tv\rangle_V = \langle u, v\rangle_U$$

for all u, $v \in U$, and thus

$$\langle u, (T^*T - \mathrm{Id}_U)v\rangle_U = 0$$

for all u, $v \in U$. This shows that $(T^*T - \mathrm{Id}_U)v \in U^\perp = \{0\}$ for all $v \in U$, which in turn shows that $T^*T = \mathrm{Id}_U$. □

PROPOSITION 2.5.11. *Assume that U and V are inner product spaces and that $T\colon U \to V$ is unitary. Then*

$$TT^* = \pi_{\operatorname{ran} T}.$$

PROOF. Let $v \in V$. Then we can decompose v as $v = v_1 + v_2$ with $v_1 \in \operatorname{ran} T$ and $v_2 \in (\operatorname{ran} T)^\perp$. We have to show that $TT^*v = v_1$.

Since $v_1 \in \operatorname{ran} T$, there exists $u_1 \in U$ such that $v_1 = Tu_1$. Since $T^*T = \mathrm{Id}_U$, we then have that

$$\begin{aligned} TT^*v = TT^*(v_1 + v_2) &= TT^*Tu_1 + TT^*v_2 \\ &= T(T^*Tu_1) + TT^*v_2 = Tu_1 + TT^*v_2 = v_1 + TT^*v_2. \end{aligned}$$

Thus it remains to show that $TT^*v_2 = 0$.

By the definition of the adjoint, we have for every $w \in V$ that

$$\langle TT^*v_2, w\rangle_V = \langle T^*v_2, T^*w\rangle_U = \langle v_2, (T^*)^*T^*w\rangle_V = \langle v_2, TT^*w\rangle_V.$$

Since $TT^*w \in \operatorname{ran} T$ and $v_2 \in (\operatorname{ran} T)^\perp$, we conclude that $\langle TT^*v_2, w\rangle_V = 0$. Since $w \in V$ was arbitrary, this shows that $TT^*v_2 = 0$, which concludes the proof. □

Adjoints in matrix form.

Finally, we investigate how the adjoint of a linear transformation looks in matrix form:

LEMMA 2.5.12. *Assume that the linear transformation $T\colon U \to V$ has the matrix representation $A \in \mathbb{K}^{m\times n}$ with respect to bases of U and V, and denote by G_U and G_V the Gram matrices of U and V. Then the matrix A^* of the adjoint transformation T^* is*

$$A^* = G_U^{-1}A^H G_V.$$

PROOF. The adjoint T^* is uniquely characterised by the condition $\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U$ for all $u \in U$ and $v \in V$. Translated to coordinates, this means that

$$y^H G_V(Ax) = (A^*y)^H G_U x$$

for all $x \in \mathbb{K}^n$ and $y \in \mathbb{K}^m$. We can rewrite this to the equation

$$yHG_V Ax = y^H (A^*)^H G_U x.$$

Since this has to hold for all $x \in \mathbb{K}^n$ and $y \in \mathbb{K}^m$, we obtain that

$$G_V A = (A^*)^H G_U,$$

or

$$(A^*)^H = G_V A G_U^{-1}.$$

Finally, we take the Hermitian conjugate on both sides and obtain

$$A^* = (G_U^{-1})^H A^H G_V^H = G_U^{-1} A^H G_V,$$

as the matrices G_U and G_V, and thus also their inverses (which exist as they are positive definite), are Hermitian. $\square$

REMARK 2.5.13. If $A \in \mathbb{K}^{m\times n}$ is the matrix of linear transformation with respect to *orthonormal bases* on the involved spaces, then

$$A^* = A^H.$$

6. Singular Value Decomposition Revisited

Assume now again that U and V are inner product spaces, and that the linear transformation $T\colon U \to V$ has the singular value decomposition

$$Tu = \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle v_i.$$

We can alternatively interpret this as a composition of three different mappings:

- The first mapping is the transformation

$$\tilde{P}\colon U \to \mathbb{K}^p,$$
$$u \mapsto \begin{pmatrix} \langle u, u_1\rangle \\ \vdots \\ \langle u, u_p\rangle \end{pmatrix}.$$

- The second mapping is the transformation

$$\Sigma\colon \mathbb{K}^p \to \mathbb{K}^p,$$
$$\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \mapsto \begin{pmatrix} \sigma_1 x_1 \\ \vdots \\ \sigma_p x_p \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_p \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

- The third and final mapping is the transformation

$$Q\colon \mathbb{K}^p \to V,$$
$$\begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \mapsto \sum_{i=1}^{p} y_i v_i.$$

Moreover, if we consider the standard (Euclidean) inner product on $\mathbb{K}^p$, then the orthonormality of the set $\{v_1, \dots, v_p\}$ implies that the mapping $Q\colon \mathbb{K}^p \to V$ is unitary.

The mapping $\tilde{P}\colon U \to \mathbb{K}^p$ typically is not unitary itself (it actually will not be injective, if T is not injective). However, it turns out that its adjoint is. More precisely, consider the mapping

$$P\colon \mathbb{K}^p \to U,$$
$$\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \mapsto \sum_{i=1}^{p} x_i u_i.$$

Then P is unitary and $\tilde{P} = P^*$.

In addition, we have that $\operatorname{ran} Q = \operatorname{span}\{v_1, \ldots, v_p\} = \operatorname{ran} T$, and $\operatorname{ran} P = \operatorname{span}\{u_1, \ldots, u_p\} = (\ker T)^{\perp}$.

All in all, we obtain the following re-interpretation of the singular value decomposition:

Theorem 2.6.1 (Singular Value Decomposition, again). *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Denote by p the rank of T. There exist unitary mappings $P\colon \mathbb{K}^p \to U$ and $Q\colon \mathbb{K}^p \to V$ and a diagonal mapping $\Sigma\colon \mathbb{K}^p \to \mathbb{K}^p$ with diagonal entries $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_p > 0$ such that*

$$T = Q \circ \Sigma \circ P^*.$$

Moreover, $\operatorname{ran} T = \operatorname{ran} Q$ *and* $\ker T = (\operatorname{ran} P)^{\perp}$.

Finally, we can rewrite this decomposition in terms of matrix products, once we have chosen orthonormal bases on the spaces U and V.

Theorem 2.6.2 (Singular Value Decomposition in Matrix Form). *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear, and that $A \in \mathbb{K}^{m\times n}$ is the matrix of T with respect to orthonormal bases on U and V, respectively. Then there exist unitary matrices $Q \in \mathbb{K}^{m\times p}$ and $P \in \mathbb{K}^{n\times p}$ as well as a diagonal matrix $\Sigma \in \mathbb{R}^{p\times p}$ with positive entries $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_p > 0$ such that*

$$A = Q\Sigma P^H.$$

7. The Moore–Penrose Inverse

Assume that we want to solve a linear system $Tu = v$. If T is bijective, then the solution is simply $v = T^{-1}u$. Also, this solution can be actually computed numerically (provided we have sufficient time and memory available), and we can use any linear solver of our choice to do so. If, however, T is not injective, then solutions of the equation will not be unique, and if T is not surjective, they might not exist at all. In that case, we can use the singular value decomposition of T to define a *generalised inverse* that comes, in a sense, as close as possible to an actual inverse of T.

Definition 2.7.1. Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ has the singular value decomposition

$$Tu = \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle v_i.$$

The *Moore–Penrose inverse* or *pseudo-inverse* of T is the mapping $T^\dagger\colon V \to U$,

$$T^\dagger v = \sum_{i=1}^{p} \frac{1}{\sigma_i} \langle v, v_i\rangle u_i.$$

Remark 2.7.2. Alternatively, if we use the decomposition

$$T = Q \circ \Sigma \circ P^*$$

as in Theorem 2.6.1, we can write the Moore–Penrose inverse as

$$T^\dagger = P \circ \Sigma^{-1} \circ Q^*.$$

THEOREM 2.7.3. *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Then the Moore–Penrose inverse $T^\dagger$ of T has the following properties:*

(1) $\ker T^\dagger = (\operatorname{ran} T)^\perp$.
(2) $\operatorname{ran} T^\dagger = (\ker T)^\perp$.
(3) $TT^\dagger = \pi_{\operatorname{ran} T}$.
(4) $T^\dagger T = \pi_{(\ker T)^\perp}$.
(5) $T^\dagger T T^\dagger = T^\dagger$.
(6) $TT^\dagger T = T$.

PROOF. Write $T = Q\Sigma P^*$ with $Q\colon \mathbb{K}^p \to U$ and $P\colon \mathbb{K}^p \to V$ unitary, and $\Sigma\colon \mathbb{K}^p \to \mathbb{K}^p$ diagonal with positive diagonal entries. Then $T^\dagger = P\Sigma^{-1}Q^*$.

From the results of the singular value decomposition we now obtain that $\operatorname{ran} T = \operatorname{ran} Q$ and $\ker T = (\operatorname{ran} P)^\perp$. Now note that the decomposition $T^\dagger = P\Sigma^{-1}Q^*$ is actually a singular value decomposition of $T^\dagger$ (up to a necessary reordering of the singular values). Thus we also have that $\operatorname{ran} T^\dagger = \operatorname{ran} P$ and $\ker T^\dagger = (\operatorname{ran} Q)^\perp$. Therefore $\operatorname{ran} T^\dagger)^\perp = \operatorname{ran} P = (\ker T)^\perp$ and $\ker T^\dagger = (\operatorname{ran} Q)^\perp = (\operatorname{ran} T)^\perp$.

Recall now that P and Q are unitary, and therefore $P^*P = \mathrm{Id}_p$ and $Q^*Q = \mathrm{Id}_p$. Thus

$$TT^\dagger = Q\Sigma P^* P\Sigma^{-1}Q^* = Q\Sigma\Sigma^{-1}Q^* = QQ^*.$$

This in turn shows that

$$TT^\dagger = QQ^* = \pi_{\operatorname{ran} Q} = \pi_{\operatorname{ran} T}.$$

Next we have that

$$T^\dagger T = P\Sigma^{-1}Q^*Q\Sigma P^* = P\Sigma^{-1}\Sigma P^* = PP^* = \pi_{\operatorname{ran} P} = \pi_{(\ker T)^\perp}.$$

Moreover,

$$T^\dagger T T^\dagger = PP^*P\Sigma^{-1}Q^* = P\Sigma^{-1}Q^* = T^\dagger$$

and

$$TT^\dagger T = QQ^*Q\Sigma P^* = Q\Sigma P^* = T,$$

which concludes the proof. □

THEOREM 2.7.4. *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Let $v \in V$. Then*

$$u^\dagger := T^\dagger v$$

solves the (bi-level) optimisation problem

$$\min_u \|u\|_U^2 \qquad \textit{s.t.} \qquad u \textit{ solves } \min_u \|Tu - v\|_V^2. \tag{15}$$

In other words, the Moore–Penrose inverse looks at the least squares problem $\min_u \|Tu - v\|^2$ and selects from all solutions of this problems the one with the smallest norm.

PROOF. The vector $u \in U$ solves the problem $\min_u \|Tu - v\|_V^2$, if and only if the vector $\hat{v} = Tu$ solves the problem

$$\min_{\hat{v}} \|\hat{v} - v\|_V^2 \qquad \text{s.t. } \hat{v} \in \operatorname{ran} T.$$

In other words, $Tu = \hat{v} = \pi_{\operatorname{ran} T} v$. Thus we can reformulate (15) equivalently as

$$\min_u \|u\|_U^2 \qquad \text{s.t.} \qquad Tu = \pi_{\operatorname{ran} T} v. \tag{16}$$

We will show in the following that $u^\dagger = T^\dagger v$ solves (16).

For that, we note first that $Tu^\dagger = TT^\dagger v = \pi_{\operatorname{ran} T} v$, which shows that $u^\dagger$ satisfies the constraint in (16). Now assume that $\tilde{u} \in U$ is an arbitrary vector satisfying $T\tilde{u} = \pi_{\operatorname{ran} T} v$. Then $T(\tilde{u} - u^\dagger) = 0$ and thus $\tilde{u} - u^\dagger \in \ker T$. Since $u^\dagger = T^\dagger v^\dagger \in \operatorname{ran} T^\dagger = (\ker T)^\perp$, it follows that

$$\|\tilde{u}\|_U^2 = \|\tilde{u} - u^\dagger\|_U^2 + \|u^\dagger\|_U^2 \geq \|u^\dagger\|_U^2.$$

Since this inequality holds for all $\tilde{u}$ satisfying $T\tilde{u} = \pi_{\operatorname{ran} T} v$, it follows that $u^\dagger$ solves (16), which concludes the proof. $\square$

8. Singular Value Decomposition of Matrices

We consider now specifically the case where $U = \mathbb{K}^n$, $V = \mathbb{K}^m$, both regarded with the standard inner product, and $T\colon U \to V$ is a linear mapping, which we can identify with a matrix $A \in \mathbb{K}^{m \times n}$. According to Theorem 2.6.2, we can decompose A as

$$\begin{array}{ccccc} A & = & Q & \Sigma & P^H \\ \Cap & & \Cap & \Cap & \Cap \\ \mathbb{K}^{m\times n} & & \mathbb{K}^{m\times p} & \mathbb{K}^{p\times p} & \mathbb{K}^{p\times n} \end{array}$$

with Q and P unitary matrices, that is $Q^H Q = \mathrm{Id}_p$ and $P^H P = \mathrm{Id}_p$, and Σ a diagonal matrix with non-increasing positive diagonal entries.

Now consider the matrices $A^H A \in \mathbb{K}^{n\times n}$ and $AA^H \in \mathbb{K}^{m \times m}$. We have

$$A^H A = (Q\Sigma P^H)^H Q \Sigma P^H = P\Sigma^H Q^H Q \Sigma P^H = P\Sigma^2 P^H,$$

and similarly

$$AA^H = Q\Sigma P^H (Q\Sigma P^H)^H = Q\Sigma P^H P \Sigma^H Q^H = Q\Sigma^2 Q^H.$$

Next we will have a closer look at the decomposition of $A^H A$ and provide a different interpretation of it. The matrix $P \in \mathbb{K}^{n\times p}$ is unitary, and thus its columns form an orthonormal system in $\mathbb{K}^n$ (the columns of P are precisely the coordinates of the vectors $u_1, \ldots, u_p$ in the SVD). We can now use Gram–Schmidt orthogonalisation in order to extend this orthonormal system to an orthonormal basis of $\mathbb{K}^n$ and then collect this basis in a square matrix. That is, we can find a matrix $P' \in \mathbb{K}^{n \times (n-p)}$ such that

$$\hat{P} := \begin{pmatrix} P & P' \end{pmatrix} \in \mathbb{K}^{n\times n}$$

is unitary. In addition, we can extend the matrix $\Sigma \in \mathbb{K}^{p\times p}$ to a matrix $\tilde{\Sigma} \in \mathbb{K}^{n \times n}$ by adding zeroes as necessary. That is, we define the matrix

$$\tilde{\Sigma} := \begin{pmatrix} \Sigma & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & 0_{(n-p)\times(n-p)} \end{pmatrix} \in \mathbb{K}^{n\times n}.$$

Then

$$\hat{P}\tilde{\Sigma}^2\hat{P}^H = \begin{pmatrix} P & P' \end{pmatrix} \begin{pmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P^H \\ (P')^H \end{pmatrix} = \begin{pmatrix} P & P' \end{pmatrix} \begin{pmatrix} \Sigma^2 P^H \\ 0 \end{pmatrix} = P\Sigma^2 P^H = A^H A.$$

Now note that $\hat{P} \in \mathbb{K}^{n\times n}$ satisfies $\hat{P}^H \hat{P} = \mathrm{Id}_n$. Since $\hat{P}$ is a square matrix, this implies that $\hat{P}$ is actually invertible with

$$\hat{P}^{-1} = \hat{P}^H.$$

Thus we actually have the decomposition

$$A^H A = \hat{P}\tilde{\Sigma}^2\hat{P}^{-1}.$$

This means that $\tilde{\Sigma}^2$ is the matrix of the mapping $A^H A$ in the basis of $\mathbb{K}^n$ given by the columns of $\hat{P}$. Since moreover $\tilde{\Sigma}^2 \in \mathbb{K}^{n\times n}$ is diagonal, this implies that $A^H A$ is diagonalisable with non-zero eigenvalues $\sigma_1^2, \ldots, \sigma_p^2$ and the eigenvalue 0 with multiplicity $(n-p)$. In addition, the columns of $\hat{P}$ form an (orthonormal) eigenbasis for A^HA. Moreover, the eigenvectors of A^HA that correspond to non-zero eigenvalues are precisely the columns of the matrix P.

We can now apply the same idea to the matrix $AA^H \in \mathbb{K}^{m\times m}$. That is, we find a matrix $Q' \in \mathbb{K}^{m\times(m-p)}$ such that $\hat{Q} := (QQ') \in \mathbb{K}^{m\times m}$ is unitary, and then obtain that

$$AA^H = \hat{Q}\begin{pmatrix}\Sigma^2 & 0\\ 0 & 0\end{pmatrix}\hat{Q}^{-1},$$

where we have again filled up the matrix $\Sigma^2 \in \mathbb{K}^{p\times p}$ with zeroes to obtain an $(m\times m)$-matrix. Again, we obtain that $\sigma_1^2,\ldots,\sigma_p^2$ are the non-zero eigenvalues of AA^H, and the columns of Q are corresponding eigenvectors.

The considerations above yield the following result:

THEOREM 2.8.1. *Let $A \in \mathbb{K}^{n\times m}$. The singular values of A are precisely the square roots of the non-zero eigenvalues of the matrices A^HA or A^HA, and the singular vectors of A are eigenvalues of A^HA and A^HA, respectively.*

Reduced and full SVD.

The decomposition of a matrix $A = Q\Sigma P^H$ with $Q \in \mathbb{K}^{m\times p}$, $\Sigma \in \mathbb{K}^{p\times p}$, and $P \in \mathbb{K}^{n\times p}$ is in the literature often called the *reduced singular value decomposition* of the matrix A. The *full singular value decomposition* then is of the form $A = \hat{Q}\hat{\Sigma}\hat{P}^H$ with $\hat{Q} \in \mathbb{K}^{m\times m}$, $\hat{\Sigma}^{m\times n}$, and $\hat{P} \in \mathbb{K}^{n\times n}$. Here the matrices $\hat{Q}$ and $\hat{P}$ are unitary, and $\hat{\Sigma}$ is a *rectangular* diagonal matrix with ordered non-negative diagonal entries. That is, $\hat{\Sigma}_{11} \ge \hat{\Sigma}_{22} \ge \ldots \ge \hat{\Sigma}_{kk} \ge 0$, where $k = \min\{m,n\}$, and all other entries of $\hat{\Sigma}$ are equal to zero.

In order to compute the (full) SVD of a matrix $A \in \mathbb{K}^{m\times n}$ one can proceed as follows: If $n \le m$, we start by computing $A^HA \in \mathbb{K}^{n\times n}$ and compute the eigenvalues as well as an orthonormal eigenbasis of A^HA. That is, we write

$$A^HA = \hat{P}\tilde{\Sigma}^2\hat{P}^2$$

with $\hat{P} \in \mathbb{K}^{n\times n}$ unitary and $\tilde{\Sigma}^2$ diagonal (with necessarily non-negative diagonal entries).

We then define the matrix $\hat{\Sigma} \in \mathbb{K}^{m\times n}$ by adding zeroes, that is,

$$\hat{\Sigma} = \begin{pmatrix}\tilde{\Sigma}\\ 0_{(m-n)\times n}\end{pmatrix},$$

and we write

$$A = \hat{Q}\hat{\Sigma}\hat{P}^H$$

with yet to be determined $\hat{Q} \in \mathbb{K}^{m\times m}$. By multiplying this equation from the right with $\hat{P}$, we obtain that

$$A\hat{P} = \hat{Q}\hat{\Sigma} = \begin{pmatrix}\hat{\sigma}_1\hat{q}_1 & \hat{\sigma}_2\hat{q}_2 & \ldots & \hat{\sigma}_n\hat{q}_n\end{pmatrix},$$

where $\hat{q}_1,\ldots,\hat{q}_m$ are the columns of $\hat{Q}$. Thus we obtain the j-th column of $\hat{Q}$ by dividing the j-th column of $A\hat{P}$ by $\hat{\sigma}_j$ whenever possible (that is, whenever $\hat{\sigma}_j \ne 0$). Finally, we obtain the remaining columns of $\hat{Q}$ by completing the computed ones to an orthonormal basis of $\mathbb{K}^m$.

If $m < n$, we can in theory use the same approach, but it is usually simpler to start with the matrix $AA^H = \hat{Q}\tilde{\Sigma}^2\hat{Q}^H$ instead, as this matrix is smaller than A^HA. Thus one computes first an orthonormal eigenvalue decomposition $AA^H = \hat{Q}\tilde{\Sigma}^2\hat{Q}^H$. Then the j-th column of $\hat{P}$ can be obtained by dividing the j-th column of $A^H\hat{Q}$ by $\hat{\sigma}_j$ if possible. The remaining columns of $\hat{P}$ can again be obtained by orthogonalisation.

EXAMPLE 2.8.2. Consider the matrix

$$A = \begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$$

with $c \neq 0$.

We first compute a Jordan decomposition of A in order to demonstrate the difference to the SVD. This matrix has a single geometric eigenvalue $\lambda_1 = 1$ with corresponding eigenvector $(1,0)$. A possible Jordan decomposition of A reads

$$A = \begin{pmatrix} c^{-1} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix}.$$

Next we will compute a singular value decomposition of A. To that end, we will compute first the eigenvalues and eigenvectors of AA^H (doing the computations with A^HA would work fine as well[4]). We have

$$AA^H = \begin{pmatrix} 1+c^2 & c \\ c & 1 \end{pmatrix}$$

with eigenvalues

$$\lambda_{1,2} = 1 + \frac{c^2}{2} \pm \sqrt{c^2 + \frac{c^4}{4}}.$$

As a consequence, the singular values of A are $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, or

$$\sigma_1 = \sqrt{1 + \frac{c^2}{2} + \sqrt{c^2 + \frac{c^4}{4}}} \qquad \text{and} \qquad \sigma_2 = \sqrt{1 + \frac{c^2}{2} - \sqrt{c^2 + \frac{c^4}{4}}}.$$

In the particular case $c = 8/3$ (which noticeable simplifies all the following calculations) we have

$$\sigma_1 = \sqrt{\lambda_1} = 3 \qquad \text{and} \qquad \sigma_2 = \sqrt{\lambda_2} = 1/3.$$

Moreover, the eigenvalues of AA^H corresponding to λ_1 and λ_2 are

$$q_1 = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad \text{and} \qquad q_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} -1 \\ 3 \end{pmatrix}.$$

Thus we can write

$$AA^H = Q\Sigma^2 Q^H = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 1/9 \end{pmatrix} \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}.$$

Moreover, the matrix Q in this decomposition of AA^T can be chosen to be precisely the matrix Q in the singular value decomposition $A = Q\Sigma P^H$ of A. Now the equation $A = Q\Sigma P^H$ implies that

$$P = A^H Q \Sigma^{-1} = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 & -3 \\ 3 & 1 \end{pmatrix}.$$

We therefore obtain the singular value decomposition

$$\begin{pmatrix} 1 & 8/3 \\ 0 & 1 \end{pmatrix} = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ 0 & 1/3 \end{pmatrix} \frac{1}{\sqrt{10}} \begin{pmatrix} 1 & 3 \\ -3 & 1 \end{pmatrix}.$$

---

[4]The decision of whether to start with AA^H or A^HA should be based on which of these matrices is expected to look simpler. In this case, there is no real difference between them, so it does not matter. In some cases, one can make one's life significantly harder, though, by starting the computations the wrong way.

9. Self-adjoint and Positive Definite Transformations

We now turn back to the more general case where T is a linear transformation between the finite dimensional inner product spaces U and V, and try to emulate the results of the previous section also in this matrix free situation. Since the matrix A^H corresponds to the operator T^*, this indicates that we should look at the eigenvalues and eigenvectors of the transformations T^*T and TT^*, respectively. For that, we will first discuss some properties of these types of transformations independent of the singular value decomposition.

DEFINITION 2.9.1. A linear transformation $T\colon U \to U$ on an inner product spaces U is called *self-adjoint* or *Hermitian* if $T^* = T$.

PROPOSITION 2.9.2. *Assume that U is a finite dimensional inner product spaces and that $T\colon U \to U$ is self-adjoint. Then the eigenvalues of T are real. In addition, if $\lambda_1 \neq \lambda_2$ are different eigenvalues of T with corresponding eigenvectors u_1 and u_2, then $\langle u_1, u_2\rangle = 0$.*

PROOF. Assume that $\lambda \in \mathbb{K}$ is an eigenvalue of T with eigenvector $u \in U$. Then

$$\lambda\|u\|^2 = \langle \lambda u, u\rangle = \langle Tu, u\rangle = \langle u, T^*u\rangle = \langle u, Tu\rangle = \langle u, \lambda u\rangle = \overline{\lambda}\|u\|^2.$$

Since $\|u\| \neq 0$, it follows that $\lambda = \overline{\lambda}$, and thus $\lambda \in \mathbb{R}$.

Next assume that $\lambda_1 \neq \lambda_2$ are distinct eigenvalues of T and that u_1, u_2 are corresponding eigenvectors. Then

$$\begin{aligned}\lambda_1\langle u_1, u_2\rangle = \langle \lambda_1 u_1, u_2\rangle = \langle Tu_1, u_2\rangle = \langle u_1, T^*u_2\rangle = \langle u_1, Tu_2\rangle \\ = \langle u_1, \lambda_2 u_2\rangle = \overline{\lambda_2}\langle u_1, u_2\rangle = \lambda_2\langle u_1, u_2\rangle.\end{aligned}$$

Thus

$$(\lambda_1 - \lambda_2)\langle u_1, u_2\rangle = 0,$$

which shows that $\langle u_1, u_2\rangle = 0$. □

Even more, it is actually possible to show that self-adjoint transformations are diagonalisable:

THEOREM 2.9.3. *Assume U is a finite dimensional inner product spaces and that $T\colon U \to U$ is self-adjoint. Then U has an orthonormal basis consisting of eigenvectors of T. In particular, T is diagonalisable.*

REMARK 2.9.4. One can show that a more general class of linear transformations has an orthonormal eigenbasis. We say that a linear transformation $T\colon U \to U$ is *normal*, if $TT^* = T^*T$. One can then show that T is normal, if and only if U has an orthonormal basis consisting of eigenvectors of T.

DEFINITION 2.9.5. Assume that U is a finite dimensional inner product space and $T\colon U \to U$ is self-adjoint. We say that T is positive semi-definite, if

$$\langle Tu, u\rangle \geq 0$$

for all $u \in U$. We say that T is positive definite, if $\langle Tu, u\rangle > 0$ for all $u \in U$, $u \neq 0$.

LEMMA 2.9.6. *Assume that U is a finite dimensional inner product space and $T\colon U \to U$ is Hermitian. Then T is positive semi-definite, if and only if all eigenvalues of T are non-negative.*

Moreover, T is positive definite, if and only if all eigenvalues of T are positive.

PROOF. Assume first that T is positive semi-definite and that λ is an eigenvalue of T with eigenvector u. Then

$$\lambda\|u\|^2 = \langle \lambda u, u\rangle = \langle Tu, u\rangle \geq 0,$$

which shows that $\lambda \geq 0$.

Now assume that all eigenvalues of T are non-negative. Denote by $\lambda_1, \ldots, \lambda_n \geq 0$ the eigenvalues of T, and let $\{u_1, \ldots, u_n\}$ be a corresponding orthonormal basis of eigenvectors. Let moreover $u \in U$. Then we can write

$$u = x_1 u_1 + \ldots + x_n u_n$$

for some $x_i \in \mathbb{K}$, $i = 1, \ldots, n$. Since the vectors u_i are eigenvectors of T for the eigenvalues λ_i, we have that

$$Tu = \lambda_1 x_1 u_1 + \ldots + \lambda_n x_n u_n.$$

Since $\{u_1, \ldots, u_n\}$ is an orthonormal basis of U and $\lambda_i \geq 0$ for all i, we obtain that

$$\langle Tu, u\rangle = \sum_{i=1}^{n} \lambda_i |x_i|^2 \geq 0,$$

which shows that T is positive semi-definite.

The proof for positive definiteness is similar. □

LEMMA 2.9.7. *Assume that U and V are finite dimensional inner product spaces and $T: U \to V$ is a linear transformation. Then $T^*T: U \to U$ and $TT^*: V \to V$ are self-adjoint positive semi-definite.*

PROOF. We note first that $(T^*T)^* = T^*(T^*)^* = T^*T$ and $(TT^*)^* = (T^*)^*T^* = TT^*$, which shows that T^*T and TT^* are self-adjoint.

Next we have for all $u \in U$ that

$$\langle u, T^*Tu\rangle_U = \langle Tu, Tu\rangle_V = \|Tu\|_V^2 \geq 0,$$

and for all $v \in V$ that

$$\langle TT^*v, v\rangle_V = \langle T^*v, T^*v\rangle_U = \|T^*v\|_U^2 \geq 0,$$

which proves the positive semi-definiteness of T^*T and TT^*. □

In order to relate the singular value decomposition of T to the properties of T^*T and TT^*, we need an additional result that shows that the transformations T and T^*T have the same kernel.

LEMMA 2.9.8. *Assume that U and V are finite dimensional inner product spaces and $T: U \to V$ is a linear transformation. Then $\ker(T^*T) = \ker T$.*

PROOF. Assume first that $u \in \ker T$. Then $Tu = 0$ and thus also $T^*Tu = T^*0 = 0$. This shows that $\ker T \subset \ker(T^*T)$.

Conversely, assume that $u \in \ker(T^*T)$. Then $T^*Tu = 0$, and thus

$$\langle T^*Tu, u\rangle_U = \langle Tu, Tu\rangle_V = \|Tu\|_V^2 = 0,$$

which shows that $Tu = 0$ and thus $u \in \ker T$. □

THEOREM 2.9.9. *Assume that U and V are finite dimensional inner product spaces and that $T: U \to V$ is linear. Let $\{u_1, \ldots, u_n\}$ be an orthonormal eigenbasis of T^*T for the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$.*

Denote $p = \operatorname{rank} T$, and define

$$\sigma_i = \sqrt{\lambda_i} \qquad \text{and} \qquad v_i = \frac{Tu_i}{\sigma_i}$$

for $k = 1, \ldots, p$. *Then*

$$Tu = \sum_{i=1}^{p} \sigma_i \langle u, u_i \rangle v_i \tag{17}$$

is a singular value decomposition of T.

In particular, the singular values of T *are precisely the non-zero eigenvalues of the mapping* T^*T.

Proof. We have to show that the numbers σ_i and the vectors v_i are well-defined, and that all the conditions for a singular value decomposition are satisfied. That is, that the sets $\{u_1, \ldots, u_p\}$ and $\{v_1, \ldots, v_p\}$ are orthonormal, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$, and that Tu actually has the form given in (17).

Since $\ker(T^*T) = \ker T$, it follows that

$$\begin{aligned}\operatorname{rank}(T^*T) = \dim(\operatorname{ran} T^*T) &= \dim U - \dim(\ker T^*T) \\ &= \dim U - \dim(\ker T) = \dim(\operatorname{ran} T) = \operatorname{rank} T.\end{aligned}$$

Thus the number of non-zero eigenvalues of T^*T (with multiplicity) is precisely p. Since T^*T is positive semi-definite, all its non-zero eigenvalues are strictly positive, and thus the numbers $\sigma_i > 0$ and the vectors v_i are well-defined. Since $\lambda_1 \geq \lambda_2 \geq \ldots$, we also have that $\sigma_1 \geq \sigma_2 \geq \ldots$.

Since $\{u_1, \ldots, u_n\}$ is an orthonormal basis of U, the set $\{u_1, \ldots, u_p\}$ is orthonormal as well. Moreover, we have that

$$u = \sum_{i=1}^{n} \langle u, u_i \rangle u_i$$

for all $u \in U$. Now we note that $u_i \in \ker(T^*T) = \ker T$ for all $i = p+1, \ldots, n$, since the corresponding eigenvalue λ_i of T^*T is by assumption equal to 0. Thus $Tu_i = 0$ for $i = p+1, \ldots, n$, and thus

$$Tu = \sum_{i=1}^{n} \langle u, u_i \rangle Tu_i = \sum_{i=1}^{p} \langle u, u_i \rangle Tu_i = \sum_{i=1}^{p} \sigma_i \langle u, u_i \rangle v_i$$

for all $u \in U$, which shows that the representation of T given in (17) is valid.

It remains to show that the set $\{v_1, \ldots, v_p\}$ is orthonormal as well. For that, we compute for $1 \leq i, j \leq p$ the inner product

$$\begin{aligned}\langle v_i, v_j \rangle_V = \frac{1}{\sigma_i \sigma_j} \langle Tu_i, Tu_j \rangle_V &= \frac{1}{\sigma_i \sigma_j} \langle u_i, T^*Tu_j \rangle_U \\ &= \frac{1}{\sigma_i \sigma_j} \langle u_i, \sigma_j^2 u_j \rangle_U = \frac{\sigma_j}{\sigma_i} \langle u_i, u_j \rangle_U = \begin{matrix} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j, \end{matrix}\end{aligned}$$

as the set $\{u_1, \ldots, u_p\}$ is orthonormal. This concludes the proof. $\square$

CHAPTER 3

Banach and Hilbert Spaces

Until now, we have treated vector spaces and linear mappings in a purely algebraic way. That is, all[1] the constructions we have performed in the proofs required finitely many well defined steps, involving only linear operations and at times also polynomials. Although we were sometimes dealing with infinite dimensional vector spaces, we were never required to rely on infinite sums of vectors.

This restriction to pure algebra is fine for solving some mathematical problems. The vast majority of problems that occur in practice in physics and engineering, however, cannot be reasonably treated in that way. Instead, it is necessary to bring in methods from analysis as well, in particular the ideas of convergence of sequences and continuity of mappings.[2] In the following, we will introduce the main ideas in that direction.

1. Metric Spaces

We start by defining the notion of a distance or metric on arbitrary sets. With that, we can define notions of convergence of sequences and continuity of functions between sets. We will return to vector spaces a bit later.

Definition 3.1.1 (Metric space). Assume that X is a set. A *metric* on X is a function $d\colon X \times X \to \mathbb{R}$ with the following properties:

- Positivity: $d(x,y) \geq 0$ for all x, $y \in X$, and $d(x,y) = 0$ if and only if $x = y$.
- Symmetry: $d(x,y) = d(y,x)$ for all x, $y \in X$.
- Triangle inequality: $d(x,z) \leq d(x,y) + d(y,z)$ for all x, y, $z \in X$.

A set X together with a metric d is called a *metric space*.

Example 3.1.2. The following are some examples of metric spaces, some of which you might be familiar with:

(1) The space $\mathbb{C}^n$ with the distance $d(x,y) := \|x - y\|$ is a metric space.
(2) On the n-dimensional unit sphere $\mathbb{S}^n := \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$ we can define the metric $d(x,y) := \arccos(\langle x, y\rangle)$. This metric measures the length of the shortest path between the points x and y that does not leave the sphere (this path is actually part of the *great circle* connecting the points x and y).
(3) If U is an inner product space, we can define a metric on U by $d(u,v) := \|u - v\|$ with $\|\cdot\|$ being the norm induced by the inner product.
(4) Every set X can be made a metric space with the *discrete metric* $d\colon X \times X \to \mathbb{R}_{\geq 0}$ defined by $d(x,y) = 0$ if $x = y$ and $d(x,y) = 1$ if $x \neq y$. This metric is pretty useless for all practical applications, though.

[1]Arguably with the exception of the existence of the singular value decomposition, where we actually employed a result from calculus concerning the existence of solutions of optimisation problems. Also, the existence of a basis of an arbitrary vector space requires a more complicated argumentation, but then we didn't actually discuss this proof.

[2]Differentiability of mappings would be the next important idea, but, alas, we won't have time in this course to go that far.

(5) On the space $\mathbb{C}^n$ we can define the metric $d(x,y) = \#\{k : x_k \neq y_k\}$, which counts the number of components in which the vectors x and y differ.

(6) The space $C([0,1];\mathbb{C}^n)$ of continuous functions $f\colon [0,1] \to \mathbb{C}^n$ can be made a metric space with the metric

$$d(f,g) = \max_{x\in[0,1]} \|f(x) - g(x)\|.$$

There are (reasonable) alternatives to this metric, though. For instance, we can define the metric as

$$d(f,g) = \int_0^1 \|f(x) - g(x)\|\,dx.$$

We will see later that these two metrics have pretty different properties.

(7) Denote by X the set of all non-empty, bounded and closed subsets of $\mathbb{R}^n$ (where $n \in \mathbb{N}$ is fixed). We can define a metric d on X setting

$$d(A,B) := \max_{x\in A} \min_{y\in B} \|x-y\| + \max_{y\in B} \min_{x\in A} \|x-y\|.$$

That is, we compute first for each $x \in A$ the minimal distance to the set B, and then take the maximum of all those values. Then we do the same for the set B, and add the resulting two numbers. (In contrast to the other examples here, it is quite a bit harder to verify that this defines actually a metric.)

Definition 3.1.3 (Convergence and limits). Assume that (X,d) is a metric space and that $\{x_n\}_{n\in\mathbb{N}} \subset X$ is a sequence in X. We say that the sequence *converges* to $x \in X$, if there exists for every $\varepsilon > 0$ some index $N \in \mathbb{N}$ such that

$$d(x_n, x) < \varepsilon \qquad \text{whenever } n \geq N.$$

We call x the *limit* of the sequence $\{x_n\}_{n\in\mathbb{N}}$ and denote the convergence by

$$x = \lim_{n\to\infty} x_n.$$

Remark 3.1.4. The statement that for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(x_n,x) < \varepsilon$ whenever $n \geq \mathbb{N}$ is precisely the same as the convergence of the sequence of *real numbers* $d(x_n,x)$ to 0. That is, we have that

$$\lim_{n\to\infty} x_n = x \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} d(x_n,x) = 0.$$

Example 3.1.5. Consider the space $C([0,1])$ of continuous real valued functions on the interval $[0,1]$ with the metric

$$d_1(f,g) = \|f-g\|_1 = \int_0^1 |f(x) - g(x)|\,dx,$$

and define the sequence $\{f_n\}_{n\in\mathbb{N}}$ by

$$f_n(x) := x^n.$$

Then this sequence converges to the function $f(x) = 0$. Indeed, we have that

$$d_1(f_n, f) = \int_0^1 |x^n - 0|\,dx = \int_0^1 x^n\,dx = \frac{1}{n+1},$$

which shows that

$$\lim_{n\to\infty} d_1(f_n, f) = 0.$$

However, if we consider the same space with the metric

$$d_\infty(f,g) = \|f-g\|_\infty := \max_{x\in[0,1]} |f(x) - g(x)|,$$

then the same sequence does not converge to 0 (in fact, it does not converge at all). In this case, we have that

$$d_\infty(f_n, 0) = \max_{x\in[0,1]} |x^n - 0| = 1$$

for all $n \in \mathbb{N}$. This shows that the notion of convergence of a sequence can change drastically if we change the metric on the space.

LEMMA 3.1.6 (Uniqueness of the limit). *Assume that (X, d) is a metric space and that $\{x_n\}_{n\in\mathbb{N}} \subset X$ is a sequence in X. If the limit $\lim_{n\to\infty} x_n$ exists, it is unique.*

PROOF. Let $x = \lim_{n\to\infty} x_n$ and $y = \lim_{n\to\infty} x_n$. Assume to the contrary that $x \neq y$. Then $d(x, y) > 0$. Define now $\varepsilon := d(x, y)/2$. Then there exist N_1, $N_2 \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ whenever $n \geq N_1$ and $d(x_n, y) < \varepsilon$ whenever $n \geq N_2$. Let now $n := \max\{N_1, N_2\}$. Then the properties of the metric d imply that

$$2\varepsilon = d(x, y) \leq d(x, x_n) + d(x_n, y) = d(x_n, x) + d(x_n, y) < \varepsilon + \varepsilon = 2\varepsilon,$$

which is clearly a contradiction. This shows that $x = y$. □

DEFINITION 3.1.7 (Continuity). Assume that (X, d_X) and (Y, d_Y) are metric spaces and that $f\colon X \to Y$ is a mapping. We say that f is *continuous at a point* $x \in X$ if for every sequence $\{x_k\}_{k\in\mathbb{N}} \subset X$ with $\lim_{k\to\infty} x_k = x$ we have that $\lim_{k\to\infty} f(x_k) = f(x)$.

We say that f is *continuous everywhere* or simply *continuous*, if it is continuous at every point $x \in X$.

Alternatively to this definition of continuity, we have the ε-δ-characterisation that everyone loves and likes:

THEOREM 3.1.8. *Let (X, d_X) and (Y, d_Y) be metric spaces. The function $f\colon X \to Y$ is continuous at $x \in X$ if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$d_Y\big(f(x), f(\hat{x})\big) < \varepsilon \qquad \textit{whenever} \qquad d_X(x, \hat{x}) < \delta. \tag{18}$$

PROOF. Assume that (18) holds for all $\varepsilon > 0$ and $\delta > 0$. Now let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence converging to x. We have to show that $\{f(x_k\}_{k\in\mathbb{N}}$ converges to $f(x)$. That is, we have to show that there exists for every $\varepsilon > 0$ some $K \in \mathbb{N}$ such that $d_Y(f(x_k), f(x)) < \varepsilon$ whenever $k \geq K$.

Let therefore $\varepsilon > 0$ and let $\delta > 0$ be such that (18) holds. Since $x = \lim_{k\to\infty} x_k$, there exists $K \in \mathbb{N}$ such that $d_X(x_k, x) < \delta$ for all $k \geq K$. According to (18), this implies that $d_Y(f(x_k), f(x)) < \varepsilon$ for all $k \geq K$. This proves that f is continuous at x.

Now assume that (18) does not hold. That is, assume that there exists some $\varepsilon > 0$ such that for every $\delta > 0$ there exists $\hat{x} \in X$ with $d_X(x, \hat{x}) < \delta$ and $d_Y\big(f(x), f(\hat{x})\big) \geq \varepsilon$. Applying this with $\delta = 1/k$, this means that we can find a sequence $\{x_k\}_{k\in\mathbb{N}}$ such that $d_X(x, x_k) < 1/k$ and $d_Y(f(x), f(x_k)) \geq \varepsilon$. Since $d_X(x, x_k) < 1/k$, we then have that $x = \lim_{k\to\infty} x_k$. On the other hand, since $d(f(x), f(x_k)) \geq \varepsilon$, the sequence $\{f(x_k)\}_{k\in\mathbb{N}}$ does not converge to $f(x)$. This shows that f is not continuous at x. □

There is an alternative characterisation of continuity that can be much easier to work with in certain cases. For that, however, we will need to introduce some additional notation, or, rather, generalise some well known notation from $\mathbb{R}^n$ to arbitrary metric spaces.

Open and closed sets.

Definition 3.1.9 (Open ball). Assume that (X, d) is a metric space, let $x \in X$ and $r > 0$. By

$$B_r(x) := \{y \in X : d(x, y) < r\}$$

we denote the (open) ball of radius r around x.

Definition 3.1.10 (Open set). A subset $U \subset X$ of a metric space (X, d) is *open* if there exists for every $x \in U$ some $r > 0$ such that $B_r(x) \subset U$.

That is, a set U is open if we can find for each point x in X a small ball around x that is still completely contained in U.

Example 3.1.11. Consider the space $C([0, 1])$ with the metric

$$d_\infty(f, g) = \|f - g\|_\infty = \max_{x \in [0,1]} |f(x) - g(x)|.$$

Then the set

$$U := \{f \in C([0, 1]) : f(x) > 0 \text{ for all } x \in [0, 1]\}$$

is open.

In order to see that this is actually true, assume that $f \in U$. Since f is continuous, the function f admits its minimum on the interval $[0, 1]$. That is, the optimisation problem $\min_{x \in [0,1]} f(x)$ has at least one solution $x^* \in [0, 1]$. Define now $r := f(x^*)$. Since $f \in U$, we have that $r > 0$. Also, from the definition of r it follows that $f(x) \geq r$ for all $x \in [0, 1]$.

Assume to that end that $g \in B_r(f)$. Then

$$\begin{aligned} g(x) \geq f(x) - |f(x) - g(x)| \geq f(x) - \max_{y \in [0,1]} |f(y) - g(y)| \\ = f(x) - d_\infty(f, g) > f(x^*) - r = 0 \end{aligned}$$

for all $x \in [0, 1]$, which shows that $g \in U$. Since this holds for every $g \in B_r(f)$, it follows that $B_r(f) \subset U$, which in turn shows that U is open.

We now revisit the previous example with a different metric.

Example 3.1.12. Consider the space $C([0, 1])$ with the metric

$$d_1(f, g) = \|f - g\|_1 = \int_0^1 |f(x) - g(x)|\, dx.$$

Then the set

$$U := \{f \in C([0, 1]) : f(x) > 0 \text{ for all } x \in [0, 1]\}$$

is *not* open.

Take for instance the function $f(x) = 1$, which is clearly contained in U. For every $r > 0$ we can then define the function

$$f_r(x) := \begin{cases} -1 + \dfrac{4x}{r} & \text{if } 0 \leq x < \dfrac{r}{2}, \\ 1 & \text{if } \dfrac{r}{2} \leq x \leq 1. \end{cases}$$

It is easy to see that this function is continuous and not contained in U (since $f_r(0) = -1$). In addition, we have that

$$d_1(f_r, f) = \int_0^1 |f_r(x) - f(x)|\, dx = \int_0^{r/2} \frac{4x}{r}\, dx = \frac{r}{2} < r.$$

Thus $f_r \in B_r(f)$ and $f_r \notin U$. Since r was arbitrary, this shows that U is not open.

Definition 3.1.13 (Closed set). A subset $K \subset X$ of a metric space (X, d) is *closed*, if every convergent sequence in K has its limit in K. That is, if $\{x_k\}_{k \in \mathbb{N}}$ is a convergent sequence with $x_k \in K$ for all $k \in \mathbb{N}$, then $\lim_{k \to \infty} x_k \in K$.

PROPOSITION 3.1.14 (Relation between open and closed sets). *A set U in a metric space (X, d) is open, if and only if its complement $U^C := X \setminus U$ is closed.*

PROOF. Assume that U is open. Let moreover $\{x_k\}_{k\in\mathbb{N}}$ be a convergent sequence with $x_k \notin U$ for all $k \in \mathbb{N}$. We have to show that $x := \lim_{k\to\infty} x_k \notin U$.

Assume to the contrary that $x \in U$. Then there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subset U$. Next, since $x = \lim_{k\to\infty} x_k$, there exists $K \in \mathbb{N}$ such that $d(x_k, x) < \varepsilon$ for all $k \geq K$. In other words, $x_k \in B_\varepsilon(x) \subset U$ for all $k \geq K$. This is clearly a contradiction to the assumption that $x_k \notin U$ for all $k \in \mathbb{N}$. Thus $x \notin U$, which shows that U^C is closed.

Now assume that U is not open. Then there exists $x \in U$ such that for every $\varepsilon > 0$ we have that $B_\varepsilon(x) \not\subset U$. As a consequence, we can find for each $k \in \mathbb{N}$ some $x_k \in U^C$ with $d(x_k, x) < 1/k$. The resulting sequence $\{x_k\}_{k\in\mathbb{N}}$ then by construction converges to x and satisfies that $x_k \in U^C$ for all $k \in \mathbb{N}$. However, $\lim_{k\to\infty} x_k = x \notin U^C$, which shows that U^C is not closed. □

LEMMA 3.1.15. *Assume that (X, d) is a metric space, let $x \in X$ and $r > 0$. Then the open ball $B_r(x) = \{y \in X : d(x, y) < r\}$ is an open set, and the closed ball $\overline{B}_r(x) = \{y \in X : d(x, y) \leq r\}$ is a closed set.*

PROOF. Exercise! □

Now we can formulate an alternative characterisation of continuity:

THEOREM 3.1.16 (Continuity). *Let (X, d_X) and (Y, d_Y) be metric spaces and let $f\colon X \to Y$ be a function. The following are equivalent:*

(1) The function f is continuous.
(2) For every open set $V \subset Y$ the set $f^{-1}(V) = \{x \in X : f(x) \in V\}$ is open.
(3) For every closed set $K \subset Y$ the set $f^{-1}(K) = \{x \in X : f(x) \in K\}$ is closed.

PROOF. To be added when I find the time. □

REMARK 3.1.17. Assume that (X, d) is a metric space and that $f\colon X \to \mathbb{R}$ is continuous. Then the sets

$$\{x \in X : f(x) = 0\} = f^{-1}(0)$$

and

$$\{x \in X : f(x) \geq 0\} = f^{-1}([0, \infty))$$

are closed, and the set

$$\{x \in X : f(x) > 0\} = f^{-1}((0, \infty))$$

is open.

We now investigate open and closed sets a bit further. We show first that arbitrary unions of open sets are open, and that arbitrary intersections of closed sets are closed.

PROPOSITION 3.1.18. *Assume that (X, d) is a metric space and that U_i, $i \in I$, are open subsets of X. Then the set $U := \bigcup_{i\in I} U_i$ is open.*

Similarly, if K_i, $i \in I$, are closed subsets of X, then the set $K := \bigcap_{i\in I} K_i$ is closed.

PROOF. Assume that $x \in U = \bigcup_{i\in I} U_i$. Then there exists an index $i \in I$ such that $x \in U_i$. Since U_i is open, there exists $r > 0$ such that $B_r(x) \subset U_i$. This implies that

$$B_r(x) \subset U_i \subset \bigcup_{i\in I} U_i = U,$$

which shows that U is open.

In order to show that K is closed, we note that

$$K^C = \Big(\bigcap_{i\in I} K_i\Big)^C = \bigcup_{i\in I} K_i^C.$$

Since K_i is closed for each i, the set K_i^C is open. Thus K^C is the union of open sets and therefore open. As a consequence K is closed. □

For intersections of open sets, the situation is different. It is still possible to show that *finite* intersections of open sets are open, but for infinite intersections, this is usually not the case.

PROPOSITION 3.1.19. *Assume that (X,d) is a metric space and that $U_1,\ldots,U_n$ are finitely many open sets. Then the sets $U := \bigcap_{k=1}^n U_k$ is open.*

Similarly, if $K_1,\ldots,K_n$ are finitely many closed sets, then $K := \bigcup_{k=1}^n K_k$ is closed.

PROOF. Let $x \in \bigcap_{k=1}^n U_k$. Then $x \in U_k$ for each k. Since each set U_k is open, there exist $r_k > 0$, $k = 1,\ldots,n$, such that $B_{r_k}(x) \subset U_k$. Now define $r := \min\{r_1,\ldots,r_n\}$. Then $B_r(x) \subset B_{r_k}(x) \subset U_k$ for each k, and thus $B_r(x) \subset \bigcap_{k=1}^n U_k = U$. This shows that U is open.

For showing that K is closed, we write again

$$K^C = \Big(\bigcup_{k=1}^n K_k\Big)^C = \bigcap_{k=1}^n K_k^C.$$

Each set K_k^C is open, and thus K^C is the finite intersection of open sets and therefore itself open. Thus K is closed. □

DEFINITION 3.1.20 (Interior, closure, and boundary). Let (X,d) be a metric space, and let $A \subset X$ be a set.

- The *interior* A° of A is the union of all open sets contained in A.
- The *closure* $\overline{A}$ of A is the intersection of all closed sets containing A.
- The *boundary* ∂A of A is defined as $\partial A = \overline{A} \setminus A^\circ$.

REMARK 3.1.21. Since the union of open sets is open, and the intersection of closed sets is closed, it follows that A° is open, and $\overline{A}$ is closed. Moreover, the set ∂A is the intersection of the closed set $\overline{A}$ and the closed set $(A^\circ)^C$, and therefore closed as well.

In addition, we can characterise A° as the largest set contained in A, and $\overline{A}$ as the smallest set containing A.

PROPOSITION 3.1.22. *Let (X,d) be a metric space, and let $A \subset X$ be a set. Then $x \in A^\circ$ if and only if there exists $r > 0$ such that $B_r(x) \subset A$.*

PROOF. Assume first that $x \in A^\circ$. Then there exists an open set $U \subset A$ such that $x \in U$. Since U is open, there exists $r > 0$ such that $B_r(x) \subset U$. As a consequence, $B_r(x) \subset U \subset A$.

Conversely, assume that there exists $r > 0$ such that $B_r(x) \subset A$. Since the set $B_r(x)$ is open and contained in A, it follows that $B_r(x) \subset A^\circ$. In particular, this shows that $x \in A^\circ$. □

LEMMA 3.1.23. *Let (X,d) be a metric space, and let $A \subset X$ be a set. Then $(A^\circ)^C = \overline{A^C}$ and $(\overline{A})^C = (A^C)^\circ$.*

PROOF. The set A° is open and $A^\circ \subset A$. Thus $(A^\circ)^C$ is closed and $A^C \subset (A^\circ)^C$. This shows that $\overline{A^C} \subset (A^\circ)^C$, as $\overline{A^C}$ is the smallest closed set containing A^C.

Similarly, $\overline{A}$ is closed and $A \subset \overline{A}$. Thus $(\overline{A})^C$ is open and $(\overline{A})^C \subset A^C$. This shows that $(\overline{A})^C \subset (A^C)^\circ$, as $(A^C)^\circ$ is the largest open set contained in A^C.

Now denote $B := A^C$. Applying the first result we have shown in this proof, we obtain that $\overline{B^C} \subset (B^\circ)^C$, which can be rewritten as $\overline{A} \subset ((A^C)^\circ)^C$, or $(A^C)^\circ \subset (\overline{A})^C$.

Similarly, applying the second result we have shown in this proof to $B = A^C$, we obtain that $(\overline{B})^C \subset (B^C)^\circ$, or $(\overline{A^C})^C \subset A^\circ$, which shows that $(A^\circ)^C \subset \overline{A^C}$. □

Lemma 3.1.24. *Let (X,d) be a metric space, and let $A \subset X$ be a set. Then $\partial A = \overline{A} \cap \overline{A^C}$.*

Proof. By definition, $\partial A = \overline{A} \setminus A^\circ$. However, by Lemma 3.1.23 we have that $A^\circ = \overline{A^C}^C$. Thus

$$\partial A = \overline{A} \setminus A^\circ = \overline{A} \cap (A^\circ)^C = \overline{A} \cap \overline{A^C}.$$

□

Lemma 3.1.25. *Let (X,d) be a metric space, and let $A \subset X$ be a set. Then $x \in \overline{A}$, if and only if for all $r > 0$ we have that $B_r(x) \cap A \neq \emptyset$.*

Proof. By Lemma 3.1.23, we have that $x \in \overline{A}$, if and only if $x \notin (A^C)^\circ$. By Proposition 3.1.22, this is the case if and only if for all $r > 0$ we have that $B_r(x) \not\subset A^C$. This, in turn, is the case if and only if for all $r > 0$ we have that $B_r(x) \cap A \neq \emptyset$, which proves the assertion. □

Lemma 3.1.26. *Let (X,d) be a metric space, and let $A \subset X$ be a set. Then $x \in \partial A$, if and only if for all $r > 0$ we have that $B_r(x) \cap A \neq \emptyset$ and $B_r(x) \cap \overline{A} \neq \emptyset$.*

Proof. By Lemma 3.1.24, we have that $x \in \partial A$, if and only if $x \in \overline{A}$ and $x \in \overline{A^C}$. By applying Lemma 3.1.25 to both $\overline{A}$ and $\overline{A^C}$, we thus obtain the claimed result. □

Proposition 3.1.27. *Let (X,d) be a metric space, and let $A \subset X$ be a set. The closure $\overline{A}$ of A consists of all limits of convergent sequences in A.*

Proof. Since $\overline{A}$ is closed, is contains the limits of all convergent sequences in $\overline{A}$, and thus *a forteriori* the limits of all convergent sequences in A.

Conversely, assume that $x \in \overline{A}$. By Lemma 3.1.25, we have for all $k \in \mathbb{N}$ that $B_{1/k}(x) \cap A \neq \emptyset$. In other words, for every $k \in \mathbb{N}$ there exists $x_k \in A$ such that $d(x_k, x) < 1/k$. Thus $x = \lim_{k\to\infty} x_k$, which proves the assertion. □

Definition 3.1.28 (Dense set). Assume that (X,d) is a metric space, that $A \subset B \subset X$ are subsets of X and that B is closed. We say that A is *dense* in B, if $\overline{A} = B$.

Equivalently, we have that A is dense in B, if there exists for every $x \in B$ some sequence $\{x_k\}_{k\in\mathbb{N}}$ with $x = \lim_{k\to\infty} x_k$ and $x_k \in A$ for all k.

Example 3.1.29. The set $\mathbb{Q}$ of rational numbers is dense in $\mathbb{R}$, as we can approximate every real number up to arbitrary (finite) accuracy by a rational number.

The following theorems show some non-trivial (and fairly hard to prove) examples of dense subsets:

Theorem 3.1.30 (Stone–Weierstraß). *The space $\mathcal{P}$ of polynomials is dense in $C([0,1])$ with respect to the metric $d_\infty(f,g) := \max_{x\in[0,1]}|f(x) - g(x)|$.*

Theorem 3.1.31 (Weierstraß). *The space of* trigonometric polynomials*, that is, the set of functions of the form $f(x) = \sum_{k=-N}^{N} c_k e^{\imath kx}$ with $N \in \mathbb{N}$ and $c_k \in \mathbb{C}$, is dense in the space of complex 2π-periodic continuous functions on $\mathbb{R}$ with respect to the metric $d_\infty(f,g) = \max_{x\in\mathbb{R}}|f(x) - g(x)|$.*

REMARK 3.1.32. The previous result does *not* state that the Fourier series of an arbitrary continuous function f converges with respect to d_∞ to f. In fact, there exists examples of continuous functions, where the Fourier series does not converge. However, it is possible to show that running averages (so called "Cesàro means") of the Fourier series converge.

2. Cauchy–sequences and Completeness

The set $\mathbb{Q}$ of the rational numbers is a metric space with the metric $d(x, y) = |x - y|$. However, it can be very difficult to treat sequences and functions on the rational numbers with analytic means, as many of the well-known results from calculus fail to hold there. For instance, the intermediate value theorem that states that a continuous function f on an interval $[a, b]$ takes as function values all the numbers between $f(a)$ and $f(b)$ fails to hold over the rational numbers.

A major problem is that many sequences of rational numbers that appear to be converging, actually don't. Or, to be more precise, they only converge if we allow the limit to lie outside the rational numbers. In fact, this is the main reason for the introduction of the real numbers. One intuition here is that $\mathbb{Q}$ is *incomplete*, but has "holes" filled with the irrational numbers. Take for instance the sequence $\{x_k\}_{k\in\mathbb{N}}$ of rational numbers defined by $x_0 = 1$ and

$$x_{k+1} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{x_k^2 + 2}{2x_k}. \tag{19}$$

This is the sequence one obtains by applying Newton's method to the equation $x^2 - 2 = 0$, and thus the sequence converges to $\sqrt{2}$, which, however, is not a real number. Thus, if we restrict ourselves only to the real numbers, the sequence does not converge at all, even though it should.

In the case of the rational numbers, we got around that problem by introducing the real numbers and working with them instead when we want to do calculus. However, something similar happens, when we are investigating different metric spaces. Take for instance the space $C([0, 1])$ of continuous real functions on the interval $[0, 1]$ with the metric $d_1(f, g) = \int_0^1 |f(x) - g(x)|\, dx$, and consider the sequence of functions

$$f_k(x) = \begin{cases} 0 & \text{if } x < \frac{1}{2}, \\ k\Big(x - \frac{1}{2}\Big) & \text{if } \frac{1}{2} \le x \le \frac{1}{2} + \frac{1}{k}, \\ 1 & \text{if } x > \frac{1}{2} + \frac{1}{k}. \end{cases} \tag{20}$$

This sequence appears to converge to the function

$$f(x) = \begin{cases} 0 & \text{if } x \le \frac{1}{2}, \\ 1 & \text{if } x > \frac{1}{2}. \end{cases}$$

This function, however, is not continuous and thus not contained in the space $C([0, 1])$ we are interested in.

In the cases discussed above, it appears pretty clear to us that the sequence we consider should converge, and it is possible for us to find a limit after turning to a larger space: the space of real numbers in the first case, the space of piecewise continuous functions in the second case. A question now is, if this is possible in general: If we are given a metric space (X, d), can we find a larger space Y in which every sequence in X that should converge actually converges? In order to begin to tackle this question, we first have to clarify what we mean when we say that a sequence "should converge". Even more, we need to find a suitable definition that does not rely on the limit of the sequence (which as of now might not exist).

For that, we recall the definition of convergence of a sequence in a metric space: The sequence $\{x_n\}_{n\in\mathbb{N}}$ converges in the metric space (X,d) to $x \in X$, if there exists for each $\varepsilon > 0$ some $n \in \mathbb{N}$ such that $d(x_n, x) < \varepsilon$ for all $n \geq N$. Now assume that this is the case, and take n, $m \geq N$. Then the triangle inequality implies that

$$d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < 2\varepsilon.$$

Note here that the inequality $d(x_n, x_m) < 2\varepsilon$ makes no mention of the limit x of the sequence. Thus we can use this in a definition of sequences that should converge.

Definition 3.2.1 (Cauchy sequence). A sequence $\{x_n\}_{n\in\mathbb{N}}$ in a metric space (X,d) is a *Cauchy sequence*, if there exists for each $\varepsilon > 0$ some $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < \varepsilon \qquad \text{whenever } n,\ m \geq N.$$

From our discussion before the definition, we immediately get the following result:

Lemma 3.2.2. *Every convergent sequence in a metric space is a Cauchy sequence.*

Conversely, the sequence defined in (19) is a Cauchy sequence in $\mathbb{Q}$, but does not converge in that space. Moreover, within the real numbers, there is no difference between convergent sequences and Cauchy sequences.

Theorem / Axiom 3.2.3. *Every Cauchy sequence in the real numbers $\mathbb{R}$ converges.*

Discussion. What makes this result somewhat awkward to even state is that we have never said what the real numbers actually *are*.

One possible definition of the real numbers is as the *completion* of the rational numbers (we will discuss the idea of completions of metric spaces later, I guess). In that case, the completeness of the real numbers is a direct consequence of their definition.

However, there are other possibilities for defining $\mathbb{R}$ that use a different approach. In that case, the completeness of the real numbers actually requires a proof. $\square$

Remark 3.2.4. In the definition of a Cauchy sequence, it is crucial that we take the distances between *all* elements in the sequence after the given index. If we were to restrict ourselves to only the consecutive differences $d(x_n, x_{n+1})$, the definition would not be helpful: The harmonic series $S_n := \sum_{k=1}^{n} 1/k$ diverges, but $|S_{n+1} - S_n| = 1/n \to 0$.

Definition 3.2.5 (Completeness). A metric space (X,d) is *complete*, if every Cauchy sequence in X converges.

Thus our Theorem/Axiom above can be rephrased as saying that the real numbers are complete.

Lemma 3.2.6. *The space $C([0,1])$ with the distance $d_1(f,g) = \int_0^1 |f(x) - g(x)|\,dx$ is* not *complete.*

Proof. Let $f_k \colon [0,1] \to \mathbb{R}$ be as in (20). We will show explicitly that the sequence $\{f_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence with respect to d_1. For that, let $\varepsilon > 0$ and choose some $N > 1/\varepsilon$. Let moreover n, $m \geq N$. Then

$$f_n(x) = f_m(x) \qquad \text{for all } x \notin \Big[\frac{1}{2}, \frac{1}{2} + \frac{1}{N}\Big].$$

Moreover, for $1/2 \leq x \leq 1/N$ we have that $|f_n(x) - f_m(x)| \leq 1$. Thus

$$d_1(f_n, f_m) = \int_{1/2}^{1/2+1/N} |f_n(x) - f_m(x)| \, dx \leq \int_{1/2}^{1/2+1/N} 1 \, dx = \frac{1}{N} < \varepsilon.$$

This shows that the sequence $\{f_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence.

Now we will show that the sequence $\{f_n\}_{n\in\mathbb{N}}$ does not converge in $C([0,1])$. Assume to the contrary that $\lim_{n\to\infty} f_n = f \in C([0,1])$. Then we have for each interval $[a,b] \subset [0,1]$ that

$$0 = \lim_{n\to\infty} \int_0^1 |f_n(x) - f(x)| \, dx \geq \lim_{n\to\infty} \int_a^b |f_n(x) - f(x)| \, dx.$$

We now consider specifically intervals $[a,b] = [0,1/2]$ and $[a,b] = [1/2+1/N, 1]$ for some fixed $N \in \mathbb{N}$.

- Since $f_n(x) = 0$ for all $x \in [0,1/2]$ and all $n \in \mathbb{N}$, we have that

$$0 = \lim_{n\to\infty} \int_0^{1/2} |f_n(x) - f(x)| \, dx = \lim_{n\to\infty} \int_0^{1/2} |f(x)| \, dx = \int_0^{1/2} |f(x)| \, dx.$$

  Since f is continuous, this is only possible if $f(x) = 0$ for all $x \in [0,1/2]$.
- Since $f_n(x) = 1$ for all $x \in [0, 1/2+1/N]$ and all $n \geq N$, we have that

$$\begin{aligned} 0 = \lim_{n\to\infty} \int_{1/2+1/N}^1 |f_n(x) - f(x)| \, dx &= \lim_{n\to\infty} \int_{1/2+1/N}^1 |f(x) - 1| \, dx \\ &= \int_{1/2+1/N}^1 |f(x) - 1| \, dx. \end{aligned}$$

  Since f is continuous, this is only possible if $f(x) = 1$ for all $x \in [1/2 + 1/N, 1]$.

The above considerations hold for each $N \in \mathbb{N}$, and thus it follows that $f(x) = 0$ for all $0 \leq x \leq 1/2$, and $f(x) = 1$ for all $1/2 < x \leq 1$. Therefore the possible limit of the sequence $\{f_n\}_{n\in\mathbb{N}}$ is not continuous, and thus the sequence $\{f_n\}_{n\in\mathbb{N}}$ does not converge in $C([0,1])$. Thus the space is incomplete. ☐

The situation looks different when we regard the space $C([0,1])$ with a different metric.

Theorem 3.2.7. *The space $C([0,1])$ with the metric*

$$d_\infty(f,g) = \max_{x\in[0,1]} |f(x) - g(x)|$$

is complete.

Proof. Let $\{f_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in $C([0,1])$. We have to show that the sequence converges to some function $f \in C([0,1])$. We start by identifying a potential candidate for the limit function f. For each $x \in [0,1]$ we have that

$$|f_n(x) - f_m(x)| \leq \max_{y\in[0,1]} |f_n(y) - f_m(y)| = d_\infty(f_n, f_m).$$

Since $\{f_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence, this implies that the sequence of real numbers $\{f_n(x)\}_{n\in\mathbb{N}}$ is a Cauchy sequence as well. Because $\mathbb{R}$ is complete, there exists some $f(x) \in \mathbb{R}$ with $\lim_{n\to\infty} f_n(x) = f(x)$.

We will now show that the resulting function $f\colon [0,1] \to \mathbb{R}$ is the limit of the sequence of functions $\{f_n\}_{n\in\mathbb{N}}$. For that, we have to show that f is continuous and that $d_\infty(f_n, f) \to 0$ as $n \to \infty$.

We will first show that f is continuous. Let therefore $\varepsilon > 0$. Since $\{f_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that

$$d_\infty(f_n, f_m) = \max_{x\in[0,1]} |f_n(x) - f_m(x)| < \varepsilon/3$$

for all n, $m \in \mathbb{N}$. Since $f(x) = \lim_{m\to\infty} f_m(x)$, we have that

$$|f_n(x) - f(x)| = \lim_{m\to\infty} |f_N(x) - f_m(x)| \le \varepsilon/3 \qquad \text{for all } x \in [0,1] \text{ and } n \ge N. \tag{21}$$

Now let $y \in [0,1]$ be arbitrary. The function f_N is continuous, and thus there exists $\delta > 0$ such that $|f_N(y) - f_N(z)| < \varepsilon/3$ whenever $|y - z| < \delta$. By applying (21) with $n = N$ we thus obtain for every z with $|y - z|$ that

$$|f(y) - f(z)| \le |f(y) - f_N(y)| + |f_N(y) - f_N(z)| + |f_N(z) - f(z)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

Since ε was arbitrary, this shows that f is continuous at $y \in [0,1]$. Since y was arbitrary, it follows that f is continuous.

Now we use (21) again to conclude that

$$d_\infty(f_n, f) = \max_{x\in[0,1]} |f_n(x) - f(x)| \le \varepsilon/3 \qquad \text{for all } n \ge N.$$

Since ε was arbitrary, this shows that $f = \lim_{n\to\infty} f_n$.

Thus we have shown that every Cauchy sequence in $C([0,1])$ converges, and thus $C([0,1])$ is complete with respect to d_∞. □

Remark 3.2.8. Assume that $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of functions $f_n : [0,1] \to \mathbb{R}$, and assume that $f : [0,1] \to \mathbb{R}$.

- The sequence $\{f_n\}_{n\in\mathbb{N}}$ converges *pointwise* to f, if

$$\lim_{n\to\infty} |f_n(x) - f(x)| = 0 \qquad \text{for all } x \in [0,1].$$

- The sequence $\{f_n\}_{n\in\mathbb{N}}$ converges *uniformly* to f, if

$$\lim_{n\to\infty} \sup_{x\in[0,1]} |f_n(x) - f(x)| = 0.$$

That is, uniform convergence of a sequence of functions is precisely convergence with respect to the metric d_∞ (though it makes sense to use the expression "uniform convergence" also for discontinuous functions). Thus Theorem 3.2.7 in particular implies that the uniform limit of a sequence of continuous functions is again continuous.

However, from the pointwise convergence of a sequence $\{f_n\}_{n\in\mathbb{N}}$ of continuous to some $f : [0,1] \to \mathbb{R}$ we cannot conclude that f is continuous as well. An example is the sequence of continuous functions from (20) that converges pointwise to the discontinuous function $f : [0,1] \to \mathbb{R}$, $f(x) = 0$ for $0 \le x \le 1/2$ and $f(x) = 1$ for $1/2 < x \le 1$.

Remark 3.2.9. In Theorem 3.2.7, we have used the completeness of $\mathbb{R}$ to show that the space $C([0,1])$ of continuous $\mathbb{R}$-valued functions is complete as well with respect to the metric d_∞.

Now assume that (Y, d_Y) is an arbitrary metric space. Then we can consider the space $C([0,1], Y)$ of continuous functions $f : [0,1] \to Y$ with the metric

$$d_\infty(f, g) := \max_{x\in[0,1]} d_Y\big(f(x), g(x)\big).$$

With essentially the same proof, we can show that the space $C([0,1], Y)$ with this metric is complete as well. Also, we can change the interval $[0,1]$ to an arbitrary closed and bounded interval $I \subset \mathbb{R}$ (or a closed and bounded set $\Omega \subset \mathbb{R}^n$).[3]

[3]The reason for requiring the interval (or set) to be closed and bounded is that we want the maximum actually to exist and, in particular, be finite. If we allow non-closed or unbounded

Since it is possible to show that the space $\mathbb{R}^d$ is complete respect to the Euclidean distance, this implies in particular that the space $C(I;\mathbb{R}^d)$ is complete with respect to d_∞ whenever $I \subset \mathbb{R}$ is a closed and bounded interval. This is something we will make use of in Section 4 below.

THEOREM 3.2.10. *Assume that (X,d) is a complete metric space and that $Y \subset X$ is a subset. Then Y is complete (with respect to the restriction of d to Y) if and only if Y is closed (as a subspace of X).*

PROOF. Assume that Y is complete as a metric space. Let moreover $\{x_n\}_{n\in\mathbb{N}}$ be a sequence with $x_n \in Y$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = x$ for some $x \in X$. We have to show that $x \in Y$.

Since the sequence $\{x_n\}_{n\in\mathbb{N}}$ converges in X, it is a Cauchy sequence in X. This, however, implies that it is a Cauchy sequence in Y, which is complete. Thus the sequence $\{x_n\}_{n\in\mathbb{N}}$ converges to some $y \in Y$. Because of the uniqueness of the limit of convergent sequences in metric spaces, it follows that $x = y \in Y$. Thus Y is closed.

Now assume that Y is closed, and let $\{y_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence in Y. Then this is a Cauchy sequence in X as well, and thus the completeness of X implies that it converges to some $y \in X$. Because Y is closed and $y = \lim_{n\to\infty} y_n$, it follows that $y \in Y$. Thus the Cauchy sequence $\{y_n\}_{n\in\mathbb{N}}$ converges in Y, and thus Y is complete. □

3. Banach's Fixed Point Theorem

DEFINITION 3.3.1 (Lipschitz continuity). Let (X,d_X) and (Y,d_Y) be metric spaces and let $f\colon X \to Y$ be a function. We say that f is *Lipschitz continuous*, if there exists $L \geq 0$ such that

$$d_Y\bigl(f(x),f(y)\bigr) \leq L d_X(x,y) \tag{22}$$

for all x, $y \in X$. A constant L with this property is called a *Lipschitz constant* of f.

REMARK 3.3.2. Sometimes one defines *the Lipschitz constant of f* as the smallest constant for which (22) holds. We will refrain from doing so, as the smallest such constant is often difficult to find, whereas finding any such constant is often feasible in practice.

LEMMA 3.3.3 (Continuity of Lipschitz functions). *Every Lipschitz continuous function is continuous.*

PROOF. Assume that (X,d_X) and (Y,d_Y) are metric spaces and let $f\colon X \to Y$ be a Lipschitz continuous function and let $L > 0$ be a (strictly positive) Lipschitz constant for f. Let $\varepsilon > 0$ and define $\delta := \varepsilon/L$. Let moreover x, $y \in X$ be such that $d_X(x,y) < \delta$. Then

$$d_Y\bigl(f(x),f(y)\bigr) \leq L d_X(x,y) < L\varepsilon/L = \varepsilon.$$

Thus f is continuous. □

REMARK 3.3.4. Assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that there exists $L \geq 0$ such that $\|\nabla f(x)\| \leq L$ for all $x \in \mathbb{R}^n$. Then f is Lipschitz continuous with Lipschitz constant L. In order to see that, note that

$$f(y) = f(x) + \int_0^1 \bigl\langle \nabla f\bigl(x + t(y-x)\bigr), y - x\bigr\rangle\, dt$$

intervals, this is no longer the case, and it could happen that the distance between two functions is $+\infty$. (We could remedy this problem, though, by restricting ourselves to *bounded* continuous functions, but we won't go that far.)

and therefore

$$\begin{aligned}\|f(y)-f(x)\| &= \Big\|\int_0^1 \big\langle \nabla f\big(x+t(y-x)\big), y-x\big\rangle\,dt\Big\| \\ &\le \int_0^1 \big|\big\langle \nabla f\big(x+t(y-x)\big), y-x\big\rangle\big|\,dt \\ &\le \int_0^1 \|\nabla f\big(x+t(y-x)\big)\|\|y-x\|\,dt \le L\|y-x\|\end{aligned}$$

for all x, $y \in \mathbb{R}^n$.

DEFINITION 3.3.5 (Contraction). Let (X, d_X) and (Y, d_Y) be metric spaces and let $f\colon X \to Y$ be a function. We say that f is a contraction, if f is Lipschitz continuous with a Lipschitz constant L satisfying $0 \le L < 1$.

In other words, f is a contraction, if there exists $0 \le L < 1$ such that

$$d_Y\big(f(x), f(y)\big) \le L d_X(x, y)$$

for all x, $y \in X$.

REMARK 3.3.6. Note that it is not sufficient for a function f to be a contraction that it satisfies

$$d_Y\big(f(x), f(y)\big) < d_X\big(x, y\big)$$

for all $x \neq y$. Instead, we require the Lipschitz constant to be strictly smaller than 1.

For instance, the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x - \arctan x$ is not a contraction, although

$$f'(x) = 1 - \frac{1}{x^2+1} = \frac{x^2}{x^2+1} < 1$$

for all $x \in \mathbb{R}$.

THEOREM 3.3.7 (Banach's Fixed Point Theorem). *Assume that (X, d) is a complete metric space and that $f\colon X \to X$ is a contraction. Then there exists a unique element $x^* \in X$ (a* fixed point of f*) such that*

$$f(x^*) = x^*. \tag{23}$$

PROOF. Let $0 \le L < 1$ be a Lipschitz constant of f.

We start by showing that the solution of (23), *if it exists*, is unique. Assume therefore that x, $y \in X$ satisfy $f(x) = x$ and $f(y) = y$. Then

$$d(x, y) = d\big(f(x), f(y)\big) \le L d(x, y).$$

Since $L < 1$, this is only possible if $d(x, y) = 0$ and thus $x = y$. This shows uniqueness of a fixed point of f. It remains to show that a fixed point actually exists.

Let now $x_0 \in X$ be arbitrary and consider the sequence given by

$$x_{n+1} = f(x_n).$$

We will show that this sequence converges to a fixed point of f.

We start by showing that this is actually a Cauchy sequence. For that, we note that for all $n \in \mathbb{N}$,

$$d(x_{n+1}, x_n) = d\big(f(x_n), f(x_{n-1})\big) \le L d(x_n, x_{n-1}).$$

By induction over n we thus obtain that

$$d(x_{n+1}, x_n) \le L^n d(x_1, x_0).$$

Thus we have for all $n \in \mathbb{N}$ and $k \in \mathbb{N}$ that

$$
\begin{aligned}
(24)\quad d(x_{n+k}, x_n) &\le \sum_{j=1}^{k} d(x_{n+j}, x_{n+j-1}) \le \sum_{j=1}^{k} L^{n+j-1} d(x_1, x_0) \\
&= L^n d(x_1, x_0) \sum_{j=1}^{k} L^{j-1} \le L^n d(x_1, x_0) \sum_{j=0}^{\infty} L^j = \frac{L^n}{1-L} d(x_1, x_0).
\end{aligned}
$$

Now let $\varepsilon > 0$ and let $N \in \mathbb{N}$ be such that

$$\frac{L^N}{1-L} d(x_1, x_0) < \varepsilon.$$

Let moreover n, $m \ge N$. Without loss of generality, assume that $m \ge n$. Then we can apply (24) with $k = m - n$ and obtain that

$$d(x_m, x_n) \le \frac{L^n}{1-L} d(x_1, x_0) \le \frac{L^N}{1-L} d(x_1, x_0) < \varepsilon.$$

Thus the sequence $\{x_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence. Since (X, d) is complete, there exists $x^* \in X$ such that $x^* = \lim_{n\to\infty} x_n$.

It remains to show that x^* is a fixed point of f. Since f is a contraction, it is in particular Lipschitz continuous and thus continuous. Since $x^* = \lim_{n\to\infty} x_n$ be definition, we thus have that

$$f(x^*) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x^*,$$

which concludes the proof. □

REMARK 3.3.8. It is important to note here that the proof is constructive in that it provides us with an explicit algorithm, namely *fixed point iteration*, to approximate the fixed point x^* of f. In fact, we have shown that the sequence

$$x_{n+1} = f(x_n)$$

converges for each starting point $x_0 \in X$ to x^*. In fact, we have therefore formulated a numerical algorithm for the solution of the fixed point equation under the assumption that f is a contraction.

Moreover, if we refine some of the analysis of the proof, we can obtain some estimates for how close we are to x^* after a finite number of steps:

By letting k tend to infinity in (24), we obtain the *a-priori estimate*

$$d(x^*, x_n) \le \frac{L^n}{1-L} d(x_1, x_0).$$

This allows us to estimate already after the first step how close we will be to x^* after n steps.

In addition, we can show the *a-posteriori* estimate

$$d(x^*, x_n) \le \frac{L}{1-L} d(x_n, x_{n-1}).$$

This gives us an upper bound for how close we are to the solution by only looking at the size of the last step we made.

COROLLARY 3.3.9. *Assume that (X, d) is a complete metric space, that $f\colon X \to X$ is a function, and that $K \subset X$ is a closed subset. Assume moreover that*

$$f(x) \in K \qquad \textit{whenever } x \in K$$

and that there exists $0 \le L < 1$ such that

$$d\big(f(x), f(y)\big) \le L d(x, y) \qquad \textit{for all } x,\, y \in K.$$

Then there exists a unique $x^* \in K$ such that $f(x^*) = x^*$.

PROOF. Since K is a closed subset of a complete metric space, it is a complete metric space itself. Moreover, by assumption the restriction $f|_K : K \to K$ is a contraction on K. Thus the result follows from Banach's fixed point Theorem. $\square$

Application to Integral Equations. Assume that $k\colon [0,1] \times [0,1] \to \mathbb{R}$ is a continuous function satisfying

$$c := \max_{(x,y)\in[0,1]^2} |k(x,y)| < 1,$$

and let $h \in C([0,1])$ be a given function. We want to solve the integral equation

$$f(x) + \int_0^1 k(x,y)\, f(y)\, dy = h(x). \tag{25}$$

That is, we want to find a function $f \in C([0,1])$ such that (25) holds for each $x \in [0,1]$. Using Banach's fixed point theorem, we can show that this equation has a unique solution:

Since k is continuous, it follows that the function $x \mapsto \int_0^1 k(x,y)f(y)\,dy$ is continuous. Since h is assumed to be continuous as well, we can therefore the mapping $T\colon C([0,1]) \to C([0,1])$,

$$Tf(x) := h(x) - \int_0^1 k(x,y)\, f(y)\, dy.$$

Then $f \in C([0,1])$ solves (25) if and only if f is a fixed point of T.

Now we show that T is a contraction on the space $C([0,1])$ with respect to the metric d_∞. For that

$$\begin{aligned}
d_\infty(Tf, Tg) &= \max_{x\in[0,1]} \bigl|Tf(x) - Tg(x)\bigr| \\
&= \max_{x\in[0,1]} \Bigl|\int_0^1 k(x,y)\bigl(f(y) - g(y)\bigr)\, dy\Bigr| \\
&\le \max_{x\in[0,1]} \int_0^1 \bigl|k(x,y)\bigr|\bigl|f(y) - g(y)\bigr|\, dy \\
&\le \max_{x\in[0,1]} \int_0^1 c\bigl|f(y) - g(y)\bigr|\, dy \\
&= c\int_0^1 \bigl|f(y) - g(y)\bigr|\, dy \\
&\le c \max_{y\in[0,1]} \bigl|f(y) - g(y)\bigr| \\
&= c d_\infty(f,g).
\end{aligned}$$

Thus T is a contraction on the complete metric space $\bigl(C([0,1]), d_\infty\bigr)$, and therefore T has a unique fixed point. Put differently, the equation (25) has, under the given assumptions on k and h, a unique continuous solution f.

4. Application of the Fixed Point Theorem: Existence for ODEs

We now consider the application of Banach's fixed point Theorem to the question of existence and uniqueness of solutions of ODEs. For that, we consider an initial value problem of the form

$$\begin{aligned} y' &= f(t,y), \\ y(t_0) &= y_0, \end{aligned} \tag{26}$$

where $f\colon \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is some continuous function, $t_0 \in \mathbb{R}$, and $y_0 \in \mathbb{R}^d$ is a given initial value.

THEOREM 3.4.1 (Picard–Lindelöf). *Let $a > 0$ and $b > 0$, and define the cylinder*

$$Z := \{(t, y) \in \mathbb{R} \times \mathbb{R}^d : |t - t_0| \le a \text{ and } \|y - y_0\| \le b\}.$$

Assume that the function $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is continuous on Z and that it satisfies a Lipschitz condition *with respect to its second component in the sense that there exists $L \ge 0$ such that*

$$\|f(t, y) - f(t, z)\| \le L\|y - z\|$$

for all (t, y), $(t, z) \in Z$. Assume moreover that

$$M := \max_{(t,y)\in Z} \|f(t, y)\| \le \frac{b}{a}.$$

Then the initial value problem has a unique solution on the interval $I = [t_0 - a, t_0 + a]$.

PROOF. Since the proof is somewhat longish, we split it up into separate steps.

Step 1: Rewrite the ODE as an integral equation.

Assume that y solves the initial value problem (26). Then we have that

$$y(t) = y_0 + \int_{t_0}^{t} y'(s)\, ds = y_0 + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds$$

for all t. Conversely, if y is a continuous function that satisfies the equation

$$y(t) = y_0 + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds, \tag{27}$$

then y is necessarily differentiable (because the right hand side is) and it satisfies (26).

Define therefore the mapping $T : C(I; \mathbb{R}^d) \to C(I; \mathbb{R}^d)$,

$$Ty(t) := y_0 + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds.$$

Then y solves the initial value problem 26 if and only if it satisfies the fixed point equation $Ty = y$.

Step 2: Restrict T to a suitable subset of $C(I; \mathbb{R}^d)$.

Our goal is to apply Banach's fixed point theorem to the equation $Ty = y$. For that, we need to find a suitable subset $K \subset C(I; \mathbb{R}^d)$ such that T maps K into itself and T is a contraction on K. Define therefore

$$K := \{y \in C(I; \mathbb{R}^d) : \|y(t) - y_0\| \le b \text{ for all } t \in I\}$$

and assume that $y \in K$. Then

$$\|f\big(s, y(s)\big)\| \le M$$

for all $s \in I$ and thus

$$\|Ty(t) - y_0\| = \Big\|\int_{t_0}^{t} f\big(s, y(s)\big)\, ds\Big\| \le \Big|\int_{t_0}^{t} \|f\big(s, y(s)\big)\|\, ds\Big| \le |t - t_0|M \le a\frac{b}{a} = b$$

for all $t \in I$. Since Ty is continuous, it follows that $Ty \in K$ as well.

Step 3: Show that we do not lose any possible solutions.

Assume now that $y \in C(I; \mathbb{R}^d)$ is any solution of the equation (27) (or, equivalently, of the ODE (26)). We want to show that, necessarily, $y \in K$. Assume therefore to the contrary that $y \notin K$. Then there exists $t \in I$ such that $\|y(t) - y_0\| > b$. Since $y(t) = y_0$ by assumption, it follows that $t \ne t_0$.

Assume now that $t > t_0$. Since the mapping $t \mapsto \|y(t) - y_0\|$ is continuous, there exists a minimal value $t_0 < t_1 < t_0 + a$ such that $\|y(t_1) - y_0\| = b$. In particular, $\|y(t_1) - y_0\| < b$ for all $t_0 < s < t_1$. As a consequence

$$b = \|y(t_1) - y_0\| = \Big\|\int_{t_0}^{t_1} f\big(s, y(s)\big)\, ds\Big\| \le M|t_1 - t_0| < Ma = b,$$

which is an obvious contradiction.

We obtain a contradiction in a similar way if we assume that $t < t_0$. This shows that such a t cannot exist, and thus $y \in K$.

Step 4: Define a suitable metric on $C(I;\mathbb{R}^d)$.

We now need to define a metric on $C(I;\mathbb{R}^d)$ with respect to which the mapping T is a contraction on K. For that we define

$$d(y, z) := \max_{t \in I} e^{-2L|t-t_0|}\|y(t) - z(t)\|. \tag{28}$$

Assume now that y, $z \in K$. Then

$$\begin{aligned}
d(Ty, Tz) &= \max_{t\in I} e^{-2L|t-t_0|}\|Ty(t) - Tz(t)\| \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big\|y_0 + \int_{t_0}^t f\big(s, y(s)\big)\, ds - y_0 - \int_{t_0}^t f\big(s, z(s)\big)\, ds\Big\| \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big\|\int_{t_0}^t f\big(s, y(s)\big) - f\big(s, z(s)\big)\, ds\Big\| \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big|\int_{t_0}^t \big\|f\big(s, y(s)\big) - f\big(s, z(s)\big)\big\|\, ds\Big| \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big|\int_{t_0}^t L\|y(s) - z(s)\|\, ds\Big| \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big|\int_{t_0}^t Le^{2L|s-t_0|}e^{-2L|s-t_0|}\|y(s) - z(s)\|\, ds\Big| \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big|\int_{t_0}^t Le^{2L|s-t_0|}\, ds\Big| \Big(\max_{s\in I} e^{-2L|s-t_0|}\|y(s) - z(s)\|\Big) \\
&= \Big(\max_{t\in I} e^{-2L|t-t_0|}\Big|\int_{t_0}^t Le^{2L|s-t_0|}\, ds\Big|\Big) d(y, z) \\
&= \Big(\max_{t\in I} e^{-2L|t-t_0|}\frac{e^{2L|t-t_0|}}{2}\Big) d(y, z) \\
&= \frac{1}{2} d(y, z).
\end{aligned}$$

This shows that T is a contraction with respect to the metric defined in (28).

Step 5: Show that $C(I;\mathbb{R}^d)$ is complete and K closed.

Denote by d_∞ the standard metric on $C(I;\mathbb{R}^d)$, that is,

$$d_\infty(y, z) := \max_{t\in I}\|y(t) - z(t)\|.$$

Since

$$1 \ge e^{-2L|t-t_0|} \ge e^{-2La}$$

for all $t \in I$, we have that

$$e^{2La} d_\infty(y, z) \le d(y, z) \le d_\infty(y, z)$$

for all y, $z \in C(I;\mathbb{R}^d)$. This, however, implies that a sequence converges with respect to d_∞ if and only if it converges with respect to d. Moreover, a sequence is a Cauchy sequence with respect to d_∞, if and only if it is a Cauchy sequence with respect to d. Since $C(I;\mathbb{R}^d)$ with the metric d_∞ is complete, this implies that it is also complete with the metric d.

Also, a set is closed with respect to d if and only if it is closed with respect to d_∞. Since K is precisely the closed ball of radius b around the constant function y_0 with respect to d_∞, it is closed for that metric. Thus K is closed for the metric d as well.

Step 6: Collect everything and call it a day.

We have shown that K is a closed subset of the complete metric space $C(I;\mathbb{R}^d)$ (Step 5), that $Ty \in K$ whenever $y \in K$ (Step 2), and that the restriction of T to K is a contraction (Step 4). Thus Banach's fixed point theorem implies that the mapping T has a unique fixed point y^* in K. Since solving the ODE (26) is equivalent to finding a fixed point of T in $C(I;\mathbb{R}^d)$ (Step 1), we can conclude that the solution (26) has at least one solution (namely y^*). Moreover, if z is another solution, it cannot be contained in K. This, however, is not possible by Step 3. $\square$

Remark 3.4.2. It is possible to show that the ODE (26) has a solution if the function f is continuous; the Lipschitz condition we have assumed in the proof is actually not necessary for that (this is the statement of the *Peano existence theorem* for ODEs). However, without the Lipschitz condition, we can no longer guarantee that the solution is unique. A classical example for this is the ODE

$$y' = y^{1/3}, \qquad\qquad y(0) = 0.$$

An obvious solution of the initial value problem is the constant function $y(t) = 0$. However, the functions

$$y_\pm(t) = \pm(2t/3)^{3/2}$$

for $t \ge 0$ are also solutions of the same ODE. Even worse, for any $T > t_0$ we can define the functions

$$y_{\pm,T}(t) := \begin{cases} 0 & \text{if } t < T, \\ \pm(2(t-T)/3)^{3/2} & \text{if } t \ge T. \end{cases}$$

All of these functions solve the same initial value problem.

This does not contradict the Picard–Lindelöf Theorem, though, as the function $y \mapsto y^{1/3}$ is not Lipschitz continuous in a neighbourhood of 0.

Remark 3.4.3 (Picard iteration). Since we have proved Theorem 3.4.1 by showing that the Banach fixed point Theorem is applicable, we obtain in addition that the fixed point iteration (the *Picard iteration*)

$$y_{k+1}(t) = Ty_k(t) = y_0 + \int_{t_0}^t f\bigl(s, y_k(s)\bigr)\,ds \tag{29}$$

(where we use the initialisation $y_0(t) = y_0$, although this is not necessary for the convergence) actually converges to a solution of the ODE (26).

If we apply this to the ODE

$$y' = y, \qquad\qquad y(0) = 1,$$

we obtain the iterates

$$\begin{aligned}
y_0(t) &= 1,\\
y_1(t) &= 1 + \int_0^t 1\,ds = 1 + t,\\
y_2(t) &= 1 + \int_0^t 1 + s\,ds = 1 + t + \frac{t^2}{2},\\
y_3(t) &= 1 + \int_0^t 1 + s + \frac{s^2}{2}\,ds = 1 + t + \frac{t^2}{2} + \frac{t^3}{6},
\end{aligned}$$

and, in general,

$$y_n(t) = \sum_{k=0}^{n} \frac{t^k}{k!},$$

which happens to be the truncated Taylor series expansion of the exponential function.

More generally, one can in principle use the iteration (29) in order to define a numerical solution method for ODEs. A difficulty here is that the integrals that need to be evaluated in each step of the iteration can in general not be solved analytically. This problem can, however, be solved in practice by falling back to numerical quadratures. Still, methods based on Picard iteration appear to be limited to some niche applications, and in general one uses either some Runge–Kutte method or some multi-step method for the numerical solution of initial value problems.

5. Normed Spaces

We now come back to the study of vector spaces and linear transformations, but now with the additional tools we have developed during our discussion of metrics and metric spaces. Since we have discussed metrics in the much more general setting of arbitrary sets, however, we might want to add some further conditions in order to make a metric a useful notion on a vector space.

A first reasonable assumption for a metric on a vector space U is that the distance between two vectors u and v does not change, if we add the same vector w to both of them. That is, it makes sense to require that

$$d(u, v) = d(u + w, v + w)$$

for all u, v, $w \in U$. This in particular implies that

$$d(u, v) = d(u - v, 0)$$

for all u, $v \in U$. That is, the distance between u and v only depends on their difference $u - v$.

Next, we might find it useful to require that the distance behaves reasonably under scaling. If we multiply two vectors u and v with the same scalar λ, then their distance should scale with $|\lambda|$. That is, we probably want to require that

$$d(\lambda u, \lambda v) = |\lambda| d(u, v)$$

for all u, $v \in U$ and $\lambda \in \mathbb{K}$.

We now introduce the notion of a norm, which allows us to define a metric on a vector space with these properties:

Definition 3.5.1 (Norm). A *norm* on a vector space U over $\mathbb{K}$ is a function $\|\cdot\| \colon U \to \mathbb{R}$ with the following properties:

(1) *Positivity:* $\|u\| \geq 0$ for all $u \in U$, and $\|u\| = 0$ if and only if $u = 0$.
(2) *Homogeneity:* $\|\lambda u\| = |\lambda| \|u\|$ for all $u \in U$ and $\lambda \in \mathbb{K}$.

(3) *Triangle inequality:* $\|u + v\| \le \|u\| + \|v\|$ for all u, $v \in U$.

A vector space together with a norm is called a *normed space*.

The next result shows that these properties are sufficient for allowing us to define a metric on U. Thus we are able to use all results on metric space also for normed spaces.

LEMMA 3.5.2. *If* $\|\cdot\|$ *is a norm on the vector space* U*, then* $d(u, v) := \|u - v\|$ *is a metric on* U*.*

PROOF. We have that $d(u, v) = \|u - v\| \ge 0$ for all u, $v \in U$, and $d(u, v) = 0$ if and only if $\|u - v\| = 0$, which is the case if and only if $u = v$. Next we have that

$$d(v, u) = \|v - u\| = \|(-1)(u - v)\| = |-1|\|u - v\| = d(u, v).$$

Finally,

$$d(u, w) = \|u - w\| = \|u - v + v - w\| \le \|u - v\| + \|v - w\| = d(u, v) + d(v, w)$$

for all u, v, $w \in U$. Thus d satisfies all requirements for a metric. □

REMARK 3.5.3. If $(U, \|\cdot\|)$ is a normed space, then a sequence $\{u_n\}_{n\in\mathbb{N}}$ converges to u, if and only if

$$\|u_n - u\| = d(u_n, u) \to_{n\to\infty} 0.$$

Thus we can (and will) formulate convergence of sequences, and thus also continuity of functions, completely in terms of the norm without mentioning the induced metric.

We will now state a consequence of the triangle inequality that turns out to be useful in a large number of estimates.

LEMMA 3.5.4 (Reverse triangle inequality). *Assume that* $(U, \|\cdot\|)$ *is a normed space and that* u, $v \in U$*. Then*

$$\big|\|u\| - \|v\|\big| \le \|u - v\|.$$

PROOF. We have that

$$\|u\| = \|u - v + v\| \le \|u - v\| + \|v\|$$

and therefore

$$\|u\| - \|v\| \le \|u - v\|.$$

Similarly,

$$\|v\| = \|v - u + u\| \le \|v - u\| + \|u\|$$

and thus

$$\|v\| - \|u\| \le \|u - v\|.$$

Together, these inequalities imply that

$$\big|\|u\| - \|v\|\big| \le \|u - v\|,$$

which proves the assertion. □

A particular class of normed spaces is that of inner product spaces. There we have already defined a normed induced by the inner product. We now show that that "norm" indeed satisfies the requirements put forward in Definition 3.5.1.

LEMMA 3.5.5. *Assume that* U *is an inner product space with inner product* $\langle\cdot,\cdot\rangle$*. Then*

$$\|u\| := \langle u, u\rangle^{1/2}$$

defines a norm on U*.*

PROOF. Since the inner product is positive definite, it follows that $\|u\| \geq 0$ for all $u \in U$ with equality if and only if $u = 0$. Next, if $u \in U$ and $\lambda \in \mathbb{K}$, then

$$\|\lambda u\| = \langle \lambda u, \lambda u\rangle^{1/2} = \big(\lambda\overline{\lambda}\langle u, u\rangle\big)^{1/2} = |\lambda|\langle u, u\rangle^{1/2} = |\lambda|\|u\|,$$

which proves the positive homogeneity. Finally, the triangle inequality has been shown in Theorem 2.1.12. $\square$

While inner products give us one possibility for defining norms, there are alternatives that can be useful in certain contexts. We present some examples here, though there will be a more detailed discussion about related norms and spaces in the next section.

EXAMPLE 3.5.6. On the space $\mathbb{K}^n$ we define

$$\|u\|_1 := |u_1| + \ldots + |u_n|$$

and

$$\|u\|_\infty := \max\big\{|u_1|, \ldots, |u_n|\big\}$$

for $u \in \mathbb{K}^n$. Both of these are norms on $\mathbb{K}^n$, as positivity and homogeneity are clearly satisfied and the triangle inequality follows from the triangle inequality on the real/complex numbers.

Actually, we have used similar definitions previously when discussing metrics on the space of continuous functions:

EXAMPLE 3.5.7. One the space $C([0, 1])$ we define

$$\|f\|_1 := \int_0^1 |f(x)|\, dx$$

and

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

for $f \in C([0, 1])$. Both of these are norms on $C([0, 1])$, and the corresponding metrics are d_1 and d_∞, respectively.

It is obvious that the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on the space $\mathbb{K}^n$ (with $n \geq 2$) are different from the Euclidean norm that is induced by the standard inner product. However, it might be still possible that there is another inner product on $\mathbb{K}^n$ that induces these norms. We will show in the following that this is not the case by deducing a property that all norms induced by an inner product need to satisfy.

LEMMA 3.5.8 (Parallelogram law). *Assume that U is an inner product space and that $\|u\| = \langle u, u\rangle^{1/2}$ is the induced norm on U. Then*

$$\|u - v\|^2 + \|u + v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

for all u, $v \in U$.

PROOF. We write

$$\begin{aligned}\|u - v\|^2 + \|u + v\|^2 &= \langle u - v, u - v\rangle + \langle u + v, u + v\rangle \\ &= \langle u, u\rangle - \langle v, u\rangle + \langle v, v\rangle - \langle u, v\rangle + \langle u, u\rangle + \langle v, u\rangle + \langle v, v\rangle + \langle u, v\rangle \\ &= 2\langle u, u\rangle + 2\langle v, v\rangle = 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

$\square$

EXAMPLE 3.5.9. The norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{K}^n$ with $n \ge 2$ are *not* induced by an inner product. We show this by demonstrating that the parallelogram law fails to hold in both cases. For this, we take $u = e_1$ and $v = e_2$, the first and second standard basis vector in $\mathbb{K}^n$. Then

$$\|e_1 - e_2\|_1^2 + \|e_1 + e_2\|_1^2 = 4 + 4 = 8,$$

whereas

$$2\|e_1\|_1^2 + 2\|e_2\|_1^2 = 2 + 2 = 4.$$

Similarly,

$$\|e_1 - e_2\|_\infty^2 + \|e_1 + e_2\|_\infty^2 = 1 + 1 = 2,$$

whereas

$$2\|e_1\|_\infty^2 + 2\|e_2\|_\infty^2 = 2 + 2 = 4.$$

Interestingly, the parallelogram law provides a complete characterisation of norms induced by inner products:

THEOREM 3.5.10 (Jordan–von Neumann). *Assume that $(U, \|\cdot\|)$ is a normed space and that the parallelogram law*

$$\|u - v\|^2 + \|u + v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

holds for all u, $v \in U$. Then there exists an inner product $\langle\cdot,\cdot\rangle$ on U such that $\|u\| = \langle u, u\rangle^{1/2}$ for all $u \in U$.

- *If U is a real vector space, then $\langle\cdot,\cdot\rangle$ is given as*

$$\langle u, v\rangle = \frac{1}{4}\big(\|u + v\|^2 - \|u - v\|^2\big).$$

- *If U is a complex vector space, then $\langle\cdot,\cdot\rangle$ is given as*

$$\langle u, v\rangle = \frac{1}{4}\sum_{k=1}^{4} i^k \|u + i^k v\|^2.$$

PROOF. I will add the proof for the real case when I find time. (It is not difficult to check that the requirements for an inner product are satisfied in the real case, but it is somewhat tedious. In the complex case, the tediousness rises significantly.) $\square$

We will now show that norms work well together with the linear structure of a vector space in that the basic operations of addition and scalar multiplication will always be continuous on a normed space. Afterwards we will show that this is not always the case for (reasonable) metrics on vector spaces.

THEOREM 3.5.11. *Assume that U is a normed space. Then the following hold:*

(1) Continuity of the norm: If $u_n \to_{n\to\infty} u$, then $\|u_n\| \to_{n\to\infty} \|u\|$.
(2) Continuity of addition: If $u_n \to_{n\to\infty} u$ and $v_n \to_{n\to\infty} v$, then $(u_n + v_n) \to_{n\to\infty} u + v$.
(3) Continuity of scalar multiplication: If $u_n \to_{n\to\infty} u$ and $\lambda_n \to_{n\to\infty} \lambda$, then $(\lambda_n u_n) \to_{n\to\infty} \lambda u$.

PROOF. Assume that $u = \lim_{n\to\infty} u_n$, that is, $\|u - u_n\| \to 0$ as $n \to \infty$. Then the reverse triangle inequality implies that

$$\big|\|u\| - \|u_n\|\big| \le \|u - u_n\| \to 0 \qquad \text{as } n \to \infty,$$

which proves that $\|u_n\| \to \|u\|$.

Now assume that $u = \lim_{n\to\infty} u_n$ and $v = \lim_{n\to\infty} v_n$. Then

$$\|u + v - (u_n + v_n)\| = \|u - u_n + v - v_n\| \le \|u - u_n\| + \|v - v_n\| \to 0 \qquad \text{as } n \to \infty,$$

which shows that $u + v = \lim_{n\to\infty}(u_n + v_n)$.

Finally, assume that $u = \lim_{n\to\infty} u_n$ and $\lambda = \lim_{n\to\infty} \lambda_n$. Then

$$\begin{aligned}\|\lambda u - \lambda_n u_n\| = \|\lambda u - \lambda u_n + \lambda u_n - \lambda_n u_n\| &\le \|\lambda(u - u_n)\| + \|(\lambda - \lambda_n)u_n\| \\ &= |\lambda|\|u - u_n\| + |\lambda - \lambda_n|\|u_n\| \to 0 \qquad \text{as } n \to \infty,\end{aligned}$$

as the sequence $\{\|u_n\|\}_{n\in\mathbb{N}}$ converges to $\|u\|$ and therefore is bounded. $\square$

EXAMPLE 3.5.12. On $\mathbb{K}^n$ we define the 0-"norm" $\|\cdot\|_0$ by

$$\|u\|_0 := \#\{k : u_k \neq 0\}.$$

That is, $\|u\|_0$ is the number of non-zero components of the vector u. We stress that this is not a norm on $\mathbb{K}^n$, because the homogeneity is not satisfied: We have for instance that $\|e_1\| = 1$, but also $\|2e_1\| = 1$.

Still, it is straightforward to show that

$$d_0(u, v) := \|u - v\|_0$$

defines a metric on $\mathbb{K}^n$: Positivity and symmetry are clear from the definition ($d(u, v) = 0$ if and only if $\|u - v\|_0 = 0$, that is, if and only if the vector $u - v$ has no non-zero components; and we obviously have $\|u - v\|_0 = \|v - u\|_0$). Also, the number of non-zero components of a vector $u - w$ cannot be larger the sum of non-zero components of $u - v$ and $v - w$, which proves the triangle inequality.

However, scalar multiplication is not a continuous operation on $(\mathbb{K}^n, \|\cdot\|_0)$. Take for instance the constant sequence $u_n = e_1$ for all n and $\lambda_n = \frac{1}{n}$. Then $u_n \to e_1$ and $\lambda_n \to 0$ as $n \to \infty$, but

$$\|\lambda_n u_n - 0\|_0 = \left\|\frac{1}{n} e_1\right\| = 1 \text{ for all } n,$$

that is, the sequence $\{\lambda_n u_n\}_{n\in\mathbb{N}}$ does not converge to 0.

Finally, we introduce the notion of a bounded set in a normed space:

DEFINITION 3.5.13 (Bounded sets). Assume that $(U, \|\cdot\|)$ is a normed space and that $A \subset U$ is some subset. We say that A is bounded, if it is contained in some ball centered at 0, that is, there exists $r > 0$ such that $A \subset B_r(0)$.

In particular, we have that Cauchy sequences (and consequently also convergent sequences) in a normed space are necessarily bounded.

LEMMA 3.5.14. *Assume that $(U, \|\cdot\|)$ is a normed space and that $\{u_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in U. Then the set $\{u_n\}_{n\in\mathbb{N}}$ is bounded.*

PROOF. Since the sequence $\{u_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that $\|u_n - u_m\| < 1$ whenever $n, m \ge N$. Define now

$$r := \max\{\|u_1\|, \|u_2\|, \dots, \|u_N\|\} + 1.$$

Then by construction $u_n \in B_r(0)$ for all $1 \le n \le N$. Also, for $n \ge N$ we have that

$$\|u_n\| \le \|u_n - u_N\| + \|u_N\| < \|u_N\| + 1 \le r.$$

Thus we have that $u_n \in B_r(0)$ for all $n \in \mathbb{N}$, which shows that the sequence is bounded. $\square$

6. p-norms

A particularly interesting and useful class of norms on $\mathbb{K}^n$ are the so called p-norms, which are generalisations of the Euclidean norm as well as the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$.

DEFINITION 3.6.1 (p-norm on $\mathbb{K}^n$). Let $1 \le p < \infty$. We define the mapping $\|\cdot\|_p \colon \mathbb{K}^n \to \mathbb{R}$,

$$\|x\|_p := \bigl(|x_1|^p + \ldots + |x_n|^p\bigr)^{1/p}$$

for $x \in \mathbb{K}^n$.

Similarly, we define $\|\cdot\|_\infty \colon \mathbb{K}^n \to \mathbb{R}$ by

$$\|x\|_\infty := \max\bigl\{|x_1|, \ldots, |x_n|\bigr\}$$

for $x \in \mathbb{K}^n$.

We will show in the following that $\|\cdot\|_p$ is indeed a norm on $\mathbb{K}^n$. The symmetry and positivity are fairly obvious, but the triangle inequality is not, unless we are in the simpler cases $p = 1$ or $p = \infty$. Also, for $p = 2$, we simply obtain the Euclidean norm, which we have previously considered in the context of inner products.

For proving the triangle inequality in all other cases, we will need some additional results.

DEFINITION 3.6.2 (Conjugate). Let $1 < p < \infty$. We define the *conjugate* (or *Hölder conjugate*) of p as the solution $1 < p_* < \infty$ of the equation

$$\frac{1}{p} + \frac{1}{p_*} = 1.$$

Moreover, for $p = 1$ we define the conjugate as $p_* := \infty$, and for $p = \infty$ as $p_* := 1$.

Alternatively, we can write the conjugate of $1 < p < \infty$ explicitly as

$$q = \frac{p}{p-1}.$$

LEMMA 3.6.3 (Young's inequality). *Let $1 < p < \infty$, and let $1 < p_* < \infty$ be the conjugate of p. Then for all a, $b \ge 0$ the inequality*

$$ab \le \frac{1}{p}a^p + \frac{1}{p_*}b^{p_*} \tag{30}$$

holds.

PROOF. For $b = 0$, this inequality holds obviously. Now assume that $b > 0$ and consider the function $f \colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$,

$$f(a) := ab - \frac{1}{p}a^p.$$

Then (30) is equivalent to the statement that $f(a) \le \frac{1}{p_*}b^{p_*}$ for all $a \ge 0$. In order to show this, we compute the maximum of f. Since $\lim_{a\to\infty} f(a) = +\infty$, this maximum actually exists in $\mathbb{R}_{\ge 0}$. Moreover, since $f'(0) = b > 0$, the point 0 cannot be the maximum of f, and thus the function f attains the maximum at some point $0 < a < \infty$. In particular, we can therefore find this maximum by computing the derivative of f and setting it to 0. We have that

$$f'(a) = b - a^{p-1},$$

which is equal to 0 for $a = b^{1/(p-1)}$. Thus $a = b^{1/(p-1)}$ is the unique maximiser of f, and therefore $f(a) \le f(b^{1/(p-1)})$ for all $a \ge 0$. Thus

$$\begin{aligned} ab - \frac{1}{p}a^p = f(a) \le f(b^{1/(p-1)}) &= bb^{1/(p-1)} - \frac{1}{p}b^{p/(p-1)} \\ &= b^{p/(p-1)} - \frac{1}{p}b^{p/(p-1)} = \frac{p-1}{p}b^{p/(p-1)} = \frac{1}{p_*}b^{p_*}, \end{aligned}$$

which proves (30). □

LEMMA 3.6.4 (Hölder's inequality). *Assume that* $1 < p < \infty$. *Then we have for all* $x, y \in \mathbb{K}^n$ *that*

$$\Big|\sum_{k=1}^n x_k y_k\Big| \le \sum_{k=1}^n |x_k||y_k| \le \Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p}\Big(\sum_{k=1}^n |y_k|^{p_*}\Big)^{1/p_*} = \|x\|_p \|y\|_{p_*}. \tag{31}$$

PROOF. Assume without loss of generality that x, $y \neq 0$, else the claim is trivial.

The estimate

$$\Big|\sum_{k=1}^n x_k y_k\Big| \le \sum_{k=1}^n |x_k||y_k|$$

is simply the triangle inequality on $\mathbb{K}$. Now define

$$a_k = \frac{|x_k|}{\|x\|_p} \qquad \text{and} \qquad b_k = \frac{|y_k|}{\|y\|_{p_*}}.$$

Then Young's inequality implies that

$$\begin{aligned}\frac{1}{\|x\|_p\|y\|_{p_*}}\sum_{k=1}^n |x_k||y_k| = \sum_{k=1}^n a_k b_k &\le \frac{1}{p}\sum_{k=1}^n a_k^p + \frac{1}{p_*}\sum_{k=1}^n b_k^p \\ &= \frac{1}{p}\sum_{k=1}^n \frac{|x_k|^p}{\|x\|_p^p} + \frac{1}{p_*}\sum_{k=1}^n \frac{|y_k|^{p_*}}{\|y\|_{p_*}^{p_*}} = \frac{1}{p} + \frac{1}{p_*} = 1.\end{aligned}$$

□

LEMMA 3.6.5 (Minkowski's inequality). *Assume that* $1 < p < \infty$. *Then*

$$\|x+y\|_p = \Big(\sum_{k=1}^n |x_k+y_k|^p\Big)^{1/p} \le \Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^n |y_k|^p\Big)^{1/p} = \|x\|_p + \|y\|_p. \tag{32}$$

for all $x, y \in \mathbb{K}^n$.

PROOF. Without loss of generality, assume that $x + y \neq 0$, else the claim is trivial.

We start by estimating the p-th power of the left hand side by

$$\begin{aligned}\sum_{k=1}^n |x_k+y_k|^p &\le \sum_{k=1}^n |x_k+y_k|^{p-1}\big(|x_k|+|y_k|\big) \\ &= \sum_{k=1}^n |x_k+y_k|^{p-1}|x_k| + \sum_{k=1}^n |x_k+y_k|^{p-1}|y_k|.\end{aligned} \tag{33}$$

Now we apply Hölder's inequality and obtain that

$$\begin{aligned}\sum_{k=1}^n |x_k+y_k|^{p-1}|x_k| &\le \Big(\sum_{k=1}^n |x_k+y_k|^{(p-1)p_*}\Big)^{1/p_*}\Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p} \\ &= \Big(\sum_{k=1}^n |x_k+y_k|^p\Big)^{(p-1)/p}\Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p},\end{aligned}$$

where we have used that the conjugate p_* of p satisfies

$$p_* = \frac{p}{p-1}.$$

Similarly, we get that

$$\sum_{k=1}^n |x_k+y_k|^{p-1}|y_k| \le \Big(\sum_{k=1}^n |x_k+y_k|^p\Big)^{(p-1)/p}\Big(\sum_{k=1}^n |y_k|^p\Big)^{1/p}.$$

Inserting these estimates in (33), we obtain

$$\sum_{k=1}^{n}|x_k+y_k|^p \le \Big(\sum_{k=1}^{n}|x_k+y_k|^p\Big)^{(p-1)/p}\Big(\Big(\sum_{k=1}^{n}|x_k|^p\Big)^{1/p}+\Big(\sum_{k=1}^{n}|y_k|^p\Big)^{1/p}\Big).$$

Now dividing both sides of this inequality by $\Big(\sum_{k=1}^{n}|x_k+y_k|^p\Big)^{(p-1)/p}$ results in (32). □

THEOREM 3.6.6. *For all $1 \le p \le \infty$, the function $\|\cdot\|_p\colon \mathbb{K}^n \to \mathbb{R}$ is a norm on $\mathbb{K}^n$.*

PROOF. For $p = 1$ and $p = \infty$, we have already discussed this earlier. Next, for $1 < p < \infty$, the symmetry and positivity of $\|\cdot\|_p$ are obvious from the definition. Finally, the triangle inequality for $\|\cdot\|_p$ is precisely Minkowski's inequality, Lemma 3.6.5. □

Similarly, we can define the p-norm of a continuous function:

DEFINITION 3.6.7. Let $1 \le p < \infty$. We define the function $\|\cdot\|_p\colon C([0,1]) \to \mathbb{R}$ by

$$\|f\|_p := \Big(\int_0^1 |f(x)|\,dx\Big)^{1/p}$$

for all $f \in C([0,1])$.

Moreover, for $p = \infty$ we define

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

We now show that this defines a norm on $C([0,1])$. The steps towards this results are really the same as for the case $(\mathbb{K}^n, \|\cdot\|_p)$. We will first show Hölder's inequality for the integral, and then Minkowski's inequality. Also, the proofs of these inequalities are essentially the same as for $\mathbb{K}^n$. All we have to do is to replace every sum by an integral.[4]

LEMMA 3.6.8 (Hölder's inequality). *Assume that $1 < p < \infty$. Then we have for all f, $g \in C([0,1])$ that*

$$\begin{aligned}(34)\qquad \Big|\int_0^1 f(x)g(x)\,dx\Big| &\le \int_0^1 |f(x)g(x)\,dx| \\ &\le \Big(\int_0^1 |f(x)|^p\,dx\Big)^{1/p}\Big(\int_0^1 |g(x)|^{p_*}\,dx\Big)^{1/p_*} = \|f\|_p\|g\|_{p_*}.\end{aligned}$$

PROOF. Without loss of generality assume that f, $g \ne 0$, else the claim is trivial. Moreover, the estimate

$$\Big|\int_0^1 f(x)g(x)\,dx\Big| \le \int_0^1 |f(x)g(x)\,dx|$$

follows from standard properties of the integral. Define now

$$\tilde f(x) := \frac{f(x)}{\|f\|_p} \qquad \text{and} \qquad \tilde g(x) := \frac{g(x)}{\|g\|_{p_*}}.$$

[4] The fact that we can use essentially the same proof in two seemingly different situations should give us a hint that these situations are not that different after all. In fact, both cases can be seen as special cases of a result that holds on general so called "measure spaces." For (many) more results on this topic, I refer to the course TMA4225 – Foundations of Analysis.

Then Young's inequality implies that

$$\begin{aligned}\frac{1}{\|f\|_p\|g\|_p}\int_0^1 |f(x)g(x)|\,dx &= \int_0^1 |\tilde f(x)\tilde g(x)|\,dx \le \int_0^1 \frac{1}{p}|\tilde f(x)|^p + \frac{1}{p_*}|\tilde g(x)|^{p_*}\,dx \\ &= \frac{1}{p}\int_0^1 \frac{|f(x)|^p}{\|f\|_p^p}\,dx + \frac{1}{p_*}\int_0^1 \frac{|g(x)|^{p_*}}{\|g\|_{p_*}^{p_*}}\,dx = \frac{1}{p}+\frac{1}{p_*} = 1.\end{aligned}$$

□

Lemma 3.6.9 (Minkowski's inequality). *Assume that* $1 < p < \infty$. *Then*

$$\begin{aligned}\|f+g\|_p &= \Big(\int_0^1 |f(x)+g(x)|^p\,dx\Big)^{1/p} \\ &\le \Big(\int_0^1 |f(x)|^p\,dx\Big)^{1/p} + \Big(\int_0^1 |g(x)|^p\,dx\Big)^{1/p} = \|f\|_p + \|g\|_p.\end{aligned}$$

for all f, $g \in C([0,1])$.

Proof. We estimate

$$\begin{aligned}\|f+g\|_p^p &= \int_0^1 |f(x)+g(x)|^p\,dx \le \int_0^1 |f(x)+g(x)|^{p-1}\big(|f(x)|+|g(x)|\big)\,dx \\ &\le \Big(\int_0^1 |f(x)+g(x)|^p\,dx\Big)^{(p-1)/p}\Big(\Big(\int_0^1 |f(x)|^p\,dx\Big)^{1/p} + \Big(\int_0^1 |g(x)|^p\,dx\Big)^{1/p}\Big),\end{aligned}$$

where we have used Hölder's inequality (34) and the fact that $p_* = p/(p-1)$ in the last step. Then we obtain the result by dividing by $\Big(\int_0^1 |f(x)+g(x)|^p\,dx\Big)^{(p-1)/p}$. □

Theorem 3.6.10. *For all* $1 \le p \le \infty$, *the function* $\|\cdot\|_p\colon C([0,1]) \to \mathbb{R}$ *is a norm on* $C([0,1])$.

Proof. For $p = 1$ and $p = \infty$, we have already discussed this earlier. Next, for $1 < p < \infty$, the symmetry and positivity of $\|\cdot\|_p$ are obvious from the definition. Finally, the triangle inequality for $\|\cdot\|_p$ is precisely Minkowski's inequality, Lemma 3.6.9. □

7. Sequence Spaces

Assume that $f\colon \mathbb{R} \to \mathbb{Z}$ is a periodic function with period 2π. Assuming enough regularity of f, we can then represent it as a Fourier series of the form

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{inx},$$

where

$$c_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-inx}\,dx$$

is the n-th Fourier coefficient of f. This Fourier coefficient indicates the contribution of oscillations of angular frequency n to the whole function f. For many practical problems it then can make sense to work directly with the Fourier coefficients instead of working with the function f itself. In other words, it makes sense to identify the function f with its Fourier series, which is a sequence $\{c_n\}_{n\in\mathbb{Z}}$. In this section, we will thus develop a theory of spaces of sequences and norms on such spaces.

Definition 3.7.1 (ℓ^p-spaces). Let $1 \le p < \infty$. By ℓ^p (or also $\ell^p(\mathbb{N})$), we denote the set of all sequences $(x_k)_{k\in\mathbb{N}}$ such that

$$\|x\|_p := \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} < \infty.$$

Moreover, we denote by ℓ^∞ (or $\ell^\infty(\mathbb{N})$) the set of all bounded sequences, that is, all sequences for which

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

Remark 3.7.2. There is a large conceptual difference between the constructions of the p-norms on $\mathbb{K}^n$ and the construction of the ℓ^p-spaces. In the former case we started with the vector space $\mathbb{K}^n$ and then defined a norm on this space. In the latter case the construction is in some sense reversed: We start by defining a "p-norm" on the set of all sequences, but then say that the space ℓ^p consists of all those sequences for which this "norm" is actually finite.

In the following we will show that these are indeed vector spaces and that the mapping $\|\cdot\|_p \colon \ell^p \to \mathbb{R}$ is a norm on ℓ^p.

Remark 3.7.3. One or both sides of the inequality (35) may in principle by infinite. However, the inequality in this case still is satisfied. That is, if the right hand side of (35) is finite, then so is the left hand side. In other words, if $x \in \ell^p$ and $y \in \ell^{p_*}$, then the pointwise product z defined by $z_k = x_k y_k$ satisfies $z \in \ell^1$.

Theorem 3.7.4 (ℓ^p as normed space). *Let $1 \le p \le \infty$. Then $(\ell^p, \|\cdot\|_p)$ is a normed space.*

Proof. We only prove the claim for $1 \le p < \infty$ and leave the case $p = \infty$ to the interested reader.

We have to show that ℓ^p is a vector space and that $\|\cdot\|_p$ is a norm on ℓ^p. For the former, we will show that ℓ^p is a subspace of the vector space of all sequences in $\mathbb{K}$. Following Lemma 1.1.12 we thus have to show that $0 \in \ell^p$, that $x + y \in \ell^p$ if $x, y \in \ell^p$, and that $\lambda x \in \ell^p$ if $x \in \ell^p$ and $\lambda \in \mathbb{K}$.

The inclusion $0 \in \ell^p$ holds trivially. Moreover, it is clear that $\|x\|_p \ge 0$ for all $x \in \ell^p$ with $\|x\|_p = 0$ if and only if $x = 0$.

Now assume that $x \in \ell^p$ and $\lambda \in \mathbb{K}$. Then

$$\|\lambda x\|_p = \Big(\sum_{k=1}^{\infty} |\lambda x_k|^p\Big)^{1/p} = \Big(|\lambda|^p \sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} = |\lambda|\|x\|_p < \infty.$$

This shows that $\lambda x \in \ell^p$ and also that $\|\cdot\|_p$ is homogeneous.

Finally, assume that $x, y \in \ell^p$. By Minkowski's inequality in $\mathbb{K}^n$ (Lemma 3.6.5) we have that

$$\Big(\sum_{k=1}^{n} |x_k + y_k|^p\Big)^{1/p} \le \Big(\sum_{k=1}^{n} |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{n} |y_k|^p\Big)^{1/p}$$

for all $n \in \mathbb{N}$. Taking the limit $n \to \infty$ and using that $\|x\|_p < \infty$ and $\|y\|_p < \infty$, it follows that

$$\|x+y\|_p = \Big(\sum_{k=1}^{\infty} |x_k + y_k|^p\Big)^{1/p} \le \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{\infty} |y_k|^p\Big)^{1/p} = \|x\|_p + \|y\|_p < \infty.$$

This shows that $x + y \in \ell^p$, and also that $\|\cdot\|_p$ satisfies the triangle inequality.

Thus ℓ^p is a subspace of the space of all sequences and thus a vector space itself, and $\|\cdot\|_p$ is a norm on ℓ^p. □

In addition, we can also show that Hölder's inequality holds on ℓ^p:

LEMMA 3.7.5 (Hölder's inequality on ℓ^p). *Let $1 < p < \infty$. For all sequences $(x_k)_{k\in\mathbb{N}}$, $(y_k)_{k\in\mathbb{N}} \subset \mathbb{K}$ we have that*

$$\Big|\sum_{k=1}^{\infty} x_k y_k\Big| \le \sum_{k=1}^{\infty}\big|x_k y_k\big| \le \Big(\sum_{k=1}^{\infty}|x_k|^p\Big)^{1/p}\Big(\sum_{k=1}^{\infty}|y_k|^{p_*}\Big)^{1/p_*}. \tag{35}$$

PROOF. The first inequality in (35) is obvious.

Moreover, we obtain from Hölder's inequality in $\mathbb{K}^n$ (see Lemma 3.6.4) that

$$\sum_{k=1}^{n}\big|x_k y_k\big| \le \Big(\sum_{k=1}^{n}|x_k|^p\Big)^{1/p}\Big(\sum_{k=1}^{n}|y_k|^{p_*}\Big)^{1/p_*}$$

for all $n \in \mathbb{N}$. Taking the limit $n \to \infty$ now results in (35). □

The space ℓ^∞ is the space of all bounded sequence, whereas ℓ^1 is the space of all absolutely summable sequences. Since every summable sequence necessary has to be bounded (and, in fact, converge to 0), it follows that ℓ^1 is contained in ℓ^∞. However, the converse inclusion does not hold, as the constant sequence $x = (1, 1, 1, \ldots)$ is contained in ℓ^∞ but certainly not in ℓ^1. Thus we have that $\ell^1 \subsetneq \ell^\infty$. This type of inclusion can be generalised to all ℓ^p spaces:

LEMMA 3.7.6. *Assume that $1 < p < q < \infty$. Then*

$$\ell^1 \subsetneq \ell^p \subsetneq \ell^q \subsetneq \ell^\infty.$$

PROOF. We show first that all the inclusions hold. Here we note first that all sequences in ℓ^q necessarily have to converge to 0 and thus are bounded, which proves the last inclusion. Next assume that $x \in \ell^p$. Then

$$\|x\|_q^q = \sum_{k=1}^{\infty}|x_k|^q = \sum_{k=1}^{\infty}|x_k|^{q-p}|x_k|^p \le \Big(\sup_{k\in\mathbb{N}}|x_k|^{q-p}\Big)\sum_{k=1}^{\infty}|x_k|^p = \Big(\sup_{k\in\mathbb{N}}|x_k|^{q-p}\Big)\|x\|_p^p.$$

Since $q > p$ and the sequence x is bounded, it follows that $\|x\|_q < \infty$ and thus $x \in \ell^q$. This shows that $\ell^p \subset \ell^q$. Moreover, the inclusion $\ell^1 \subset \ell^p$ can be shown using a similar argument.

Now we show that the inclusions are actually strict. First we note that the constant sequence $x = (1, 1, \ldots)$ is contained in ℓ^∞ but not in ℓ^q, which shows that $\ell^1 \subsetneq \ell^\infty$. Next consider the sequence $x = (x_1, x_2, \ldots)$ given by

$$x_k = \frac{1}{k^{1/p}}.$$

Then

$$\|x_k\|_q^q = \sum_{k=1}^{\infty}\frac{1}{k^{q/p}} < \infty$$

as $q > p$ and therefore $x \in \ell^q$. However,

$$\|x_k\|_p^p = \sum_{k=1}^{\infty}\frac{1}{k} = +\infty,$$

which shows that $x \notin \ell^p$. Thus $\ell^p \subsetneq \ell^q$. Moreover, the proof of the strict inclusion $\ell^1 \subsetneq \ell^p$ is similar. □

LEMMA 3.7.7 (Inner product on ℓ^2). *The norm $\|\cdot\|_2$ on the space ℓ^2 is induced by the inner product*

$$\langle x, y\rangle := \sum_{k=1}^{\infty} x_k\overline{y_k}.$$

PROOF. It is clear from the definition that $\|x\|_2 = (\langle x, x\rangle)^{1/2}$. Thus we only have to show that $\langle\cdot,\cdot\rangle$ is actually an inner product. However, it is straightforward to verify that all the requirements for an inner product, that is, linearity in the first component, conjugate symmetry, and positive definiteness, are satisfied. □

8. Banach and Hilbert spaces

We now discuss the completeness of the different normed and inner product spaces we have defined.

DEFINITION 3.8.1 (Banach space). A *Banach space* is a complete normed space.

DEFINITION 3.8.2 (Hilbert space). A *Hilbert space* is a complete inner product space.

EXAMPLE 3.8.3. The space $(\mathbb{K}^n, \|\cdot\|_p)$ is a Banach space for all $1 \le p \le \infty$, and a Hilbert space for $p = 2$.

THEOREM 3.8.4. *The space* $(C([0,1]), \|\cdot\|_\infty)$ *is a Banach space.*

PROOF. This has in fact been shown in Theorem 3.2.7. □

EXAMPLE 3.8.5. The space $(C[0,1], \|\cdot\|_1)$ is *not* a Banach space, as it is not complete (see Lemma 3.2.6).

THEOREM 3.8.6. *The space* $(\ell^p, \|\cdot\|_p)$ *is a Banach space for all* $1 \le p \le \infty$*, and a Hilbert space for* $p = 2$.

PROOF. We only have to show that $(\ell^p, \|\cdot\|_p)$ is complete. For simplicity, we will restrict ourselves to the case $p = 1$. (The case $1 < p < \infty$ is similar; the case $p = \infty$ is simpler, and actually a problem on the current exercise sheet.)

Assume that $\{x^{(n)}\}_{n\in\mathbb{N}} \subset \ell^1$ is a Cauchy sequence. We have to show that this sequence convergences with respect to $\|\cdot\|_1$ to some $x \in \ell^1$. We start by identifying a candidate for x.

Since $\{x^{(n)}\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists for all $\varepsilon > 0$ some $N \in \mathbb{N}$ such that

$$\|x^{(n)} - x^{(m)}\|_1 = \sum_{k=1}^{\infty} |x_k^{(n)} - x_k^{(m)}| < \varepsilon$$

whenever n, $m \ge N$. In particular, this shows that

$$|x_k^{(n)} - x_k^{(m)}| < \varepsilon$$

for all $k \in \mathbb{N}$ and all n, $m \ge N$, and thus $\{x_k^{(n)}\}_{n\in\mathbb{N}} \subset \mathbb{K}$ is for all $k \in \mathbb{N}$ a Cauchy sequence in $\mathbb{K}$. Since $\mathbb{K}$ is complete, it follows that $x_k := \lim_{n\to\infty} x_k^{(n)}$ exists for all $k \in \mathbb{N}$. Define now $x := (x_1, x_2, \ldots)$. We will show that $x \in \ell^1$ and that $\lim_{n\to\infty} \|x^{(n)} - x\|_1 = 0$, that is, $x = \lim_{n\to\infty} x^{(n)}$ with respect to $\|\cdot\|_1$.

Since $\{x^{(n)}\}_{n\in\mathbb{N}}$ is a Cauchy sequence, it is bounded by Lemma 3.5.14. Thus there exists $R > 0$ such that

$$\|x^{(n)}\|_1 < R \qquad \text{for all } n \in \mathbb{N}.$$

For all $K \in \mathbb{N}$ we have that

$$\begin{aligned}(36)\qquad \sum_{k=1}^{K} |x_k| = \sum_{k=1}^{K} \Big|\lim_{n\to\infty} x_k^{(n)}\Big| &= \sum_{k=1}^{K} \lim_{n\to\infty} |x_k^{(n)}| = \lim_{n\to\infty} \sum_{k=1}^{K} |x_k^{(n)}| \\ &\le \sup_{n\in\mathbb{N}} \sum_{k=1}^{K} |x_k^{(n)}| \le \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} |x_k^{(n)}| = \sup_{n\in\mathbb{N}} \|x^{(n)}\|_1 \le R.\end{aligned}$$

Note here that we were able to exchange the order of the limit and the sum, as we were only dealing with a finite sum. The estimate (36) holds for all $K \in \mathbb{N}$. Thus it still holds in the limit $K \to \infty$, which shows that

$$\|x\|_1 = \lim_{K\to\infty} \sum_{k=1}^{K} |x_k| \leq R,$$

and, in particular, $x \in \ell^1$.

Finally we show that $\lim_{n\to\infty} \|x^{(n)} - x\|_1 = 0$. Let therefore $\varepsilon > 0$. Since $\{x^{(n)}\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists $N \in \mathbb{N}$ such that $\|x^{(n)} - x^{(m)}\|_1 < \varepsilon$ whenever $n, m \geq N$. In particular, we have for every $K \in \mathbb{N}$ that

$$\sum_{k=1}^{K} |x_k^{(n)} - x_k^{(m)}| < \varepsilon \qquad \text{for all } n,\, m \geq N.$$

Taking the limit $m \to \infty$, this implies that

$$\sum_{k=1}^{K} |x_k^{(n)} - x_k| \leq \varepsilon$$

for all $K \in \mathbb{N}$ and all $n \geq N$. Now we can again take the limit $K \to \infty$ and obtain that

$$\|x^{(n)} - x\|_1 = \sum_{k=1}^{\infty} |x_k^{(n)} - x_k| \leq \varepsilon$$

for all $n \geq N$. Since $\varepsilon > 0$ was arbitrary, this proves that $\|x^{(n)} - x\|_1 \to 0$. Thus $(\ell^1, \|\cdot\|_1)$ is complete. □

9. Completions

We have seen that there exist incomplete normed spaces like, for instance, the space $C([0,1])$ with the norm $\|\cdot\|_1$. Now one can ask the question, whether it is possible to enlarge this space in such a way that it becomes complete. The basic intuition for this is that we only need to "add all limits of the non-convergent Cauchy sequences" to the space in order to achieve this goal. In this section, we will discuss this idea in an somewhat more mathematical manner. For that, we have to introduce some more notation.

Definition 3.9.1 (Isomorphism). Let U and V be vector spaces (over the same field $\mathbb{K}$). An *isomorphism* between U and V is a bijective linear transformation $T\colon U \to V$. The spaces U and V are called *isomorphic*, if an isomorphism between U and V exists.

Example 3.9.2. The spaces $\mathbb{K}^{n+1}$ and $\mathcal{P}_n(\mathbb{K})$ (of polynomials of degree $\leq n$) are isomorphic. An example of an isomorphism between $\mathbb{K}^{n+1}$ and $\mathcal{P}_n(\mathbb{K})$ is the mapping that maps a vector $(c_0, c_1, \ldots, c_n)$ to the polynomial $p(x) = c_0 + c_1 x + \ldots + c_n x^n$.

Example 3.9.3. The spaces $\mathbb{K}^{n^2}$ and $\mathbf{Mat}_n(\mathbb{K})$ are isomorphic. One example of an isomorphism between $\mathbb{K}^{n^2}$ and $\mathbf{Mat}_n(\mathbb{K})$ is the mapping that maps a vector $x \in \mathbb{K}^{n^2}$ to the matrix $A = (a_{ij})_{1\leq i,j\leq n} \in \mathbf{Mat}_n(\mathbb{K})$ with entries given as $a_{ij} = x_{ni+j}$ for $1 \leq i, j \leq n$.

Example 3.9.4. Generally, every n-dimensional vector space U over $\mathbb{K}$ is isomorphic to $\mathbb{K}^n$. If $\{u_1, \ldots, u_n\}$ is a basis of U, we can define an isomorphism $T\colon \mathbb{K}^n \to U$ by

$$T(\alpha_1, \ldots, \alpha_n) = \alpha_1 u_1 + \ldots + \alpha_n u_n.$$

Essentially, we say that the spaces U and V are isomorphic if they "look the same" as vector spaces. However, we ignore all additional structures like norms or inner product in this definition.

DEFINITION 3.9.5 (Isometry). Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and let $T\colon U \to V$ be linear. We say that T is an *isometry* or an *embedding*, if $\|Tu\|_V = \|u\|_U$ for all $u \in U$.

We say that $(U, \|\cdot\|_U)$ is *embedded in* $(V, \|\cdot\|_V)$, if there exists an embedding $T\colon U \to V$.

REMARK 3.9.6. If $T\colon U \to V$ is an isometry, then it is necessarily injective: If $Tu = 0$, then $\|Tu\|_V = 0$ and thus also $\|u\|_U = 0$, which can only happen if $u = 0$. This shows that $\ker T = \{0\}$, which implies the injectivity of T.

EXAMPLE 3.9.7. The space $(\mathbb{K}^n, \|\cdot\|_p)$ is embedded in $(\mathbb{K}^{n+1}, \|\cdot\|_p)$. One example of an embedding is the mapping $T\colon \mathbb{K}^n \to \mathbb{K}^{n+1}$ defined by $(x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n, 0)$.

In general, however, the space $(\mathbb{K}^n, \|\cdot\|_p)$ is *not* embedded in $(\mathbb{K}^{n+1}, \|\cdot\|_q)$ if $q \neq p$. Whereas there exist many linear mappings $T\colon \mathbb{K}^n \to \mathbb{K}^{n+1}$, in general it is impossible to find any one that satisfies $\|Tu\|_q = \|u\|_p$ for all $u \in \mathbb{K}^n$.[5]

THEOREM 3.9.8 (Completion). *Assume that $(U, \|\cdot\|_U)$ is a normed space. Then there exists a Banach space $(V, \|\cdot\|_V)$ and an embedding $i\colon U \to V$ such that* $\operatorname{ran} i$ *is dense in V.*

Moreover, the space V is unique in the following sense: If W is another Banach space with the same properties, then the spaces V and W are isometrically isomorphic. (That is, there exists a bijective isometry $T\colon V \to W$.)

DEFINITION 3.9.9. The space $(V, \|\cdot\|_V)$ from Theorem 3.9.8 is called the *completion* of $(U, \|\cdot\|_U)$.

Intuitively, we can say that U is a dense subspace of V and that the mapping $i\colon U \to V$ is just the inclusion operator that maps $u \in U$ to itself, but now seen as an element of V. Because of the density of U in V, it now follows that every element $v \in V$ can be approached by a sequence $\{u_k\}_{k\in\mathbb{N}} \subset U$ such that $v = \lim_{k\to\infty} u_k$. We thus can say informally that the completion V of U consists of the "limits" of all Cauchy sequences in the space U.

EXAMPLE 3.9.10. Denote by

$$c_{\text{fin}} := \bigl\{x = (x_k)_{k\in\mathbb{N}} : \text{ there exists } K \in \mathbb{N} \text{ such that } x_k = 0 \text{ for all } k > K\bigr\}$$

the set of all finite sequences. Then it is easy to see that c_{fin} is a vector space. Moreover, for each $1 \le p \le \infty$ we can consider the p-norm $\|\cdot\|_p$ on ℓ^p.

We will show in the following that for $1 \le p < \infty$ the completion of $(c_{\text{fin}}, \|\cdot\|_p)$ is equal to ℓ^p. For that, we recall first that ℓ^p is indeed a Banach space. Moreover, the inclusion

$$i\colon c_{\text{fin}} \to \ell^p, \qquad ix = x$$

is an embedding of c_{fin} into ℓ^p. We therefore only have to show that c_{fin} is dense in ℓ^p.

Let therefore $x = (x_k)_{k\in\mathbb{N}} \in \ell^p$ be arbitrary. We have to show that there exist $x^{(n)} \in c_{\text{fin}}$ such that $\lim_{n\to\infty} \|x^{(n)} - x\|_p = 0$. Define therefore

$$x^{(n)} := (x_1, \ldots, x_n, 0, 0, \ldots) \in c_{\text{fin}}.$$

[5]There are some exceptions to this, though. Try to find them!

Let moreover $\varepsilon > 0$. Since $x \in \ell^p$, we have that

$$\|x\|_p = \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} < \infty,$$

and thus there exists $N \in \mathbb{N}$ such that

$$\Big(\sum_{k=N+1}^{\infty} |x_k|^p\Big)^{1/p} < \varepsilon.$$

Let now $n \geq N$. Then

$$\|x - x^{(n)}\|_p = \Big(\sum_{k=n+1}^{\infty} |x_k|^p\Big)^{1/p} \leq \Big(\sum_{k=N+1}^{\infty} |x_k|^p\Big)^{1/p} < \varepsilon.$$

Since this holds for all $n \geq N$ and $\varepsilon > 0$ was arbitrary, this shows that $x = \lim_{n\to\infty} x^{(n)}$. Thus every vector $x \in \ell^p$ can be approximate (in the p-norm) by a sequence of vectors in c_{fin}, and thus c_{fin} is dense in ℓ^p. Together, this implies that ℓ^p is the completion of c_{fin} with respect to $\|\cdot\|_p$ for $1 \leq p < \infty$.

We now consider the case $p = \infty$. In this case, it is easy to see that c_{fin} is *not* dense in ℓ^∞. Indeed, take $x = (1, 1, \ldots) \in \ell^\infty$. Then, if $y = (y_1, \ldots, y_K, 0, \ldots) \in c_{\text{fin}}$ is any finite sequence, it follows that $\|x - y\|_\infty \geq 1$. Thus it is impossible to find any sequence of finite sequences that converges to x.

In fact, one can show that the completion of c_{fin} with respect to $\|\cdot\|_\infty$ is the space c_0 of all sequences that converge to 0. For a proof, I have to refer to the exercises.

Remark 3.9.11. One important example in practice is the case of the space $C([0,1])$ with the norm $\|\cdot\|_p$, $1 \leq p < \infty$. Here the completion of $(C([0,1]), \|\cdot\|_p)$ is the space $L^p([0,1])$ of *Lebesgue p-integrable functions*. A detailed discussion of the construction of these spaces can be found in the course TMA4225 - Foundations of Analysis.

Remark 3.9.12. If U is an inner product space, then its completion V will necessarily be a Hilbert space. The reason for this is that the parallelogram law holds on the dense subspace $U \subset V$ because U is a Hilbert space, and therefore also on the whole space.

CHAPTER 4

Bounded Linear Operators

1. Continuity of linear mappings

In the following, we will study linear mappings between normed spaces. To start with, we have to address the question of continuity. It is not that difficult to see that linear mappings $T\colon \mathbb{K}^n \to \mathbb{K}^m$ are necessarily continuous. In the case of linear mappings between arbitrary (infinite dimensional) normed spaces, however, the continuity need not necessarily hold. As an example, consider the mapping $T\colon C([0,1]) \to \mathbb{R}$, $f \mapsto Tf := f(1)$. This mapping is not continuous with respect to the norm $\|\cdot\|_1$ on $C([0,1])$. This can be seen by considering the sequence $f_n(x) := x^n$. We have that

$$\|f_n\|_1 = \int_0^1 |x^n|\,dx = \frac{1}{n+1} \to_{n\to\infty} 0,$$

which shows that $0 = \lim_{n\to\infty} f_n$. However, $T(f_n) = f_n(1) = 1$ for all $n \in \mathbb{N}$ and thus

$$T\bigl(\lim_{n\to\infty} f_n\bigr) = T(0) = 0 \neq 1 = \lim_{n\to\infty} Tf_n.$$

Now assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are normed spaces and that $T\colon U \to V$ is an arbitrary (not necessarily linear) mapping. Then T is continuous at a point $u \in U$ if one of the following equivalent conditions hold:

(1) For all sequences $\{u_n\}_{n\in\mathbb{N}}$ with $\|u_n - u\|_U \to_{n\to\infty} 0$ we have that $\|T(u_n) - T(u)\|_V \to_{n\to\infty} 0$.
(2) For all $\varepsilon > 0$ there exists $\delta > 0$ such that $\|T(u) - T(v)\|_V < \varepsilon$ whenever $v \in U$ satisfies $\|u - v\|_U < \delta$.

In the case of linear mappings, though, these conditions can be simplified significantly.

Definition 4.1.1 (Bounded linear mapping). Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and let $T\colon U \to V$ be linear. We say that T is *bounded*, if there exists $C \geq 0$ such that

$$\|Tu\|_V \leq C\|u\|_U \qquad \text{for all } u \in U. \tag{37}$$

Example 4.1.2. Let $A = (a_{ij}) \in \mathbb{K}^{m\times n}$ be a matrix and consider the linear mapping $A\colon \mathbb{K}^n \to \mathbb{K}^m$, $x \mapsto Ax$. We will show that A is bounded with respect to $\|\cdot\|_1$ on both $\mathbb{K}^n$ and $\mathbb{K}^m$. For that we estimate

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m \Bigl|\sum_{j=1}^n a_{ij}x_j\Bigr| \leq \sum_{i=1}^m\sum_{j=1}^n |a_{ij}||x_j| \\ &= \sum_{j=1}^n\Bigl(\sum_{i=1}^m |a_{ij}|\Bigr)|x_j| \leq \Bigl(\max_{1\leq j\leq n}\sum_{i=1}^m |a_{ij}|\Bigr)\Bigl(\sum_{j=1}^n |x_j|\Bigr) = \Bigl(\max_{1\leq j\leq n}\sum_{i=1}^m |a_{ij}|\Bigr)\|x\|_1.\end{aligned}$$

Thus (37) holds with

$$C = \max_{1\leq j\leq n}\sum_{i=1}^m |a_{ij}|$$

and thus the mapping A is bounded.

EXAMPLE 4.1.3. Let $k\colon [0,1]\times[0,1]\to\mathbb{K}$ be continuous and define

$$T\colon C([0,1])\to C([0,1]),$$
$$Tf(x) := \int_0^1 k(x,y)f(y)\,dy.$$

Then T is bounded with respect to $\|\cdot\|_\infty$ on both spaces. Indeed, we can estimate

$$\begin{aligned}\|Tf\|_\infty &= \max_{x\in[0,1]}|Tf(x)| = \max_{x\in[0,1]}\Bigl|\int_0^1 k(x,y)f(y)\,dy\Bigr|\\ &\le \max_{x\in[0,1]}\Bigl(\max_{y\in[0,1]}|k(x,y)|\int_0^1|f(y)|\,dy\Bigr) = \Bigl(\max_{(x,y)\in[0,1]^2}|k(x,y)|\Bigr)\Bigl(\int_0^1|f(y)|\,dy\Bigr)\\ &\le \Bigl(\max_{(x,y)\in[0,1]^2}|k(x,y)|\Bigr)\|f\|_\infty.\end{aligned}$$

Thus (37) holds with

$$C = \max_{(x,y)\in[0,1]^2}|k(x,y)|$$

and T is bounded.

EXAMPLE 4.1.4. Let $T\colon C([0,1])\to\mathbb{K}$, $Tf = f(1)$. Then T is bounded with respect to $\|\cdot\|_\infty$, because

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

for all $f\in C([0,1])$.

However, the mapping T is *unbounded* with respect to $\|\cdot\|_1$. Consider e.g. the functions $f_n(x) := (n+1)x^n$. Then

$$\|f_n\|_1 = \int_0^1 |(n+1)x^n|\,dx = 1$$

for all $n\in\mathbb{N}$, but

$$|Tf_n| = |f_n(1)| = n+1 = (n+1)\|f_n\|_1.$$

Thus it is not possible to find a finite number $C\ge 0$ such that (37) holds, and thus T is unbounded with respect to $\|\cdot\|_1$.

It is not by accident that we found the same mapping T to be both unbounded and discontinuous, as the following important theorem shows:

THEOREM 4.1.5 (Boundedness and continuity). *Let $(U,\|\cdot\|_U)$ and $(V,\|\cdot\|_V)$ be normed spaces and let $T\colon U\to V$ be linear. The following are equivalent:*

(1) T is bounded.
(2) T is continuous.
(3) There exists $u_0\in U$ such that T is continuous at u_0.

PROOF. We start by showing the implication (1) $\Longrightarrow$ (2). Assume therefore that T is bounded, that is, that there exists $C\ge 0$ such that

$$\|Tu\|_V \le C\|u\|_U \qquad \text{for all } u\in U.$$

Assume moreover that $u\in U$ and that $\{u_n\}_{n\in\mathbb{N}}$ is a sequence in U satisfying $\lim_{n\to\infty}u_n = u$. Then

$$\|T(u_n)-T(u)\|_V = \|T(u_n-u)\|_V \le C\|u_n-u\|_U \to_{n\to\infty} 0,$$

and thus $Tu = \lim_{n\to\infty}Tu_n$, which proves that T is continuous at u. Since u was arbitrary, this proves the continuity of T.

The implication (2) $\implies$ (3) is trivial.

For the implication (3) $\implies$ (1), assume that $u_0 \in U$ is such that T is continuous at u_0. Then there exists $\delta > 0$ such that

$$\|Tu_0 - Tv\|_V < 1 \qquad \text{whenever } v \in U \text{ satisfies } \|u_0 - v\|_U < \delta.$$

Now let $w \in U$ be arbitrary with $w \neq 0$. Define moreover

$$v = u + \frac{\delta}{2\|w\|_U} w.$$

Then

$$\|u - v\| = \Big\|\frac{\delta}{2\|w\|_U} w\Big\|_U = \frac{\delta}{2\|w\|_U}\|w\|_U = \frac{\delta}{2} < \delta,$$

and thus

$$\begin{aligned}1 > \|Tu - Tw\|_V = \Big\|Tu - T\Big(u + \frac{\delta}{2\|w\|_U} w\Big)\Big\|_V \\ = \Big\|T\Big(\frac{\delta}{2\|w\|_U} w\Big)\Big\|_V = \frac{\delta}{2\|w\|_U}\|Tw\|_V.\end{aligned}$$

This shows that

$$\|Tw\|_V \le \frac{2}{\delta}\|w\|_U.$$

Since w was arbitrary and the constant on the right hand side is independent of w, it follows that T is bounded. □

We have thus shown that, for linear mappings on normed spaces, continuity is precisely the same as boundedness and continuous linear mappings are the same as bounded linear mappings. Because of the relatively simpler definition of boundedness, one usually the term "bounded linear mapping" instead of "continuous linear mapping."

Remark 4.1.6. The equivalence of boundedness and continuity does not hold for general non-linear mappings. There are some generalisation to e.g. bilinear mappings available, that turn out to be extremely useful for the analysis of partial differential equations.

Example 4.1.7. We now revisit the examples from the start of this section.

- Every mapping $A\colon \mathbb{K}^n \to \mathbb{K}^m$ (defined by a matrix) is bounded and thus continuous with respect to $\|\cdot\|_1$ on both spaces. More general, it is possible to show that every linear mapping between finite dimensional normed spaces is continuous. (We might actually prove this later in the course.)
- The mapping $T\colon C([0,1]) \to \mathbb{K}$, $Tf = f(1)$, is bounded and thus continuous with respect to $\|\cdot\|_\infty$, but unbounded and discontinuous with respect to $\|\cdot\|_1$.
- If $k\colon [0,1]^2 \to \mathbb{K}$ is continuous, then the mapping $T\colon C([0,1]) \to C([0,1])$, $Tf(x) = \int_0^1 k(x,y)f(y)\,dy$ is bounded and thus continuous with respect to $\|\cdot\|_\infty$.

We now discuss some further examples on sequence spaces:

Example 4.1.8. Let $1 \le p \le \infty$. The *left shift operator* $L\colon \ell^p \to \ell^p$ is the mapping defined as

$$L(x_1, x_2, x_3, \ldots) := (x_2, x_3, x_4, \ldots).$$

The *right shift operator* $R\colon \ell^p \to \ell^p$ is the mapping defined as

$$R(x_1, x_2, x_3, \ldots) := (0, x_1, x_2, \ldots).$$

Both of these operators are bounded. Specifically, we have that $\|Rx\|_p = \|x\|_p$ for all $x \in \ell^p$, and $\|Lx\|_p \le \|x\|_p$ for all $x \in \ell^p$.

EXAMPLE 4.1.9. Assume that $y = (y_1, y_2, \dots) \in \ell^\infty$ is some fixed sequence, and define the mapping $T\colon \ell^1 \to \mathbb{K}$,

$$Tx := \sum_{k=1}^\infty x_k y_k.$$

This mapping is well-defined (in the sense that the infinite series on the right hand side converges for all $x \in \ell^1$, as

$$\Big|\sum_{k=1}^\infty x_k y_k\Big| \le \big(\sup_k |y_k|\big) \sum_k |x_k| = \|y\|_\infty \|x\|_1.$$

In fact, this last inequality actually shows that T is bounded, the constant C in (37) being $\|y\|_\infty$.

More general, let $1 \le p \le \infty$ and let p_* be the Hölder conjugate of p. Let moreover $y \in \ell^{p_*}$ be fixed and define $T\colon \ell^p \to \mathbb{K}$,

$$Tx = \sum_{k=1}^\infty x_k y_k.$$

Then Hölder's inequality implies that

$$|Tx| = \Big|\sum_{k=1}^\infty x_k y_k\Big| \le \|x\|_p \|y\|_{p_*},$$

which shows that T is a well-defined bounded linear mapping.

2. Norms of bounded linear operators

Assume that U and V are vector spaces. Let moreover T, $S\colon U \to V$ be linear operators. Then we can define a new operator $T + S\colon U \to V$ by

$$(T+S)(u) := Tu + Su.$$

Moreover, it is easy to verify that this operator $T + S$ again is linear. For instance, we have that

$$(T+S)(u+v) = T(u+v)+S(u+v) = Tu+Tv+Su+Sv = (T+S)(u)+(T+S)(v).$$

Moreover, for $\lambda \in \mathbb{K}$ we can define the operator $\lambda T\colon U \to V$ by

$$(\lambda T)(u) := \lambda(Tu).$$

Again, the linearity of T implies that also λT is linear. Now we can see that the set of all linear operators between two vector spaces U and V forms itself a vector space (with the just defined addition and scalar multiplication).

DEFINITION 4.2.1. Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. By $L(U, V)$ we denote the set of bounded linear operators $T\colon U \to V$.

LEMMA 4.2.2. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. The set $L(U, V)$ is a vector space.*

PROOF. We have to show that $L(U, V)$ is a subspace of the space of all linear operators from U to V. For that, we have to verify that 0 is bounded linear, that the sum of two bounded linear operators is again bounded, and that scalar multiples of bounded linear operators are bounded.

First, the 0-operator is obviously bounded, as

$$\|0u\|_V = \|0\|_V = 0 = 0\|u\|_U$$

for all $u \in U$. Thus $0 \in L(U,V)$.

Next, assume that T, $S \in L(U,V)$ are bounded linear operators. Then there exist C, $D \geq 0$ such that

$$\|Tu\|_V \leq C\|u\|_U \qquad \text{and} \qquad \|Su\|_V \leq D\|u\|_U$$

for all $u \in U$. Now the triangle inequality implies that

$$\|(T+S)u\|_V = \|Tu+Su\|_V \leq \|Tu\|_V + \|Su\|_V \leq C\|u\|_U + D\|u\|_U = (C+D)\|u\|_U,$$

which shows that $T+S$ is bounded as well.

Finally, assume that $T \in L(U,V)$ and $\lambda \in \mathbb{K}$. Because of the boundedness of T, there exists $C \geq 0$ such that

$$\|Tu\|_V \leq C\|u\|_U$$

for all $u \in U$. Now the homogeneity of the norm and the linearity of T imply that

$$\|(\lambda T)u\|_V = \|\lambda(Tu)\|_V = \|T(\lambda u)\|_V \leq C\|\lambda u\|_U = C|\lambda|\|u\|_U$$

for all $u \in U$, and thus λT is bounded. □

Next, we will use the definition of boundedness of linear operators in order to define a norm on $L(U,V)$. By definition, we can find for every $T \in L(U,V)$ some number $C \geq 0$ such that $\|Tu\|_V \leq C\|u\|_U$ for all $u \in U$. The *smallest* such number C can be used as a norm.

DEFINITION 4.2.3. Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. We define the mapping $\|\cdot\|_{L(U,V)} \colon L(U,V) \to \mathbb{R}$ by

$$\|T\|_{L(U,V)} := \inf\bigl\{C > 0 : \|Tu\|_V \leq C\|u\|_U \text{ for all } u \in U\bigr\}. \tag{38}$$

The following result provides an alternative characterisation of $\|\cdot\|_{L(U,V)}$ that is more useful in practice.

LEMMA 4.2.4. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces with $U \neq \{0\}$. Then*

$$\|Tu\|_{L(U,V)} = \sup_{\substack{u \in U,\\ u \neq 0}} \frac{\|Tu\|_V}{\|u\|_U} = \sup_{\substack{u \in U,\\ \|u\|_U = 1}} \|Tu\|_V. \tag{39}$$

PROOF. We note first that, for all $u \in U$ with $u \neq 0$, we have

$$\frac{\|Tu\|_V}{\|u\|_U} = \Bigl\|T\Bigl(\frac{u}{\|u\|_U}\Bigr)\Bigr\|_V \qquad \text{and} \qquad \Bigl\|\frac{u}{\|u\|_U}\Bigr\|_U = 1.$$

This shows the second equality in (39).

Next, assume that $C \geq 0$ is such that $\|Tu\|_V \leq C\|u\|_U$ for all $u \in U$. For $u \neq 0$ it then follows that

$$C \geq \frac{\|Tu\|_V}{\|u\|_U} \geq \sup_{\substack{v \in U,\\ v \neq 0}} \frac{\|Tv\|_V}{\|v\|_U}.$$

Taking the infimum over all $C \geq 0$, we obtain that

$$\|T\|_{L(U,V)} \geq \sup_{\substack{v \in U,\\ v \neq 0}} \frac{\|Tv\|_V}{\|v\|_U}.$$

Conversely, we have for all $u \neq 0$ that

$$\|Tu\|_V = \frac{\|Tu\|_V}{\|u\|_U}\|u\|_U \leq \Bigl(\sup_{\substack{v \in U,\\ v \neq 0}} \frac{\|Tv\|_V}{\|v\|_U}\Bigr)\|u\|_U,$$

which shows that $\sup_{\substack{v\in U,\\ v\neq 0}} \frac{\|Tv\|_V}{\|v\|_U}$ can be used as a constant C in (37). As a consequence,

$$\|T\|_{L(U,V)} \le \sup_{\substack{v\in U,\\ v\neq 0}} \frac{\|Tv\|_V}{\|v\|_U}.$$

□

REMARK 4.2.5. An immediate consequence of Lemma 4.2.4 is the inequality

$$\|Tu\|_V \le \|T\|_{L(U,V)}\|u\|_U$$

for all $u \in U$ and all $T \in L(U,V)$.

THEOREM 4.2.6. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. Then $\|\cdot\|_{L(U,V)}$ as defined in* (38) *defines a norm on $L(U,V)$.*

PROOF. Exercise sheet 9! □

LEMMA 4.2.7. *Assume that $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$, and $(W, \|\cdot\|_W)$ are normed spaces and that $T: U \to V$ and $S: V \to W$ are bounded linear. The the composition $S \circ T: U \to W$ is bounded linear and*

$$\|S \circ T\|_{L(U,W)} \le \|S\|_{L(V,W)}\|T\|_{L(U,V)}. \tag{40}$$

PROOF. For every $u \in U$ we have

$$\|(S\circ T)u\|_W = \|S(Tu)\|_W \le \|S\|_{L(V,W)}\|Tu\|_V \le \|S\|_{L(V,W)}\|T\|_{L(U,V)}\|u\|_U.$$

This shows that $S \circ T$ is bounded. Moreover, we can use $C = \|S\|_{L(V,W)}\|T\|_{L(U,V)}$ as a constant in the definition (37). Now (40) follows from the definition of $\|S \circ T\|_{L(U,W)}$ as the smallest such constant. □

EXAMPLE 4.2.8. We now consider specifically the case where $A \in \mathbb{K}^{m\times n}$ is a matrix defining a linear operator $A: \mathbb{K}^n \to \mathbb{K}^m$. Moreover, we consider on $\mathbb{K}^n$ and $\mathbb{K}^m$ the p-norm for some $1 \le p \le \infty$. For simplicity, we denote

$$\|A\|_p := \sup_{\|x\|_p=1} \|Ax\|_p$$

the norm of the matrix (or operator) A with respect to the p-norm on both spaces. We have four different cases:

(1) The case $p = 2$, that is,

$$\|A\|_2 := \sup_{\|x\|_2=1} \|Ax\|_2.$$

This is a problem that we have already discussed previously in the chapter about the singular value decomposition, and we seen in the proof of Theorem 2.4.1 that the solution of the optimisation problem on the right hand side is precisely the largest singular value of A. That is, we have that

$$\|A\|_2 = \sigma_1.$$

(2) The case $p = 1$, that is,

$$\|A\|_1 := \sup_{\|x\|_1=1} \|Ax\|_1.$$

This is a case we have discussed previously in Example 4.1.2. Although we have not tried to find the optimal constant there, it turns out that we have actually found it. That is, one can show that

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|.$$

(3) The case $p = \infty$, that is,

$$\|A\|_\infty := \sup_{\|x\|_\infty = 1} \|Ax\|_\infty.$$

Here one can show that

$$\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|.$$

(4) All other cases, that is, $p \neq 1, 2, \infty$. Here no analytic formulas for the p-norm of a general matrix $A \in \mathbb{K}^{m \times n}$ exist, but it is possible to approximate solutions of the optimisation problem

$$\max_{x \in \mathbb{K}^n} \|Ax\|_p \qquad \text{subject to } \|x\|_p = 1$$

numerically.

Example 4.2.9. More general, we can use different norms on the domain $\mathbb{K}^n$ and the target space $\mathbb{K}^m$, say the p-norm on $\mathbb{K}^n$ and the q-norm on $\mathbb{K}^m$ with $p \neq q$. That is, we interpret the matrix A as a mapping between the normed spaces $(\mathbb{K}^n, \|\cdot\|_p)$ and $(\mathbb{K}^m, \|\cdot\|_q)$. The resulting operator norm is defined as

$$\|A\|_{p \mapsto q} := \sup_{\|x\|_p = 1} \|Ax\|_q.$$

In the specific case $p = 1$, we again have a simple explicit formula for the resulting norm, namely

$$\|A\|_{1 \mapsto q} = \max_{1 \le j \le n} \Big(\sum_{i=1}^{m} |a_{ij}|^q \Big)^{1/q} \tag{41}$$

for $1 \le q < \infty$, and

$$\|A\|_{1 \mapsto \infty} = \max_{\substack{1 \le j \le n, \\ 1 \le i \le m}} |a_{ij}|$$

for $q = \infty$.

We will show now that (41) holds for $1 \le q < \infty$. Denote by $e_j \in \mathbb{K}^n$ the j-th standard basis vector. Then we have for all $x \in \mathbb{K}^n$ with $\|x\|_1 = 1$ that

$$\begin{aligned} \|Ax\|_q = \Big\| \sum_{j=1}^{n} A(x_j e_j) \Big\|_q &= \Big\| \sum_{j=1}^{n} x_j (Ae_j) \Big\|_q \le \sum_{j=1}^{n} |x_j| \|Ae_j\|_q \\ &\le \|x\|_1 \big(\max_{1 \le j \le n} \|Ae_j\|_q \big) = \max_{1 \le j \le n} \Big(\sum_{i=1}^{m} |a_{ij}|^q \Big)^{1/q}. \end{aligned}$$

Moreover, if $1 \le k \le n$ is an index where the maximum in (41) is attained, then

$$\|Ae_k\|_q = \Big(\sum_{i=1}^{m} |a_{ik}|^q \Big)^{1/q} = \max_{1 \le j \le n} \Big(\sum_{i=1}^{m} |a_{ij}|^q \Big)^{1/q}.$$

For the case $q = \infty$, the proof is similar.

It turns out, however, that the cases $p \neq 1$ are much more difficult to treat, and there no longer exist any analytical formulas (unless of course $p = q = 2$ and $p = q = \infty$). Theoretically, the norm $\|A\|_{\infty \mapsto 1}$ can be computed exactly, but the computation involves the solution of an NP-hard optimisation problem. For somewhat larger dimensions, this is not feasible in practice.

EXAMPLE 4.2.10. Let $1 \le p \le \infty$ and consider again the left and right shift operators L, $R\colon \ell^p \to \ell^p$,

$$L(x_1, x_2, \ldots) = (x_2, x_3, \ldots) \qquad \text{and} \qquad R(x_1, x_2, \ldots) = (0, x_1, x_2, \ldots).$$

We will show that $\|L\| = \|R\| = 1$. For simplicity, we only show this in the case $1 \le p < \infty$, though the proof for $p = \infty$ is essentially the same (but the definition of the p-norm looks different).

First we note that

$$\|Rx\|_p = \Big(\sum_{k=1}^{\infty} |(Rx)_k|^p\Big)^{1/p} = \Big(0 + \sum_{k=2}^{\infty} |x_{k-1}|^p\Big)^{1/p} = \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} = \|x\|_p.$$

This shows that

$$\frac{\|Rx\|_p}{\|x\|_p} = 1$$

for all $x \in \ell^p$, and thus also $\|R\| = 1$.

Concerning the left shift operator L, we have that

$$\|Lx\|_p = \Big(\sum_{k=1}^{\infty} |(Lx)_k|^p\Big)^{1/p} = \Big(\sum_{k=1}^{\infty} |x_{k+1}|^p\Big)^{1/p} = \Big(\sum_{k=2}^{\infty} |x_k|^p\Big)^{1/p} \le \|x\|_p,$$

and thus

$$\|L\| = \sup_{x \neq 0} \frac{\|Lx\|_p}{\|x\|_p} \le 1.$$

Moreover, if we consider for instance the sequence $x = (0, 1, 0, 0, \ldots)$, we see that $\|x\|_p = 1$ and $\|Lx\|_p = \|(1, 0, \ldots)\|_p = 1$ as well. Thus

$$\|L\| \ge \frac{\|(1, 0, \ldots)\|_p}{\|(0, 1, 0, \ldots)\|_p} = 1.$$

Together, we obtain again that $\|L\| = 1$.

EXAMPLE 4.2.11. Let $k\colon [0,1] \times [0,1] \to \mathbb{K}$ be continuous and define the linear mapping $T\colon C([0,1]) \to C([0,1])$,

$$Tf(x) = \int_0^1 k(x,y) f(y)\, dt.$$

We have previously seen that T is bounded with respect to $\|\cdot\|_\infty$ on both copies of $C([0,1])$, and we have also derived the estimate

$$\|Tf\|_\infty \le \Big(\max_{(x,y) \in [0,1]^2} |k(x,y)|\Big) \|f\|_\infty$$

for all $f \in C([0,1])$. This shows that

$$\|T\| \le \max_{(x,y) \in [0,1]^2} |k(x,y)|.$$

However, we cannot conclude that we have equality in this estimate, and, in fact, in general we have not.

In fact, we can obtain a better estimate of the norm of T as follows:

$$\begin{aligned} \|Tf\|_\infty &= \max_{x \in [0,1]} |Tf(x)| = \max_{x \in [0,1]} \Big| \int_0^1 k(x,y) f(y)\, dy \Big| \\ &\le \max_{x \in [0,1]} \Big(\Big(\int_0^1 |k(x,y)|\, dy \Big) \Big(\max_{y \in [0,1]} |f(y)| \Big) \Big) = \Big(\max_{x \in [0,1]} \int_0^1 |k(x,y)|\, dy \Big) \|f\|_\infty. \end{aligned} \tag{42}$$

We therefore see that

$$\|T\| \le \max_{x \in [0,1]} \int_0^1 |k(x,y)|\, dy. \tag{43}$$

In fact, we have equality in (43).

If $k(x, y) \geq 0$ for all $(x, y) \in [0, 1]^2$, this is easy to see: We simply can use the function $f(x) = 1$ in (42) and obtain an equality at all points. The argumentation is a bit more involved, though, if k changes sign.

3. Extension of Linear Mappings

Assume that U and V are *finite dimensional* vector spaces and that $T : U \to V$ is a linear mapping. Then T is uniquely determined by the action of T on a basis of U. That is, if $\{u_1, \ldots, u_n\}$ is a basis of U, then we know T as soon as we know the values of $Tu_1, \ldots, Tu_n \in V$: Because $\{u_1, \ldots, u_n\}$ is a basis, we can write each $u \in U$ uniquely in the form

$$u = \alpha_1 u_1 + \ldots + \alpha_n u_n.$$

The linearity of T then implies that

$$Tu = \alpha_1(Tu_1) + \ldots + \alpha_n(Tu_n).$$

Conversely, if we are given vectors $v_1, \ldots, v_n \in V$ and a basis $\{u_1, \ldots, u_n\}$, then there exists a unique linear mapping $T : U \to V$ with $Tu_i = v_i$ for all $1 \leq i \leq n$, and this mapping is given by

$$Tu = \sum_{i=1}^{n} \alpha_i v_i \qquad \text{if} \qquad u = \sum_{i=1}^{n} \alpha_i u_i.$$

Now a question is, whether this also works in infinite dimensional vector spaces. At first, the answer appears to be yes: First of all, we note that the dimension of V does not play any role in the consideration above. Now assume that U is infinite dimensional and that $\{u_i : i \in I\}$ is a basis of U (with some infinite index set I). Then we can again write each $u \in U$ uniquely in the form

$$u = \alpha_1 u_{i_1} + \ldots + \alpha_N u_{i_N}$$

for some $N \in \mathbb{N}$, indices $i_1, \ldots, i_N \in I$, and coefficients $\alpha_1, \ldots, \alpha_N \in \mathbb{K}$. Thus, if we know the values of $Tu_i \in V$ for all $i \in I$, then we have that

$$Tu = \alpha_1(Tu_{i_1}) + \ldots + \alpha_N(Tu_{i_N}).$$

There is, however, a pretty significant catch: All of this only works if we actually have a basis in the sense of linear algebra, that is, if we have vectors u_i, $i \in I$, such that we can write each element $u \in U$ as a *finite* linear combination of the vectors u_i. In most situations we encounter in practice, this is not the case. Consider for instance the case $U = \ell^2$ of all square summable sequences. Similar to the case of $\mathbb{K}^n$ we can define there "unit sequences"

$$e_i = (0, \ldots, 0, \underbrace{1}_{i\text{-th position}}, 0, \ldots) \tag{44}$$

for $i \in \mathbb{N}$. However, these sequences do not form a basis of ℓ^2 in the sense above, as it is impossible to write an infinite sequence as a finite linear combination of the unit sequences e_i.

Note, however, that these unit sequences actually *do* form a basis of c_{fin}, as we can write any *finite* sequence $x = (x_1, \ldots, x_N, 0, \ldots)$ as the *finite* linear combination

$$x = \sum_{i=1}^{N} x_i e_i.$$

Now assume that V is a vector space and that we are given vectors $v_i \in V$, $i \in \mathbb{N}$. Then we can define a linear mapping

$$T\colon c_{\mathrm{fin}} \to V,$$
$$(x_1, \ldots, x_N, 0, \ldots) \mapsto x_1 v_1 + \ldots + x_N v_N.$$

The question now becomes, whether it is possible to *extend* this mapping from c_{fin} to the whole space ℓ^2 in a meaningful way.

Definition 4.3.1 (Extension). Assume that X and Y are sets and that $S \subset X$ is a subset. Assume moreover that $T\colon S \to Y$ is some mapping. An *extension* of T to X is a mapping

$$\bar{T}\colon X \to Y$$

such that

$$\bar{T}(x) = T(x) \qquad \text{for all } x \in S.$$

The next result states that extensions of bounded linear operators exist.

Theorem 4.3.2. *Assume that $(U, \|\cdot\|_U)$ is a normed space and that $(V, \|\cdot\|_V)$ is a Banach space. Assume moreover that $M \subset U$ is a dense subspace and that $T\colon M \to V$ is a bounded linear operator. Then there exists a* unique bounded linear extension *$\bar{T}$ of T to U. In addition, we have that*

$$\|\bar{T}\|_{L(U,V)} = \|T\|_{L(M,V)}.$$

Proof. We start by constructing the mapping $\bar{T}$.

Assume that $u \in U$. Because $M \subset U$ is dense, there exists a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in M$ for all $n \in \mathbb{N}$ and $u = \lim_{n\to\infty} u_n$. Now the boundedness of T implies that

$$\|Tu_n - Tu_m\|_V = \|T(u_n - u_m)\|_V \le \|T\|_{L(M,V)} \|u_n - u_m\|_U.$$

Since $\{u_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence (as it is convergent), this implies that the sequence $\{Tu_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in V. Now the completeness of V implies that the limit $\lim_{n\to\infty} Tu_n$ exists in V. Define therefore

$$\bar{T}u := \lim_{n\to\infty} Tu_n.$$

We now show that the value $\bar{T}u \in V$ does not depend on the choice of the sequence $\{u_n\}_{n\in\mathbb{N}}$ converging to u. Indeed, assume that $\{v_n\}_{n\in\mathbb{N}}$ is another sequence with $u = \lim_{n\to\infty} v_n$ and $v_n \in M$ for all $n \in \mathbb{N}$. Then we have in particular that

$$\|u_n - v_n\|_U \le \|u_n - u\|_U + \|u - v_n\|_U \to_{n\to\infty} 0.$$

Thus it follows that

$$\begin{aligned} \|\bar{T}u - Tv_n\|_V &\le \|\bar{T}u - Tu_n\|_V + \|Tu_n - Tv_n\|_V \\ &\le \|\bar{T}u - Tu_n\|_V + \|T\| \|u_n - v_n\|_U \to_{n\to\infty} 0, \end{aligned}$$

that is $\bar{T}u = \lim_{n\to\infty} Tv_n$.

In particular, for $u \in M$ we can choose $u_n = u$ for all n, which implies that $\bar{T}u = Tu$ for all $u \in M$. Thus $\bar{T}$ is indeed an extension of T.

Next, we will show that $\bar{T}$ is in fact linear. For that, assume that $u, v \in U$, and let $\{u_n\}_{n\in\mathbb{N}}$ and $\{v_n\}_{n\in\mathbb{N}}$ be sequences in M converging to u and v, respectively. Then

$$\bar{T}u = \lim_{n\to\infty} Tu_n \qquad \text{and} \qquad \bar{T}v = \lim_{n\to\infty} Tv_n.$$

Now the continuity of addition in a normed space implies that

$$\lim_{n\to\infty}(u_n + v_n) = \lim_{n\to\infty} u_n + \lim_{n\to\infty} v_n = u + v$$

and thus, by the definition of $\bar{T}$ and the linearity of T,

$$\bar{T}(u+v) = \lim_{n\to\infty} T(u_n+v_n) = \lim_{n\to\infty} Tu_n + \lim_{n\to\infty} Tv_n = \bar{T}u + \bar{T}v.$$

Next, assume that $u \in U$ and $\lambda \in \mathbb{K}$. Let again $\{u_n\}_{n\in\mathbb{N}}$ be a sequence in M converging to u. Then the continuity of the scalar multiplication implies that

$$\lim_{n\to\infty} \lambda u_n = \lambda \lim_{n\to\infty} u_n = \lambda u$$

and thus

$$\bar{T}(\lambda u) = \lim_{n\to\infty} T(\lambda u_n) = \lim_{n\to\infty} \lambda T u_n = \lambda \lim_{n\to\infty} Tu_n = \lambda \bar{T} u.$$

This shows the linearity of T.

Next we will show that $\bar{T}$ is bounded and that $\|\bar{T}\|_{L(U,V)} = \|T\|_{L(M,V)}$. For that, assume that $u \in U$. Then there exists a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in M$ for all n and $u = \lim_{n\to\infty} u_n$. Then by construction $\bar{T}u = \lim_{n\to\infty} Tu_n$ and thus the continuity of the norm implies that

$$\|\bar{T}u\|_V = \lim_{n\to\infty} \|Tu_n\|_M \le \|T\|_{L(M,V)} \lim_{n\to\infty} \|u_n\|_U = \|T\|_{L(M,V)}\|u\|_U.$$

This shows that $\bar{T}$ is bounded linear and $\|\bar{T}\|_{L(U,V)} \le \|T\|_{L(M,V)}$. On the other hand, we have that

$$\|\bar{T}\|_{L(U,V)} = \sup_{\substack{u\in U,\\ \|u\|_U=1}} \|\bar{T}u\|_U \ge \sup_{\substack{u\in M,\\ \|u\|_U=1}} \|Tu\|_U = \|T\|_{L(M,V)},$$

which shows that, in fact, $\|\bar{T}\|_{L(U,V)} = \|T\|_{L(M,V)}$.

Finally we show the uniqueness of $\bar{T}$. Assume therefore that $S\colon U \to V$ is another bounded linear mapping such that $Su = Tu$ for all $u \in M$. Assume moreover that $u \in U$. The density of M in U again implies that there exists a sequence $\{u_n\}_{n\in\mathbb{N}}$ in M such that $\lim_{n\to\infty} u_n = u$. Since the mappings S and $\bar{T}$ are bounded linear, they are continuous. Thus we have that

$$Su = \lim_{n\to\infty} Su_n = \lim_{n\to\infty} Tu_n = \lim_{n\to\infty} \bar{T}u_n = \bar{T}u.$$

This proves the uniqueness of $\bar{T}$, which concludes that proof. □

Corollary 4.3.3. *Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are Banach spaces, $M \subset U$ is some subspace of U, and that $T\colon M \to V$ is a bounded linear operator. Denote by $\bar{M} \subset U$ the closure of M. Then there exists a unique bounded linear extension $\bar{T}\colon \bar{M} \to V$.*

Proof. The closure $\bar{M}$ of M by definition closed and thus complete, which implies that $\bar{M}$ is a Banach space. Moreover, M is by definition dense in its closure. Thus we can apply Theorem 4.3.2. □

Remark 4.3.4. The construction in the proof of Theorem 4.3.2 relies heavily on both the linearity of T and its boundedness. For unbounded linear operators, it is still possible to show that extensions exist, but they will in general not be unique. Moreover, the existence proof relies on the axiom of choice, which means that there is no possibility for actually constructing such an extension.

Example 4.3.5. Let $1 \le p < \infty$. Then c_{fin} is dense in ℓ^p and the unit sequences e_i, $i \in \mathbb{N}$, as defined in (44) form a basis of c_{fin}. Assume now that V is a Banach space and let v_i, $i \in \mathbb{N}$, be such that the sequence $(\|v_1\|_V, \|v_2\|, \|v_3\|, \ldots)$ is contained in ℓ^{p_*}, where p_* is the Hölder conjugate of p. Define now the mapping $T\colon c_{\text{fin}} \to V$ by

$$T(x_1, \ldots, x_N, 0, \ldots) = x_1 v_1 + \ldots + x_N v_N.$$

Then we have for all $x = (x_1, \dots, x_N, 0, \dots) \in c_{\text{fin}}$ that

$$\|Tx\|_V = \|x_1 v_1 + \dots + x_N v_N\|_V \le |x_1|\|v_1\|_V + \dots + |x_N|\|v_N\|_V.$$

Now we can use Hölder's inequality and obtain that

$$\begin{aligned} |x_1|\|v_1\|_V + \dots + |x_N|\|v_N\|_V &\le \Big(\sum_{n=1}^N |x_1|^p\Big)^{1/p} \Big(\sum_{n=1}^N \|v_n\|_V^{p_*}\Big)^{1/p_*} \\ &\le \|x\|_p \,\|(\|v_1\|_V, \|v_2\|_V, \dots)\|_{p_*}. \end{aligned}$$

Thus $T\colon c_{\text{fin}} \to V$ is bounded linear. Since $c_{\text{fin}} \subset \ell^p$ is dense, it follows that T can be extended to a bounded linear mapping $\bar{T}\colon \ell^p \to V$. Moreover, following the construction in the proof of that result, we obtain that

$$\bar{T}x = \sum_{n=1}^\infty x_n v_n := \lim_{N\to\infty} \sum_{n=1}^N x_n v_n$$

for all $x = (x_1, x_2, \dots) \in \ell^p$.

Example 4.3.6. Denote by $C_K(\mathbb{R})$ the set of continuous functions with *compact support*. That is, $f \in C_K(\mathbb{R})$, if f is continuous and there exists $R > 0$ such that $f(x) = 0$ for all $x \in \mathbb{R}$ with $|x| \ge R$. On $C_K(\mathbb{R})$ we can consider the norm $\|\cdot\|_1$ defined by

$$\|f\|_1 := \int_{-\infty}^\infty |f(x)|\,dx.$$

Note here that $\|f\|_1$ is finite for all $f \in C_K(\mathbb{R})$: If $f \in C_K(\mathbb{R})$, then there exists $R > 0$ such that $f(x) = 0$ for all $x \notin [-R, R]$. Thus

$$\|f\|_1 = \int_{-\infty}^\infty |f(x)|\,dx = \int_{-R}^R |f(x)|\,dx \le 2R \max_{x\in[-R,R]} |f(x)| < \infty,$$

as the function $|f|$ is continuous and thus admits its maximum on the bounded and closed interval $[-R, R]$.

Denote now by $C_b(\mathbb{R})$ the space of *bounded* continuous mappings on $\mathbb{R}$. If is possible to show that $C_b(\mathbb{R})$ is a Banach space with the norm

$$\|g\|_\infty := \sup_{x\in\mathbb{R}} |g(x)|.$$

Consider now the Fourier transform

$$\begin{aligned} \mathcal{F}\colon C_K(\mathbb{R}) &\to C_b(\mathbb{R}), \\ f &\mapsto \mathcal{F}f =: \hat{f}, \end{aligned}$$

where $\hat{f}\colon \mathbb{R} \to \mathbb{C}$ is defined as

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty f(x) e^{-i\omega x}\,dx.$$

Then we can estimate

$$\begin{aligned} \|\hat{f}\|_\infty = \sup_{\omega\in\mathbb{R}} |\hat{f}(\omega)| &\le \frac{1}{\sqrt{2\pi}} \sup_{\omega\in\mathbb{R}} \int_{-\infty}^\infty |f(x) e^{-i\omega x}|\,dx \\ &= \frac{1}{\sqrt{2\pi}} \sup_{\omega\in\mathbb{R}} \int_{-\infty}^\infty |f(x)|\,dx = \frac{1}{\sqrt{2\pi}} \|f\|_1. \end{aligned}$$

This shows that $\mathcal{F}\colon C_K(\mathbb{R}) \to C_b(\mathbb{R})$ is bounded linear with $\|\mathcal{F}\| \le \frac{1}{\sqrt{2\pi}}$. As a consequence, there exists a bounded linear extension of $\mathcal{F}$ to the completion of

$C_K(\mathbb{R})$, which can be shown to be the space $L^1(\mathbb{R})$ of *Lebesgue summable* functions.[1] That is, we have a natural definition of the Fourier transform of arbitrary summable functions. Moreover, the Fourier transform $\hat{f}$ of a summable function $f \in L^1(\mathbb{R})$ is necessarily a (bounded) continuous function.

4. Range and Kernel

Consider again the left and right shift operators L, $R\colon \ell^2 \to \ell^2$, that is

$$L(x_1, x_2, x_3, \ldots) = (x_2, x_3, x_4, \ldots),$$
$$R(x_1, x_2, x_3, \ldots) = (0, x_1, x_2, \ldots).$$

We have already seen that these are bounded linear mappings with $\|L\| = \|R\| = 1$. Moreover, it is easy to see that L is surjective but not injective, whereas R is injective but not surjective.

Now compare this to the finite dimensional case: In Theorem 1.5.6 we have shown that, for a linear mapping $T\colon V \to V$ on a *finite dimensional* vector space, injectivity, surjectivity, and bijectivity are all equivalent. As the example of the shift operator shows, this no longer holds in infinite dimensional spaces. Now we can ask the question, whether we have other possibilities for deciding whether a linear mapping on an infinite dimensional space is invertible. In this section we will give some partial answers to that question. For that, we have to discuss the kernel and the range of bounded linear mappings.

Recall here that, for a linear mapping $T\colon U \to V$ between vector spaces U and V, the kernel and range are defined as

$$\ker T = \{u \in U : Tu = 0\},$$
$$\operatorname{ran} T = \{v \in V : \text{ there exists } u \in U \text{ with } Tu = v\}.$$

LEMMA 4.4.1. *Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are normed spaces and that $T\colon U \to V$ is bounded linear. Then $\ker T$ is a closed linear subspace of U.*

PROOF. Since T is bounded linear, it is continuous. Moreover we can write $\ker T = T^{-1}(\{0_V\})$ as the pre-image of the closed set $\{0_V\} \subset V$ containing only the zero vector $0_V \in V$. Thus $\ker T$ is closed, being the pre-image of a closed set under a continuous mapping.[2] □

We have now seen that the kernel of a bounded linear mapping is always a closed linear space. Interestingly, and unfortunately, this is not the case for the range. That is, if $T\colon U \to V$ is bounded linear, then there may exist a sequence $\{v_n\}_{n\in\mathbb{N}} \subset V$ converging to some $v \in V$ such that $v_n \in \operatorname{ran} T$ for all $n \in \mathbb{N}$, whereas $v \notin \operatorname{ran} T$. Put differently, for every $n \in \mathbb{N}$, the equation $Tu = v_n$ has a solution, but the "limit equation" $Tu = v$ is no longer solvable.

EXAMPLE 4.4.2. Consider the mapping $T\colon C([0,1]) \to C([0,1])$,

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

Then $\operatorname{ran} T$ consists of all definite integrals of continuous function f on the unit interval $[0,1]$. The fundamental theorem of analysis now implies that $\operatorname{ran} T$ consists

[1] A concrete construction of this space relies on measure theory and is discussed in the course "Foundations of Analysis." Essentially, $L^1(\mathbb{R})$ consists of all functions f for which the integral $\int_{-\infty}^{\infty} |f(x)|\,dx$ can be reasonably defined and is finite.

[2] Alternatively, if $\{u_n\}_{n\in\mathbb{N}} \subset \ker T$ is a convergent sequence with $u = \lim_{n\to\infty} u_n$, then $Tu_n = 0$ for all $n \in \mathbb{N}$ (as $u_n \in \ker T$ for all n), and thus the continuity of T implies that $Tu = \lim_{n\to\infty} Tu_n = 0$ as well, which implies that also $u \in \ker T$.

of all continuously differentiable functions $g \in C^1([0,1])$ with $g(0) = 0$. Now consider both copies of $C([0,1])$ with the ∞-norm $\|f\|_\infty = \max_{x\in[0,1]}|f(x)|$. Then

$$\|Tf\|_\infty = \max_{x\in[0,1]}\Big|\int_0^x f(y)\,dy\Big| \le \int_0^1 |f(y)|\,dy \le \max_{x\in[0,1]}|f(x)| = \|f\|_\infty,$$

which shows that T is bounded. However, the range of T is not closed. Take for instance the sequence of functions

$$g_n(x) := \sqrt{x+\frac{1}{n}} - \frac{1}{\sqrt{n}}.$$

For all $n \in \mathbb{N}$ we have that g_n is continuously differentiable on the interval $[0,1]$ and $g_n(0) = 0$. Thus $g_n \in \operatorname{ran} T$ for all $n \in \mathbb{N}$. In addition, we have $\lim_{n\to\infty} g_n = g$ with

$$g(x) = \sqrt{x}.$$

Since the derivative $g'(0)$ does not exist, the function g is not contained in $\operatorname{ran} T$. That is, the equation

$$\int_0^x f(y)\,dy = \sqrt{y}$$

has no solution $f \in C([0,1])$. This shows that $\operatorname{ran} T$ is not a closed linear subspace of $C([0,1])$.

Theorem 4.4.3. *Assume that* $(U, \|\cdot\|_U)$ *and* $(V, \|\cdot\|_V)$ *are Banach spaces and that* $T\colon U \to V$ *is bounded linear. Assume that there exists* $C > 0$ *such that*

$$\|Tu\|_V \ge C\|u\|_U \tag{45}$$

for all $u \in U$. *Then* $\operatorname{ran} T \subset V$ *is closed and thus a Banach space. Moreover,* T *has a bounded linear inverse*

$$T^{-1}\colon \operatorname{ran} T \to U$$

and we can estimate

$$\|T^{-1}\|_{L(\operatorname{ran} T, U)} \le \frac{1}{C},$$

where $C > 0$ *is the constant from* (45).

Proof. We show first that T is injective. Assume to that end that $u \in \ker T$, that is, $Tu = 0$. Then

$$0 = \|Tu\|_V \ge C\|u\|_U,$$

which shows that $\|u\|_U = 0$ and thus $u = 0$. Thus $\ker T = \{0\}$, which shows the injectivity of T.

Next we show that $\operatorname{ran} T$ is closed. Assume to that end that $\{v_n\}_{n\in\mathbb{N}}$ is a sequence such that $v_n \in \operatorname{ran} T$ for all $n \in \mathbb{N}$ and that $v = \lim_{n\to\infty} v_n$ exists in V. We have to show that $v \in \operatorname{ran} T$. That is, we need to show that there exists $u \in U$ such that $Tu = v$.

Because $v_n \in \operatorname{ran} T$, there exist $u_n \in U$, $n \in \mathbb{N}$, such that $Tu_n = v_n$. By (45), we can now estimate

$$\|Tu_n - Tu_m\|_V = \|T(u_n - u_m)\|_V \ge C\|u_n - u_m\|_U$$

for all n, $m \in \mathbb{N}$, or

$$\|u_n - u_m\|_U \le \frac{1}{C}\|Tu_n - Tu_m\|_V = \frac{1}{C}\|v_n - v_m\|. \tag{46}$$

By assumption, the sequence $\{v_n\}_{n\in\mathbb{N}}$ converges, which implies that it is a Cauchy sequence. Now (46) implies that the sequence $\{u_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence as well. Because U is a Banach space, it is complete, and thus the Cauchy sequence $\{u_n\}_{n\in\mathbb{N}}$ converges to some $u \in U$. Now the boundedness of T implies that

$$Tu = T\Big(\lim_{n\to\infty} u_n\Big) = \lim_{n\to\infty} Tu_n = \lim_{n\to\infty} v_n = v.$$

Thus $v \in \operatorname{ran} T$, which implies that $\operatorname{ran} T$ is closed.

By definition, the mapping $T\colon U \to \operatorname{ran} T$ is surjective. Because is it is also injective, as we have already shown above, it is bijective, and thus the inverse $T^{-1}\colon \operatorname{ran} T \to U$ exists and is linear. It thus remains to show that T^{-1} is bounded linear with norm bounded by $1/C$. Let now $v \in \operatorname{ran} T$. Then we can write $v = T(T^{-1}v)$ and thus

$$\|v\|_V = \|T(T^{-1}v)\|_V \geq C\|T^{-1}v\|_U$$

by (45). Thus

$$\|T^{-1}v\|_U \leq \frac{1}{C}\|v\|_V$$

for all $v \in \operatorname{ran} T$, which shows that $T^{-1}\colon \operatorname{ran} T \to U$ is bounded and that we can estimate $\|T^{-1}\|_{L(\operatorname{ran} T, U)} \leq 1/C$. □

We now show that the inequality (45) is also a necessary condition for T to have a bounded inverse.

LEMMA 4.4.4. *Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are normed spaces and that $T\colon U \to V$ is bounded linear and bijective. Assume in addition that $T^{-1}\colon V \to U$ is bounded. Then there exists $C \geq 0$ such that*

$$\|Tu\|_V \geq C\|u\|_U \qquad \text{for all } u \in U. \tag{47}$$

PROOF. Without loss of generality assume that U, $V \neq \{0\}$. Then $T^{-1} \neq 0$ and thus $\|T^{-1}\|_{L(V,U)} > 0$. Now let $u \in U$ be arbitrary. Then $u = T^{-1}Tu$, and thus we can estimate

$$\|u\|_U = \|T^{-1}Tu\|_U \leq \|T^{-1}\|_{L(V,U)}\|Tu\|_V.$$

This shows that (47) holds with $C = 1/\|T^{-1}\|_{L(U,V)}$. □

REMARK 4.4.5. Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are Banach spaces and that $T\colon U \to V$ is bounded linear. Then Theorem 4.4.3 implies that T has a bounded linear inverse $T^{-1}\colon V \to U$ provided that $\operatorname{ran} T \subset V$ is dense and the estimate (45) holds. (The density of $\operatorname{ran} T$ implies that the closure of $\operatorname{ran} T$ is equal to V; because $\operatorname{ran} T$ already is closed by Theorem 4.4.3 it is equal to its closure and thus $\operatorname{ran} T = V$.)

Conversely, assume that $T\colon U \to V$ is bounded linear and that T is bijective. Then it is possible to show that (45) holds for some $C > 0$ and thus T^{-1} is actually a *bounded* linear mapping as well. This result, however, is far from trivial and goes beyond the scope of this class.[3]

5. Spaces of Bounded Linear Mappings

We have seen previously that the space $L(U, V)$ is a normed space whenever U and V are normed spaces. The next question would then be whether $L(U, V)$ is complete and thus a Banach space. As we show in the next result, this is actually the case provided that the space V is a Banach space itself. Note, though, that the space U need not be complete.

THEOREM 4.5.1. *Assume that $(U, \|\cdot\|_U)$ is a normed space and that $(V, \|\cdot\|_V)$ is a Banach space. Then $L(U, V)$ is a Banach space with the norm $\|\cdot\|_{L(U,V)}$.*

[3] I refer to the class TMA4230 – Functional Analysis for more details.

PROOF. We have to show that the normed space $L(U,V)$ is complete, that is, that every Cauchy sequence in $L(U,V)$ converges. Assume thus that $\{T_n\}_{n\in\mathbb{N}} \subset L(U,V)$ is a Cauchy sequence. We have to show that there exists $T \in L(U,V)$ such that $\|T_n - T\|_{L(U,V)} \to 0$ as $n \to \infty$.

Let $\varepsilon > 0$. Because $\{T_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists some $N \in \mathbb{N}$ such that

$$\|T_n - T_m\|_{L(U,V)} < \varepsilon \qquad \text{whenever } n,\, m \geq N. \tag{48}$$

Let now $u \in U$. Then

$$\|T_n u - T_m u\|_U \leq \|T_n - T_m\|_{L(U,V)} \|u\|_U < \varepsilon \|u\|_U \text{ whenever } n,\, m \geq N.$$

Thus for every fixed $u \in U$, the sequence $\{T_n u\}_{n\in\mathbb{N}} \subset V$ is a Cauchy sequence. Because V is complete, it follows that $\lim_{n\to\infty} Tu_n$ exists in V. Define therefore

$$Tu := \lim_{n\to\infty} Tu_n.$$

With this definition, we obtain a mapping $T\colon U \to V$. We still have to show that T is bounded linear, and that $\|T_n - T\|_{L(U,V)} \to 0$ as $n \to \infty$.

For the linearity of T we observe that

$$T(u+v) = \lim_{n\to\infty} T_n(u+v) = \lim_{n\to\infty} (T_n u + T_n v) = \lim_{n\to\infty} T_n u + \lim_{n\to\infty} T_n v = Tu + Tv$$

for all u, $v \in U$, and that

$$T(\lambda u) = \lim_{n\to\infty} T_n(\lambda u) = \lambda \lim_{n\to\infty} T_n u = \lambda T_n u$$

for all $u \in U$ and $\lambda \in \mathbb{K}$.

Since $\{T_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in a normed space, it is bounded. Thus there exists $C \geq 0$ such that $\|T_n\|_{L(U,V)} \leq C$ for all $n \in \mathbb{N}$. Consequently, we can estimate

$$\|Tu\|_U = \lim_{n\to\infty} \|T_n u\|_U \leq C\|u\|_U$$

for all $u \in U$. This shows that T is bounded linear.

It remains to show that $\|T_n - T\|_{L(U,V)} \to 0$ as $n \to \infty$. Let again $\varepsilon > 0$, and let $N \in \mathbb{N}$ be such that (48) holds. Then we have for all $u \in U$ and all $n \geq N$ that

$$\begin{aligned} \|(T_n - T)u\|_V = \|T_n u - Tu\|_V &= \lim_{m\to\infty} \|T_n u - T_m u\|_V \\ &\leq \lim_{m\to\infty} \|T_n - T_m\|_{L(U,V)} \|u\|_U \leq \varepsilon \|u\|_U. \end{aligned}$$

Thus

$$\|T_n - T\|_{L(U,V)} \leq \varepsilon \qquad \text{whenever } n \geq N,$$

which shows that $\|T_n - T\|_{L(U,V)} \to 0$ as $n \to \infty$. □

In particular, we may apply the previous result to the case $V = \mathbb{K}$.

DEFINITION 4.5.2 (functional). Let $(U, \|\cdot\|_U)$ be a normed space. A *bounded linear functional* on U is a bounded linear mapping $f\colon U \to \mathbb{K}$.

EXAMPLE 4.5.3. We consider some examples of bounded linear functionals:

- We consider the Banach space $(C([0,1]), \|\cdot\|_\infty)$. For $x \in [0,1]$ define $\delta_x\colon C([0,1]) \to \mathbb{K}$,

  $$\delta_x(f) := f(x).$$

  Then

  $$|\delta_x(f)| = |f(x)| \leq \sup_{y\in[0,1]} |f(y)| = \|f\|_\infty,$$

which shows that δ_x is a bounded linear functional. The mapping δ_x is called the *point evaluation* at x, or also the *Dirac δ-distribution.*[4]

- Assume that U is an inner product space and that $v \in U$ is fixed. Then the mapping $u \mapsto \langle u, v\rangle_U$ is bounded linear and thus a bounded linear functional. For a proof of this statement, I refer to the exercise sheet.
- Consider again the Banach space $(C([0,1]), \|\cdot\|_\infty)$, and let $g \in C([0,1])$ be a fixed function. Then the mapping $T_g : C([0,1]) \to \mathbb{K}$,
$$T_g(f) := \int_0^1 f(x)g(x)\,dx$$
is a bounded linear functional on $C([0,1])$, as
$$|T_g(f)| \le \int_0^1 |f(x)||g(x)|\,dx \le \|f\|_\infty \|g\|_\infty$$
for all $f \in C([0,1])$.

Definition 4.5.4 (dual space). Assume that U is a normed apce. The *dual space* of U is defined as
$$U^* := L(U, \mathbb{K}) = \bigl\{ f : U \to \mathbb{K} : f \text{ is bounded linear } \bigr\}.$$

Theorem 4.5.5. *Let $(U, \|\cdot\|_U)$ be a normed space. Then U^* is a Banach space with the norm*
$$\|f\|_{U^*} = \sup_{\substack{u \in U,\\ \|u\|_U \le 1}} |f(u)|.$$

Proof. This immediately follows from Theorem 4.5.1 using the completeness of $\mathbb{K}$. ☐

Example 4.5.6. We consider the case $U = \mathbb{K}^n$ with the norm $\|\cdot\|_p$ for some $1 \le p \le \infty$. Then every linear mapping $T : \mathbb{K}^n \to \mathbb{K}$ can be written as a matrix-vector produce $Tx = Ax$ for some matrix $A \in \mathbb{K}^{1\times n} \sim \mathbb{K}^n$. Moreover, one can see that every linear mapping $T : \mathbb{K}^n \to \mathbb{K}$ is actually bounded. Thus we can "identify" the dual space $(\mathbb{K}^n)^*$ with the space $\mathbb{K}^n$. The norm on the dual space $(\mathbb{K}^n)^*$, however, in general is different from the norm on $\mathbb{K}^n$. Indeed, since we are considering the p-norm $\|\cdot\|_p$ on $\mathbb{K}^n$, we have
$$\|A\|_{(K^n)^*} = \sup_{\|x\|_p = 1} |Ax| = \sup_{\|x\|_p=1} \Bigl|\sum_{i=1}^n a_i x_i\Bigr|.$$
By Hölder's inequality, we can now estimate
$$\Bigl|\sum_{i=1}^n a_i x_i\Bigr| \le \|A\|_{p_*}\|x\|_p,$$
and thus
$$\|A\|_{(K^n)^*} \le \|A\|_{p_*}.$$
We will show now that we have in fact an equality here, that is, that $\|A\|_{(K^n)^*} = \|A\|_{p_*}$. For this, we have to find for each $A \in \mathbb{K}^n$ some $x \in \mathbb{K}^n$ with $x \ne 0$ such that $|Ax| = \|A\|_{p_*}\|x\|_p$. For simplicity, we will restrict ourselves to the case $1 < p < \infty$.

For $A = 0$ this holds for all $x \in \mathbb{K}^n$. Let therefore $A \in \mathbb{K}^n \setminus \{0\}$. Define $x \in \mathbb{K}^n$ by setting
$$x_i := \begin{cases} \overline{a_i}|a_i|^{p_*-2} & \text{if } a_i \ne 0,\\ 0 & \text{if } a_i = 0.\end{cases}$$

[4]For more information on distributions and specifically their importance in the solution of partial differential equations, I refer to the class TMA4305 - Partial Differential Equations.

Then we have that

$$|x_i| = |a_i|^{p_*-1} \qquad \text{and} \qquad a_i x_i = |a_i|^{p_*}$$

for all i. Using the identities $p_* = p/(p-1)$ and $p_*/p = p_* - 1$ we then can compute

$$\|x\|_p = \Big(\sum_{i=1}^n |a_i|^{(p_*-1)p}\Big)^{1/p} = \Big(\sum_{i=1}^n |a_i|^{p_*}\Big)^{1/p} = \|A\|_{p_*}^{p_*/p} = \|A\|_{p_*}^{p_*-1}.$$

Thus

$$|Ax| = \Big|\sum_{i=1}^n a_i x_i\Big| = \sum_{i=1}^n |a_i|^{p_*} = \|A\|_{p_*}^{p_*} = \|A\|_{p_*}\|A\|_{p_*}^{p_*-1} = \|A\|_{p_*}\|x\|_p.$$

Together with the previous estimate, this shows that $\|A\|_{(\mathbb{K}^n)^*} = \|A\|_{p_*}$. Thus the dual space of $(\mathbb{K}^n, \|\cdot\|_p)$ can be identified with the space $(\mathbb{K}^n, \|\cdot\|_{p_*})$.

THEOREM 4.5.7. *Let $1 \le p < \infty$ and let $f \colon \ell^p \to \mathbb{K}$ be linear. The following are equivalent:*

(1) $f \in (\ell^p)^*$.

(2) There exists a sequence $y \in \ell^{p_*}$ such that

$$f(x) = \sum_{k=1}^\infty x_k y_k$$

for all $x \in \ell^p$.

In addition, the sequence $y \in \ell^{p_*}$ from (2) is unique and we have

$$\|f\|_{(\ell^p)^*} = \|y\|_{p_*}.$$

PROOF. (2) $\implies$ (1): By Hölder's inequality, we can estimate

$$|f(x)| = \Big|\sum_{k=1}^\infty x_k y_k\Big| \le \|x\|_p \|y\|_{p_*},$$

which shows that f is bounded linear and that

$$\|f\|_{(\ell^p)^*} \le \|y\|_{p_*}. \tag{49}$$

(1) $\implies$ (2): Define the sequence $y = (y_1, y_2, \ldots)$ by

$$y_k = f(e_k)$$

with $e_k \in \ell^p$ being the k-th unit sequence. Then we have for all finite sequences $x = (x_1, \ldots, x_N, 0, \ldots) \in c_{\text{fin}}$ that

$$f(x) = \sum_{k=1}^N x_k f(e_k) = \sum_{k=1}^N x_k y_k.$$

Moreover, we can estimate

$$\Big|\sum_{k=1}^N x_k y_k\Big| = |f(x)| \le \|f\|_{(\ell^p)^*}\|x\|_p.$$

We will show now that $y \in \ell^{p_*}$. Again, we assume for simplicity that $1 < p < \infty$. As in Example 4.5.6 we define a sequence $x = (x_1, x_2, \ldots)$ setting

$$x_k = \operatorname{sgn}(y_k)|y_k|^{p_*-1}.$$

Then $x^{(N)} := (x_1, \ldots, x_N, 0, \ldots) \in c_{\mathrm{fin}} \subset \ell^p$ for all $N \in \mathbb{N}$. Define now $y^{(N)} := (y_1, \ldots, y_N, 0, \ldots)$. Similarly as in Example 4.5.6 we have that

$$\|x^{(N)}\|_p = \Big(\sum_{k=1}^{N} |y_k|^{p_*}\Big)^{1/p} = \|y^{(N)}\|_{p_*}^{p_*-1}.$$

Thus

$$\begin{aligned}\|y^{(N)}\|_{p_*}^{p_*} = \sum_{k=1}^{N} |y_k|^{p_*} &= \Big|\sum_{k=1}^{N} x_k y_k\Big| \\ &= |f(x^{(N)})| \le \|f\|_{(\ell^p)^*} \|x^{(N)}\|_p = \|f\|_{(\ell^p)^*} \|y^{(N)}\|_{p_*}^{p_*-1}.\end{aligned}$$

This shows that

$$\|y^{(N)}\|_{p_*} \le \|f\|_{(\ell^p)^*}$$

for all $N \in \mathbb{N}$. Taking the limit $N \to \infty$, this implies that

$$\|y\|_{p_*} \le \|f\|_{(\ell^p)^*} \tag{50}$$

and thus $y \in \ell^{p_*}$.

Since $c_{\mathrm{fin}} \subset \ell^p$ is dense, it follows from Theorem 4.3.2 that the mapping

$$x \mapsto \sum_{k=1}^{\infty} x_k y_k$$

is a well-defined bounded linear mapping on ℓ^p. Moreover, by construction it coincides with f on the dense linear subspace c_{fin} of ℓ^p. From the uniqueness in Theorem 4.3.2 we can thus conclude that

$$f(x) = \sum_{k=1}^{\infty} x_k y_k \qquad \text{for all } x \in \ell^p.$$

Assume now that $z \in \ell^{p_*}$ is another sequence such that

$$f(x) = \sum_{k=1}^{\infty} x_k z_k$$

for all $x \in \ell^p$. Then we have in particular that

$$y_k = f(e_k) = z_k.$$

This proves the uniqueness of the sequence $y \in \ell^{p_*}$.

Finally, the equality $\|f\|_{(\ell^p)^*} = \|y\|_{p_*}$ follows from the two estimates (49) and (50). □

Similar as in the finite dimensional case, this implies that we can "identify" for $1 \le p < \infty$ the dual space $(\ell^p)^*$ with the space ℓ^{p_*}.

Remark 4.5.8. For $p = \infty$ the statement of the previous theorem no longer holds. It is still straightforward to show that every sequence $y \in \ell^1$ defines a bounded linear mapping $f \in (\ell^\infty)^*$ as

$$f(x) := \sum_{k=1}^{\infty} x_k y_k.$$

Thus we can say that $\ell^1 \subset \ell^\infty$. However, there exist in addition bounded linear functionals on ℓ^∞ that *cannot* be written in this form. Even more, there exist non-zero bounded linear functionals $f \in (\ell^\infty)^*$, $f \neq 0$, such that $f(e_k) = 0$ for all $k \in \mathbb{N}$. This is due to fact that the space of finite sequences c_{fin} is not dense in ℓ^∞, and thus we cannot make use of Theorem 4.3.2.

CHAPTER 5

Hilbert Spaces

1. Best approximation and projections

Previously we have introduced the notion of the orthogonal projection onto a linear subspace in the context of finite dimensional inner product spaces. In the following, we will discuss to which extent this can be generalised to infinite dimensional Hilbert spaces. More general, we will discuss the idea of a projection onto closed convex subsets of Hilbert spaces.

Definition 5.1.1 (convex set). Assume that U is a vector space and that $K \subset U$ is a subset. We say that K is *convex* if for all u, $v \in K$ and all $0 < \lambda < 1$ we have that $\lambda u + (1-\lambda)v \in K$.

Put differently, a set K is convex if for all vectors u, $v \in K$ the whole line segment connecting u and v is contained in K.

Example 5.1.2. Assume $K \subset U$ is a linear subspaces. Then K is convex. Indeed, assume that u, $v \in K$. Then $\lambda u + \mu v \in K$ for all λ, $\mu \in \mathbb{K}$. With $0 < \lambda < 1$ and $\mu = 1 - \lambda$ we obtain the convexity of K.

Example 5.1.3. Assume $(U, \|\cdot\|_U)$ is a normed space and let $R > 0$. Then the ball

$$B_R(0) = \{u \in U : \|u\|_U < R\}$$

is convex. Indeed, assume that u, $v \in B_R(0)$, that is, $\|u\|_U < R$ and $\|v\|_U < R$, and let $0 < \lambda < 1$. Then

$$\|\lambda u + (1-\lambda)v\|_U \le |\lambda|\|u\|_U + |1-\lambda|\|v\|_U = \lambda\|u\|_U + (1-\lambda)\|v\|_U < \lambda R + (1-\lambda)R = R,$$

which shows that $\lambda u + (1-\lambda)v \in B_R(0)$.

Definition 5.1.4 (distance to a set). Assume that $(U, \|\cdot\|_U)$ is a normed space and that $K \subset U$ is a non-empty subset. For $u \in U$ we denote by

$$\operatorname{dist}(u, K) := \inf_{v \in K} \|u - v\|_U$$

the distance between u and K.

Assume now that $K \subset U$ is some non-empty subset and that $u \notin K$. One interesting question is whether there exists a closest point to u in K. Put differently, can we find some point $z \in K$ such that $\|u - z\|_U = \operatorname{dist}(u, K)$? If yes, is that point unique? The next result gives a positive answer to both of these questions in the case where U is a Hilbert space.

Theorem 5.1.5 (approximation theorem). *Assume that U is a Hilbert space and that $K \subset U$ is a non-empty closed and convex subset. There there exists for each $u \in U$ a unique point $z = \pi_K(u) \in K$ called the* projection of u onto K *such that*

$$\|u - z\|_U = \operatorname{dist}(u, K).$$

PROOF. By definition of the distance, there exists for each $n \in \mathbb{N}$ some $z_n \in K$ such that

$$\|u - z_n\|_K^2 \leq \operatorname{dist}(u, K)^2 + \frac{1}{n}.$$

Now let n, $m \in \mathbb{N}$. Then we can apply the parallelogram law (see Lemma 3.5.8) to $z_n - z_m$ and $2u - z_n - z_m$ and obtain that

$$\begin{aligned}
2\|z_n - z_m\|_U^2 &+ 2\|2u - z_n - z_m\|_U^2 \\
&= \|z_n - z_m + 2u - z_n - z_m\|_U^2 + \|z_n - z_m - (2u - z_n - z_m)\|_U^2 \\
&\qquad\qquad\qquad\qquad = 4\|u - z_m\|_U^2 + 4\|u - z_n\|_U^2.
\end{aligned}$$

Now the convexity of K and the assumption z_n, $z_m \in K$ imply that $(z_n + z_m)/2 \in K$ and thus

$$\|2u - z_n - z_m\|_U^2 = 4\left\|u - \frac{z_n + z_m}{2}\right\|_U^2 \geq 4 \operatorname{dist}(u, K)^2.$$

Thus we can estimate

$$\begin{aligned}
\|z_n - z_m\|_U^2 &\leq 2\|u - z_m\|_U^2 + 2\|u - z_n\|_U^2 - 4 \operatorname{dist}(u, K)^2 \\
&\leq 2 \operatorname{dist}(u, K)^2 + \frac{2}{m} + 2 \operatorname{dist}(u, K)^2 + \frac{2}{n} - 4 \operatorname{dist}(u, K)^2 = \frac{2}{m} + \frac{2}{n}.
\end{aligned}$$

This, however, shows that the sequence $\{z_n\}_{n \in \mathbb{N}}$ is a Cauchy sequence.

Since U is a Hilbert space and thus complete, it follows that the sequence $\{z_n\}_{n \in \mathbb{N}}$ converges to some $z \in U$. Now the closedness of K and the assumption $z_n \in K$ for all $n \in \mathbb{N}$ imply that $z \in K$. As a consequence, we have that

$$\|u - z\|_U^2 \geq \operatorname{dist}(u, K)^2.$$

On the other hand, the continuity of the norm implies that

$$\|u - z\|_U^2 = \lim_{n \to \infty} \|u - z_n\|_U^2 \leq \operatorname{dist}(u, K).$$

Thus $z \in K$ satisfies

$$\|u - z\|_U = \operatorname{dist}(u, K).$$

We still have to show that z is unique. Assume therefore that y, $z \in K$ satisfy

$$\|u - y\|_U = \operatorname{dist}(u, K) = \|u - z\|_U.$$

The convexity of K implies again that $(y - z)/2 \in K$. With the same argumentation as above, we thus obtain that

$$\|y - z\|_U^2 \leq 2\|u - y\|_U^2 + 2\|u - z\|_U^2 - 4 \operatorname{dist}(u, K) = 0.$$

This, however, implies that $y = z$, which proves the uniqueness of z. ☐

THEOREM 5.1.6 (characterisation of the projection). *Assume that U is a Hilbert space, that $K \subset U$ is a non-empty closed and convex subset, and that $u \in U$. Then $z = \pi_K(u)$ is the projection of u onto K, if and only if $z \in K$ and the* variational inequality

$$\Re(\langle u - z, v - z \rangle) \leq 0 \qquad \textit{for all } v \in K \tag{51}$$

holds.

PROOF. Assume first that $z = \pi_K(u)$ is the projection of u onto K. Then we have for all $v \in K$ that $\|u - z\|_U^2 \leq \|u - v\|_U^2$. Even more, because of the convexity of K we have that

$$\|u - z\|_U^2 \leq \|u - \lambda v - (1 - \lambda) z\|_U^2 \text{ for all } v \in K \text{ and } 0 \leq \lambda \leq 1. \tag{52}$$

Now we can write

$$\|u - z\|_U^2 = \|u\|_U^2 - 2\Re(\langle u, z\rangle) + \|z\|_U^2$$

and

$$\begin{aligned}\|u - \lambda v - (1-\lambda)z\|_U^2 = \|u\|_U^2 &+ \lambda^2\|v\|_U^2 + (1-\lambda)^2\|z\|_U^2 \\ &- 2\lambda\Re(\langle u, v\rangle) - 2(1-\lambda)\Re(\langle u, z\rangle) + 2\lambda(1-\lambda)\Re(\langle v, z\rangle).\end{aligned}$$

Define therefore

$$f_v(\lambda) := \lambda^2\|v\|_U^2 - \lambda(2-\lambda)\|z\|_U^2 - 2\lambda\Re(\langle u, v\rangle) + 2\lambda\Re(\langle u, z\rangle) + 2\lambda(1-\lambda)\Re(\langle v, z\rangle).$$

Thus (52) can be reformulated as

$$f_v(\lambda) \ge 0 \qquad \text{for all } v \in K \text{ and } 0 \le \lambda \le 1. \tag{53}$$

Now note that $f_v(\lambda)\colon \mathbb{R} \to \mathbb{R}$ is a quadratic function with non-negative leading coefficient and that $f_v(0) = 0$. Thus (53) holds if and only if $f_v'(0) \ge 0$ for all $v \in K$. Now we compute

$$f_v'(\lambda) = 2\lambda\|v\|_U^2 - 2(1-\lambda)\|z\|_U^2 - 2\Re(\langle u, v\rangle) + 2\Re(\langle u, z\rangle) + 2(1-2\lambda)\Re(\langle v, z\rangle),$$

and thus

$$f_v'(0) = -2\|z\|_U^2 - 2\Re(\langle u, v\rangle) + 2\Re(\langle u, z\rangle) + 2\Re(\langle v, z\rangle) = -2\Re(\langle z - u, z - v\rangle).$$

Thus (53) and therefore (52) is equivalent to the variational inequality

$$-f_v'(0) = \Re(\langle z - u, z - v\rangle) \le 0 \qquad \text{for all } v \in K.$$

This shows that the assumption $z = \pi_K(u)$ implies (51).

Conversely, assume that $z \in K$ and that (51) holds. As seen above, this is equivalent to z satisfying (52). Together with the assumption $z \in K$, this implies that z solves the problem $\min_{v\in K}\|u - v\|_U$. Because of the uniqueness of the projection, this shows that $z = \pi_K(u)$. □

Corollary 5.1.7. *Assume that U is a Hilbert space and that $W \subset U$ is a closed subspace. Then there exists a unique point $z = \pi_W(u)$ such that $z \in W$ and*

$$\|z - u\| = \operatorname{dist}(u, W) = \inf_{w\in W}\|u - w\|.$$

Moreover, z is the unique point satisfying

$$z \in W \qquad \text{and} \qquad u - z \in W^\perp,$$

where

$$W^\perp = \{v \in U : \langle v, w\rangle = 0 \text{ for all } w \in W\}$$

is the orthogonal complement of W in U.

Proof. The existence and uniqueness follows from Theorem 5.1.5. Moreover, Theorem 5.1.6 implies that $z = \pi_W(u)$ is the unique point satisfying $z \in W$ and

$$\Re(\langle u - z, v - z\rangle) \le 0 \qquad \text{for all } v \in W. \tag{54}$$

We thus have to show that (54) holds if and only if $u - z \in W^\perp$.

Assume therefore that (54) holds and let $w \in W$. Since W is a subspace of U and $z \in W$ it follows that $\lambda w + z \in W$ for all $\lambda \in \mathbb{K}$. Thus (54) implies that

$$0 \ge \Re(\langle u - z, \lambda w + z - z\rangle) = \Re(\langle u - z, \lambda w\rangle) \qquad \text{for all } \lambda \in \mathbb{K}.$$

Now

$$\Re(\langle u - z, \lambda w\rangle) = \Re(\bar\lambda\langle u - z, w\rangle) = \Re(\lambda)\Re(\langle u - z, w\rangle) + \Im(\lambda)\Im(\langle u - z, w\rangle).$$

This can only be non-positive for all $\lambda \in \mathbb{K}$ if $\langle u - z, w\rangle = 0$ and thus $u - z \in W^\perp$.

Conversely, assume that $u - z \in W^\perp$, that is $\langle u - z, w\rangle = 0$ for all $w \in W$, and let $v \in W$. Then $v-z \in W$ because W is a subspace of U, and thus $\langle u-z, v-z\rangle = 0$. In particular (54) holds. □

Remark 5.1.8. We have stated Theorem 5.1.6 and Corollary 5.1.7 in the setting of closed convex subsets (subspaces) of Hilbert spaces. However, the characteristions for the projections that we have derived there remain valid for arbitrary convex subsets (subspaces) of inner product spaces. That is, if U is an inner product space and $K \subset U$ is convex and non-empty, then $z = \pi_K(u)$ solves the optimisation problem

$$\min_{v\in K} \|v - u\|,$$

if and only if

$$z \in K \qquad \text{and} \qquad \Re(\langle u - z, v - z\rangle) \le 0 \text{ for all } v \in K.$$

Moreover, if $W \subset U$ is a subspaces, then $z = \pi_W(u)$ solves the problem

$$\min_{w\in W} \|u - w\|$$

if and only if

$$z \in W \qquad \text{and} \qquad u - z \in W^\perp.$$

Note, however, that such a vector need not exist.

In the following example we make use of the characterisation of the projection by means of a variational inequality in order to compute projections on balls in Hilbert spaces.

Example 5.1.9. Let U be a Hilbert space, let $R > 0$ and let $K = \bar{B}_R(0) \subset U$ be the closed unit ball of radius R in U. Then K is a closed and convex set in a Hilbert space, and thus the projection $\pi_K(u)$ exists for every $u \in U$. We will now show that we can explicitly compute this projection as

$$\pi_K(u) = \left\{\begin{matrix} u & \text{if } \|u\| \le R \\ R\dfrac{u}{\|u\|} & \text{if } \|u\| > R \end{matrix}\right\} = \frac{Ru}{\max\{\|u\|, R\}}. \tag{55}$$

First assume that $u \in K$, that is, $\|u\| \le R$. Then obviously $\pi_K(u) = u$. Thus (55) holds in this case.

Now assume that $u \notin K$, that is, $\|u\| > R$, and denote

$$z = R\frac{u}{\|u\|}.$$

Then $\|z\| = R$ and thus $z \in K$. In view of Theorem 5.1.6 we thus have to show that

$$\Re(\langle u - z, v - z\rangle) \le 0 \qquad \text{for all } v \in K. \tag{56}$$

Assume therefore that $v \in K$. We can compute

$$\begin{aligned}\langle u - z, v - z\rangle &= \langle u, v\rangle - \langle z, v\rangle - \langle u, z\rangle + \langle z, z\rangle \\ &= \langle u, v\rangle - \frac{R}{\|u\|}\langle u, v\rangle - R\|u\| + R^2 = \big(\|u\| - R\big)\Big(\frac{\langle u, v\rangle}{\|u\|} - R\Big).\end{aligned}$$

By assumption, $\|u\| > R$. Moreover, the Cauchy–Schwarz–Bunyakovsky inequality implies that

$$\Re(\langle u, v\rangle) \le |\langle u, v\rangle| \le \|u\|\|v\|.$$

Since $\|v\| \le R$ it follows that

$$\Re(\langle u - z, v - z\rangle) = \big(\|u\| - R\big)\Big(\frac{\Re(\langle u, v\rangle)}{\|u\|} - R\Big) \le \big(\|u\| - R\big)\big(\|v\| - R\big) \le 0.$$

Thus (56) holds, which in turn shows that $\pi_K(u) = z = Ru/\|u\|$.

Remark 5.1.10. The existence and uniqueness proof for the projection onto a convex set relied strongly on the fact that U is a Hilbert space: During the proof we relied twice on the parallelogram law, both for proving the existence and for proving the uniqueness. Indeed, in general Banach spaces neither existence nor uniqueness need to hold.

Consider first the case of $U = \mathbb{R}^2$ with the ∞-norm. In this case, the projection need not be unique. That is, given a closed and convex set $K \subset U$ and a point $u \notin K$ there might be different points $z \in K$ such that $\|u - z\| = \operatorname{dist}(u, K)$. Take, for instance, $K = \mathbb{R}e_1$ the x-axis, and let $u = (0, 1)$. Then $\operatorname{dist}(u, K) = 1$, but every point $(a, 0) \in K$ with $|a| \le 1$ satisfies $\|(a, 0) - u\|_\infty = 1 = \operatorname{dist}(u, K)$. That is, every point $(a, 0) \in K$ with $|a| \le 1$ can be considered a projection of $(0, 1)$ onto K. In particular, the projection is not unique.

Now we consider the case $U = \ell^1$ is the space of all real-valued summable sequences, and we define

$$K = \Big\{x \in \ell^1 : \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)x_n \ge 1\Big\}.$$

Then one can show that K is a non-empty, closed and convex set. However, the projection of 0 onto K does not exist. That is, there exists no $x \in K$ such that $\|x - 0\|_1 = \operatorname{dist}(0, K)$. In order to see this, we note first that for all $x \in \ell^1$ we have

$$1 \le \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)x_n \le \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)|x_n| < \sum_{n=1}^{\infty}|x_n| = \|x\|_1.$$

That is, $\|x\|_1 > 1$ for all $x \in \ell^1$, which also implies that $\operatorname{dist}(0, K) \ge 1$. On the other hand, for all $n \ge 2$ we have that

$$\frac{n}{n-1}e_n \in K,$$

which shows that

$$\operatorname{dist}(0, K) \le \Big\|\frac{n}{n-1}e_n\Big\|_1 = \frac{n}{n-1}$$

for all $n \ge 2$. Taking the infimum over all $n \ge 2$, it follows that $\operatorname{dist}(0, K) \le 1$. Together, these estimates show that $\operatorname{dist}(0, K) = 1$. However, as shown above, $\|x\|_1 > 1$ for all $x \in K$. Thus there exists no $x \in K$ such that $\|x\|_1 = \operatorname{dist}(0, K)$.

Proposition 5.1.11. *Assume that U is a Hilbert space and that $K \subset U$ is non-empty, closed and convex. Then we have for all u, $v \in U$ that*

$$\|\pi_K(u) - \pi_K(v)\| \le \|u - v\|.$$

Proof. If $\pi_K(u) = \pi_K(v)$, the claim is trivial, as the left hand side is equal to zero. Assume therefore that $\pi_K(u) \ne \pi_K(v)$. We can write

$$\begin{aligned}\|\pi_K(u) - \pi_K(v)\|^2 &= \langle \pi_K(u) - \pi_K(v), \pi_K(u) - u + u - v + v - \pi_K(v)\rangle \\ &= \langle \pi_K(v) - \pi_K(u), u - \pi_K(u)\rangle + \langle \pi_K(u) - \pi_K(v), u - v\rangle \\ &\qquad + \langle \pi_K(u) - \pi_K(v), v - \pi_K(v)\rangle.\end{aligned}$$

Since $\pi_K(v) \in K$ and $\pi_K(u) \in K$, we can now use the variational inequality (51), which shows that

$$\begin{aligned}\Re(\langle \pi_K(v) - \pi_K(u), u - \pi_K(u)\rangle) &\le 0, \\ \Re(\langle \pi_K(u) - \pi_K(v), v - \pi_K(v)\rangle) &\le 0.\end{aligned}$$

Moreover, using the Cauchy–Schwarz–Bunyakovsky inequality, we can estimate

$$\begin{aligned}\Re(\langle \pi_K(v) - \pi_K(u), u - v\rangle) &\le |\langle \pi_K(v) - \pi_K(u), u - v\rangle| \\ &\le \|\pi_K(v) - \pi_K(u)\|\|u - v\|.\end{aligned}$$

Thus

$$\begin{aligned}\|\pi_K(u) - \pi_K(v)\|^2 &= \Re\big(\|\pi_K(u) - \pi_K(v)\|\big)^2 \\ &= \Re\big(\langle \pi_K(v) - \pi_K(u), u - \pi_K(u)\rangle \\ &\qquad + \langle \pi_K(u) - \pi_K(v), u - v\rangle \\ &\qquad + \langle \pi_K(u) - \pi_K(v), v - \pi_K(v)\rangle\big) \\ &\le \Re(\langle \pi_K(v) - \pi_K(u), v - u\rangle) \\ &\le \|\pi_K(v) - \pi_K(u)\|\|v - u\|.\end{aligned}$$

Dividing by $\|\pi_K(u) - \pi_K(v)\|$, we obtain the claimed estimate. □

2. Projections on subspaces

We now study in more details the properties of projections onto closed subspaces. For this we will exploit the characterisation of the projection derived in Corollary 5.1.7.

The results to be derived will be a generalisations of our discussion in Section 3 of the finite dimensional case.

PROPOSITION 5.2.1. *Assume that U is a Hilbert space and that $W \subset U$ is a closed subspace. Then the projection $\pi_W : U \to U$ is a bounded linear mapping. Moreover,* $\ker \pi_W = W^\perp$ *and* $\operatorname{ran} \pi_W = W$.

PROOF. *Linearity:* Assume that $u, v \in U$. Then $\pi_W(u) \in W$ and $\pi_W(v) \in W$, which implies that also $\pi_W(u) + \pi_W(v) \in W$. Moreover, $u - \pi_W(u) \in W^\perp$ and $v - \pi_W(v) \in W^\perp$, and thus $(u + v) - (\pi_W(u) + \pi_W(v)) \in W^\perp$. This shows that $z = \pi_W(u) + \pi_W(v)$ satisfies the conditions of Corollary 5.1.7 and thus $\pi_W(u) + \pi_W(v) = \pi_W(u + v)$.

Similarly, if $u \in U$ and $\lambda \in \mathbb{K}$, then $\lambda\pi_W(u) \in W$ and

$$\lambda u - \lambda\pi_W(u) = \lambda(u - \pi_W(u)) \in W^\perp,$$

which shows that $\lambda\pi_W(u) = \pi_W(\lambda u)$.

Boundedness: This is a direct consequence of Proposition 5.1.11.

Kernel and range: We have that $\pi_W(u) = 0$ if and only if $0 \in W$ and $u - 0 \in W^\perp$. This is the case, if and only if $u \in W^\perp$, which shows that $\ker \pi_W = W^\perp$. Next, $\pi_W(u) \in W$ for all $u \in U$, which shows that $\operatorname{ran} \pi_W \subset W$. Moreover, $\pi_W(u) = u$ for all $u \in W$, which proves that, actually, $\operatorname{ran} \pi_W = W$. □

LEMMA 5.2.2. *Let U be a Hilbert space and let $W \subset U$ be a closed subspace. Then*

$$U = W \oplus W^\perp.$$

PROOF. We can write every $u \in U$ as $u = \pi_W u + (u - \pi_W u)$ with $\pi_W u \in W$ and $u - \pi_W u \in W^\perp$. Thus $U = W + W^\perp$. Now assume that $u \in W \cap W^\perp$. Since $u \in W$, we have $u = \pi_W u$. Thus

$$\|u\|^2 = \langle u, u\rangle = \langle u, \pi_W u\rangle = 0,$$

the last equality following from the fact that $u \in W^\perp$ and $\pi_W u \in W$. This shows that $W \cap W^\perp = \{0\}$. Thus the sum of W and $W^\perp$ is direct. □

For the next results, we need some topological information about orthogonal complements.

LEMMA 5.2.3. *Let U be an inner product space and let $v \in U$ be fixed. Then the mapping $f\colon U \to \mathbb{K}$, $f(u) := \langle u, v\rangle$ is bounded linear.*

PROOF. The linearity of f is simply the linearity of the inner product in the first component. Moreover, the boundedness follows from the estimate
$$|f(u)| = |\langle u, v\rangle| \le \|u\|\|v\|.$$
□

LEMMA 5.2.4. *Assume that U is an inner product space and that $S \subset U$ is non-empty. Then $S^\perp \subset U$ is a closed subspace.*

PROOF. We have already shown previously that $S^\perp$ is a subspace. Thus we only have to show that $S^\perp$ is closed. For that assume that $\{u_n\}_{n\in\mathbb{N}} \subset S^\perp$ is a sequence such that $u := \lim_{n\to\infty} u_n$ exists in U. Since $u_n \in S^\perp$ for all $n \in \mathbb{N}$, we have that $\langle u_n, v\rangle = 0$ for all $v \in S$. Because of the boundedness of the functional $w \mapsto \langle u_n, v\rangle$ it follows that
$$\langle u, v\rangle = \lim_{n\to\infty} \langle u_n, v\rangle = 0$$
for all $v \in S$, which proves that $u \in S^\perp$. □

LEMMA 5.2.5. *Let U be a Hilbert space and let $W \subset U$ be a closed subspace. Then*
$$(W^\perp)^\perp = W.$$

PROOF. Assume first that $w \in W$. Then $\langle w, z\rangle = 0$ for all $z \in W^\perp$. This, however, implies that $w \in (W^\perp)^\perp$. Thus $W \subset (W^\perp)^\perp$.

Now assume that $u \in (W^\perp)^\perp$. Since W is a closed subspace, we can write $u = w + z$ with $w = \pi_W u \in W$ and $z = u - \pi_W u \in W^\perp$. By our previous considerations, we obtain that $w \in W \subset (W^\perp)^\perp$. Since $(W^\perp)^\perp$ is a subspace of U, this implies that $z = u - w \in (W^\perp)^\perp$. Now, if we apply Lemma 5.2.2 to the closed subspace $W^\perp$, we obtain that
$$U = W^\perp \oplus (W^\perp)^\perp.$$
In particular, $W^\perp \cap (W^\perp)^\perp = \{0\}$. Thus
$$z \in W^\perp \cap (W^\perp)^\perp = \{0\},$$
which implies that $z = 0$ and consequently $u = w \in W$. Thus $(W^\perp)^\perp \subset W$. □

Next we will consider the question as to what happens, if the subspace W is not closed.

PROPOSITION 5.2.6. *Assume that U is a Hilbert space and that W is a (not necessarily closed) subspace. Then*
$$(W^\perp)^\perp = \overline{W}.$$

PROOF. Since $W \subset \overline{W}$, it follows that $\overline{W} \perp\subset W^\perp$, and thus, as $\overline{W}$ by definition is closed,
$$(W^\perp)^\perp \subset (\overline{W}^\perp)^\perp = \overline{W}.$$
Conversely, we obtain as in the proof of the previous Lemma that $W \subset (W^\perp)^\perp$. Since $(W^\perp)^\perp$ is closed and $\overline{W}$ is the smallest closed subspace of U containing W, this implies that actually $\overline{W} \subset (W^\perp)^\perp$. □

COROLLARY 5.2.7. *Assume that U is a Hilbert space and W is a subspace. Then*
$$U = \overline{W} \oplus W^\perp.$$

Proof. Since $W^\perp$ is close, we can write

$$U = W^\perp \oplus (W^\perp)^\perp = W^\perp \oplus \overline{W}.$$

□

Corollary 5.2.8. *Assume that U is a Hilbert space and $S \subset U$ is a non-empty subset. Then*

$$S^\perp = (\operatorname{span} S)^\perp = (\overline{\operatorname{span} S})^\perp.$$

Proof. Since $S \subset \operatorname{span} S$, we have that $(\operatorname{span} S)^\perp \subset S^\perp$. Now assume that $u \in S^\perp$. Then $\langle u, w\rangle = 0$ for all $w \in S$. Thus $\langle u, w\rangle = 0$ whenever w is a finite linear combination of elements in S. In other words, $u \in (\operatorname{span} S)^\perp$. This shows that $S^\perp = (\operatorname{span} S)^\perp$. Now it follows from Proposition 5.2.6 that

$$(\overline{\operatorname{span} S})^\perp = (((\operatorname{span} S)^\perp)^\perp)^\perp = (\operatorname{span} S)^\perp.$$

□

Corollary 5.2.9. *Assume that U is a Hilbert space and $S \subset U$ a non-empty subset. Then* $\operatorname{span} S$ *is dense in U if and only if $S^\perp = \{0\}$.*

Proof. Recall that $\operatorname{span} S$ is dense if and only if $\overline{\operatorname{span} S} = U$. Now we can write

$$U = \overline{\operatorname{span} S} \oplus (\overline{\operatorname{span} S})^\perp = \overline{\operatorname{span} S} \oplus S^\perp.$$

Thus $U = \overline{\operatorname{span} S}$ if and only if $S^\perp = \{0\}$, which proves the claim. □

3. Riesz' Representation Theorem

Recall that the dual of a normed space U is the space $U^* = L(U, \mathbb{K})$ of all bounded linear functionals on U. One of the results we have obtained when we were discussing the dual of a normed space was that the Hilbert space $(\ell^2)^*$ can be identified with the space ℓ^2. We will now show that this identification of a Hilbert space with its dual works in general.

Theorem 5.3.1 (Riesz' Representation Theorem). *Assume that U is a Hilbert space. Then $f \in U^*$ if and only if there exists $v_f \in U$ such that*

$$f(u) = \langle u, v_f\rangle \qquad \text{for all } u \in U.$$

Moreover, v_f is unique, and we have that

$$\|f\|_{U^*} = \|v_f\|.$$

Proof. The "if" part follows directly from the fact that the mapping $u \mapsto \langle u, v_f\rangle$ is bounded linear, as discussed previously.

For the "only if" part, assume that $f \in U^*$. If $f = 0$, then we can choose $v_f = 0$. Assume therefore that $f \neq 0$, and denote $W := \ker f$. Then W is a closed subspace of U, and thus we can write $U = W \oplus W^\perp$. Moreover, since $f \neq 0$, we have that $W = \ker f \subsetneq U$, which implies that $W^\perp \neq \{0\}$. Choose now some $z \in W^\perp$ with $z \neq 0$. Since $z \notin W = \ker f$, it follows that $f(z) \neq 0$. Moreover, we have for all $u \in U$ that

$$f\Big(u - \frac{f(u)}{f(z)} z\Big) = f(u) - \frac{f(u)}{f(z)} f(z) = 0,$$

and thus

$$u - \frac{f(u)}{f(z)} z \in \ker f = W.$$

Since $z \in W^\perp$, this implies that

$$\Big\langle u - \frac{f(u)}{f(z)} z, z\Big\rangle = 0$$

for all $u \in U$. This can be rewritten as

$$\langle u, z\rangle = \frac{f(u)}{f(z)}\|z\|^2,$$

or, as $\|z\| \neq 0$,

$$f(u) = \frac{f(z)}{\|z\|^2}\langle u, z\rangle = \Big\langle u, \frac{\overline{f(z)}}{\|z\|^2} z\Big\rangle$$

for all $u \in U$. Thus, if we define

$$v_f = \frac{\overline{f(z)}}{\|z\|^2} z,$$

we obtain that $f(u) = \langle u, v_f\rangle$ for all $u \in U$.

Uniqueness: Assume now that w, $z \in U$ are such that

$$\langle u, w\rangle = f(u) = \langle u, z\rangle$$

for all $u \in U$. Then we have in particular that

$$\langle u, w - z\rangle = 0 \qquad \text{for all } u \in U,$$

which implies that $w - z = 0$ or $w = z$. This proves the uniqueness of v_f.

In order to compute the norm of v_f, we note first that

$$|f(u)| = |\langle u, v_f\rangle| \leq \|u\|\|v_f\|$$

for all $u \in U$ by the Cauchy–Schwarz–Bunyakovsky inequality, which shows that $\|f\|_{U^*} \leq \|v_f\|_U$. Conversely, for $u = v_f$ we obtain that

$$|f(v_f)| = |\langle v_f, v_f\rangle| = \|v_f\|^2,$$

and thus $\|f\|_{U^*} \geq \|v_f\|_U$. □

4. Adjoint operators on Hilbert spaces

In Section 5 we have introduced the notion of the adjoint of a linear transformation T between finite dimensional inner product spaces. There, we have first defined the adjoint T^* based on the singular value decomposition of the mapping T, and then we have shown that T^* can also be characterised by the property that $\langle Tu, v\rangle = \langle u, T^*v\rangle$ for all vectors u and v. We now will generalise the notion of an adjoint to arbitrary Hilbert spaces. In contrast to the finite dimensional case, however, we will do so by using the characterisation by $\langle Tu, v\rangle = \langle u, T^*v\rangle$ as the definition. This is necessary, as we have not developed a theory of the singular value decomposition in infinite dimensional Hilbert spaces.[1]

In the proofs of the theorems to come, we repeatedly make use of the following result, which thus deserves its own Lemma:

LEMMA 5.4.1. *Assume that U is an inner product space and that $v \in U$ is some vector. Then $v = 0$ if and only if*

$$\langle u, v\rangle = 0 \qquad \textit{for all } u \in U.$$

PROOF. If $v = 0$, then also $\langle u, v\rangle = 0$ for all $u \in U$.

Conversely, assume that $\langle u, v\rangle = 0$ for all $u \in U$. Then in particular $\|v\|^2 = \langle v, v\rangle = 0$, which proves that $v = 0$. □

[1] It is actually possible to generalise the singular value decomposition to operators between Hilbert spaces. This generalisation is quite complex, though, and requires some knowledge of measure and integration theory.

THEOREM 5.4.2 (Adjoint operator). *Assume that U and V are Hilbert spaces and that $T\colon U \to V$ is a bounded linear transformation. Then there exists a unique bounded linear transformation $T^*\colon V \to U$ such that*

$$\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U \qquad \text{for all } u \in U \text{ and } v \in V.$$

The operator T^* is called the adjoint *of T.*

PROOF. *Construction of T^*:* Define for fixed $v \in V$ the mapping $f_v\colon U \to \mathbb{K}$,

$$f_v(u) := \langle Tu, v\rangle_V.$$

The linearity of T together with the linearity of the inner product imply that f_v is linear. Moreover we have the estimate

$$|f_v(u)| = |\langle Tu, v\rangle_V| \le \|Tu\|_V\|v\|_V \le \|T\|_{L(U,V)}\|v\|_V\|u\|_U,$$

which shows that f_v is a bounded linear functional on U. The Riesz Representation Theorem now implies that there exists some $w \in U$ such that

$$f_v(u) = \langle u, w\rangle_U$$

for all $u \in U$. Define thus $T^*v := w$.

Repeating this construction for each $v \in V$ yields a mapping $T^*\colon V \to U$ with the property that

$$\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U$$

for all $u \in U$ and $v \in V$. We will now show that T^* is bounded linear.

Linearity: Let v, $w \in V$. Then

$$\begin{aligned}\langle u, T^*(v+w)\rangle_U = \langle Tu, v+w\rangle_V &= \langle Tu, v\rangle_V + \langle Tu, w\rangle_V\\ &= \langle u, T^*v\rangle_U + \langle u, T^*w\rangle_U = \langle u, T^*v + T^*w\rangle_U\end{aligned}$$

for all $u \in U$. Thus

$$\langle u, T^*(v+w) - T^*v - T^*w\rangle_U = 0$$

for all $u \in U$, which implies that $T^*(v+w) = T^*v + T^*w$.

Similarly, if $v \in V$ and $\lambda \in \mathbb{K}$, then

$$\langle u, T^*(\lambda v)\rangle_U = \langle Tu, \lambda v\rangle_U = \overline{\lambda}\langle Tu, v\rangle_U = \overline{\lambda}\langle u, T^*v\rangle_U = \langle u, \lambda T^*v\rangle_U$$

for all $u \in U$, which shows that $T^*(\lambda v) = \lambda T^*v$. This proves the linearity of T^*.

Boundedness: For all $v \in V$ we can estimate

$$\|T^*v\|_U^2 = \langle T^*v, T^*v\rangle_U = \langle TT^*v, v\rangle_V \le \|TT^*v\|_V\|v\|_V \le \|T\|_{L(U,V)}\|T^*v\|_U\|v\|_V.$$

This shows that

$$\|T^*v\|_U \le \|T\|_{L(U,V)}\|v\|_V, \tag{57}$$

which proves the boundedness of T^*.

Uniqueness: Assume now that $S \in L(V,U)$ is another bounded linear transformation such that $\langle Tu, v\rangle_V = \langle u, Sv\rangle_U$ for all $u \in U$ and $v \in V$. Then

$$\langle u, Sv\rangle_U = \langle Tu, v\rangle_V = \langle u, T^*v\rangle_V$$

for all $u \in U$ and $v \in V$, and thus

$$\langle u, Sv - T^*v\rangle_U = 0$$

for all $u \in U$ and $v \in V$. This implies that $Sv = T^*v$ for all $v \in V$, which proves that $S = T^*$. □

DEFINITION 5.4.3. Let U be a Hilbert space and let $T\colon U \to U$ be bounded linear.

- We say that T is *self-adjoint*, if $T^* = T$.

- We say that T is *normal*, if $TT^* = T^*T$.

PROPOSITION 5.4.4 (Properties of adjoints). *Assume that U and V are Hilbert spaces and that $T\colon U \to V$ is bounded linear. Then the following hold:*

- $(T^*)^* = T$.
- $\|T^*\|_{L(V,U)} = \|T\|_{L(U,V)}$.
- $\|T^*T\|_{L(U,U)} = \|TT^*\|_{L(V,V)} = \|T\|^2_{L(U,V)}$.

PROOF. Let $u \in U$ and $v \in V$. Then

$$\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U = \overline{\langle T^*v, u\rangle_U} = \overline{\langle v, (T^*)^*u\rangle_U} = \langle (T^*)^*u, v\rangle_V.$$

Since this holds for all $v \in V$, it follows that $Tu = (T^*)^*u$ for all $u \in U$, which in turn implies that $T = (T^*)^*$.

We next show that $\|T^*\|_{L(V,U)} = \|T\|_{L(U,V)}$. For that, we note first that the estimate $\|T^*\|_{L(V,U)} \le \|T\|_{L(U,V)}$ follows from (57). Applying the same estimate to T^*, we obtain moreover that

$$\|T\|_{L(U,V)} = \|(T^*)^*\|_{L(U,V)} \le \|T^*\|_{L(V,U)},$$

which proves the assertion.

For showing that $\|T^*T\|_{L(U,U)} = \|T\|^2_{L(U,V)}$, we note first that

$$\|T^*T\|_{L(U,U)} \le \|T^*\|_{L(V,U)}\|T\|_{L(U,V)} = \|T\|^2_{L(U,V)},$$

as $\|T^*\|_{L(V,U)} = \|T\|_{L(U,V)}$. For the opposite inequality we note that, for all $u \in U$,

$$\|Tu\|_V^2 = \langle Tu, Tu\rangle_V = \langle u, T^*Tu\rangle_U \le \|u\|_U\|T^*Tu\|_U \le \|u\|_U^2\|T^*T\|_{L(U,U)}.$$

Thus

$$\|T^*Tu\|_U \ge \sup_{\substack{u\in U\\ u\neq 0}} \frac{\|Tu\|_V^2}{\|u\|_U^2} = \|T\|^2_{L(U,V)}.$$

This shows that $\|T^*T\|_{L(U,U)} = \|T\|^2_{L(U,V)}$.

Finally, we have that

$$\|TT^*\|_{L(V,V)} = \|(T^*)^*T^*\|_{L(V,V)} = \|T^*\|^2_{L(V,U)} = \|T\|^2_{L(U,V)},$$

which concludes the proof. □

LEMMA 5.4.5. *Assume that U, V, and W are Hilbert spaces and that $T\colon U \to V$ and $S\colon V \to W$ are bounded linear. Then*

$$(S \circ T)^* = T^* \circ S^*.$$

PROOF. We have that

$$\begin{aligned}\langle u, (S\circ T)^*w\rangle_U = \langle (S\circ T)u, w\rangle_W = \langle S(Tu), w\rangle_W \\ = \langle Tu, S^*w\rangle_V = \langle u, T^*(S^*w)\rangle_U = \langle u, (T^*\circ S^*)w\rangle_U\end{aligned}$$

for all $u \in U$ and $w \in W$, which proves the assertion. □

We now look at some concrete examples of adjoints:

EXAMPLE 5.4.6 (Right and left shift). Let $U = \ell^2$ and let $T = R\colon U \to U$ be the right shift operator, that is,

$$Rx = (0, x_1, x_2, \dots)$$

for all $x \in \ell^2$. Then $R^*\colon \ell^2 \to \ell^2$ is defined by the condition

$$\langle Rx, y\rangle = \langle x, R^*y\rangle \qquad \text{for all } x,\ y \in \ell^2.$$

Now we can write

$$\langle Rx, y\rangle = \sum_{n=1}^{\infty} (Rx)_n \overline{y_n} = \sum_{n=2}^{\infty} x_{n-1}\overline{y_n} = \sum_{n=1}^{\infty} x_n \overline{y_{n+1}}$$

and

$$\langle x, R^* y\rangle = \sum_{n=1}^{\infty} x_n \overline{(R^* y)_n}.$$

Thus we have the condition that

$$\sum_{n=1}^{\infty} x_n \overline{y_{n+1}} = \sum_{n=1}^{\infty} x_n \overline{(R^* y)_n}$$

for all x, $y \in \ell^2$. From this, we can conclude that

$$y_{n+1} = (R^* y)_n$$

for all $y \in \ell^2$ and $n \in \mathbb{N}$, which implies that $R^* = L$ the left shift operator

$$L(y_1, y_2, y_3, \ldots) = (y_2, y_3, \ldots).$$

Obviously, $R \neq L = R^*$ and thus R is not self-adjoint. In addition, we note that $(L \circ R)x = x$ for all $x \in \ell^2$, whereas

$$(R \circ L)x = (0, x_2, x_3, \ldots).$$

This shows that $R \circ R^* = R \circ L \neq L \circ R = R^* \circ R$, and thus R is not normal.

EXAMPLE 5.4.7. Recall that $L^2([0,1])$ is the completion of the space $C([0,1])$ with respect to the norm $\|\cdot\|_2$ defined by

$$\|f\|_2 := \Big(\int_0^1 |f(x)|^2\,dx\Big)^{1/2}.$$

This norm is induced by the inner product $\langle \cdot, \cdot\rangle$ defined by

$$\langle f, g\rangle := \int_0^1 f(x)\overline{g(x)}\,dx,$$

which implies that $L^2([0,1])$ is a Hilbert space.

Now assume that $k\colon [0,1] \times [0,1] \to \mathbb{K}$ is continuous and define the mapping $T\colon C([0,1]) \to L^2([0,1])$,

$$Tf(x) := \int_0^1 k(x,y) f(y)\,dy.$$

Then

$$\begin{aligned}\|Tf\|_2^2 = \int_0^1 |Tf(x)|^2\,dx &= \int_0^1 \Big|\int_0^1 k(x,y)f(y)\,dy\Big|^2\,dx \\ &\leq \int_0^1 \Big(\int_0^1 |k(x,y)|^2\,dy\Big)\Big(\int_0^1 |f(y)|^2\,dy\Big)\,dx \\ &= \Big(\int_0^1 \int_0^1 |k(x,y)|^2\,dx\,dy\Big)\|f\|_2^2,\end{aligned}$$

which proves that T is bounded linear. As a consequence, T has a unique bounded linear extension (which we denote for simplicity again by T) to $L^2([0,1])$.

We now compute the adjoint of T. For all f and g we have that

$$\begin{aligned}
\langle f, T^*g\rangle &= \langle Tf, g\rangle \\
&= \int_0^1 Tf(x)\overline{g(x)}\,dx \\
&= \int_0^1 \Big(\int_0^1 k(x,y)f(y)\,dy\Big)\,\overline{g(x)}\,dx \\
&= \int_0^1 \int_0^1 k(x,y)f(y)\overline{g(x)}\,dy\,dx \\
&= \int_0^1 \int_0^1 k(x,y)f(y)\overline{g(x)}\,dx\,dy \\
&= \int_0^1 f(y)\Big(\int_0^1 k(x,y)\overline{g(x)}\,dx\Big)\,dy \\
&= \langle f, h\rangle,
\end{aligned}$$

where

$$h(y) := \int_0^1 \overline{k(x,y)}g(x)\,dx.$$

This shows that T^* is the operator given by

$$T^*g := \int_0^1 \overline{k(x,y)}g(x)\,dx.$$

That is, T^* is again an integral functional with kernel $k^*(x,y) := \overline{k(y,x)}$.

From this we see in particular that T is self-adjoint if k is conjugate symmetric in the sense that $k(x,y) = \overline{k(y,x)}$ for all x, $y \in [0,1]$.

EXAMPLE 5.4.8. Consider the linear initial value problem with varying coefficients

$$\begin{aligned}
y'(t) &= A(t)y(t), \\
y(0) &= x,
\end{aligned} \tag{58}$$

where $x \in \mathbb{R}^n$ is some given initial value and $A\colon [0,1] \to \mathbb{R}^{n\times n}$ is a continuous function. Then we have for each initial value $x \in \mathbb{R}^n$ a unique solution $y(t;x)\colon [0,1]\times\mathbb{R}^n \to \mathbb{R}^n$. Moreover, the solution $y(t;x)$ depends for every $t \in [0,1]$ linearly on the initial value x.

Denote know by $T\colon \mathbb{R}^n \to \mathbb{R}^n$ the mapping

$$Tx := y(1;x)$$

that maps the initial value x to the *final value* of the corresponding solution of (58). In the following, we will discuss the adjoint mapping $T^*\colon \mathbb{R}^n \to \mathbb{R}^n$.

For that, define the *adjoint equation* to (58) as the *final value problem*

$$\begin{aligned}
v'(t) &= -A(t)^*v(t), \\
v(1) &= w,
\end{aligned} \tag{59}$$

where $w \in \mathbb{R}^n$ is some given final value and $A(t)^*$ denotes the adjoint of $A(t)$. Then we have again for each final value $w \in \mathbb{R}^n$ a unique solution $v(t;w)\colon [0,1]\times\mathbb{R}^n \to \mathbb{R}^n$ of (59).

Now the product rule implies that

$$\begin{aligned}
\langle y(t;x), v(t;w)\rangle' &= \langle y(t;x)', v(t;w)\rangle + \langle y(t;x), v(t;w)'\rangle \\
&= \langle A(t)y(t;x), v(t;w)\rangle + \langle y(t;x), -A(t)^*v(t;w)\rangle \\
&= \langle A(t)y(t;x), v(t;w)\rangle - \langle A(t)y(t;x), v(t;w)\rangle = 0
\end{aligned}$$

for all $t \in [0,1]$ and all x, $w \in \mathbb{R}^n$. Thus

$$0 = \int_0^1 \langle y(t;x), v(t;w)\rangle'\,dt = \langle y(1;x), v(1;w)\rangle - \langle y(0;x), v(0;w)\rangle$$
$$= \langle Tx, w\rangle - \langle x, v(0;w)\rangle$$

for all x, $w \in \mathbb{R}^n$. This shows that

$$T^*w = v(0;w)$$

for all $w \in \mathbb{R}^n$. That is, the adjoint of T maps the final value w to the initial value of the corresponding solution of the adjoint equation (59).

In the finite dimensional case, we have shown the relations $\ker T = (\operatorname{ran} T^*)^\perp$ and $\operatorname{ran} T = (\ker T^*)^\perp$ (see Lemma 2.5.9). The next result provides a generalisation to the infinite dimensional case.

PROPOSITION 5.4.9. *Assume that U and V are Hilbert spaces and that $T\colon U \to V$ is bounded linear. Then*

$$\overline{\operatorname{ran} T} = (\ker T^*)^\perp \qquad \textit{and} \qquad \ker T = (\operatorname{ran} T^*)^\perp.$$

PROOF. We show first that $\overline{\operatorname{ran} T} = (\ker T^*)^\perp$.

Assume for that first that $v \in \operatorname{ran} T$. Then $v = Tu$ for some $u \in U$. Now let $w \in \ker T^*$. Then $T^*w = 0$ and thus

$$\langle v, w\rangle_V = \langle Tu, w\rangle_V = \langle u, T^*w\rangle_U = 0.$$

This shows that $v \in (\ker T^*)^\perp$. Since this holds for all $v \in \operatorname{ran} T$, it follows that $\operatorname{ran} T \subset (\ker T^*)^\perp$. Since $(\ker T^*)^\perp$ is closed, we can conclude that also $\overline{\operatorname{ran} T} \subset (\ker T^*)^\perp$.

Now assume that $v \in (\operatorname{ran} T)^\perp$, that is, that $\langle w, v\rangle = 0$ for all $w \in \operatorname{ran} T$. Then we have for all $u \in U$ that

$$\langle u, T^*v\rangle_U = \langle Tu, v\rangle_V = 0,$$

which implies that T^*v and thus $v \in \ker T^*$. This shows that

$$(\operatorname{ran} T)^\perp \subset \ker T^*.$$

Thus we can conclude that

$$(\ker T^*)^\perp \subset ((\operatorname{ran} T)^\perp)^\perp = ((\overline{\operatorname{ran} T})^\perp)^\perp = \overline{\operatorname{ran} T}.$$

Together, these results show that

$$\overline{\operatorname{ran} T} = (\ker T^*)^\perp.$$

Applying this result to T^* shows furthermore that

$$\overline{\operatorname{ran} T^*} = (\ker (T^*)^*)^\perp = (\ker T)^\perp,$$

which is equivalent to

$$\ker T = (\overline{\operatorname{ran} T^*})^\perp = (\operatorname{ran} T^*)^\perp.$$

□

As a consquence of this, we obtain the following results:

$$\overline{\operatorname{ran} T} = (\ker T^*)^\perp, \qquad \ker T^* = (\operatorname{ran} T)^\perp,$$
$$\overline{\operatorname{ran} T^*} = (\ker T)^\perp, \qquad \ker T = (\operatorname{ran} T^*)^\perp.$$

In addition, these imply the decompositions

$$U = \ker T \oplus \overline{\operatorname{ran} T^*} \qquad \text{and} \qquad V = \ker T^* \oplus \overline{\operatorname{ran} T}.$$

Apart from the necessity to take the closure of the ranges of T and T^*, these are the same results that we have derived in the finite dimensional case.

COROLLARY 5.4.10. *Assume that* U *and* V *are Hilbert spaces and that* $T: U \to V$ *is bounded linear. Then* T *is injective, if and only if* $\operatorname{ran} T^* \subset U$ *is dense. Similarly,* T^* *is injective, if and only if* $\operatorname{ran} T \subset V$ *is dense.*

PROOF. We have that T is injective, if and only if $\ker T = \{0\}$. Using the decomposition $U = \ker T \oplus \overline{\operatorname{ran} T^*}$, this is equivalent to the equality $U = \overline{\operatorname{ran} T^*}$, which is precisely the density of $\operatorname{ran} T^*$ in U. The second equivalence is similar. □

5. Hilbert space bases

Recall that a basis in a vector space U is a subset $\{u_i : i \in I\} \subset U$ with the property that every vector $u \in U$ can be uniquely written as a finite linear combination of the form

$$u = \sum_{n=1}^{N} \lambda_n u_{i_n}$$

for some $N \in \mathbb{N}$, $\lambda_n \in \mathbb{K}$, and $i_n \in I$ for $1 \le n \le N$. In the context of infinite dimensional spaces, such a basis is often called a *Hamel basis*, in order to distinguish from other basis concepts, some of which we will discuss in the following.

Before introducing alternative basis definitions, we need to discuss some concepts related to measuring the "size" of different infinite sets.

DEFINITION 5.5.1 (Countable set). An infinite set S is called *countable*, if there exists a surjective mapping $f: \mathbb{N} \to S$. If an infinite set S is not countable, it is called *uncountable*.

That is, a set S is countable, if we can write it as a sequence

$$S = \{s_1, s_2, s_3, \ldots\}.$$

EXAMPLE 5.5.2. The set $\mathbb{N}$ of natural numbers is obviously countable. So is the set $\mathbb{Z}$ of integers, where we can use the ordering $\mathbb{Z} = \{0, +1, -1, +2, -2, \ldots\}$.

EXAMPLE 5.5.3. If S is a countable set, then so is the Cartesian product S^n for every $n \ge 1$. Indeed, assume that $S = \{s_1, s_2, s_3, \ldots\}$, and take for simplicity $n = 2$. Then we can write

$$S \times S = \{(s_1, s_1), (s_1, s_2), (s_2, s_1), (s_3, s_1), (s_2, s_2), (s_1, s_3), \ldots\}.$$

A similar construction works for $n \ge 3$.

More general, if $S_1, \ldots, S_n$ are countable sets, then so is the product $S_1 \times S_2 \times \cdots \times S_n$.

EXAMPLE 5.5.4. The set $\mathbb{Q}$ of rational numbers is countable. In order to see this, note that we can write every rational number x as a fraction $x = p/q$ with $p \in \mathbb{Z}$ and $q \in \mathbb{N}$. Since the Cartesian product $\mathbb{Z} \times \mathbb{N}$ is countable, it follows that $\mathbb{Q}$ is countable as well.

EXAMPLE 5.5.5. The set $\mathbb{R}$ of real numbers is uncountable. In order to show this, we will employ a technique known as *Cantor's diagonal argument*.

Assume to the contrary that $\mathbb{R}$ were countable. Then we could write $\mathbb{R} = \{s_1, s_2, s_3, \ldots\}$. We will now construct a real number that is different from each of the numbers $s_1, s_2, \ldots$. For this, note that we can write each number s_n in decimal expansion in the form

$$s_n = \pm a^{(n)}_{-N_n} a^{(n)}_{-(N_n-1)} \cdots a^{(n)}_0 , a^{(n)}_1 a^{(n)}_2 a^{(n)}_3 \ldots$$

with digits $a^{(n)}_j \in \{0, 1, \ldots, 9\}$. That is,

$$s_n = \sum_{j=-N_n}^{\infty} a_j \cdot 10^{-j}.$$

Define now the number

$$x = \sum_{j=1}^{\infty} b_j \cdot 10^{-j},$$

where the digits of x are chosen such that $b_j \neq a_j^{(j)}$. Then x differs from s_n at least in its n-th digit after the comma, and thus it is different from s_n.[2] This, however, implies that x is not contained in the set $\{s_1, s_2, s_3, \ldots\}$, which contradicts the assumption that this set includes all real numbers.

One problem with Hamel bases in infinite dimensional Banach spaces is that one can show that they are necessarily uncountable, which makes them essentially impossible to use in practice.

PROPOSITION 5.5.6. *Let U be an infinite dimensional Banach space. Then every Hamel basis of U is uncountable.*

REMARK 5.5.7. For incomplete normed spaces, the situation is different as we have already seen. For instance, the set $\{e_1, e_2, \ldots, \ldots\}$ of unit sequences is a Hamel basis in the space c_{fin} of finite sequences. However, c_{fin} is incomplete with respect to all the norms we have discussed.

We now introduce an alternative notion of a basis of a normed space.

DEFINITION 5.5.8 (Schauder basis). Let $(U, \|\cdot\|_U)$ be an infinite dimensional normed space. A countable set $\{u_1, u_2, u_3, \ldots\} \subset U$ is a *Schauder basis* of U, if there exists for every $u \in U$ a unique sequence $\lambda_1, \lambda_2, \ldots$, such that

$$\lim_{N\to\infty} \Big\| u - \sum_{n=1}^{N} \lambda_n u_n \Big\| = 0.$$

Put differently, if $\{u_1, u_2, \ldots\}$ is a Schauder basis of U, then we can write each $u \in U$ uniquely as an *infinite series*

$$u = \sum_{n=1}^{\infty} \lambda_n u_n.$$

EXAMPLE 5.5.9. Let $1 \le p < \infty$ and consider the set of unit sequences $\{e_1, e_2, \ldots\}$ in ℓ^p. Then this set is a Schauder basis of ℓ^p. In order to see this, assume that $x = (x_1, x_2, \ldots) \in \ell^p$ and define $\lambda_n = x_n$. Then

$$\Big\| x - \sum_{n=1}^{N} \lambda_n e_n \Big\|_p = \Big(\sum_{n=N+1}^{\infty} |x_n|^p \Big)^{1/p}.$$

Since $x \in \ell^p$, it follows that

$$\lim_{N\to\infty} \Big\| x - \sum_{n=1}^{N} \lambda_n e_n \Big\|_p = 0.$$

Moreover, if we choose any other sequence $\lambda_1, \lambda_2, \ldots$ of coefficients, then the series $\sum_n \lambda_n e_n$ does not converge to x. Thus the representation

$$x = \sum_{n=1}^{\infty} \lambda_n e_n$$

is unique, which shows that $\{e_1, e_2, \ldots\}$ is a Schauder basis of ℓ^p.

---

[2] One has to be slightly careful here, as the decimal expansion of a real number is not always unique (for instance we have $1 = 0.99999\ldots$). However, this non-uniqueness issue only occurs for numbers with a finite decimal expansion, and we can easily avoid any potential issues by for instance setting $b_j = 5$ if $a_j^{(j)} = 0$ or $a_j^{(j)} = 9$.

EXAMPLE 5.5.10. The set of unit sequences $\{e_1, e_2, \dots\}$ is *not* a Schauder basis of ℓ^∞. Take for instance $x = (1, 1, \dots)$. Then

$$\Big\| x - \sum_{n=1}^{N} \lambda_n e_n \Big\| \geq 1$$

for all sequences $\lambda_1, \lambda_2, \dots$, and thus we cannot represent x in the desired form.

DEFINITION 5.5.11 (Separability). A normed space U is called *separable*, if there exists a countable dense subset of U.

EXAMPLE 5.5.12. The set $\mathbb{R}$ of real numbers is separable, as it contains $\mathbb{Q}$ as a countable dense subset. Similarly, for every $n \in \mathbb{N}$ the sets $\mathbb{R}^n$ and $\mathbb{C}^n$ are separable.

LEMMA 5.5.13. *Assume that the normed space U has a Schauder basis. Then U is separable.*

PROOF. Exercise! □

REMARK 5.5.14. A long standing problem in functional analysis was the question, whether every separable Banach space has a Schauder basis. In the seventies, this question was answered negatively. That is, there exist separable Banach spaces, which do not have any Schauder basis.

EXAMPLE 5.5.15. The space ℓ^∞ is not separable.
The basic idea is to consider the sequences $e^{(I)} \in \ell^\infty$ defined by

$$e_n^{(I)} = \begin{cases} 1 & \text{if } n \in I, \\ 0 & \text{if } n \notin I, \end{cases}$$

where $I \subset \mathbb{N}$ is any subset. One can show[3] that the set $\big\{I : I \subset \mathbb{N}\big\}$ (and thus also the set $\big\{e^{(I)} : I \subset \mathbb{N}\big\}$) is uncountable. Moreover, we have that $\|e^{(I)} - e^{(J)}\|_\infty = 1$ whenever $I \neq J$.

Thus we have an uncountable *discrete* subset of ℓ^∞, which makes it impossible to find a countable dense subset: Assume that $\{s_1, s_2, \dots\} \subset \ell^\infty$ is dense. Then there exists in particular for each $I \subset \mathbb{N}$ some index n_I such that $\|s_{n_I} - e^{(I)}\|_\infty < 1/4$. Then, however, we have for all $I \neq J$ that

$$\begin{aligned} 1 = \|e^{(I)} - e^{(J)}\|_\infty \leq \|e^{(I)} - s_{n_I}\|_\infty + \|s_{n_I} - s_{n-J}\|_\infty + \|s_{n_J} - e^{(J)}\|_\infty \\ \leq \frac{1}{2} + \|s_{n_I} + s_{n_J}\|_\infty, \end{aligned}$$

which shows that

$$\|s_{n_I} + s_{n_J}\|_\infty \geq \frac{1}{2}.$$

In particular, we have that $n_I \neq n_J$ whenever $I \neq J$. Since we have uncountably many subsets $I \subset \mathbb{N}$, this implies that we also need uncountably many indices n_I, which contradicts the assumption that $\{s_1, s_2, \dots\}$ is countable.

We will now discuss in more detail the case where U is a Hilbert space. Recall to that end that a sequence $\{u_1, u_2, \dots\}$ in an inner product space is

- *orthogonal*, if $\langle u_i, u_j \rangle = 0$ whenever $i \neq j$,
- *orthonormal*, if it is orthogonal and $\|u_i\| = 1$ for all i.

[3]Try it yourself! This is really only Cantor's diagonal argument again.

THEOREM 5.5.16. *Let U be an inner product space. Assume moreover that $\{u_1, u_2, \ldots, u_n\}$ is a finite orthonormal system in U and denote*

$$W = \operatorname{span}(\{u_1, u_2, \ldots, u_n\}).$$

Then W is a closed subspace of U and we have

$$\sum_{j=1}^{n} \langle u, u_j \rangle u_j = \pi_W(u)$$

and

$$\|u - \pi_W(u)\|^2 = \|u\|^2 - \sum_{j=1}^{n} |\langle u, u_j \rangle|^2$$

for all $u \in U$.

PROOF. Exercise! □

LEMMA 5.5.17 (Bessel's inequality). *Assume that U is an infinite dimensional inner product space and $\{u_1, u_2, \ldots\}$ is an infinite orthonormal system. Then*

$$\sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2 \leq \|u\|^2$$

for all $u \in U$.

PROOF. Applying the result of Theorem 5.5.16 to the spaces

$$W_n = \operatorname{span}(\{u_1, u_2, \ldots, u_n\}),$$

we obtain that

$$\sum_{j=1}^{n} |\langle u, u_j \rangle|^2 = \|u\|^2 - \|u - \pi_W(u)\|^2 \leq \|u\|^2$$

for all $u \in U$. Taking the limit $n \to \infty$, we obtain the claimed estimate. □

THEOREM 5.5.18. *Assume that U is an infinite dimensional Hilbert space and that $\{u_1, u_2, \ldots\}$ is an orthonormal system in U. Assume moreover that $\lambda = (\lambda_1, \lambda_2, \ldots)$ is a sequence in $\mathbb{K}$. Then the sequence $\{v_n\}_{n \in \mathbb{N}} \subset U$ defined by*

$$v_n = \sum_{j=1}^{n} \lambda_j u_j$$

converges in U, if and only if $\lambda \in \ell^2$.

PROOF. Assume first that the sequence $\{v_n\}_{n \in \mathbb{N}}$ converges to some $v \in U$. Then it is necessarily bounded and thus there exists $R > 0$ such that $\|v_n\| \leq R$ for all $n \in \mathbb{N}$. Now we note that, by Pythagoras' theorem, using the orthonormality of the system $\{u_1, u_2, \ldots\}$,

$$R^2 \geq \|v_n\|^2 = \sum_{j=1}^{n} |\lambda_j|^2 \|u_j\|^2 = \sum_{j=1}^{n} |\lambda_j|^2.$$

Since this holds for all $n \in \mathbb{N}$, it follows that $\lambda \in \ell^2$.

Assume now that $\lambda \in \ell^2$ and let $\varepsilon > 0$. Then there exists $N \in \mathbb{N}$ such that

$$\sum_{j=N+1}^{\infty} |\lambda_j|^2 < \varepsilon^2.$$

Now assume that n, $m \geq N$. Without loss of generality assume that $m \geq n$. Then

$$\|v_n - v_m\|^2 = \Big\| \sum_{j=n+1}^{m} \lambda_j u_j \Big\|^2 = \sum_{j=n+1}^{m} |\lambda_j|^2 \leq \sum_{j=N+1}^{\infty} |\lambda_j|^2 \leq \varepsilon^2.$$

This shows that $\{v_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence. Since U is a Hilbert space and thus complete, it follows that the sequence $\{v_n\}_{n\in\mathbb{N}}$ converges. □

THEOREM 5.5.19. *Assume that U is an infinite dimensional Hilbert space and that $\{u_1, u_2, \dots\} \subset U$ is an orthonormal system. Denote moreover*

$$W = \overline{\operatorname{span}\{u_1, u_2, \dots\}}.$$

Then

$$\pi_W(u) = \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j$$

for all $u \in U$.

PROOF. By Bessel's inequality we have that

$$\sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2 \leq \|u\|^2,$$

which shows that the sequence $\{\langle u, u_j \rangle\}_{j\in\mathbb{N}}$ is contained in ℓ^2. Thus

$$z := \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j = \lim_{n\to\infty} \sum_{j=1}^{n} \langle u, u_j \rangle u_j$$

exists in U. Moreover, by definition z is the limit of a sequence in $\operatorname{span}\{u_1, u_2, \dots\}$, which implies that $z \in W$.

It thus remains to show that $u - z \in W^\perp$. Now let $k \in \mathbb{N}$ be fixed. Then

$$\begin{aligned} \langle z, u_k \rangle &= \lim_{n\to\infty} \sum_{j=1}^{n} \big\langle \langle u, u_j \rangle u_j, u_k \big\rangle \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} \langle u, u_j \rangle \langle u_j, u_k \rangle = \langle u, u_k \rangle, \end{aligned}$$

as $\langle u_j, u_k \rangle = 0$ whenever $j \neq k$ and $\langle u_k, u_k \rangle = 1$. Thus

$$\langle u - z, u_k \rangle = \langle u, u_k \rangle - \langle u, u_k \rangle$$

for all $k \in \mathbb{N}$ and thus

$$u - z \in \{u_1, u_2, \dots\}^\perp = \overline{\operatorname{span}\{u_1, u_2, \dots\}}^\perp = W^\perp.$$

Thus we have shown that $z \in W$ and $u - z \in W^\perp$, which proves that $z = \pi_W(u)$. □

REMARK 5.5.20. Assume that U is a Hilbert space and that $\{u_1, u_2, \dots\}$ is an orthonormal system in U. The numbers $\langle u, u_j \rangle$ are sometimes called the *(generalised) Fourier coefficients* of u with respect to the system $\{u_1, u_2, \dots\}$.

DEFINITION 5.5.21 (total subset). An orthonormal system $\{u_1, u_2, \dots\}$ in a Hilbert space U is *total* or *maximal*, if

$$\overline{\operatorname{span}(\{u_1, u_2, \dots\})} = U,$$

or, equivalently, if

$$\{u_1, u_2, \dots\}^\perp = \{0\}.$$

DEFINITION 5.5.22 (Hilbert basis). An orthonormal system $\{u_1, u_2, \ldots\}$ in an infinite dimensional Hilbert space U is an *orthonormal basis* or a *Hilbert basis* of U, if

$$u = \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j$$

for all $u \in U$.

THEOREM 5.5.23. *Assume that U is an infinite dimensional Hilbert space and $S = \{u_1, u_2, \ldots\} \subset U$ is an orthonormal system. The following are equivalent:*

(1) S is total.
(2) S is a Hilbert basis of U.
(3) For every $u \in U$, Parseval's identity

$$\sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2 = \|u\|^2 \tag{60}$$

holds.

PROOF. Assume that S is total. Then $S^\perp = \{0\}$ and thus $U = \overline{\operatorname{span} S}$. Thus we have for all $u \in U$ that $\pi_{\overline{\operatorname{span} S}}(u) = u$. Thus

$$u = \pi_{\overline{\operatorname{span} S}}(u) = \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j.$$

Since this holds for all $u \in U$, it follows that S is a Hilbert basis of U.

Assume now that S is a Hilbert basis of u. Then

$$u = \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j$$

and thus

$$\|u\|^2 = \Big\| \lim_{n\to\infty} \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j \Big\|^2 = \lim_{n\to\infty} \Big\| \sum_{j=1}^{\infty} \langle u, u_j \rangle u_j \Big\|^2$$
$$= \lim_{n\to\infty} \sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2 = \sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2.$$

Thus Parseval's identity holds.

Finally assume that Parseval's identity holds and let $u \in S^\perp$. Then $\langle u, u_j \rangle = 0$ for all $j \in \mathbb{N}$ and thus

$$\|u\|^2 = \sum_{j=1}^{\infty} |\langle u, u_j \rangle|^2 = 0,$$

showing that $u = 0$. Thus $S^\perp = \{0\}$ and therefore S is total. □

LEMMA 5.5.24. *Every infinite dimensional Hilbert space contains an infinite orthonormal system.*

IDEA OF PROOF. Choose an infinite, linear independent sequence in U. By applying Gram–Schmidt orthogonalisation to this sequence, we obtain an infinite orthonormal system. □

THEOREM 5.5.25. *Every infinite dimensional separable Hilbert space has a Hilbert basis.*

IDEA OF PROOF. Apply Gram–Schmidt orthogonalisation to a countable dense subset. □

In particular, we obtain from this result that every separable Hilbert space "looks like" the sequence space ℓ^2.

COROLLARY 5.5.26 (Riesz–Fischer). *Every infinite dimensional separable Hilbert space is isometrically isomorphic to* ℓ^2.

PROOF. Let U be an infinite dimensional separable Hilbert space, and let $\{u_1, u_2, \ldots\}$ be a Hilbert space basis of U. Define the mapping $T\colon U \to \ell^2$,

$$Tu = \big(\langle u, u_j\rangle\big)_{j\in\mathbb{N}}.$$

Then every vector $u \in U$ can uniquely be written as

$$u = \sum_{j\in\mathbb{N}} \langle u, u_j\rangle u_j,$$

which implies in particular that T is injective and surjective. Moreover, by Parseval's identity we have that

$$\|u\|_U^2 = \sum_{j\in\mathbb{N}} |\langle u, u_j\rangle|^2 = \|Tu\|_{\ell^2}$$

for all $u \in U$, which proves that T is an isometry. □

EXAMPLE 5.5.27. In the space ℓ^2 the set $\{e_j : j \in \mathbb{N}\}$ of unit sequences is a Hilbert basis.

EXAMPLE 5.5.28. Let $L^2([0, 2\pi], \mathbb{C})$ denote the space of square integrable complex valued functions on $[0, 2\pi]$. That is, $L^2([0, 2\pi], \mathbb{C})$ is the completion of the space $C([0, 2\pi], \mathbb{C})$ of continuous complex values functions on $[0, 2\pi]$ with respect to the norm

$$\|f\|_2 = \int_0^{2\pi} |f(x)|^2\, dx,$$

which is induced by the inner product

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

Then the set of functions

$$f_n(x) := \frac{1}{\sqrt{2\pi}} e^{inx}$$

for $n \in \mathbb{Z}$ is a Hilbert basis (the *Fourier basis*) of $L^2([0, 2\pi], \mathbb{C})$.

The structure of general Banach spaces is much more complicated, and there exists no result similar to the Riesz–Fischer theorem for Banach spaces. However, the following result states that at least finite dimensional normed spaces cannot be "too different" from each other.

THEOREM 5.5.29. *Assume that* $(U, \|\cdot\|_U)$ *is a* finite dimensional *normed space and that* $\|\cdot\|_a$ *and* $\|\cdot\|_b$ *are two norms on* U. *Then there exist constants* $0 < c < C$ *such that*[4]

$$c\|u\|_a \le \|u\|_b \le C\|u\|_a$$

for all $u \in U$. *In particular, a sequence* $\{u_n\}_{n\in\mathbb{N}} \subset U$ *is a Cauchy-sequence (converges) with respect to* $\|\cdot\|_a$, *if and only if it is a Cauchy-sequence (converges) with respect to* $\|\cdot\|_b$.

COROLLARY 5.5.30. *Every finite dimensional normed space is complete.*

REMARK 5.5.31. Assume that U is a (not necessarily complete) normed space and that $W \subset U$ is a finite dimensional subspace. Then it is In particular, this implies that finite dimensional subspaces of normed spaces are necessarily closed.

[4]Norms with such a property are said to be *equivalent*.

6. Galerkin's Method

We will now provide a glimpse into the theory of numerical functional analysis, that is, the (approximate) numerical solution of equations in infinite dimensional spaces. One central idea there, which builds on everything that we did in this course, is the so called *Galerkin method*. The basic idea in this method is that one tries to approximate an infinite dimensional problem by a sufficiently high dimensional finite dimensional one. As a preparation, we return once to least squares solutions of linear systems.

THEOREM 5.6.1. *Assume that U and V are Hilbert spaces, that $T: U \to V$ is bounded linear, and that $v \in V$ is given. Then $u \in U$ solves the* least squares problem

$$\min_{u \in U} \|Tu - v\|_V^2 \tag{61}$$

if and only if u solves the normal equation

$$T^*Tu = T^*v. \tag{62}$$

PROOF. The least squares problem 61 is equivalent to the problem

$$\min_{w \in \operatorname{ran}(T)} \|w - v\|_V^2 \tag{63}$$

in the sense that w solves (63) if and only if $w = Tu$ for some solution u of (61).

Since (63) is just a projection problem, we know that $w \in V$ is a solution of (63) if and only if

$$w \in \operatorname{ran} T \qquad \text{and} \qquad w - v \in (\operatorname{ran} T)^\perp = \ker(T^*).$$

That is, $w = Tu$ for some $u \in U$ and $T^*(w - v) = 0$. This, however, is in turn equivalent to the normal equation $T^*(Tu - v) = 0$. $\square$

REMARK 5.6.2. Note that it can happen that neither of the problems (61) or (62) has a solution. A part of the claim of the theorem is, however, that if one of the problems has a soultion, then so has the other.

Assume now that U and V are infinite dimensional Hilbert spaces, that $T: U \to V$ is bounded linear, and that $v \in V$ is given. Our goal is the solution of the equation $Tu = v$. Since we cannot really do so in this infinite dimensional setting, we have to apply some form of discretisation. The idea of *Galerkin's method* is to do this by "projecting" the problem $Tu = v$ to a finite dimensional subspace of U.

Assume therefore that $W \subset U$ is some finite dimensional subspace and denote by $S := T|_W : W \to V$ the restriction of T to W. Then we can consider the least squares problem

$$\min_{u \in W} \|Su - v\|_V^2.$$

That is, we try to find a "best solution" of the infinite dimensional problem within the finite dimensional subspace W. In view of the previous result, this is equivalent to solving the normal equation

$$S^*Su = S^*v, \tag{64}$$

where $S^*: V \to W$ is the adjoint of $S = T|_W$. Note moreover that this problem necessarily has a solution, because W, and therefore also $\operatorname{ran}(S)$, is finite dimensional.

Assume now that $\{u_1, u_2, \ldots, u_N\}$ is a basis of W. Then $u \in W$ solves (64), if and only if

$$\langle S^*Su, u_i \rangle = \langle S^*v, u_i \rangle$$

for all $1 \le i \le N$. By the definition of the adjoint S^* of S, this is equivalent to the system of equations

$$\langle Su, Su_i\rangle = \langle v, Su_i\rangle$$

for all $1 \le i \le N$. Moreover, by definition $Su = Tu$ for all $u \in W$ and thus we obtain the system of equations

$$\langle Tu, Tu_i\rangle = \langle v, Tu_i\rangle \qquad \text{for all } 1 \le i \le N. \tag{65}$$

Finally, we will rewrite this system in matrix-vector form. Since $\{u_1, \ldots, u_N\}$ is a basis of W, we can write every solution u of (65) as

$$u = \sum_{j=1}^{N} x_j u_j$$

with unique coordinates $x_1, \ldots, x_N \in \mathbb{K}$. Inserting this form of u in (65), we obtain the system of equations

$$\sum_{j=1}^{N} x_j \langle Tu_j, Tu_i\rangle = \langle v, Tu_i\rangle \qquad \text{for all } 1 \le i \le N \tag{66}$$

for the coordinates $x_1, \ldots, x_N$ of u. Now define the matrix

$$A = \big(\langle Tu_j, Tu_i\rangle\big)_{i,j=1,\ldots,N}$$

and the vector

$$b = \big(\langle v, Tu_i\rangle\big)_{i=1,\ldots,N}.$$

Then the system (66) can be written in matrix-vector form as

$$Ax = b.$$

Note moreover that, by construction, $A = A^H$ is a Hermitian matrix and that $b \in \operatorname{ran} A$. In particular, the equation has a solution in $\mathbb{K}^n$. Moreover, if T is injective, then the solution is unique.

Now an important question is, whether this actually yields an approximation of the solution of the infinite dimensional problem. That is, if we increase the dimension of the discretisation space W, do we get closer to solving the original problem? The following results provides an answer to this question:

THEOREM 5.6.3 (Convergence of Galerkin methods). *Assume that U and V are infinite dimensional Hilbert spaces, that $T\colon U \to V$ is bounded linear and invertible with bounded linear inverse $T^{-1}\colon V \to U$. Assume that $v \in V$ is given and denote by $u^\dagger \in U$ the unique solution of the equation $Tu = v$.*

Assume moreover that $\{u_1, u_2, \ldots\}$ is a Hilbert basis of U. For $N \in \mathbb{N}$ denote

$$W_N := \operatorname{span}\{u_1, u_2, \ldots, u_N\} \subset U$$

and denote by $u^{(N)} \in W_N$ the (unique) solution of the problem

$$\min_{u \in W_N} \|Tu - v\|_V^2. \tag{67}$$

Then

$$u^\dagger = \lim_{N\to\infty} u^{(N)}.$$

PROOF. Since $T^{-1}\colon V \to U$ is bounded linear, we can estimate

$$\|u^{(N)} - u^\dagger\|_U \le \|T^{-1}\|_{L(V,U)} \|Tu^{(N)} - Tu^\dagger\|_V = \|T^{-1}\|_{L(V,U)} \|Tu^{(N)} - v\|_V.$$

Thus it is sufficient to show that $\|Tu^{(N)} - v\|_V \to 0$. Since $u^{(N)}$ solves the least squares problem (67) we have that

$$\|Tu^{(N)} - v\|_V \le \|T\pi_{W_N}(u^\dagger) - v\|_V$$

for all $N \in \mathbb{N}$.

Since $\{u_1, u_2, \dots\}$ is a Hilbert basis of U we have that

$$\pi_{W_N}(u^\dagger) = \sum_{j=1}^{N} \langle u^\dagger, u_j \rangle u_j$$

and thus

$$\lim_{N\to\infty} \pi_{W_N}(u^\dagger) = \lim_{N\to\infty} \sum_{j=1}^{N} \langle u^\dagger, u_j \rangle u_j = u^\dagger.$$

Because of the boundedness of T, this implies that

$$\lim_{N\to\infty} T\pi_{W_N}(u^\dagger) = Tu^\dagger = v.$$

Thus

$$\lim_{N\to\infty} \|Tu^{(N)} - v\|_V \leq \lim_{N\to\infty} \|T\pi_{W_N}(u^\dagger) - v\|_V = 0.$$

□

REMARK 5.6.4. In the proof of the previous result, we actually obtain the concrete estimate

$$\|u^{(N)} - u^\dagger\|_U \leq \|T\|_{L(U,V)} \|T^{-1}\|_{L(V,U)} \|u^\dagger - \pi_{W_N}(u^\dagger)\|,$$

which can be rewritten as

$$\|u^{(N)} - u^\dagger\|_U \leq \|T\|_{L(U,V)} \|T^{-1}\|_{L(V,U)} \sum_{n=N+1}^{\infty} |\langle u^\dagger, u_i \rangle|^2.$$

The approximation error thus depends on the decay properties of the generalised Fourier coefficients of the true solution $u^\dagger$; the faster these decay, the better the approximation.