Linear Methods

Markus Grasmair

Trondheim, Norway

Lecture notes for the course TMA4145 - Linear Methods, Autumn 2024, NTNU

Contents

CHAPTER 0

Mathematical Language

When you open a maths book, you will immediately note that it does not look the same as other books, or even textbooks in other disciplines. There are of course the obvious differences like the prevalent use of mathematical symbols and the frequent appearance of equations, both incorporated in the running text and as separate lines or even paragraphs. Then there is the strangely fragmented structure of the text with definitions, theorems, lemmas, proofs, remarks, and so on, but often hardly any text in between. The differences, however, go deeper than that, and you will notice that the language itself looks weird. Especially when reading through definitions and theorems, you might get the impression that they are written in an unnecessarily complicated and convoluted manner, and these complications can look even worse in the case of proofs. In the following, we will briefly discuss the differences between mathematical language and everyday language. Many of the thoughts in this chapter are based on the excellent book [**Dev12**], which intends to ease the transition from high school to university mathematics. Even though you are not the main target group of that book, having already put up with several advanced mathematic courses, it can still make sense to skim through it.

Mathematical Statements

If you use language for everyday communication, you usually do not have to be extremely precise. If something is unclear, you always can ask additional questions. In written communication, this is not always possible, but you still have some context and common sense to your disposal, which can clear up most misunderstandings. In mathematics, you no longer have this luxury, as soon as you are dealing with more complicated objects than integers or rational numbers, or with more complex subjects than spatial geometry. Because of this, we have to be very careful how to use language to talk about abstract mathematical objects.

Modern mathematics is largely concerned with investigating whether statements about mathematical objects are true or false. Examples of such statements are the following:

(1) The number 7 is a prime number. The number 6 is not a prime number.
(2) The number $\sqrt{2}$ is irrational.
(3) Every quadratic equation $x^2 + bx + c = 0$ has at least two solutions.
(4) Every even natural number greater than 2 can be expressed as the sum of two prime numbers.

The two statements in (1) are true. The statement (2) is true as well. The statement (3) is wrong. The statement (4) is *Goldbach's conjecture*; at the time of the writing of this note, it was not yet known whether this is true or false.

In order to find the truth value of a mathematical statement, it is necessary to actually know what the different expressions in the statement mean. That is, we need a sufficiently precise *definition* of all the expressions. For instance, in order to determine whether the statement that 7 is a prime number is true, we first need to know what a prime number is. One possibility for defining prime numbers is the

following: *"A natural number p is a prime number, if there do not exist numbers a and b, both strictly smaller than p, such that $p = ab$."* With this definition at hand, we can then actually verify, or *prove*, that 7 is a prime number: All we have to do is to simply compute all the different products ab with natural numbers a and b strictly smaller than 7. Since the number 7 never appears as such a product, it is prime. Conversely, one possibility for showing that 6 is not a prime number is to find numbers a and b strictly smaller than 6 such that their product is 6; for instance, the numbers $a = 3$ and $b = 2$ would do.

Of course, everyone reading this text knew in advance what a prime number was, and in particular that 7 is prime. Thus the necessity of both a precise definition of primes and a proof that 7 is a prime might look quite a bit superficial and superfluous.[1] If we are dealing with less intuitive concepts, however, the situation is different. Then, in order to avoid misunderstandings, precise definitions are crucial. An important example here, which you have seen in your first calculus class, is that of continuity. Although we might think that we have a good intuitive understanding of what it means for a function to be continuous, this intuition can get challenged by complicated functions. It is pretty clear that the function $f(x) = x^2$ is continuous, whereas the function $f(x) = \operatorname{sgn}(x)$ is not. But what about, say, the function $f(x) = \sin(1/x)$, where we define $f(0) := 0$? In order to be able to make a statement of its continuity, we need a precise definition.

Connectors and Quantifiers

Most mathematical statements are built up of simpler statements using the connectors "and", "or", "not", and "if ... then". Moreover, the *quantifiers* "for all" and "there exists" are regularly needed. While these innocent looking terms also appear in everyday language,[2] there are some notable differences between their everyday meanings and their mathematical meanings. In the following, we will briefly review these main connectors and quantifiers in terms of their effect on the truth values of the component expressions.

And. The statement (ϕ and ψ) is true precisely when both statements ϕ and ψ are true. If at least one statement (or both) are false, it is false as well.

In everyday language, the word "and" can have a slightly different meaning than the mathematical connector "and" and also convey a temporal or final relationship: In the sentence "The sun went down and it became dark," we are really talking about a sequence of events: *First*, the sun went down, and *then* it became dark. We might even imply that the first part of the sentence is the *cause* of the second one.[3] In mathematical expressions, the word "and" is never used in that sense. The mathematical statements (ϕ and ψ) and (ψ and ϕ) are precisely the same, and the order of the components ϕ and ψ does not matter. Consider in contrast the sinister meaning of the reversed sentence "It became dark and the sun went down."

Or. The statement (ϕ or ψ) is true precisely when at least one of the statements ϕ or ψ (or both) are true. It is only false, when both ϕ and ψ are false. For instance, the statement

$$(7 \text{ is a prime}) \text{ or } (7 < 0)$$

[1] Although, there is sometimes doubt about the status of the number 1: Is it a prime or not? The definition above should clarify this doubt!

[2] At least the connectors do so pretty regularly; the appearance of the quantifiers is probably much more sparse.

[3] You should not jump to conclusions that fast, though: Do not fall easy prey to the fallacy *post hoc ergo propter hoc* and always remember that correlation does not imply causation!

is true, as at least one of the component statements (namely the statement that 7 is a prime) is true; it does not matter that the statement $7 < 0$ is false.

In addition, we have to be a bit careful when augmenting an "or" statement to an "either ... or" statement. In everyday language these two phrases are often, but not always, used interchangeably, but the "either ... or" can also be understood as an *exclusive* or, which is true whenever *precisely* one of the component statements is true, but not both. *In this class, I will use the phrase "either ... or" interchangeably with "or". If I need to make use of an exclusive or, I will use more complicated expressions.*[4]

Not. The negation inverts the truth statement of an expression: If the statement ϕ is true, then the statement (not ϕ) is false; if the statement ϕ is wrong, then the statement (not ϕ) is true.

We have the following rules for combining *not* with *and* and *or* (De Morgan's laws of propositional logic):

$$(\text{not}(\phi \text{ and } \psi)) = ((\text{not } \phi) \text{ or } (\text{not } \psi)),$$
$$(\text{not}(\phi \text{ or } \psi)) = ((\text{not } \phi) \text{ and } (\text{not } \psi)).$$

If — then. The statement (if ϕ then ψ) is *false* only in the case where ϕ is true, but ψ is false. That is, in the case where the statement ϕ is false, the statement (if ϕ then ψ) is true no matter the truth value of ψ. This fact is known as *ex falso quodlibet*, roughly translating to "from the wrong, anything [follows]": If one starts with a wrong premise (that is, if ϕ is false), then one cannot conclude anything about the conclusion (that is, we don't know whether ψ is true or false), even if the deduction process (that is, the statement (if ϕ then ψ)) was completely correct.

Many mathematical theorems can be formulated in terms of "if — then" statements. As a consequence, reformulations of such statements are regularly used for the proof of theorems. The probably most important reformulations are the following two:

$$(\text{if } \phi \text{ then } \psi) = (\text{if } (\text{not } \psi) \text{ then } (\text{not } \phi)),$$
$$(\text{if } \phi \text{ then } \psi) = \text{not}(\phi \text{ and } (\text{not } \psi)).$$

The first reformulation is used in *proofs by contraposition*: In order to show the truth of the statement (if ϕ then ψ), we assume that the conclusion ψ does not hold (that is, we assume (not ψ)) and try to conclude that ϕ does not hold either (that is, we try to show that (not ϕ) holds). The second reformulation is used in *proofs by contradiction*: Again we are trying to show that the statement (if ϕ then ψ) holds. For that, we assume that ϕ and the negation of ψ are simultaneously true, and then we try to show that this implies a false statement.

Be aware that the common *meaning* of the word "implication" is different from the *truth value* of an "if ... then" expression *even in mathematics*. For instance, the statement

$$(2 + 2 = 4) \implies (7 \text{ is a prime})$$

is true as a mathematical statement: Both the antecedent $(2 + 2 = 4)$ and the consequent (7 is a prime) are true statements. However, we would not say (or write) that "the assumption that $2 + 2 = 4$ implies that 7 is a prime," but we rather reserve the notion of "implications" to the case where there is actually a connection between the two statements.

[4]Knowing myself, I will invariably use at some point during the lecture a phrase along the lines of "If a, then either b or c," even though the phrase calls for a non-exclusive or. This is intended as a safeguard against accusations of even more errors than really warranted.

For all, there exists. The statement (for all $x:\ \phi(x)$) is true precisely when the statement $\phi(x)$ is true for every single admissible argument x. Conversely, it is false when there exists at least one admissible argument such that $\phi(x)$ is false. The statement (there exists $x:\ \phi(x)$) is true when the statemet $\phi(x)$ is true for *at least one* (but possibly more) admissible arguments x. The negation of (for all $x:\ \phi(x)$) is (there exists $x:\ $ (not $\phi(x)$)).

Many mathematical statements are "for all" or "there exists" statements (or both) in disguise: For instance, the statement "The number 6 is a prime number." can be translated to (or should be interpreted as) the statement "There exist natural numbers a and b with $a < 6$ and $b < 6$, such that $ab = 6$." Breaking this down further, we may arrive at the statement "There exist natural numbers a and b with the properties $a < 6$ and $b < 6$ and $ab = 6$."

More complicated, the expression "Every quadratic equation $x^2 + bx + c = 0$ has a complex solution" can be translated to "For all numbers b and c, there exists a complex number x with the property $x^2 + bx + c = 0$."

This last example also demonstrates the importance of sentence order: The statement above is true (and you even know an explicit formula for all the solutions of the equation), but if we reverse the order of the "for all" and the "there exists" terms, we obtain the blatantly wrong statement "There exists a complex number x such that all numbers b and c have the property $x^2 + bx + c = 0$." (In order to see that this reversed statement is false, consider *for fixed* x the numbers $b = -x$ and $c = 1$.)

CHAPTER 1

Basic Notions

0. Reading Guide

In this note, definitions, remarks, examples, and exercises end with a ■; proofs in most mathematical texts including this one end with a □. Theorems, propositions, lemmas, and corollaries are written in *italics*, which makes it unnecessary to delineate their boundaries further.

Some of the sections are marked by a * to indicate that they are of (sometimes considerably) increased difficulty and/or abstraction. The content of these sections is not used in the other parts of the notes, and thus you can freely skip over these. The same convention is used for remarks, exercises, proofs, and so on. You may, of course, read these parts as well, but only at your own risk. You have been warned.

1. Sets

Although it is possible to develop a precise notion of sets, I will not do so in these notes. The reason is that such a rigorous introduction would be tedious and difficult, but at the same time would only contribute little to the main subjects of our studies in this course, namely linear mappings between vector spaces. This section is intended to ensure that we have a common understanding about how we can deal with sets. For a classical (and well written) introduction to set theory, I refer to [**Hal74**]. I, in contrast, will rather loosely follow the exposition in the first pages of [**HS75**].

Informally, a "set is any identifiable collection of objects of any sort." The objects contained in a set are called its *elements* or *members*. There are three main possibilities to introduce a new set: First, we can simply list all of its elements and, say, define a set $X := \{\text{“Norwegian”}, \text{“blue”}, \text{“parrot”}\}$. This is usually done for *small*, *finite* sets, but we can use dots (...) for larger or even infinite sets; here I assume that everybody understands what I mean with the sets $\{1, \dots, 1729\}$ or $\{3, 4, 5, \dots\}$. Second, we can define a set by a property of its elements. For instance, we could define a set

$$Y := \big\{x : x \text{ is a dairy product that can be bought in a cheese shop}\big\}.$$

Finally, we can build up a set from other sets using operations like *union*, *intersection*, or *product* of sets (see below).

The symbol $\in$ indicates that an object is a member of a set; the symbol $\notin$ indicates that is is not. For instance, with the definition of X above, we have that “parrot” $\in X$, but “Norwegian blue” $\notin X$.

Definition 1.1 (Subset and proper subset). Assume that X and Y are sets. We say that X is a *subset* of Y, denoted $X \subset Y$, if for every $x \in X$ we have that $x \in Y$; otherwise we write $X \not\subset Y$. If $X \subset Y$ and $Y \subset X$, then we write $X = Y$ (and the sets X and Y are equal); else $X \neq Y$. If $X \subset Y$ but $X \neq Y$, then we say that X is a *proper subset* of Y, denoted $X \subsetneq Y$. ■

Remark 1.2. There is an alternative school of notation that uses the symbol $\subseteq$ to denote subsets, and $\subset$ to denote proper subsets. *We do not follow this notation*

in this course. However, it is important to know that this alternative notation exists, especially when you are reading additional support material. In this case, it is important that you check what notation is used there. ■

REMARK 1.3. If we want to show that two sets, say X and Y, are equal, we have to show that they contain precisely the same elements. Thus, the basic approach for a proof is, first to take an arbitrary element $x \in X$ and prove that it is also contained in Y, and then to take an arbitrary element $y \in Y$ and prove that it is also contained in X. ■

DEFINITION 1.4 (Empty set). By $\emptyset$ we denote the empty set, which has no elements. ■

Statements about the empty set always appear somewhat awkward, and it can be difficult to formulate proofs, as the "standard start" of picking some element $x \in \emptyset$ does not really work. Thus it can be helpful to rely on proofs by contradiction. As an example, we prove the following (really trivial) statement:

LEMMA 1.5. *For every set X we have $\emptyset \subset X$.*

PROOF. Let X be an arbitrary set, and assume to the contrary that $\emptyset \not\subset X$. Then there exists some $y \in \emptyset$ such that $y \notin X$. Because of the definition of the emptyset, however, such a y cannot exist, and thus we have a contradiction. As a consequence, the assumption $\emptyset \not\subset X$ is wrong, or, put differently, $\emptyset \subset X$. □

REMARK 1.6. In many cases, the elements of a set can be sets themselves. As an example, a line is a set of point. Thus, if we are talking about the set of all lines in the plane, we are in fact talking about a set consisting of sets. ■

DEFINITION 1.7 (Union, intersection, difference, and product). Assume that X and Y are sets.

- The *union* $X \cup Y$ is the set of all elements contained in at least one of the sets X and Y. That is,
$$x \in X \cup Y :\iff x \in X \text{ or } x \in Y.$$
- The *intersection* $X \cap Y$ is the set of all elements contained in both of the sets X and Y. That is,
$$x \in X \cap Y :\iff x \in X \text{ and } x \in Y.$$
- The *difference* $X \setminus Y$ of the sets X and Y is the set of all elements contained in X but not in Y. That is,
$$x \in X \setminus Y :\iff x \in X \text{ and } x \notin Y.$$
- The *Cartesian product* (or simply *product*) of the sets X and Y is the set $X \times Y$ consisting of all ordered pairs (x, y) such that
$$(x, y) \in X \times Y :\iff x \in X \text{ and } y \in Y.$$

■

DEFINITION 1.8 (Power set). Let X be a set. By $\mathcal{P}(X)$ we denote the *power set* of X, which consists of all subsets of X. That is,

$$S \in \mathcal{P}(x) :\iff S \subset X.$$

■

EXAMPLE 1.9. Let $X := \{\text{“Norwegian”}, \text{“blue”}\}$. Then

$$\mathcal{P}(X) = \big\{\emptyset, \{\text{“Norwegian”}\}, \{\text{“blue”}\}, \{\text{“Norwegian”}, \text{“blue”}\}\big\}.$$

Note here that the elements of $\mathcal{P}(X)$ are themselves sets. In particular, the *set* $\{\text{“blue”}\}$ is *not the same* as the *element* “blue”. Note also that both statements $\emptyset \subset \mathcal{P}(X)$ and $\emptyset \in \mathcal{P}(X)$ are simultaneously true: the empty set is a subset of $\mathcal{P}(X)$ (as it is a subset of every set), but at the same time, the empty set appears amongst the elements of $\mathcal{P}(X)$. ■

EXERCISE 1.1. Assume that X is a set. Determine the following sets:

- The set $X \cup \emptyset$.
- The set $X \cap \emptyset$.
- The set $X \setminus \emptyset$.
- The set $\emptyset \setminus X$.
- The set $X \times \emptyset$.
- The set $\mathcal{P}(\emptyset)$.

■

DEFINITION 1.10 (Complement). Assume that X is a set and $U \subset X$ is a subset. By $U^C := X \setminus U$ we denote the *complement* of U (in X). ■

REMARK 1.11. When we are working sets, we do not care about the “order” in which elements appear in a set, nor the “multiplicity”. Thus the sets $\{\text{“blue”}, \text{“parrot”}\}$, $\{\text{“parrot”}, \text{“blue”}\}$, and $\{\text{“blue”}, \text{“parrot”}, \text{“blue”}\}$ are all the same (as they contain the same elements), even though they look different at first glance. ■

REMARK 1.12* (Russell’s paradox). When we specify a set by the properties of its elements, we have to be very careful not to destroy the very fabric of mathematics. In fact, if we allow for the construction of *arbitrary* sets in that way, we readily run into contradictions: Take for instance the “set”

$$X := \big\{Y : Y \text{ is a set}\big\},$$

that is, the set containing *all* possible sets (including itself) as elements. Then we can consider the subset

$$Z := \big\{Y \in X : Y \notin Y\big\}$$

of all sets that do not contain itself as an element. Now we can ask the question, whether the set Z is an element of itself, that is, whether $Z \in Z$ holds, or not. The answer to this question is that neither can logically be the case:

- Assume first that the inclusion $Z \in Z$ holds, that is, that Z contains itself as an element. By the definition of Z as the set of all sets that do *not* contain itself as an element, this would further imply that Z is not contained in Z. This, however, contradicts the initial assumption $Z \in Z$, and therefore that assumption cannot hold.
- Now assume conversely that $Z \notin Z$. Since Z consists of *all* sets that do not contain itself as an element, the statement $Z \notin Z$ then implies that Z satisfies the description of the elements of Z, and thus we can conclude that $Z \in Z$. Again, this contradicts the initial assumption $Z \notin Z$.

Thus neither of the statements $Z \in Z$ or $Z \notin Z$ can be true, and we have a contradiction.

The common solution is to deny the object X the status of a “set.” More precisely, we allow the specification of sets by the properties of their elements only in the case of subsets. Given an *already existing* set U (which may of course contain

elements that are themselves sets) and some property $p(x)$ for elements in U, we are allowed to define a set of the form

$$\big\{x \in U : \text{ the property } p(x) \text{ holds}\big\} \subset U.$$

However, the same construction is not allowed without some (explicit or implicit) specification of the superset U.

Still, one sometimes want to talk simultaneously about all different sets, or, in this course rather, simultaneously about all possible real or complex vector spaces. In order to do so, one can use the notion of a *class* instead of a set, which is (very informally) a collection of sets defined by a common property, but is not a set itself. Thus one can talk about the *class of all sets* or the *class of all vector spaces*. Now, the contradiction we have observed above is no longer there: The class of all sets that do not contain itself as elements is a perfectly fine class. Moreover, it trivially does not contain itself as an element, as it only contains sets, not (proper) classes. ■

2. Functions

Informally spoken, a function $f\colon X \to Y$ from set X to a set Y is an operation that assigns to each element $x \in X$ a unique element $y := f(x) \in Y$. Formally, within the context of set theory, we can define a function as follows:

DEFINITION 1.13 (Function). Let X and Y be sets. A *function* from X to Y, written $f\colon X \to Y$, is a subset $f \subset X \times Y$ such that there exists for each $x \in X$ precisely one $y \in Y$ with $(x, y) \in f$. Given $x \in X$, we denote by $f(x)$ the unique element in Y such that $(x, f(x)) \in f$. ■

REMARK 1.14. In this course, we will not make use of the set theoretic definition (and notation) of a function, but we will rather use the "standard" interpretation as an operation. Feel therefore free to forget or ignore the set theoretic Definition 1.13. ■

DEFINITION 1.15 (Domain and codomain). Assume that X and Y are sets and that $f\colon X \to Y$ is a function. We say that X is the *domain* of f and Y the *codomain* of f. ■

REMARK 1.16. In this course, we will use the word *function* interchangeably with *mapping*. Sometimes, we will also use the words *transformation* or *operator*, though these will be mostly reserved for linear mappings. ■

DEFINITION 1.17 (Injectivity, surjectivity, and bijectivity). Assume that X and Y are sets and that $f\colon X \to Y$ is a function.

- The function f is *injective*, if there exists for each $y \in Y$ *at most* one $x \in X$ such that $f(x) = y$.
- The function f is *surjective*, if there exists for each $y \in Y$ *at least* one $x \in X$ such that $f(x) = y$.
- The function f is *bijective*, if there exists for each $y \in Y$ *precisely* one $x \in X$ such that $f(x) = y$.

■

LEMMA 1.18 (Inverse). *Assume that X and Y are sets and that $f\colon X \to Y$ is a bijective function. Then there exists a unique function $f^{-1}\colon Y \to X$, the* inverse *of f, such that*

$$f(f^{-1}(y)) = y \qquad \text{for all } y \in Y$$

and

$$f^{-1}(f(x)) = x \qquad \text{for all } x \in X.$$

PROOF. We have to show that a function $f^{-1}\colon Y \to X$ with these properties *exists*, and that it is *unique*. We start with the existence of f^{-1}, which we prove by actually constructing f^{-1}.

Let $y \in Y$ be arbitrary. Since f is bijective, there exists a unique $x \in X$ such that $f(x) = y$. Thus we can define $f^{-1}(y) := x$. Doing so for each $y \in Y$, we obtain a function $f^{-1}\colon Y \to X$. Moreover, by construction, we have that $f(f^{-1}(y)) = y$ for each $y \in Y$.

Now let $x \in X$ be arbitrary and denote $y := f(x)$. Then

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

Now the injectivity of f implies that $x = f^{-1}(f(x))$. Thus the function f^{-1} has the desired properties.

Assume now that $g\colon Y \to X$ is another function such that $f(g(y)) = y$ for all $y \in Y$ and $g(f(x)) = x$ for all $x \in X$. Let moreover $y \in Y$. Then $f(g(y)) = y = f(f^{-1}(y))$. Now the injectivity of f implies that $g(y) = f^{-1}(y)$. Since y was arbitrary, this proves that $g = f^{-1}$. Thus f^{-1} is the unique function with these properties. □

DEFINITION 1.19 (Composition). Assume that X, Y, Z are sets, and that $f\colon X \to Y$ and $g\colon Y \to Z$ are functions. We define the *composition* of g and f as the function $g \circ f\colon X \to Z$,

$$g \circ f(x) := g(f(x)) \qquad \text{for all } x \in X.$$

■

LEMMA 1.20. *Assume that X, Y, Z are sets, and that $f\colon X \to Y$ and $g\colon Y \to Z$ are functions.*

- *If $g \circ f$ is injective, then f is injective.*
- *If $g \circ f$ is surjective, then g is surjective.*
- *If g and f are injective, then $g \circ f$ is injective.*
- *If g and f are surjective, then $g \circ f$ is surjective.*
- *If g and f are bijective, then g and f are bijective, and we have that*

$$(g \circ f)^{-1} = f^{-1} \circ g^{-1}.$$

PROOF. Exercise! □

EXERCISE 1.2. In Lemma 1.20, some implications are missing—for the excellent reason that they do not hold. In this exercise, you are tasked with finding counterexamples to these missing implications. That is, find sets X, Y, Z and functions $f\colon X \to Y$ and $g\colon Y \to Z$ such that:

- $g \circ f$ is injective, but g is not injective.
- $g \circ f$ is surjective, but f is not surjective.
- $g \circ f$ is bijective, but neither g nor f is bijective.

■

DEFINITION 1.21 (Restriction). Assume that X and Y are sets and that $f\colon X \to Y$ is a function. Assume moreover that $U \subset X$ is a subset. By $f|_U$ we denote the *restriction of f to U*, defined as the function $f|_U\colon U \to Y$ satisfying

$$f|_U(x) := f(x) \qquad \text{for all } x \in U.$$

■

DEFINITION 1.22 (Image and pre-image). Assume that X and Y are sets and that $f\colon X \to Y$ is a function. If $A \subset X$ is a subset of X we denote by

$$f(A) := \big\{y \in Y : \text{there exists } x \in A \text{ with } f(x) = y\big\}$$

the *image of A under f*.

If $B \subset Y$ is a subset of Y we denote by

$$f^{-1}(B) := \big\{x \in X : f(x) \in B\big\}$$

the *pre-image of B under f*. ■

REMARK 1.23. Assume that $f\colon X \to Y$ is a bijective function with inverse f^{-1}. Then, if $B \subset Y$ is a subset of Y, we can consider both the *pre-image* of B under the function f and the *image* of B under the function f^{-1}, and we actually use the same notation for these two sets. Fortunately, this is no problem, because they are actually the same.

In addition, if $B = \{y\} \subset Y$ is a subset consisting of a single element $y \in Y$, we can apply both the inverse f^{-1} to the element y, and the pre-image to the set B. In this case, we obtain (after a brief consideration) that

$$f^{-1}(\{y\}) = \{f^{-1}(y)\}.$$

Note here, that the f^{-1} on the left hand side denotes the pre-image operation, whereas the f^{-1} on the right hand side denotes the inverse function.

Because of this, the fact that f^{-1} denotes two different objects poses no difficulties in practice. *However, it is important to remember that the pre-image operation is also defined for non-invertible functions.* In particular, using the notion "f^{-1}" for the pre-image does by no means imply that the function f is actually invertible. ■

PROPOSITION 1.24. *Assume that X and Y are sets and that $f\colon X \to Y$ is a function. For every subset $A \subset X$ we have*

$$A \subset f^{-1}\big(f(A)\big).$$

For every subset $B \subset Y$ we have

$$f\big(f^{-1}(B)\big) = B.$$

PROOF. Exercise! □

3. Families and Sequences

Recall that sets are *unordered* collections of objects *without repetition*. In many cases, however, we will need "sets" where the order is important and repetitions might occur. For this type of "sets", we introduce the notion of a *family*, or, informally, an *indexed set*.

DEFINITION 1.25 (Family). Assume that X is a set. A *family* in X is a function $x\colon I \to X$ for some set I. The set I is called the *index set* of the family and the values $x(i)$ are the members of the family. Usually, we denote a family in the form $(x_i)_{i\in I}$, where I is the index set and $x_i := x(i)$ are the *members* or *terms* of the family. ■

In order to differentiate between sets and families, we will use curly brackets $\{\cdot\}$ to enclose sets, but round brackets $(\cdot)$ to enclose families.

REMARK 1.26. If the index set of a family is the set $\{1, 2, \dots, n\}$ for some $n \in \mathbb{N}$, we rather speak of a *tuple* instead of a family. Also, we usually use in this case either the notation $(x_1, x_2, \dots, x_n)$ or $(x_i)_{i=1,\dots,n}$. ■

EXAMPLE 1.27. In order to see the difference between a family and a set, let us first consider the *set*

$$X := \big\{x_k \in \mathbb{R} : x_k = (-1)^k,\ k \in \mathbb{N}\big\} = \big\{(-1)^k : k \in \mathbb{N}\big\}.$$

By defining the set X in that way, we might get the impression that X has infinitely many elements, but this is in fact not the case: A set is defined only through the *distinct* elements it contains, and it does not matter how many times the same element appears in the definition; the notion of "multiplicity" of elements in a set is not defined at all. Next, we could also write the set X as

$$X = \{-1, +1, -1, +1, -1, \ldots\}.$$

Still, this might give a quite wrong impression about X as a set, as it looks like -1 would be the "first" element of X, then $+1$ would be the "second", -1 would be again the "third", and so on. However, the order in which its elements appear in the definition plays no role at all, and it does not make sense to talk about "the first element" of a set. The set X simply is

$$X = \{+1, -1\} = \{-1, +1\}.$$

Now consider instead the *family*

$$Y := ((-1)^k)_{k\in\mathbb{N}} = (-1, +1, -1, +1, \ldots).$$

Here it is actually true to say that this family has infinitely many members, although it, obviously, only has 2 *distinct* ones. Also, -1 is the first term, $+1$ is the second term, and so on. ■

In the special case where the index set are the natural numbers, we usually speak of a sequence instead of a family:

Definition 1.28 (Sequence). Assume that X is a set. A *sequence* in X is a function $x\colon \mathbb{N} \to X$. Usually, we denote a sequence in the form $(x_i)_{i\in\mathbb{N}}$, where $x_i := x(i)$ are the *terms* of the sequence. ■

Remark 1.29*. Although we won't need the notion of a subsequence in this course, we can use the definition of sequences as functions in order to obtain a precise definition of subsequences as well:

Assume that $x\colon \mathbb{N} \to X$ is a sequence in X. A *subsequence* of x is the composition $x \circ k\colon \mathbb{N} \to X$ of x with a strictly increasing function $k\colon \mathbb{N} \to \mathbb{N}$ (in the sense that $k(i) > k(j)$ whenever $i > j$). Usually, we denote a subsequence in the form $(x_{k(i)})_{i\in\mathbb{N}}$, where $(x_i)_{i\in\mathbb{N}}$ is the original sequence. ■

In a slight abuse of notation, we extend the notion of subsets to families:

Definition 1.30. Assume that X is a set and that $(x_i)_{i\in I}$ is a family in X. Assume moreover that $Y \subset X$ is a subset of X. We say that the family is contained in Y, denoted $(x_i)_{i\in I} \subset Y$, if each member of the family is contained in Y, that is, if $x_i \in Y$ for all $i \in I$. ■

Definition 1.31 (Subfamily). Assume that X is a set, that $S := (x_i)_{i\in I}$ is a family in X. We say that T is a *subfamily* of S, if there exists a subset $J \subset I$ such that $T = (x_j)_{j\in J}$. In this case, we write $T \subset S$. ■

4. Real Numbers

There are different approaches to the definition of the real numbers, all of which are somewhat complicated and not necessarily intuitive. Because of that, we will abstain from providing a precise definition, but rather assume that everybody has an intuitive understanding of what we mean by the real numbers. At some point during the course (specifically, in Section 4), however, we have to make use of the *completeness* of real numbers, a concept that we will introduce only then. In this section, we will therefore introduce the main property of the real numbers that will be required then.

DEFINITION 1.32 (Lower and upper bound). Let $S \subset \mathbb{R}$.

- We say that $y \in \mathbb{R} \cup \{\pm\infty\}$ is a *lower bound* of S, if $y \leq x$ for all $x \in S$.
- We say that $z \in \mathbb{R} \cup \{\pm\infty\}$ is an *upper bound* of S is $z \geq x$ for all $x \in S$.
- The set S is *bounded*, if it has both an upper and lower bound within the real numbers $\mathbb{R}$.

■

As an example, -29 is a lower bound of the half-open interval $(0, 1]$, whereas 1 is an example of an upper bound. We have for all subsets $S \subset \mathbb{R}$ that $+\infty$ is an upper bound of S and $-\infty$ is a lower bound of S. For the particular case $S = \emptyset$, every element of $\mathbb{R} \cup \{\pm\infty\}$ is simultaneously a lower bound and an upper bound.

DEFINITION 1.33 (Infimum and supremum). Let $S \subset \mathbb{R}$.

- We say that $y \in \mathbb{R} \cup \{\pm\infty\}$ is a *supremum* of S, if y is an upper bound, and $y \leq z$ for all upper bounds z of S.
- We say that $y \in \mathbb{R} \cup \{\pm\infty\}$ is an *infimum* of S, if y is a lower bound of S and $y \geq z$ for all lower bounds z of S.

■

Put differently, the supremum of a set S is the smallest upper bound of S, whereas the infimum is the largest lower bound. In the next proposition, we show that the use of the definite article in the previous sentence was actually justified.

PROPOSITION 1.34. *The supremum and infimum of a set $S \subset \mathbb{R}$ are unique, if they exist.*

PROOF. Assume that y and z are two suprema of S. Then we have by definition that $y \leq z$ (because y is a supremum) and $z \leq y$ (because z is a supremum), and therefore $y = z$. The proof of the uniqueness of infima is similar. □

PROPOSITION 1.35. *Assume that $S \subset \mathbb{R}$ is a non-empty bounded set.*

- *We have that $x = \sup S$, if and only if $x \geq s$ for all $s \in S$ and there exists for every $\varepsilon > 0$ some $s_\varepsilon \in S$ such that $s_\varepsilon > x - \varepsilon$.*
- *We have that $y = \inf_S$, if and only if $y \leq s$ for all $s \in S$ and there exists for every $\varepsilon > 0$ some $s_\varepsilon \in S$ such that $s_\varepsilon < y + \varepsilon$.*

PROOF. We only show the first part of the claim, as the second is similar.

Assume that $x = \sup S$. Then x is an upper bound of S and therefore $x \geq s$ for every $s \in S$. Now assume to the contrary that there exists some $\varepsilon > 0$ such that $s \leq x - \varepsilon$ for all $s \in S$. Then $x - \varepsilon$ is also an upper bound of S, but $x - \varepsilon < x$. This contradicts the assumption that x is the smallest upper bound of S.

Conversely, assume that $x \in \mathbb{R}$ is such that $x \geq s$ for all $s \in S$ and there exists for every $\varepsilon > 0$ some $s_\varepsilon \in S$ such that $s_\varepsilon > x - \varepsilon$. Then clearly x is an upper bound of S. Assume now to the contrary that x is not the smallest upper bound of S. Then there exists some upper bound z of S with $z < x$. In particular, $s \leq z$ for all $s \in S$. Define now $\varepsilon := x - z$. Then there exists $s_\varepsilon \in S$ with $s_\varepsilon > x - \varepsilon = z$. This contradicts the assumption that z was an upper bound of S. Thus x is indeed the smallest upper bound of S, which proves the assertion. □

A central property of the real numbers (which we state here as an axiom) is that infimum and supremum of arbitrary sets actually exist.

AXIOM 1.36. *Every subset of $\mathbb{R}$ has a supremum and an infimum in $\mathbb{R} \cup \{\pm\infty\}$.*

DEFINITION 1.37 (Inf and sup). Let $S \subset \mathbb{R}$. By $\inf S$ and $\sup S$ we denote the infimum and supremum of S, respectively. ■

EXAMPLE 1.38. The infimum of the half-open interval $[0,1)$ is 0, and the supremum is 1. More general, if $I = (a,b)$ is any interval in $\mathbb{R}$, then $\inf I = a$ and $\sup I = b$, and the same holds if we replace the open interval by the closed interval or any half-open interval.

Moreover, we have that

$$\inf \mathbb{R} = -\infty \qquad \text{and} \qquad \sup \mathbb{R} = +\infty,$$

whereas[1]

$$\inf \emptyset = +\infty \qquad \text{and} \qquad \sup \emptyset = -\infty.$$

■

PROPOSITION 1.39. *Infimum and supremum have the following properties:*

- *If $S \subset T \subset \mathbb{R}$, then*
$$\inf T \leq \inf S \qquad \textit{and} \qquad \inf S \leq \inf T.$$
- *If S is a non-empty set and $x \in S$, then*
$$\inf S \leq x \leq \sup S.$$
- *For all sets $S \subset \mathbb{R}$ we have that*
$$\inf(-S) = -\sup S,$$
where $-S := \{-x : x \in S\}$.

PROOF. Exercise! □

5*. Choice

Modern mathematics is built upon the foundation of axiomatic set theory. However, this foundation is buried so deep under lots of familiar looking definitions and theorems that you *almost* never will come in contact with it, unless you want to specialise in set theory or in logic. There is one major exception, though, in the form of a set theoretic axiom that consistently shows itself in unexpected places, namely the axiom of choice. In this class, we will, for instance, meet this axiom when we will discuss the existence of bases of arbitrary vector spaces.

In this section, I will present a *brief* overview over a naive approach to axiomatic set theory up to and including the notorious axiom of choice. I will largely follow here the basic approach of [**Hal74**], but will skip over almost all of the details; you can find them there if you are interested. Let us thus start our journey into the mathematical underworld with the fitting "lasciate ogne speranza, voi ch'intrate." Or, on a slightly more positive note, quoting Paul Halmos [**Hal74**, p. vi]: "[G]eneral set theory is pretty trivial stuff really, but, if you want to be a mathematician, you need some, and here it is; read it, absorb it, and forget it."

In axiomatic approaches to set theory, one generally ignores the question of what sets actually *are*, but rather focusses what one can *do* with sets. Essentially, one specifies operations that are allowed with sets and that result in other sets being formed.

AXIOM 1.40* (Axiom of Extension). *Two sets are equal, if and only if the contain the same elements.*

This is just a rewording (and axiomatisation) of Definition 1.1.

AXIOM 1.41* (Axiom of Specification). *If Y is a set and $p(x)$ is a statement about the elements $x \in Y$, then there exists a set $X \subset Y$ containing precisely the elements $x \in Y$ for which the statement $p(x)$ is true.*

[1] Do verify the next statement!

In other words, the "object" $\{x \in Y : p(x) \text{ is true}\}$ is actually a set. Note here that we explicitly restrict ourselves to subsets of a given set, thereby sidestepping the issues discussed in Remark 1.12*.

AXIOM 1.42* (Axiom of Pairing). *If X and Y are sets, there exists a set Z containing both X and Y.*

Note here that the elements of the set Z are the *sets* X and Y, not the elements of these sets. The latter is only taken care of in the next axiom:

AXIOM 1.43* (Axiom of Unions). *For every collection of sets, there exists a set containing each element of each set in that collection.*

Put (slightly) differently, we can form the union of arbitrarily many sets.

AXIOM 1.44* (Axiom of Powers). *For every set X, there exists a set that contains all the subsets of X as elements.*

In particular, this means that the power set $\mathcal{P}(X)$ of any set X is again a set.

Note that all axioms until now build on already existing sets, but we have not yet stated that "sets" actually exist. We will finally do so in the next axiom:

AXIOM 1.45* (Axiom of Infinity). *The set $\mathbb{N}$ of natural numbers exists.*

REMARK 1.46*. Usually, this axiom is formulated quite a bit differently: One starts by *defining* $0 := \emptyset$, then defines $1 := \{0\} = \{\emptyset\}$, $2 := \{0, 1\} = \{\emptyset, \{\emptyset\}\}$, $3 := \{0, 1, 2\}$, and so on. The axiom of infinity then is formulated at stating that the empty set $\emptyset$ exists, as well as a set that contains all the sets 0, 1, 2, ..., constructed in that manner.

In particular, we obtain from this construction that, viewed from a set theoretical perspective, the set of natural numbers is $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, starting at 0. However, a guesstimated majority of mathematicians excludes 0 from belonging to the natural numbers and defines $\mathbb{N} = \{1, 2, 3, \ldots\}$ starting at 1. I personally belong to the latter camp as a mathematician and to the former camp when programming, but really do not bother too much. You can therefore assume that the set $\mathbb{N}$ does not contain the number 0 whenever I use it as an index set for a sequence, although it mostly should not matter. At the same time, I am not offended if you follow the alternative convention and insist that 0 is natural.[2] ■

With these axioms, it is already possible to perform most constructions you will ever need during your mathematical life (without you even ever noticing this). In particular, it is possible (though by no means simple or intuitive) to define what the Cartesian product of two sets is,[3] which then allows us to define the notions of functions and families. Now we are able to define the Cartesian product of arbitrary families of sets.

DEFINITION 1.47*. Assume that $(X_i)_{i \in I}$ is a collection of sets and denote by

$$X := \bigcup_{i \in I} X_i$$

[2]As a fine Norwegian compromise position, we could decide to take the average of the two definitions and define $\mathbb{N} = \{1/2, 3/2, 5/2, \ldots\}$, though I do doubt that this definition would gain much traction in practice.

[3]This construction is somewhat roundabout and unintuitive, though, as we do not have the notion of pairs at our disposal: If X and Y are sets, we can define the Cartesian product of X and Y as the set of all sets of the form $\{x, \{x, y\}\}$ with $x \in X$ and $y \in Y$.

the union of this collection. The *Cartesian product* of the collection of sets is the set $\bigtimes_{i\in I} X_i$ defined as

$$\bigtimes_{i\in I} X_i := \big\{x\colon I \to X : x_i \in X_i \text{ for all } i \in I\big\}.$$

■

AXIOM 1.48* (Axiom of Choice). *Assume that $(X_i)_{i\in I}$ is a family of non-empty sets X_i indexed by a non-empty index set I. Then the Cartesian product $\prod_{i\in I} X_i$ is non-empty.*

In view of the definition of the Cartesian product, we can reformulate the Axiom of Choice in terms of functions as follows: If $(X_i)_{i\in I}$ is a family of non-empty sets X_i indexed by a non-empty index set I, there exists a function $f\colon I \to \bigcup_{i\in I} X_i$ such that $f(i) \in X_i$ for all $i \in I$. Such a function is called a *choice function* for this family. Colloquially, we can say that we may (simultaneously) choose from any collection of sets one element from each of the sets in that collection.

At first glance, this axiom is obviously true, and one might think that there is nothing controversial about it: Of course, we can choose elements from sets. However, the axiom does not put any limits on the number and form of sets from which we choose, and it does not provide any concrete method for constructing a choice function. For instance, we can consider the situation where the collection of sets we are interested in is just the power set of $\mathbb{R}$ excluding $\emptyset$, that is, the set of all the non-empty subsets of $\mathbb{R}$. The Axiom of Choice then claims/postulates that it is possible to select from each non-empty subset $X \subset \mathbb{R}$ an element $x \in X$. For certain subsets, such a selection is easily possible: For bounded intervals, for instance, we could simply the midpoint. However, the assertion is that a selection is possible no matter the shape of the subset of $\mathbb{R}$.

We now look at two important equivalent reformulations of the Axiom of Choice intended to illustrate the controversial nature of said axiom, namely the Well-ordering Axiom and Zorn's Lemma. For this, we first have to introduce the notion of ordered sets.

DEFINITION 1.49* (Partial order). Let X be a non-empty set. A *partial order* on X is a subset $R \subset X \times X$ with the following properties:

- *Reflexivity:* For all $x \in X$ we have that $(x, x) \in R$.
- *Anti-symmetry:* If $(x, y) \in R$ and $(y, x) \in R$, then $x = y$.
- *Transitivity:* If $(x, y) \in R$ and $(y, z) \in R$, then $(x, z) \in R$.

Given a partial order R on X, we usually indicate the relation $(x, y) \in R$ by $x \leq_R y$.

A set X together with a partial order on X is called a *partially ordered set*. ■

DEFINITION 1.50* (Total order). Let X be a non-empty set and R be a partial order on X. We say that R is a *total order*, if for all x, $y \in X$, at least one of the relations $x \leq_R y$ or $y \leq_R x$ holds. A set X together with a total order on X is called a *totally ordered set*. ■

EXAMPLE 1.51*. On $\mathbb{R}^n$ we can consider the *standard*, or *componentwise*, partial order $x \leq_R y$, if and only if $x_i \leq y_i$ for all $1 \leq i \leq n$. This is indeed a partial order, as the following considerations show:

- If $x \in \mathbb{R}^n$, then, trivially, $x_i \leq x_i$ for each i, which implies that $x \leq_R x$.
- If $x \leq_R y$ and $y \leq_R x$, then $x_i \leq y_i$ and $y_i \leq x_i$ for each i, which implies that $x_i = y_i$ for each i. Thus $x = y$.
- If $x \leq_R y$ and $y \leq_R z$, then we have for each i that $x_i \leq y_i$ and $y_i \leq z_i$, which implies that $x_i \leq z_i$ for each i. Thus we also have that $x \leq_R z$.

In dimensions $n \geq 2$, this is not a total order. For instance, the vectors $e_1 = (1, 0, 0, \ldots)$ and $e_2 = (0, 1, 0, \ldots)$ are not comparable: Neither of the inequalities $e_1 \leq_R e_2$ or $e_2 \leq_R e_1$ holds. ■

EXAMPLE 1.52*. Let Y be any set and let $X := \mathcal{P}(Y)$ be the power set of Y consisting of all subsets of Y. On X we can define a partial order by $(S, T) \in R$ if $S \subset T$. In general, this is not a total order.[4] ■

DEFINITION 1.53*. Let X be a non-empty set with partial order R. Assume moreover that $S \subset X$ is a non-empty subset of X.

- We say that $x \in X$ is a *lower bound* of S, if $x \leq_R y$ for all $y \in S$.
- We say that $x \in X$ is a *smallest element* of S, if $x \in S$ and $x \leq_R y$ for all $y \in S$.
- We say that $x \in X$ is an *upper bound* of S, if $y \leq_R x$ for all $y \in S$.
- We say that $x \in X$ is a *largest element* of S, if $x \in S$ and $y \leq_R x$ for all $y \in S$.
- We say that $x \in X$ is a *minimal element* of S, if $x \in S$ and there does not exist any $y \in S$, $y \neq x$, such that $y \leq_R x$.
- We say that $x \in X$ is a *maximal element* of S, if $x \in S$ and there does not exist any $y \in S$, $y \neq x$, such that $x \leq_R y$.

■

REMARK 1.54*. It is easy to see (in view of the anti-symmetry of orders) that smallest elements are necessarily unique, if they exist. Thus we may as well speak of *the* smallest element of a set S instead of *a* smallest element, once we have assured that such an element actually exists. In contrast, minimal and maximal elements need not be unique. ■

EXAMPLE 1.55*. Take the set $X = \mathbb{R}^2$ with the componentwise order R and consider the subset $S := \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0,\ x_2 \geq 0\}$. Then $(0, 0)$ is the smallest element of S, as $(0, 0) \leq_R (x_1, x_2)$ for all $(x_1, x_2) \in S$. Now consider the set $T := \{(x_1, x_2) \in \mathbb{R}^2 : x_2 \geq -x_1\}$. Then it is easy to see that there does not exist any $(x_1, x_2) \in T$ such that $(x_1, x_2) \leq_R (0, 0)$. Thus $(0, 0)$ is a minimal element of T. However, (x_1, x_2) is *not* a smallest element of T, as, for instance, $(1, -1) \in T$, but $(0, 0) \not\leq_R (1, -1)$. In fact, the set T has no smallest element at all. ■

DEFINITION 1.56* (Well-ordering). Let X be a non-empty set with total order R. We say that R is a well-ordering of X, if every non-empty subset of X has a least element. A set X with a well-ordering on X is called a *well-ordered set*. ■

EXAMPLE 1.57*. The set $\mathbb{N}$ of natural numbers with the standard order is well-ordered, as every non-empty subset of positive integers has a smallest element. The set $\mathbb{Z}$ of integers is not well-ordered; for instance the whole set $\mathbb{Z}$ has no smallest element, as it is unbounded below. The set $\mathbb{R}_{\geq 0}$ of non-negative real numbers is not well ordered: For instance, the open interval $(1, 2)$ contains no smallest element. ■

THEOREM 1.58* (Well-ordering Theorem). *Every non-empty set can be well-ordered.*

PROOF. See [**Hal74**, Sec. 17]. □

REMARK 1.59*. This theorem of course does not state that every non-empty set X with a total order R *is* already well-ordered, but only that we can find another total order S on X that is a well-ordering. This order S may have nothing to do with R and need not be compatible to any other structure on X. ■

[4]Exercise: Find *all* cases, where this actually *is* a total order!

DEFINITION 1.60*. Assume that X is a partially ordered set with partial order R. A *chain* in X is a subset Y of X such that the restriction of R to Y defines a total order on Y. ■

THEOREM 1.61* (Zorn's Lemma). *Assume that X is a partially ordered set, such that every chain in X has an upper bound in X. Then X contains a maximal element.*

PROOF. See [**Hal74**, Sec. 16]. □

In fact, one can show that the Axiom of Choice, the Well-ordering Theorem, and Zorn's Lemma are all equivalent in the following sense: If we assume that the other axioms of set theory hold, then the validity of either of these three statements implies the validity of the other two. Now, the Axiom of Choice looks, at least at first sight, like a reasonable proposition, whereas the Well-ordering Theorem is far less convincing (can you think of a well-ordering on the real numbers?). The mathematician Jerry L. Bona summed this up as follows: "The axiom of choice is obviously true, the well-ordering principle obviously false, and who can tell about Zorn's lemma?"

CHAPTER 2

Linear Algebra

In this chapter, we will discuss basics of linear algebra up to and including eigendecompositions of linear transformations of vector spaces. The finite dimensional part of this chapter is largely inspired by, and follows the argumentation as, [**Axl24**], though we will deviate at different points.

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 2.1 (Vector space). 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 2.2 (Vector). An element of a vector space is called a *vector*. ■

REMARK 2.3. 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.[1] ■

EXAMPLE 2.4. 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[2] to verify that all the requirements for a vector space (that is, Items (1)–(6) in Definition 2.1) are indeed satisfied in each example.

[1] 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.

[2] 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.

(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 $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_n) \in \mathbb{K}^n$ and $\lambda \in \mathbb{K}$ we define

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

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 $\mathbf{Mat}_{m,n}(\mathbb{K}) = \mathbb{K}^{m\times n}$ 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.

■

Example 2.5 (Function spaces). Assume that S is an arbitrary set and denote by

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

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.

More general, let U be a vector space over $\mathbb{K}$ and let S be an arbitrary set. We denote by

$$U^S := \{f\colon S \to U\}$$

the set of functions on S with values in U. Similarly as above, we can define on U^S addition and scalar multiplication by

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

for u, $v \in U$, and $\lambda \in \mathbb{K}$. With these operations, the set U^S becomes again a vector space.

As particular examples, the set $\mathbb{R}^{\mathbb{R}}$ is the (real) vector space of all real valued functions on $\mathbb{R}$, and the set $\mathbf{Mat}_{m,n}(\mathbb{C})^{\mathbb{R}}$ denotes the (complex) vector space of all functions $f\colon \mathbb{R} \to \mathbf{Mat}_{m,n}(\mathbb{C})$ defined on $\mathbb{R}$ and taking values in the vector space of complex $m \times n$-matrices. ■

Remark 2.6. The usage of $+$ and $\cdot$ for the addition and scalar multiplication in Definition 2.1 is intended to be suggestive and follows the standard notation of these operations. However, it somewhat hides how restrictive these 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 $f\colon V \times V \to V$ and $g\colon \mathbb{K} \times V \to V$ 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 2.7. *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 2.8. *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 2.9. 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 2.10. *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 2.7), it follows that $0 \cdot v = 0$. □

LEMMA 2.11. *Let V be a vector space and $v \in V$. Then $-v = (-1) \cdot v$.*

PROOF. We have that

$$v + (-1) \cdot v \underbrace{=}_{(5)} 1 \cdot v + (-1) \cdot v \underbrace{=}_{(6)} (1 - 1) \cdot v = 0 \cdot v \underbrace{=}_{\text{Lemma 2.10}} 0.$$

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. Linear Transformations.

DEFINITION 2.12. Assume that U, V are vector spaces over $\mathbb{K}$ 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 2.13. For linear transformations it is common to drop the parentheses around the argument. That is, if $T\colon U \to V$ is linear, one usually writes Tu instead of $T(u)$. ■

REMARK 2.14. If the domain and co-domain of a linear transformation T are equal, say for instance that $T\colon U \to U$ is a linear transformation from U to U itself, we usually say that T is a *linear transformation on* U. ■

EXAMPLE 2.15. The following mappings are examples of linear transformations:

(1) The mapping $T\colon \mathbf{Mat}_n(\mathbb{K}) \to \mathbf{Mat}_n(\mathbb{K})$, $A \mapsto TA := A^T$, which takes a matrix and returns its transpose, 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 $D\colon \mathcal{P}(\mathbb{K}) \to \mathcal{P}(\mathbb{K})$, $p \mapsto Dp := p'$, which takes a polynomial and returns its derivative, is linear.

■

EXAMPLE 2.16 (Point evaluation). We now assume that S is a set and U a vector space over $\mathbb{K}$. For fixed $s \in S$, we define the *point evaluation* in s as the mapping $\delta_s\colon U^S \to U$ defined by

$$\delta_s(f) := f(s)$$

for $f \in U^S$. (That is, the input of the mapping is the function f, and the output is the value of f in the point s.) Then this mapping is a linear transformation. ■

The following example shows that we have to be a bit careful when dealing with complex vector spaces, as the scalars λ in the definition of linearity might also be imaginary numbers.

EXAMPLE 2.17. Consider the mapping $T\colon \mathbb{C}^n \to \mathbb{C}^n$,

$$T(x_1, x_2, \ldots, x_n) := T(\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}),$$

where $\overline{x_j}$ denotes the complex conjugate of x_j. Then T is *not* linear.

In order to demonstrate that T is not linear, it suffices to find a single example of a condition for linearity that is not satisfied. Take to that end

$$x := (1, 1, \ldots, 1) \in \mathbb{C}^n.$$

Then

$$Tx = x.$$

However,

$$T(ix) = T(i, i, \ldots, i) = (\overline{i}, \overline{i}, \ldots, \overline{i}) = (-i, -i, \ldots, -i) = -ix.$$

Thus

$$T(ix) = -ix \neq ix = i\,Tx,$$

and thus the mapping T is not linear. ■

EXERCISE 2.1. Given a matrix $A \in \mathbf{Mat}_{m,n}(\mathbb{C})$, we define its *Hermitian conjugate* as the transpose of the componentwise complex conjugate of A, that is,

$$A^H := (\overline{A})^T \in \mathbf{Mat}_{n,m}(\mathbb{C}).$$

Show that the mapping $T\colon \mathbf{Mat}_{m,n}(\mathbb{C}) \to \mathbf{Mat}_{n,m}(\mathbb{C})$,

$$TA := A^H$$

is *not* linear. ■

PROPOSITION 2.18. *Assume that U, V, and W are vector spaces over $\mathbb{K}$, and that $T\colon U \to V$ and $S\colon V \to W$ are linear transformations. Then the composition $S \circ T\colon U \to W$ is a linear transformation as well.*

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

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

□

1.4. Subspaces. We will now introduce the notion of subspaces of vector spaces. Roughly spoken, a subspace of a vector space is a subset, which is a vector itself when using the same addition and scalar multiplication.

DEFINITION 2.19 (Subspace). Assume that V is a vector space over $\mathbb{K}$ 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 over $\mathbb{K}$ with the same addition and scalar multiplication as on U.

That is, if $+_V$ and $\cdot_V$ denote addition and scalar multiplication on V, and $+_U$ and $\cdot_U$ denote addition and scalar multiplication on $U \subset V$, then $u +_V v = u +_U v$ for all u, $v \in U$, and $\lambda \cdot_V u = \lambda \cdot_U u$ for all $\lambda \in \mathbb{K}$ and $u \in U$. ■

In the definition of a subspace, we assume that we are already given a vector space structure on the set U, which, a-priori, might be independent from the structure of the set V. In most practical applications with subspaces, however, we are given the vector space V and the subset U, and we want to know whether this subset is actually a subspace when we use the same addition and scalar multiplication as on the larger space. The next result answers this question.

LEMMA 2.20. *Assume that V is a vector space and $U \subset V$ a subset. Then U is a subspace of V (with the same addition and scalar multiplication as on U), 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. Assume first that Items (1)–(3) hold. 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 for all elements of the larger space V. Thus U is actually a vector space, which implies that it is a subspace of V.

Conversely, assume that U is a subspace of V. Then, necessarily, the addition and scalar multiplication map $U \times U$ and $\mathbb{K} \times U$ onto U, and thus Items (2) and (3) hold. Next, since U is a vector space itself, it is non-empty. Choose therefore some

$u \in U$. Then Items (2) and (3) imply that $0 = u + (-1) \cdot u \in U$, and thus (1) holds. □

EXAMPLE 2.21. Consider the following examples of vector spaces and subsets, which in some cases are subspaces and in others are not:[3]

(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 $\mathbb{R}^2$.
(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 subspaces 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 subspace 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 $\mathbf{Mat}_n(\mathbb{K})$.

■

1.5. Injectivity and Surjectivity of Linear Transformations. Assume now that U and V are vector spaces over $\mathbb{K}$ and that $T\colon U \to V$ is a linear transformation. Then we have two important associated subspaces of U and V:

DEFINITION 2.22. 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 $Tu = 0$; the range of T is the set of all right hand sides v for which the equation $Tu = v$ has a solution.

LEMMA 2.23. *The range and the kernel of a linear transformation are subspaces of the respective vector spaces.*

PROOF. Exercise! □

EXAMPLE 2.24. 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.

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

- 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

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

■

EXAMPLE 2.25. Consider the mapping $D\colon \mathcal{P}(\mathbb{K}) \to \mathcal{P}(\mathbb{K})$, $p \mapsto Dp := p'$. Then

$$\ker(D) = \big\{p \in \mathcal{P}(\mathbb{K}) : p' = 0\big\} = \big\{p \in \mathcal{P}(\mathbb{K}) : p \text{ is constant}\big\}$$

and

$$\operatorname{ran}(D) = \mathcal{P}(\mathbb{K}).$$

■

PROPOSITION 2.26 (Injectivity and surjectivity). *Assume that U and V are vector spaces over $\mathbb{K}$ and that $T\colon U \to V$ is a linear transformation.*

- *T is injective if and only if $\ker(T) = \{0\}$.*
- *T is surjective if and only if $\operatorname{ran}(T) = V$.*

PROOF. In the case of surjectivity, this is a trivial consequence of the definition. Thus we only discuss the case of injectivity.

Assume first that T is injective, and let $u \in \ker(T)$. Then $Tu = 0$. Moreover, the linearity of T implies that $T0 = 0$. Thus we have that $Tu = 0 = T0$. The injectivity of T now implies that $u = 0$, which proves that $\ker(T) = \{0\}$.

Conversely, assume that $\ker(T) = \{0\}$. Moreover, let u_1, $u_2 \in U$ be such that $Tu_1 = Tu_2$. We have to show that $u_1 = u_2$. Because of the linearity of T we have that

$$0 = Tu_1 - Tu_2 = T(u_1 - u_2),$$

which shows that $u_1 - u_2 \in \ker(T)$. Since, by assumption, $\ker(T) = \{0\}$, this implies that $u_1 - u_2 = 0$ and thus $u_1 = u_2$, which proves the claim. □

PROPOSITION 2.27. *Assume that U and V are vector spaces over $\mathbb{K}$ and that $T\colon U \to V$ is a linear transformation. Assume moreover that T is bijective. Then the inverse $T^{-1}\colon V \to U$ is a linear transformation as well.*

PROOF. Let v_1, $v_2 \in V$, and λ, $\mu \in \mathbb{K}$. Denote moreover $u_1 := T^{-1}v_1$ and $u_2 := T^{-1}v_2$. The linearity of T implies that

$$\lambda v_1 + \mu v_2 = \lambda T u_1 + \lambda T u_2 = T(\lambda u_1 + \mu u_2).$$

Applying the function T^{-1} on both sides of this equation, we obtain that

$$T^{-1}(\lambda v_1 + \mu v_2) = T^{-1}(T(\lambda u_1 + \mu u_2)) = \lambda u_1 + \mu u_2 = \lambda T^{-1} v_1 + \mu T^{-1} v_2,$$

which proves the linearity of T^{-1}. □

2. Linear Independence and Bases

2.1. Linear Span.

DEFINITION 2.28. Assume that V is a vector space and that $S := (v_i)_{i \in I}$ is a (possibly infinite) family in V. We define the *linear span* $\operatorname{span}(S) \subset V$ of S as the set of all *finite* linear combinations of members of 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 $v_{i_1}, \dots, v_{i_N} \in S$, and scalars $\lambda_1, \dots, \lambda_N \in \mathbb{K}$ such that

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

■

REMARK 2.29. It is useful to be able to talk about the span of the empty family $\emptyset$, 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 2.30. *Assume that V is a vector space and $S = (v_i)_{i\in I}$ is a family in V. Then* $\operatorname{span}(S)$ *is the smallest subspace of V containing all members of 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, $M \in \mathbb{N}$, indices $i_1, \ldots, i_{N+M} \in I$, and scalars $\lambda_1, \ldots, \lambda_{N+M} \in \mathbb{K}$. As a consequence,

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

Next assume that $u \in \operatorname{span}(S)$ and $\alpha \in \mathbb{K}$. Since $u \in \operatorname{span}(S)$, we can write

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

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

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

which shows that also $\alpha u \in \operatorname{span}(S)$. According to Lemma 2.20, 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 $N \in \mathbb{N}$, indices $i_1, \ldots, i_N \in I$, 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 2.31. 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 2.32 (Linear independence). Assume that V is a vector space over $\mathbb{K}$ and $S = (v_i)_{i\in I}$ is a family in V. We say that S is *linearly independent*, if the following holds:

For each *finite* subset of indices $J \subset I$ and each family $(\lambda_j)_{j\in J} \subset \mathbb{K}$ or scalars, if

$$\sum_{j\in J} \lambda_j v_j = 0,$$

then

$$\lambda_j = 0 \text{ for all } j \in J.$$

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

REMARK 2.33. We have formulated the notions of linear dependence and independence for families, but it is straightforward to formulate them for sets: A set $S \subset V$ is linearly independent, if for each finite subset $W \subset S$ and each family $(\lambda_w)_{w\in W} \subset \mathbb{K}$ of scalars such that $\sum_{w\in W} \lambda_w w = 0$ we have that $\lambda_w = 0$ for all $w \in W$.

There is, however, a difference as to how sets and families handle "duplicate" members, which can lead to an at first glance counterintuitive result about linear independence of sets: It is true to say that a family $S = (v_i)_{i\in I}$ is linearly dependent, if it contains the same element twice, that is, if there exist $i \neq j \in I$ such that $v_i = v_j$. In particular, if V is a vector space and $v \in V \setminus \{0\}$ is any non-zero vector in V, then the family (v, v) is linearly dependent. However, the *set* $\{v, v\}$ is linearly independent, as, actually, $\{v, v\} = \{v\}$ (as a set). Because of this (and because of the possibility of actually accessing different members by means of their indices) we will most of the time employ families when talking about linear independence. ■

PROPOSITION 2.34. *Assume that V is a vector space and $S = (v_i)_{i\in I}$ is a family in 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 members of S.*

More precisely, S is linearly independent if and only if the following holds: If $u \in \operatorname{span}(S)$, then there exists a unique finite subset $J \subset I$ and a unique family $(\lambda_j)_{j\in J} \in \mathbb{K}\setminus\{0\}$ of non-zero coefficients, such that $u = \sum_{j\in J} \lambda_j v_j$. Here we define the empty sum to be equal to 0.

PROOF. Assume first that S is linearly independent, and let $u \in \operatorname{span}(S)$. From the definition of $\operatorname{span}(S)$, it follows that u can be written as a finite linear combination of members of S. Thus we only have to show that this finite linear combination is unique. Assume therefore that we can write

$$u = \sum_{j\in J} \lambda_j v_j = \sum_{k\in K} \mu_k v_k$$

for finite subsets $J, K \subset I$, and families $(\lambda_j)_{j\in J} \subset \mathbb{K}\setminus\{0\}$ and $(\mu_k)_{k\in K} \subset \mathbb{K}\setminus\{0\}$ of non-zero coefficients. Define now $L := J \cup K$, and define the family $(\nu_\ell)_{\ell\in L}$ setting

$$\nu_\ell := \begin{cases} \lambda_\ell - \mu_\ell & \text{if } \ell \in J \cap K, \\ \lambda_\ell & \text{if } \ell \in J \setminus K, \\ -\mu_\ell & \text{if } \ell \in K \setminus J. \end{cases}$$

Then

$$\sum_{\ell\in L} \nu_\ell v_\ell = \sum_{\ell\in J} \lambda_\ell v_\ell - \sum_{\ell\in K} \mu_\ell v_\ell = u - u = 0.$$

Now the linear independence of S implies that $\nu_\ell = 0$ for all $\ell \in L$. In particular, $\lambda_\ell = \mu_\ell$ for all $\ell \in J \cap K$. It remains to show that $J = K$, that is, that $J \setminus K = \emptyset$

and $K \setminus J = \emptyset$. Assume therefore that $\ell \in J \setminus K$. Then $0 = \nu_\ell = \lambda_\ell$, which contradicts the assumption that $\lambda_\ell \neq 0$ for all $\ell \in J$. Thus $J \setminus K = \emptyset$. The proof that $K \setminus J = \emptyset$ is similar.

Assume now that every vector $u \in \operatorname{span}(S)$ can be written in a unique way as a finite linear combination of members of S. We have to show that S is linearly independent. Assume to that end that $J \subset I$ and $(\lambda_j)_{j\in J} \subset \mathbb{K}$ are such that

$$\sum_{j\in J} \lambda_j v_j = 0.$$

Denote by $J' \subset J$ the subset of indices $j \in J$ for which $\lambda_j \neq 0$. We have to show that $J' = \emptyset$. By construction of the set J' we have that

$$0 = \sum_{j\in J'} \lambda_j v_j.$$

At the same time, we can write 0 as the empty sum $\sum_{j\in\emptyset} \lambda_j v_j$. Thus the uniqueness assumption of the representation of $0 \in \operatorname{span}(S)$ implies that $J' = \emptyset$, which implies that S is linearly independent. □

Example 2.35. 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 family of all monomials. Then M is linearly independent. ■

Definition 2.36 (Basis). Let V be a vector space and $B \subset V$ a family in V. We say that B is a (Hamel) *basis* of V, if B is linearly independent and $\operatorname{span}(B) = V$. ■

Remark 2.37. In view of Proposition 2.34, 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 2.38. The set M of monomials is a basis of the vector space of all polynomials. ■

Theorem 2.39. *Assume that V is a vector space and $S \subset V$ is some linearly independent family in V. Then there exists a basis B of V with $S \subset B$.*

In particular, every vector space has a basis.

The proof of this theorem is by no means simple, and it actually relies on Zorn's Lemma (and thus the Axiom of Choice).

Proof*. Let $\mathcal{L} \subset \mathcal{P}(V)$ be the set of all linearly independent subsets of V containing S. On $\mathcal{L}$ we consider the partial order R defined by inclusion, that is, $A \leq_R B$ if $A \subset B$. We will first show that $\mathcal{L}$ contains a maximal element with respect to this order. After that, we will show that such a maximal element is a basis.

In order to show that $\mathcal{L}$ contains a maximal element, we will use Zorn's Lemma (Theorem 1.61*). For this, we have to show that every chain in $\mathcal{L}$ has an upper bound. Assume therefore that $\mathcal{M}$ is a chain in $\mathcal{L}$. That is, $\mathcal{M}$ contains linearly independent subsets of V, each containing S, and for all A, $B \in \mathcal{M}$ we have that either $A \subset B$ or $B \subset A$.

Define now $M := \bigcup_{A\in\mathcal{M}} A$ the union of all the sets in $\mathcal{M}$. We will show that M is an upper bound for $\mathcal{M}$ in $\mathcal{L}$. First we note that, clearly, $A \subset M$ for every $A \in \mathcal{M}$. Thus we only have to show that $M \in \mathcal{L}$. It is obvious that $S \subset M$. Thus it remains to show that M is linearly independent.

Assume therefore that $J \subset M$ is a finite subset and that λ_u, $u \in J$, are such that

$$0 = \sum_{u \in J} \lambda_u u.$$

Since $J \subset M = \bigcup_{A \in \mathcal{M}} M$, there exists for each $u \in J$ some $A_u \in \mathcal{M}$ such that $u \in A_u$. Now the sets in $\mathcal{M}$ are totally ordered by inclusion, and thus we have for each u, $v \in J$, that at least one of the inclusions $A_u \subset A_v$ or $A_v \subset A_u$ holds. Since the set J is finite, it follows that there exists $v \in J$ such that $A_u \subset A_v$ for all $u \in J$. In particular, $u \in A_v$ for all $u \in J$. Since the set A_v is by assumption linearly independent, it follows that $\lambda_u = 0$ for all $u \in J$. This shows that M is linearly independent.

We have now shown that every chain in $\mathcal{L}$ has an upper bound. Zorn's Lemma therefore implies that there exists a maximal element $B \in \mathcal{L}$. We want to show that B is a basis containing S. By construction, we have that B is linearly independent and that $S \subset B$. Thus it remains to show that $\operatorname{span}(B) = V$.

For this, assume to the contrary that $\operatorname{span}(B) \neq V$. Then there exists $v \in V \setminus \operatorname{span}(B)$. We assert that $B \cup \{v\}$ is linearly independent as well. Assume therefore that $J \subset B \cup \{v\}$ is a finite subset and that $\lambda_u \in \mathbb{K}$, $u \in J$, are such that

$$\sum_{u \in J} \lambda_u u = 0.$$

If $J \subset B$, then the linear independence of B implies that $\lambda_u = 0$ for all $u \in J$. Assume therefore there $v \in J$. Then we can write

$$-\lambda_v v = \sum_{u \in J \setminus \{v\}} \lambda_u u,$$

which implies that $-\lambda_v v \in \operatorname{span}(B)$. Since by assumption $v \notin \operatorname{span}(B)$, this further implies that $\lambda_v = 0$. As a consequence, we have that

$$0 = \sum_{u \in J \setminus \{v\}} \lambda_u u.$$

Now the linear independence of B implies that $\lambda_u = 0$ for all $u \in J \setminus \{v\}$. In total, we therefore have that $\lambda_u = 0$ for all $u \in J$, and thus $B \cup \{v\}$ is linearly independent. This, however, is a contradiction to the maximality of B. Therefore $\operatorname{span}(B) = V$, which proves that B is a basis of V. □

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. (In fact, it is possible to show that the existence of bases of aribtrary vector spaces is equivalent to the Axiom of Choice.) 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 2.40. A vector space V is called *finite dimensional*, if V has a finite basis. Else, V is called *infinite dimensional*. ■

Theorem 2.41. *Assume that V is finite dimensional. Then all bases of V are finite and contain the same number of members.*

Definition 2.42 (Dimension). Assume that V is finite dimensional. The number of members of members of any (and therefore every) basis of V is called the *dimension* of V and denote $\dim V$. ■

The proof of Theorem 2.41 is surprisingly tricky and relies on another important result, namely the Steinitz Exchange Lemma 2.43*, which we formulate (and prove) below. Feel therefore free to skip over the proof.

PROOF*. Since V is a finite dimensional space, it has a finite basis, say $C = (u_1, \dots, u_m)$. Assume now that $B = (v_i)_{i\in I}$ is an arbitrary basis of V. By Lemma 2.43* below, we can find distinct indices $i_1, \dots, i_m \in I$, replace the vectors $v_{i_1}, \dots, v_{i_m}$ in B by the vectors $u_1, \dots, u_m$, and still obtain a basis of V, say $\tilde{B} = (\tilde{v}_i)_{i\in I}$. In particular, this implies that I has at least m members. On the other hand, if I has more than m members, then there exists some index $j \notin \{i_1, \dots, i_m\}$. Since C is a basis of V, we can write the corresponding vector v_j as a finite linear combination of the vectors $u_1, \dots, u_m$. Since $v_j = \tilde{v}_j$, this contradicts the linear independence of the family $\tilde{B}$. As a consequence, B has precisely m members, which proves the assertion. □

LEMMA 2.43* (Steinitz Exchange Lemma). *Assume that V is a vector space over $\mathbb{K}$. Then, if $B = (v_i)_{i\in I}$ is a basis of V and $(u_1, \dots, u_m)$ is a finite family of linearly independent vectors, then there exist distinct indices $i_1, \dots, i_m \in I$, such that we can replace the basis elements $v_{i_1}, \dots, v_{i_m}$ by the vectors $u_1, \dots, u_m$ and still obtain a basis of V.*

PROOF*. We prove this assertion by induction over m:

For $m = 0$, the claim is trivial.

Assume therefore that the claim is true for all bases of V and all linearly independent families with m members. Let moreover $B = (v_i)_{i\in I}$ be a basis of V, and let $(u_1, \dots, u_m, u_{m+1})$ be a family of $m+1$ linearly independent vectors in V. By our induction hypothesis, we can find distinct indices $i_1, \dots, i_m \in I$ and replace the basis elements $v_{i_1}, \dots, v_{i_m}$ by $u_1, \dots, u_m$, and still obtain a basis of V. Let us denote this new basis by $\tilde{B} = (\tilde{v}_i)_{i\in I}$. Since this is a basis of V and $u_{m+1} \neq 0$, we can in particular find a unique finite set $J \subset I$ and unique non-zero coefficients $\lambda_j \in \mathbb{K} \setminus \{0\}$, $j \in J$, such that

$$u_{m+1} = \sum_{j\in J} \lambda_j \tilde{v}_j.$$

We claim now that $J \setminus \{i_1, \dots, i_m\} \neq \emptyset$. Indeed, if it were true that $J \setminus \{i_1, \dots, i_m\} = \emptyset$ and thus $\{i_1, \dots, i_m\} \subset J$, we could write u_{m+1} as a linear combination of the vectors $u_1, \dots, u_m$, which contradicts the assumption that the family $(u_1, \dots, u_{m+1})$ was linearly independent. Choose now an element $i_{m+1} \in J \setminus \{i_1, \dots, i_m\}$, replace $v_{i_{m+1}}$ by u_{m+1}, and denote the resulting family by $\hat{B} := (\hat{v}_i)_{i\in I}$. We claim that this is a basis of V, that is, that it is linearly independent and spans V.

We first show that $\operatorname{span}(\hat{B}) = V$. Assume to that end that $v \in V$. Since $\tilde{B}$ is a basis of V, it spans V and thus there exist a finite subset $L \subset I$ and coefficients $\mu_\ell \in \mathbb{K}$, $\ell \in L$, such that

$$v = \sum_{\ell\in L} \mu_\ell \tilde{v}_\ell.$$

Now, if $i_{m+1} \notin L$ we have that $\tilde{v}_\ell = \hat{v}_\ell$ and we have thus written v as a finite linear combination of members of $\hat{B}$. On the other hand, if $i_{m+1} \in L$, we can write

$$\begin{aligned} v = \mu_{i_{m+1}} \tilde{v}_{i_{m+1}} + \sum_{\ell\in L\setminus\{i_{m+1}\}} \mu_\ell \hat{v}_\ell \\ = \frac{\mu_{i_{m+1}}}{\lambda_{i_{m+1}}} \Big(u_{m+1} - \sum_{j\in J\setminus\{i_{m+1}\}} \lambda_j \hat{v}_j \Big) + \sum_{\ell\in L\setminus\{i_{m+1}\}} \mu_\ell \hat{v}_\ell, \end{aligned}$$

which again is a finite linear combination of members of $\hat{B}$. Thus it follows that $v \in \operatorname{span}(\hat{B})$.

We now show that $\operatorname{span}(\hat{B})$ is linearly independent. Assume therefore that $L \subset I$ is finite and that $\mu_\ell \in \mathbb{K}$, $\ell \in L$, are such that

$$\sum_{\ell \in L} \mu_\ell \hat{v}_\ell = 0.$$

Denote now $L' := \big\{\ell \in L : \mu_\ell \neq 0\big\}$. We have to show that $L' = \emptyset$. We can write

$$\sum_{\ell \in L'} \mu_\ell \hat{v}_\ell = 0$$

with $\mu_\ell \neq 0$ for all $\ell \in L'$. Now, if $i_{m+1} \notin L'$, then $\hat{v}_\ell = \tilde{v}_\ell$ for all $\ell \in L'$ and thus the linear independence of $\tilde{B}$ implies that $\mu_\ell = 0$ for all $\ell \in L'$, which implies that $L' = \emptyset$. Assume therefore that $i_{m+1} \in L'$. Then we can write

$$0 = \frac{\mu_{i_{m+1}}}{\lambda_{i_{m+1}}}\Big(u_{m+1} - \sum_{j \in J \setminus \{i_{m+1}\}} \lambda_j \tilde{v}_j\Big) + \sum_{\ell \in L' \setminus \{i_{m+1}\}} \mu_\ell \tilde{v}_\ell.$$

As a consequence,

$$u_{m+1} = \sum_{j \in J \setminus \{i_{m+1}\}} \lambda_j \tilde{v}_j + \sum_{\ell \in L' \setminus \{i_{m+1}\}} \frac{\lambda_{i_{m+1}}}{\mu_{i_{m+1}}} \mu_\ell \tilde{v}_\ell. \tag{1}$$

Now recall that we also have that

$$u_{m+1} = \sum_{j \in J} \lambda_j \tilde{v}_j = \lambda_{i_{m+1}} \tilde{v}_{i_{m+1}} + \sum_{j \in J \setminus \{i_{m+1}\}} \lambda_j \tilde{v}_j$$

and that $\lambda_{i_{m+1}} \neq 0$. Since $\tilde{B}$ is a basis of V and the basis representation of every vector $v \in V$ is unique, it follows that the vector $\tilde{v}_{i_{m+1}}$ must also appear (with a non-zero coefficient) in the representation (1), which is obviously not the case. Thus $i_{m+1} \notin L'$, which in turn shows that $L' = \emptyset$. Thus the set $\hat{B}$ is linearly independent, which concludes the proof. □

3. Matrix Representations of Linear Transformations

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, $j = 1, \dots, n$, separately. Since $\mathcal{C}$ is a basis of V and $Tu_j \in V$, there exist, for each j, unique coefficients a_{ij}, $i = 1, \dots, m$, such that

$$Tu_j = \sum_{i=1}^{m} a_{ij} v_i. \tag{2}$$

Now consider the vector

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

with coordinates $(\alpha_1, \dots, \alpha_n)$. The linearity of T implies that

$$Tu = \sum_{j=1}^{n} \alpha_j \, Tu_j = \sum_{j=1}^{n} \alpha_j \Big(\sum_{i=1}^{m} a_{ij} v_i \Big) = \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_j a_{ij} v_i = \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} a_{ij} \alpha_j \Big) v_i. \tag{3}$$

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

$$Tu = v = \sum_{i=1}^{m} \beta_i v_i. \tag{4}$$

Since $(v_1, \dots, v_m)$ is a basis of V, the coefficients in the representation are unique. By comparing the right hand sides of (3) and (4), we obtain that

$$\sum_{j=1}^{n} a_{ij} \alpha_j = \beta_i \qquad \text{for } i = 1, \dots, m. \tag{5}$$

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 (5) 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 (2) represents the linear transformation T in the bases $\mathcal{B}$ and $\mathcal{C}$.

Example 2.44. 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

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

■

4. Direct Sums and Invariant Subspaces

4.1. External Direct Sums.

DEFINITION 2.45 (Direct sum). Assume that U and V are vector spaces over $\mathbb{K}$ with additions $+_U$ and $+_V$ and scalar multiplications $\cdot_U$ and $\cdot_V$, respectively. The *(external) direct sum* of U and V, denoted $U \oplus V$, is the set $U \times V$ together with the operations $+_{U\oplus V} \colon (U \times V) \times (U \times V) \to (U \times V)$ and $\cdot_{U\oplus V} \colon \mathbb{K} \times (U \times V) \to (U \times V)$, defined by

$$(u_1, v_1) +_{U\oplus V} (u_2, v_2) := (u_1 +_U u_2, v_1 +_V v_2) \quad \text{for all } (u_1, v_1), (u_2, v_2) \in U \times V$$

and

$$\lambda \cdot_{U\oplus V} (u, v) := (\lambda \cdot_U u, \lambda \cdot_V v) \quad \text{for all } (u, v) \in U \times V \text{ and } \lambda \in \mathbb{K}.$$

■

PROPOSITION 2.46. *Let U and V be vector spaces over $\mathbb{K}$. Then their direct sum $U \oplus V$ is again a vector space over $\mathbb{K}$.*

PROOF. Exercise! □

More general, we can consider several (but finitely many) vector spaces $U_1, \ldots, U_n$ and define their direct sum $U_1 \oplus U_2 \oplus \cdots \oplus U_n$ as the product $U_1 \times U_2 \times \cdots \times U_n$ together with the componentwise addition

$$(u_1, u_2, \ldots, u_n) + (v_1, v_2, \ldots, v_n) := (u_1 + v_1, u_2 + v_2, \ldots, u_n + v_n)$$

and componentwise scalar multiplication

$$\lambda \cdot (u_1, u_2, \ldots, u_n) := (\lambda u_1, \lambda u_2, \ldots, \lambda u_n).$$

EXAMPLE 2.47. The direct sum of m copies of $\mathbb{K}$ is the space $\mathbb{K}^m$. The direct sum of n copies of the space $\mathbb{K}^m$ is the space $\mathbf{Mat}_{m,n}(\mathbb{K})$. ■

EXAMPLE 2.48. Let $U := C(\mathbb{R}; \mathbb{K})$ the vector space of continuous functions on $\mathbb{R}$ with values in $\mathbb{K}$. Then the direct sum $U \oplus U$ consists of all pairs (f, g), where both f and g are $\mathbb{K}$-valued continuous functions on $\mathbb{R}$. An alternative interpretation of such a pair (f, g) of functions on $\mathbb{R}$ is as the vector-valued function that maps a point $t \in \mathbb{R}$ to the pair $(f(t), g(t)) \in \mathbb{K}^2$. Thus we can write the direct sum $U \oplus U$ alternatively as

$$C(\mathbb{R}; \mathbb{K}) \oplus C(\mathbb{R}; \mathbb{K}) = C(\mathbb{R}; \mathbb{K}^2).$$

■

EXAMPLE 2.49. In the previous examples, we only have the case where all the involved spaces in the direct sum were identical. However, this need not be the case in general. A simple, but important, example here is the case where $U = \mathbb{K}^n$ and $V = \mathbb{K}^m$ for some $n, m \in \mathbb{N}$. For this case, we obtain that

$$\mathbb{K}^n \oplus \mathbb{K}^m = \mathbb{K}^{n+m}.$$

■

DEFINITION 2.50. Assume that $U_1,\dots,U_n$ are vector spaces and $1 \le k \le n$. By

$$\begin{aligned} \imath_k \colon U_k &\to U_1 \oplus \cdots \oplus U_n, \\ u &\mapsto (0,\dots,0,\underbrace{u}_{k\text{-th component}},0\dots,0), \end{aligned}$$

and

$$\begin{aligned} \pi_k \colon U_1 \oplus \cdots \oplus U_n &\to U_k, \\ (u_1,\dots,u_n) &\mapsto u_k, \end{aligned}$$

we denote the inclusion of U_k in the direct sum $U_1 \oplus \cdots \oplus U_n$ and the projection from the direct sum $U_1 \oplus \cdots \oplus U_n$ to the k-th component U_k, respectively. ■

LEMMA 2.51. *Assume that $U_1,\dots,U_n$ are vector spaces and $1 \le k \le n$. The k-th inclusion $\imath_k \colon U_k \to U_1 \oplus \cdots \oplus U_n$ and the k-th projection $\pi_k \colon U_1 \oplus \cdots \oplus U_n \to U_k$ are linear transformations.*

PROOF. Exercise! □

Assume now that U_1, U_2, and V_1, V_2 are vector spaces and that $T \colon U_1 \oplus U_2 \to V_1 \oplus V_2$ is a linear transformation. Moreover, let $(u_1, u_2) \in U_1 \oplus U_2$ be arbitrary (that is, let $u_1 \in U_1$ and $u_2 \in U_2$ be arbitrary). Then the linearity of T implies that

$$T(u_1,u_2) = T\big((u_1,0)+(0,u_2)\big) = T(u_1,0) + T(0,u_2).$$

Now both $T(u_1,0)$ and $T(0,u_2)$ are vectors in $V_1 \oplus V_2$, and thus we can decompose both of these vectors into parts

$$T(u_1,0) =: (T_{11}u, T_{21}u) \in V_1 \oplus V_2 \quad \text{and} \quad T(0,u_2) =: (T_{21}v, T_{22}v) \in V_1 \oplus V_2.$$

Here T_{11} is the part of T that maps U_1 to V_1; then T_{21} is the part of T that maps U_1 to V_2; next, T_{12} is the part of T that maps U_2 to V_1; finally, T_{22} is the part of T that maps U_2 to V_2. Thus we can write the transformation T in "(block)-matrix-vector-form" as

$$T\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} T_{11}u_1 + T_{12}u_2 \\ T_{21}u_1 + T_{22}u_2 \end{pmatrix}.$$

The block T_{ij} can be expressed explicitly expressed in terms of a composition of the inclusion $\imath_j \colon U_j \to U_1 \oplus U_2$, the transformation T, and the projection $\pi_i \colon V_1 \oplus V_2 \to V_i$ as

$$T_{ij} = \pi_i \circ T \circ \imath_j \colon U_j \to V_i. \tag{6}$$

More general, if $U_1,\dots,U_n$ and $V_1,\dots,V_m$ are vector spaces and $T \colon U_1 \oplus \cdots \oplus U_n \to V_1 \oplus \cdots \oplus V_m$ is linear, we can write the operator T in the form

$$T\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} T_{11} & \dots & T_{1n} \\ \vdots & \ddots & \vdots \\ T_{m1} & \dots & T_{mn} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} T_{11}u_1 + \ldots + T_{1n}u_n \\ \vdots \\ T_{m1}u_1 + \ldots + T_{mn}u_n \end{pmatrix}, \tag{7}$$

where the blocks $T_{ij} \colon U_j \to V_i$ are again defined as in (6).

Conversely, if $T_{ij} \colon U_j \to V_i$ are arbitrary linear mappings, we can *define* a mapping $T \colon U_1 \oplus \cdots \oplus U_n \to V_1 \oplus \cdots \oplus V_m$ using the formula (7). Thus providing the whole mapping T is equivalent to providing the blocks T_{ij}.

REMARK 2.52. What we did just know looks pretty much like matrix–vector calculus, and up to a point it really is: The case of linear mappings from $\mathbb{K}^n$ to $\mathbb{K}^m$ written as matrices $A \in \mathbb{K}^{m\times n}$ can be regarded as a special case, since we can interpret $\mathbb{K}^n$ and $\mathbb{K}^m$ as the n-fold and m-fold direct sum of copies of $\mathbb{K}$, respectively. ■

We now consider the special case where $U_1,\dots,U_n$ are vector spaces and we are given linear transformations $T_i\colon U_i \to U_i$ for $1 \le i \le n$. Denote $U := U_1 \oplus \dots \oplus U_n$. Then we can define a *block-diagonal* transformation $\operatorname{Diag}(T_1,\dots,T_n)\colon U \to U$ by setting

$$\operatorname{Diag}(T_1,\dots,T_n) := \begin{pmatrix} T_1 & 0 & \dots & \dots & 0 \\ 0 & T_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & T_{n-1} & 0 \\ 0 & \dots & \dots & 0 & T_n \end{pmatrix}$$

with the notation from above. Note here that a 0 on the ij-th position is actually the 0-operator from U_j to U_i, which maps every element $u \in U_j$ to the 0-vector in U_i.

In a sense, block-diagonal linear transformations are the simplest possible linear transformations, both from a numerical and an analytical point of view, as all the information about the transformation is already contained in the blocks $T_1,\dots,T_n$. For instance, in order to represent the whole transformation, it is only necessary to store (or implement) these n blocks; all the off-diagonal 0-blocks can simply be ignored. Also, it is for instance straight-forward to show that a block-diagonal transformation is invertible, if and only if each of the blocks is invertible; in this case, the inverse is again block-diagonal and the blocks are the inverses T_i^{-1} of the transformations T_i.

Remark 2.53* (Infinite direct sums). In the discussion above, we have only considered direct sums of finitely many vector spaces. It is also possible, and sometimes necessary, to look at direct sums of infinitely many vector spaces. However, the construction of these infinite direct sums is slightly more complicated than that of finite direct sums.

Let us therefore assume that we are given a family $(U_i)_{i\in I}$ of vector spaces for some non-empty (and possibly infinite) index set I. Then we can the direct sum of this family as

$$\bigoplus_{i\in I} U_i := \Big\{ (u_i)_{i\in I} \in \bigtimes_{i\in I} U_i : \text{there exists a finite set } J \subset I \text{ with } u_i = 0 \text{ for } i \notin J \Big\}.$$

That is, we take all the elements in the Cartesian product of the vector space for which all but finitely many components are 0. As in the finite case, we define the addition and the scalar multiplication on this set componentwise.

All the results we indicated above concerning the interplay between linear transformations and direct sums still hold in this setting: If we are given two families of vector spaces $(U_j)_{j\in J}$ and $(V_i)_{i\in I}$ and for each pair $(i,j) \in I \times J$ a linear transformation $T_{ij}\colon U_j \to V_i$, we can define a linear transformation $T\colon \bigoplus_{j\in J} U_j \to \bigoplus_{i\in I} V_i$ by

$$T\big((u_j)_{j\in J}\big) := \Big(\sum_{j\in J} T_{ij} u_j\Big)_{i\in I}. \tag{8}$$

Note here that the sum $\sum_{j\in J} T_{ij}u_j$ is actually a *finite* sum for each $i \in I$, as only finitely many components of $(u_j)_{j\in J} \in \bigoplus U_j$ are non-zero. Conversely, given a linear transformation $T\colon \bigoplus_{j\in J} U_j \to \bigoplus_{i\in I} V_i$, we can define $T_{ij} := \pi_i \circ T \circ \imath_j\colon U_j \to V_i$. Then the equation (8) holds for all $(u_j)_{j\in J} \in \bigoplus_{j\in J} U_j$. ∎

4.2. Internal Direct Sums. In the previous section, we have seen how we can use the idea of a direct sum to combine different vector spaces to a larger space. Moreover, we have discussed the connection between a linear transformation T of

the larger space and the mappings T_{ij} between the component spaces. Of particular interest is the case where the mappings T_{ij} are 0 whenever $i \neq j$, and thus T has a block-diagonal form.

The main goal of this chapter is to investigate whether we can go the opposite direction as well. That is, we start with an arbitrary vector space U and a linear transformation $T\colon U \to U$. Is it then possible to *decompose* U as a direct sum $U = U_1 \oplus \cdots \oplus U_n$ such that the mapping T has block-diagonal form w.r.t. this decomposition? Also, to which extent are the spaces U_j unique?

As a first step, we define sums of arbitrary subsets of a vector space.

DEFINITION 2.54. Assume that V is a vector space and that $S_1, \ldots, S_N$ are subsets of V. We define

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

■

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 2.55. *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 2.56. 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 2.57. 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 *(internal) 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

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

that necessarily

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

EXAMPLE 2.58. 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 2.59. *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 $U_1 \cap U_2 = \{0\}$. □

REMARK 2.60. The result from Lemma 2.59 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. ■

EXAMPLE 2.61. Define $U_1, U_2 \subset \mathbb{R}^3$ as

$$U_1 = \big\{(x, y, z) \in \mathbb{R}^3 : y = z = 0\big\}$$

and

$$U_2 = \big\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\big\}.$$

We will show

$$\mathbb{R}^3 = U_1 \oplus U_2.$$

We start by showing that $U_1 \cap U_2 = \{0\}$. To that end let $(x, y, z) \in U_1 \cap U_2$. Because $(x, y, z) \in U_1$, it follows that $y = z = 0$. Now the fact that $(x, y, z) = (x, 0, 0) \in U_2$ implies that $x = 0$ as well. Thus $(x, y, z) = (0, 0, 0)$.

Next we need to show that $\mathbb{R}^3 = U_1 + U_2$. Assume therefore that $(x, y, z) \in \mathbb{R}$. Then we can write

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + y + z \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -y - z \\ y \\ z \end{pmatrix}. \tag{9}$$

Since $(x + y + z, 0, 0) \in U_1$ and $(-y - z, y, z) \in U_2$, it follows that we can write an arbitrary vector in $\mathbb{R}^3$ as a sum of a vector in U_1 and a vector in U_2. Thus $\mathbb{R}^3 = U_1 + U_2$. Since $U_1 \cap U_2 = \{0\}$, it follows that $\mathbb{R}^3 = U_1 \oplus U_2$. ∎

DEFINITION 2.62 (Projection). Assume that V is a vector space and that U_1 and U_2 are subspaces such that $V = U_1 \oplus U_2$. The *(oblique) projection onto U_1 along U_2* is the mapping $\pi_1 \colon V \to V$ defined by the condition that $\pi_1(v) = u_1$ if $v = u_1 + u_2$ with $u_1 \in U_1$ and $u_2 \in U_2$.

Similarly, the projection onto U_2 along U_1 is the mapping $\pi_2 \colon V \to V$ defined by the condition that $\pi_2(v) = u_2$ if $v = u_1 + u_2$ with $u_1 \in U_1$ and $u_2 \in U_2$. ∎

PROPOSITION 2.63. *Assume that V is a vector space and that U_1 and U_2 are subspaces such that $V = U_1 \oplus U_2$. Then the projections π_1 and π_2 onto U_1 along U_2 and onto U_2 along U_1, respectively, are linear, and we have that*

$$v = \pi_1(v) + \pi_2(v)$$

for all $v \in V$.

PROOF. The decomposition $v = \pi_1(v) + \pi_2(v)$ follows immediately from the definition. Assume now that v, $w \in V$. Then we can uniquely write $v = u_1 + u_2$ and $w = z_1 + z_2$ with u_1, $z_1 \in U_1$, and u_2, $z_2 \in U_2$. In particular, we have that $u_1 = \pi_1(v)$, $u_2 = \pi_2(v)$, $z_1 = \pi_1(w)$, and $z_2 = \pi_2(w)$. As a consequence, we can write

$$v + w = u_1 + u_2 + z_1 + z_2 = (u_1 + z_1) + (u_2 + z_2).$$

Since U_1 and U_2 are subspaces, it follows that $u_1 + z_1 \in U_1$ and $u_2 + z_2 \in U_2$. Thus the definition of π_1 implies that

$$\begin{aligned} \pi_1(v + w) &= u_1 + z_1 = \pi_1(v) + \pi_1(w), \\ \pi_2(v + w) &= u_2 + z_2 = \pi_2(v) + \pi_2(w). \end{aligned}$$

Next assume that $v \in V$ and $\lambda \in \mathbb{K}$. Again, we can write $v = u_1 + u_2$ with $u_1 \in U_1$ and $u_2 \in U_2$, and thus $u_1 = \pi_1(v)$ and $u_2 = \pi_2(v)$. Because U_1 and U_2 are subspaces, we have that $\lambda u_1 \in U_1$ and $\lambda u_2 \in U_2$. Thus we can decompose $\lambda v = \lambda u_1 + \lambda u_2$ with $\lambda u_1 \in U_1$ and $\lambda u_2 \in U_2$. This implies that

$$\pi_1(\lambda v) = \lambda u_1 = \lambda \pi_1(v) \qquad \text{and} \qquad \pi_2(\lambda v) = \lambda u_2 = \lambda \pi_2(v).$$

This proves the linearity of π_1 and π_2. □

EXAMPLE 2.64. We continue with Example 2.61 and consider again the two subspaces

$$U_1 = \big\{(x, y, z) \in \mathbb{R}^3 : y = z = 0\big\}$$

and

$$U_2 = \big\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\big\}.$$

We have already seen that $\mathbb{R}^3 = U_1 \oplus U_2$. Moreover, from (9) we immediately obtain that

$$\pi_1 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+y+z \\ 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \pi_2 \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -y-z \\ y \\ z \end{pmatrix}.$$

■

We now generalise the notion of an oblique projection to the case where the vector space V is the direct sum of more than two subspaces.

DEFINITION 2.65 (Projection). Assume that V is a vector space and that $U_1,\ldots,U_N$ are subspaces such that $V = U_1 \oplus \cdots \oplus U_N$. The *projections* $\pi_i \colon V \to V$, $i = 1, \ldots, N$, are defined as the unique mappings such that $\pi_i(v) \in U_i$ for each $i = 1, \ldots, N$ and $v \in V$, and

$$v = \pi_1(v) + \ldots + \pi_N(v)$$

for each $v \in V$. ■

PROPOSITION 2.66. *Assume that V is a vector space and that $U_1,\ldots,U_N$ are subspaces such that $V = U_1 \oplus \cdots \oplus U_N$. The projections $\pi_i \colon V \to V$, $i = 1, \ldots, N$, defined in Definition 2.65 are linear transformations.*

PROOF. This is essentially the same as the proof of Proposition 2.63. □

PROPOSITION 2.67. *Assume that V is a vector space and that $U_1,\ldots,U_N$ are subspaces such that $V = U_1 \oplus \cdots \oplus U_N$. Assume moreover that B_i is a basis of U_i, $i = 1, \ldots, N$. Then $B := B_1 \cup \ldots \cup B_N$ is a basis of V.*

PROOF. For $i = 1, \ldots, N$ we can write $B_i = (v_j)_{j \in J_i}$ for some index set J_i. Here we may assume without loss of generality that the index sets are pairwise disjoint, that is, that $J_i \cap J_j = \emptyset$ for $i \neq j$. Denote moreover $J := J_1 \cup \ldots \cup J_N$.

Assume now that $v \in V$ is an arbitrary vector. Then there exist unique vectors $u_i \in U_i$, $i = 1, \ldots, N$, such that

$$v = u_1 + \ldots + u_N.$$

Moreover, for each $i = 1, \ldots, N$ there exists a unique finite subset $K_i \subset J_i$ and unique non-zero scalars $\lambda_j \in \mathbb{K} \setminus \{0\}$, $j \in J_i$, such that

$$u_i = \sum_{j \in K_i} \lambda_j v_j.$$

As a consequence, we can write

$$v = \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} \sum_{j \in K_i} \lambda_j v_j = \sum_{j \in K_1 \cup \ldots \cup K_N} \lambda_j v_j.$$

Thus we can represent the vector $v \in V$ as finite linear combination of members of B. Now assume that $L \subset J$ is another finite subset and that $\mu_\ell \in \mathbb{K} \setminus \{0\}$, $\ell \in L$, are such that

$$v = \sum_{\ell \in L} \mu_\ell v_\ell.$$

Define for $i = 1, \dots, N$

$$L_i := L \cap J_i \qquad \text{and} \qquad w_i := \sum_{\ell \in L_i} \mu_\ell v_\ell.$$

Then $w_i \in U_i$ for each i and $v = \sum_{i=1}^N w_i$. Thus the assumption that the sum of the vector spaces U_i is direct implies that $w_i = u_i$ for each i. Next, the assumption that B_i is a basis of U_i and the fact that

$$\sum_{j \in K_i} \lambda_j v_j = u_i = w_i = \sum_{\ell \in L_i} \mu_\ell v_\ell$$

imply that $K_i = L_i$ and $\lambda_j = \mu_j$ for each $j \in K_i = L_i$. This further implies that $L = K := \bigcup_{i=1}^N K_i$ and that $\lambda_j = \mu_j$ for each $j \in K$. This shows that the representation of v is unique, which in turn shows that B indeed is a basis of V. □

As a particular consequence, we obtain that the dimension of a direct sum is the sum of the dimensions of the subspaces:

PROPOSITION 2.68. *Assume that V is a finite dimensional vector space and that $U_1, \dots, U_N$ 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 2.69. *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 2.39, can be extended to a basis of the whole space V. That is, there exists an index set J satisfying $J \cap I = \emptyset$ and a (linearly independent) family $\mathcal{C} := (u_j)_{j \in J} \subset V$, such that the combined family $(u_k)_{k \in I \cup J}$ 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 $v \in U \cap W$. Since $(u_k)_{k \in I \cup J}$ is a basis of V, there exists a unique way of writing v in the form

$$v = \sum_{\ell=1}^N \lambda_\ell u_{i_\ell} + \sum_{m=1}^M \mu_m v_{j_m}$$

with indices $i_1, \dots, i_N \in I$ and $j_1, \dots, j_M \in J$, and coefficients λ_ℓ, $\mu_m \in \mathbb{K}$. However, because $v \in U \cap W \subset W$, it follows that the coefficients μ_m have to be equal to 0 for all m. Similarly, because $v \in U$, it follows that the coefficients λ_ℓ 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. □

4.3. Invariant Subspaces. Assume now that V is a vector space and $T \colon V \to V$ is a linear transformation of V. Assume moreover that $U_1, \dots, U_N \subset V$ are subspaces such that $V = U_1 \oplus \cdots \oplus U_N$. Similar as in Section 4.1, we can then write the transformation T in block-matrix-form with respect to the given decomposition of V. To that end, we define the transformations

$$T_{ij} := \pi_i \circ T|_{U_j} \colon U_j \to U_i$$

for $i,\ j = 1,\dots,N$. As before, the transformation T_{ij} represents the "part of T that maps U_j to U_i." Now identify a vector $u \in V$ with the tuple

$$u = \big(u_1,\dots,u_N\big) \qquad \text{with } u_i := \pi_i u \text{ for } i = 1,\dots,N.$$

Then we can (again, see (7)) write

$$Tu = \begin{pmatrix} T_{11} & \dots & T_{1n} \\ \vdots & \ddots & \vdots \\ T_{m1} & \dots & T_{mn} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} T_{11}u_1 + \dots + T_{1n}u_n \\ \vdots \\ T_{m1}u_1 + \dots + T_{mn}u_n \end{pmatrix}$$

for every $u \in V$.

Our goal, as stated in the beginning of Section 4.2, is to discuss whether we can choose the subspaces $U_1,\dots,U_N$ in such a way that the mapping T has block-diagonal-form with respect to this decomposition. In the next definition, we will introduce a notion that is central to this goal.

Definition 2.70 (T-invariant subspace). 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 2.2. Show that $\ker(T)$ and $\operatorname{ran}(T)$ are always T-invariant subspaces. ■

Example 2.71. 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 2.72. 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. ■

Lemma 2.73. *Assume that V is a vector space, that $T\colon V \to V$ is a linear transformation, and that $U \subset V$ is a T-invariant subspace. Let $W \subset V$ be a subspace such that $V = U \oplus W$. Then the block-matrix-representation of T with respect to this decomposition is block-upper-triangular.*

Proof. We write

$$T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \qquad \text{with } T_{ij} = \pi_i \circ T|_{U_j}.$$

We have to show that $T_{21} = 0$. Let therefore $u \in U$. Then the assumption that U is T-invariant implies that $Tu \in U$. Thus $\pi_1(Tu) = Tu$, and $\pi_2(Tu) = 0$. This shows that $(\pi_2 \circ T)u = 0$ for every $u \in U$, which proves that $T_{21} = \pi_2 \circ T|_U = 0$. □

Lemma 2.74. *Assume that V is a vector space and that $U_1,\dots,U_N \subset V$ are subspaces such that $V = U_1 \oplus \dots \oplus U_N$. Let moreover $T\colon V \to V$ be a linear transformation such that U_i is a T-invariant subspace of V for every $i = 1,\dots,N$. Then the block-matrix-representation of T with respect to this decomposition is block-diagonal.*

Proof. Exercise! □

5. Eigenvalues and Eigenspaces

Recall once again our goal of decomposing a vector space V as a direct sum $V = U_1 \oplus \cdots \oplus U_N$ in such a way that the given linear transformation $T\colon V \to V$ has block-diagonal form. In the previous section, we have just seen that this requires the subspaces U_j to be T-invariant. It therefore makes sense to look at the smallest possible (non-trivial) T-invariant subspaces of V, which are precisely the eigenspaces of T.

DEFINITION 2.75. 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

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

■

EXAMPLE 2.76. 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. ■

EXAMPLE 2.77. Consider the transformation $T\colon \mathcal{P}(\mathbb{K}) \to \mathcal{P}(\mathbb{K})$, $p \mapsto Tp := p'$. Then the only eigenvalue of T is 0, and the corresponding eigenvectors are the constant polynomials $p(x) = c$ for $c \in \mathbb{K} \setminus \{0\}$.

We can, however, consider "the same" transformation over the space $C^\infty(\mathbb{R}, \mathbb{K})$ of arbitrarily differentiable $\mathbb{K}$-valued functions on $\mathbb{R}$. That is, we consider the function $S\colon C^\infty(\mathbb{R}, \mathbb{K}) \to C^\infty(\mathbb{R}, \mathbb{K})$, $f \mapsto Sf := f'$. In this case, every value $\lambda \in \mathbb{K}$ is an eigenvalue of S with corresponding eigenfunctions $f(x) = ce^{\lambda x}$ with $c \in \mathbb{K} \setminus \{0\}$. ■

LEMMA 2.78. *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

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

Thus U is T-invariant. □

We will now introduce an alternative characterisation of eigenvalues in finite dimensional spaces, which, however, requires some technical preparation.

LEMMA 2.79. *Assume that U, V are vector spaces, U is finite dimensional, and that $T\colon U \to V$ is bijective. Assume moreover that $\mathcal{B} = (u_1, \ldots, u_n)$ is a basis of U. Then $T\mathcal{B} = (Tu_1, \ldots, Tu_n)$ 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 $u = c_1u_1 + \ldots + c_nu_n$ for some $c_n \in \mathbb{K}$. As a consequence,

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

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

It remains to show that the family $(Tu_1, \ldots, Tu_n)$ is linearly independent. Assume to that end that

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

We can rewrite this as

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

that is,

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

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

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

Now the linear independence of the family $(u_1, \ldots, u_n)$ implies that $c_j = 0$ for all j, which in turn proves the linear independence of the set $(Tu_1, \ldots, Tu_n)$. □

THEOREM 2.80. *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 2.69, 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 $w \in W$. That is, $u_0 = \pi_{\ker T}u$ and $w = \pi_W u$. 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, it follows from Lemma 2.79 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 2.81. *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. □

DEFINITION 2.82. 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 2.83. *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 2.81.

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). □

REMARK 2.84. Do note that the characterisation of eigenvalues from Theorem 2.83 only holds in finite dimensions. In infinite dimensions, however, it is only true to say that λ is an eigenvalue of T if and only if $T - \lambda \mathrm{Id}$ is not injective. The other equivalences do no longer hold.

In fact, we have already seen an example of this, namely the transformation $T \colon \mathcal{P}(\mathbb{K}) \to \mathcal{P}(\mathbb{K})$, $p \mapsto Tp := p'$. Here we know that 0 is an eigenvalue, which is the same as saying that T is not injective. Nevertheless, the transformation T is surjective. ■

6. Existence of Eigenvalues

We will now show that linear transformations of *complex*, *finite dimensional* vector spaces always have at least one eigenvalue. In order to so, we need to introduce the concept of the evaluation of a polynomial on a linear transformation. Assume to that end that $T \colon V \to V$ is a linear transformation of the space V. Then we can consider the composition $T \circ T$ with itself, which is again a linear transformation of V. Thus we can further consider the composition $T \circ T \circ T$. Again, this is a linear transformation of V. In a similar way, we can consider arbitrary (but finite) compositions of T with itself. In order to simplify notation, we abbreviate in the following $T^2 := T \circ T$, and similarly $T^3 = T \circ T \circ T$, and so on.

DEFINITION 2.85. 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 2.86. *Assume that V is a vector space, $T\colon V \to V$ 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 2.87. *Assume that V is a finite dimensional complex vector space, $V \neq \{0\}$, and that $T\colon V \to V$ 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 space, which implies that they cannot be linearly independent. That is, there exist numbers $c_0, \ldots, c_n \in \mathbb{C}$, not all of them equal to 0, such that

$$c_0v + c_1Tv + c_2T^2v + \ldots + c_nT^nv = 0. \tag{10}$$

Define now the polynomial

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

Then we can write (10) 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 (10) 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 is 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 2.88. 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 2.89. *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 2.90. *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 2.89, 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 2.91. *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 2.87). 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 2.86.

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 $(u_1, \dots, u_m)$ 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 $(w_1, \dots, w_{n-m})$ of W. We now consider the matrix representation A of T with respect to the basis $(u_1, \dots, u_m, w_1, \dots, w_{n-m})$. Following our considerations from Section 4.3 and specifically Lemma 2.73, 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 $(u_1, \dots, u_m)$ 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 2.92. 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 2.93. *Assume that V is a vector space and $T\colon V \to V$ is linear. Let moreover $\lambda_1, \dots, \lambda_m \in \mathbb{K}$ be* distinct *eigenvalues (that is, $\lambda_i \neq \lambda_j$ for $i \neq j$), and let $v_1, \dots, v_m \in V$ be corresponding eigenvectors. Then the family $(v_1, \dots, v_m)$ 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 family $(v_1, \dots, v_{m-1})$ is linearly independent.

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

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

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

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

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

Since by our induction assumption, the set $(v_1, \dots, v_{m-1})$ is linearly independent, it follows that the coefficients $c_1, \dots, c_{m-1}$ 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 $j = 1, \dots, m-1$, we then obtain the equation

$$v_m = d_1 v_1 + \dots + d_{m-1} v_{m-1}. \tag{11}$$

If we apply the transformation T to (11) 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 (11) 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 family $(v_1, \ldots, v_{m-1})$ 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 2.94. *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

$$0 = (u_1 - v_1) + \ldots + (u_m - v_m) \tag{12}$$

with $(u_j - v_j) \in E(\lambda_j, T)$ for all $j = 1, \ldots, m$. Now recall that we shown in Theorem 2.93 that eigenvectors for distinct eigenvalues are linearly independent. Thus (12) 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 2.95. *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 4.3.

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 2.94. 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 2.68. Finally, assume that (5) holds. By Lemma 2.94, the sum of the eigenspaces is direct, and thus, by Proposition 2.68,

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

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 2.96. *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

$$A = \begin{pmatrix} \lambda & * & \ldots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \ldots & 0 & \lambda \end{pmatrix} \tag{13}$$

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

In this section, we will show that this 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 (13). In order to arrive at this representation, we will have to introduce some notation, though, and then do some hard work.

Nilpotent Transformations.

DEFINITION 2.97. 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 2.98. *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 2.99. *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 2.100. *Assume that V is a complex finite dimensional vector space and that $T \colon V \to V$ is nilpotent. Then*

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

PROOF. By Theorem 2.91, we can find a basis of V such that the matrix A of T is upper triangular. Next we know from Lemma 2.98 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 2.99. □

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 (13). 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 2.101. Let V be a vector space, let $T \colon V \to V$ 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 2.102. 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 2.103. 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 2.104. *Let V be a vector space, let $T \colon V \to V$ 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 2.105. *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. □

LEMMA 2.106. *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 2.105 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 $\ker T^{k+m+1} \subset \ker T^{k+m}$. □

LEMMA 2.107. *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 2.106 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 2.106 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. □

PROPOSITION 2.108. *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 2.107 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

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

Next we know from Theorem 2.80 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. □

LEMMA 2.109. *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*

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

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 2.107 we know that we may assume without loss of generality that $j \leq \dim V$. Then, however, we can use Lemma 2.105 and obtain that $v \in \ker\big((T - \lambda \operatorname{Id})^{\dim V}\big)$, which proves the assertion. □

Linear Independence of Generalised Eigenvectors.

THEOREM 2.110. *Let V be a vector space and let $T\colon V \to V$ be linear. Assume moreover that $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of T and that $v_1, \ldots, v_m$ are corresponding generalised eigenvectors. Then the family $(v_1, \ldots, v_m)$ is linearly independent.*

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

$$0 = c_1 v_1 + \ldots + c_m v_m. \tag{14}$$

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

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

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 $w \neq 0$, 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

$$\begin{aligned} Sv_1 &= (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1} \circ (T - \lambda_1 \operatorname{Id})^{k_1} v \\ &= (T - \lambda_2 \operatorname{Id})^{k_2+1} \circ \cdots \circ (T - \lambda_m \operatorname{Id})^{k_m+1} w \\ &= (\lambda_1 - \lambda_2)^{k_2+1} \cdots (\lambda_1 - \lambda_m)^{k_m+1} w. \end{aligned}$$

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

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

Now apply the operator S to both sides of the equation (14). 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 $w \neq 0$ and $\lambda_1 \neq \lambda_j$ for $j = 2, \ldots, m$, 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 family $(v_1, \ldots, v_m)$. □

Decomposition into Generalised Eigenspaces.

THEOREM 2.111. *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*

$$V = G(\lambda_1, T) \oplus \cdots \oplus G(\lambda_m, T) \tag{15}$$

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 $G(\lambda_j, T) = \ker\big((T - \lambda_j \operatorname{Id})^{\dim V}\big)$, 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 (15), 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 (15) 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 2.87. Define moreover

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

Then we can write

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

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

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

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

$$U = G(\lambda_2, S) \oplus \ldots \oplus G(\lambda_m, S), \tag{16}$$

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

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

Since $U \subset V$ 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 (15) 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 2.112 (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 2.113. 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. ■

DEFINITION 2.114. Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ is linear. Denote be $\lambda_1, \dots, \lambda_m$ the distinct eigenvalues of T, and let $d_1, \dots, d_m$ 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 2.115. 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

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

■

REMARK 2.116. Assume that the operator $T\colon V \to V$ has for some basis an upper triangular matrix with not necessarily distinct diagonal entries $\mu_1, \mu_2, \dots, \mu_n$ (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. ■

THEOREM 2.117 (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 $\lambda_1, \dots, \lambda_m$ are the distinct eigenvalues of T and $d_1, \dots, d_m$ 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 $j = 1, \dots, m$.

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

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

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 2.118. 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 2.119. *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 2.117 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

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

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

EXAMPLE 2.121. Assume that $T\colon \mathbb{C}^3 \to \mathbb{C}^3$ 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 2.122. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ 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 2.123. *Assume that V is a finite dimensional complex vector space and that $T\colon V \to V$ 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 $0 \leq s_j \leq d_j$. It remains to show that $s_j \geq 1$ 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 $s_j \geq 1$ 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 2.111, which describes the, in a sense, simplest possibility of representing an arbitrary linear transformation of a finite dimensional complex vector space.

THEOREM 2.124. *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*

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

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

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

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

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

(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 2.125. 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 3

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 3.1. 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 3.2. 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 3.3. *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

$$\begin{aligned}\langle u, \lambda v + \mu w\rangle = \overline{\langle \lambda v + \mu w, u\rangle} &= \overline{\lambda\langle v, u\rangle + \mu\langle w, u\rangle} \\ &= \overline{\lambda}\overline{\langle v, u\rangle} + \overline{\mu}\overline{\langle w, u\rangle} = \overline{\lambda}\langle u, v\rangle + \overline{\mu}\langle u, w\rangle.\end{aligned}$$

□

REMARK 3.4. 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 3.5. 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 3.6. 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 3.7. *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

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

□

DEFINITION 3.8. 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 $\langle u, v\rangle = 0$. ■

Note that the order of the inner product does not matter in the definition of orthogonality: If $\langle u, v\rangle = 0$, then the conjugate symmetry of the inner product implies that also $\langle v, u\rangle = \overline{\langle v, u\rangle} = \overline{0} = 0$.

THEOREM 3.9 (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

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

□

REMARK 3.10. 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 3.11 (Cauchy–Schwarz–Bunyakovsky). *Assume that* V *is an inner product space and that* u, $v \in V$. *Then*

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

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

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

and we have equality if and only if $w = 0$. In the latter case, however, we have that $u = \frac{\langle u,v\rangle}{\|v\|^2}v$ and thus u and v are linearly dependent.

It remains to show that the linear dependence of u and v implies that $|\langle u, v\rangle| = \|u\|\|v\|$. Assume therefore that u and v are linearly dependent. If $v = 0$, then the

claimed equality holds. Thus we may assume without loss of generality that $v \neq 0$ and that $u = \lambda v$ for some $\lambda \in \mathbb{K}$. Then

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

which concludes the proof. □

THEOREM 3.12 (Triangle inequality). *Assume that V is an inner product space and u, $v \in V$. Then*

$$\|u + v\| \leq \|u\| + \|v\|,$$

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

PROOF. We can write

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

This shows that the inequality $\|u + v\| \leq \|u\| + \|v\|$ 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 dependent, 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 $\lambda \in \mathbb{R}_{\geq 0}$, which proves the assertion. □

2. Bases of Inner Product Spaces

DEFINITION 3.13. Let V be an inner product space. A family $(v_1, \ldots, v_m) \subset V$ 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 $i \neq j$. ■

LEMMA 3.14. *Assume that V is an inner product space and that the family $(v_1, \ldots, v_m) \subset V$ 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 3.15. Let V be an inner product space. An *orthonormal basis* of V is an orthonormal family that is a basis of V. ■

THEOREM 3.16 (Gram–Schmidt Orthogonalisation). *Assume that V is a vector space and that $S = (v_1, v_2, \ldots) \subset V$ is a linearly independent family in V. Then we can construct a family $E = (e_1, e_2, \ldots)$ as follows:*

- *Start with setting*

$$e_1 = \frac{v_1}{\|v_1\|}.$$

- *Assume we have already constructed orthonormal vectors $e_1, \dots, e_m$. We define first*

$$f_{m+1} = v_{m+1} - \langle v_{m+1}, e_1\rangle e_1 - \langle v_{m+1}, e_2\rangle e_2 - \dots - \langle v_{m+1}, e_m\rangle e_m$$

and then set

$$e_{m+1} = \frac{f_{m+1}}{\|f_{m+1}\|}.$$

Then $(e_1, e_2, \dots)$ is a well-defined orthonormal family. Moreover, we have that $\operatorname{span}\{v_1, v_2, \dots, v_m\} = \operatorname{span}\{e_1, e_2, \dots, e_m\}$ for all m.

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

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

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

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

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

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

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

COROLLARY 3.17. *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 3.18. *Assume that V is a finite dimensional inner product space with orthonormal basis $(v_1, \dots, v_n)$. 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, that is,*

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

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

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

■

REMARK 3.20. We see immediately that the Gram matrix for a linear system is the identity matrix, if and only if the system is orthonormal. ■

PROPOSITION 3.21. *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 3.22. 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 3.23. A matrix $A \in \mathbf{Mat}_n(\mathbb{K})$ is called *positive semi-definite*, if

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

for all $x \in \mathbb{K}^n$. It 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 3.24. *Assume that V is a finite dimensional inner product space and that $(v_1, \dots, v_n)$ is a family in V, and denote by G the corresponding Gram matrix. Then G is Hermitian and positive semi-definite. Moreover, G is positive definite, if and only if the family $(v_1, \dots, v_n)$ is linearly independent.*

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

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

which proves that G is positive semi-definite.

Now assume that $(v_1, \dots, v_n)$ is linearly independent, let $x = (x_1, \dots, x_n) \in \mathbb{K}^n \setminus \{0\}$ and let $v = \sum_{i=1}^{n} x_i v_i$. Since $x \neq 0$ and $(v_1, \dots, v_n)$ is linearly independent, it follows that $v \neq 0$. Thus the same computation as above shows that $x^H G x = \|v\|^2 > 0$. This proves that G is positive definite.

Now assume conversely that G is positive definite, and let $(x_1, \dots, x_n) \in \mathbb{K}^n$ be such that

$$\sum_{i=1}^{n} x_i v_i = 0.$$

Then

$$0 = \|0\|^2 = x^H G x.$$

Thus the positive definiteness of G implies that $x = 0$. This proves that $(v_1, \dots, v_n)$ is linearly independent. □

REMARK 3.25. 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 3.26. *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 $x_1, \dots, x_n$ of v are given as*

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

PROOF. Let $x_1, \dots, x_n$ be the coordinates of v with respect to $(v_1, \dots, v_n)$. Denote moreover $y_j := \langle v, v_j \rangle$ for $j = 1, \dots, n$. Then

$$y_j = \langle v, v_j \rangle = \sum_{i=1}^{n} x_i \langle v_i, v_j \rangle = \sum_{i=1}^{n} x_i G_{ji}.$$

Thus the vectors $y = (y_1, \dots, y_n)$ and $x = (x_1, \dots, x_n)$ satisfy the equation

$$y = Gx.$$

Solving this equation for x yields the assertion. □

COROLLARY 3.27. *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 3.28. Assume that V is an inner product space and $U \subset V$ is a subset. We denote by

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

the *orthogonal complement* of U. ■

LEMMA 3.29. *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 3.30. *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 3.31. *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 $\mathcal{B} = (u_1, \dots, u_m)$ 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 family $\mathcal{C} = (v_1, \dots, v_{n-m})$ 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,

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

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 3.32. *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 3.33. Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. The *orthogonal projection onto U*, denoted $\pi_U \colon V \to V$, is the projection onto U along $U^\perp$. ■

THEOREM 3.34. *Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Then the orthogonal projection π_U onto U has the following properties:*

(1) $\pi_U \colon V \to V$ *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) $\|v\|^2 = \|\pi_U v\|^2 + \|v - \pi_U v\|^2$ *for all* $v \in V$.
(8) $\|\pi_U v\| \leq \|v\|$ *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 $(e_1,\dots,e_m)$ 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. The linearity of the orthogonal projection is a special case of Proposition 2.63 on general (oblique) projections. Moreover, (2)–(6) are immediate consequences of the definition of the orthogonal projection.

Now let $v \in V$. Then $\pi_U v \in U$ and $v - \pi_U v \in U^\perp$, which implies that $\langle \pi_U v, v - \pi_U v\rangle = 0$. Thus 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).

Next, we note that for all $u \in U$ we have 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).

Finally, let $(e_1,\dots,e_m)$ be an orthonormal basis of U, and let $(e_{m+1},\dots,e_n)$ be an orthonormal basis of $U^\perp$, such that $(e_1,\dots,e_n)$ 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 concludes the proof. □

Finally, we provide a method for computing the projection onto a subspace, if we have a basis of that subspace, which is not necessarily orthonormal.

PROPOSITION 3.35. *Assume that V is a finite dimensional inner product space and that $U \subset V$ is a subspace. Assume moreover that $(u_1,\dots,u_m)$ is a basis of U with associated Gram matrix G. Then we have for every $v \in V$ that*

$$\pi_U v = \sum_{k=1}^{m} x_k u_k,$$

where the vector $x \in \mathbb{K}^m$ satisfies the equation

$$Gx = y \qquad \text{with} \qquad y = \begin{pmatrix}\langle v, u_1\rangle \\ \vdots \\ \langle v, u_m\rangle\end{pmatrix} \in \mathbb{K}^m.$$

PROOF. By Proposition 3.26, we know that the coordinates $x \in \mathbb{K}^m$ of $\pi_U v$ satisfy the equation $Gx = \tilde{y}$, where

$$\tilde{y} = \begin{pmatrix} \langle \pi_U v, u_1 \rangle \\ \vdots \\ \langle \pi_U v, u_m \rangle \end{pmatrix}.$$

Thus is is sufficient to show that $\tilde{y}_k = \langle \pi_U v, u_k \rangle = \langle v, u_k \rangle = y_k$ for all k. However, we have that

$$\langle v, u_k \rangle = \langle v - \pi_U v + \pi_U v, u_k \rangle = \langle v - \pi_U v, u_k \rangle + \langle \pi_U v, u_k \rangle = \langle \pi_U v, u_k \rangle,$$

as $v - \pi_U \in U^\perp$ and therefore $\langle v - \pi_U, u_k \rangle = 0$. □

4. Singular Value Decomposition

THEOREM 3.36 (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 families $(u_1, \ldots, u_p) \subset U$ and $(v_1, \ldots, v_p) \subset V$, and real numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$ 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 families $(u_1, \ldots, u_p)$ 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 $n \geq 1$ 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

$$\max_{\|u\|_U = 1} \|Tu\|_V. \tag{17}$$

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

Now choose a solution u_1 of (17), define $\sigma_1 = \|Tu_1\|$ and $v_1 = Tu_1/\sigma_1$.[3] Then we have in particular that

$$\|Tu\|_V = \|u\|_U \Big\| T\Big(\frac{u}{\|u\|_U}\Big) \Big\|_V \leq \sigma_1 \|u\|_U \tag{18}$$

for all $u \in U$ with $u \neq 0$, and the same inequality $\|Tu\|_V \leq \sigma_1 \|u\|_U$ 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 $w \in U_2$.

[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.

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

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

As a consequence, we have that

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

On the other hand, we have by (18) 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

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

or

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

Since $\alpha \in \mathbb{K}$ was arbitrary and $\sigma_1 \neq 0$, 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 $(u_2, \dots, u_p)$ and $(v_2, \dots, v_p)$ in U_2 and V_2, respectively, and positive numbers $\sigma_2 \ge \dots \ge \sigma_p > 0$ 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 $(u_1, u_2, \dots, u_p)$ is an orthonormal system in U, and similarly $(v_1, v_2, \dots, v_p)$ 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 (18),

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

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

REMARK 3.37. 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 $\sigma_{j-1} > \sigma_j = \sigma_{j+1} = \dots = \sigma_{j+k} > \sigma_{j+k+1} \dots$, then the linear spans $\operatorname{span}\{u_j, \dots, u_{j+k}\}$ and $\operatorname{span}\{v_j, \dots, v_{j+k}\}$ are independent of the choice of the singular vectors. In particular, if all the singular values are distinct, then the singular vectors are uniquely determined up to orientation. ■

LEMMA 3.38. *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

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

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 transformations unitary instead:

DEFINITION 3.39. 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 3.40. *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 3.41. *Assume that U and V are finite dimensional inner product spaces with bases $(u_1, \dots, u_n)$ and $(v_1, \dots, v_m)$, 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 3.42. If $(u_1, \dots, u_n)$ and $(v_1, \dots, v_m)$ 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 3.43 (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 3.44. *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*

$$\langle Tu, v\rangle_V = \langle u, T^*v\rangle_U \qquad \text{for all } u \in U \text{ and } v \in V. \tag{19}$$

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

We next show that (19) 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 (19) 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 3.45. *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 3.44!) □

LEMMA 3.46. *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 3.44!) □

LEMMA 3.47. *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 3.44, 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 3.48. *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 3.49. *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 3.50. *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*

$$A^* = G_U^{-1}A^HG_V.$$

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

$$y^HG_V(Ax) = (A^*y)^HG_Ux$$

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

$$yHG_VAx = y^H(A^*)^HG_Ux.$$

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. □

REMARK 3.51. 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 family $(v_1, \dots, v_p)$ 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 3.52 (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 $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$ 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 3.53 (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 $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p > 0$ 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 3.54. 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 3.55. Alternatively, if we use the decomposition

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

as in Theorem 3.52, we can write the Moore–Penrose inverse as

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

■

THEOREM 3.56. *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 3.57. *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

$$\min_u \|u\|_U^2 \qquad \text{s.t.} \qquad u \text{ solves } \min_u \|Tu - v\|_V^2. \tag{20}$$

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 (20) equivalently as

$$\min_u \|u\|_U^2 \qquad \text{s.t.} \qquad Tu = \pi_{\operatorname{ran} T} v. \tag{21}$$

We will show in the following that $u^\dagger = T^\dagger v$ solves (21).

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 (21). 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 (21), which concludes the proof. □

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 3.53, 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 $Q^HQ = \mathrm{Id}_p$ and $P^HP = \mathrm{Id}_p$, and Σ a diagonal matrix with non-increasing positive diagonal entries.

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

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

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^HA 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

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

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

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

This means that $\tilde{\Sigma}^2$ is the matrix of the mapping A^HA 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^HA is diagonalisable with non-zero eigenvalues $\sigma_1^2, \dots, \sigma_p^2$ 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 $\sigma_1^2, \dots, \sigma_p^2$ are the non-zero eigenvalues of AA^H, and the columns of Q are corresponding eigenvectors.

The considerations above yield the following result:

THEOREM 3.58. *Let $A \in \mathbb{K}^{n\times m}$. The matrices A^HA and AA^H have the same non-zero eigenvalues with the same multiplicities. Moreover, the singular values of A are precisely the square roots of the non-zero eigenvalues of the matrices A^HA or AA^H, and the singular vectors of A are eigenvectors of A^HA and AA^H, 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, $\hat{\Sigma}_{11} \geq \hat{\Sigma}_{22} \geq \ldots \geq \hat{\Sigma}_{kk} \geq 0$, 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 $n \leq m$, 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

$$A^HA = \hat{P}\tilde{\Sigma}^2\hat{P}^H$$

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 $\hat{q}_1, \dots, \hat{q}_m$ 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 $\hat{\sigma}_j \neq 0$). 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 3.59. 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 eigenvectors 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

$$AA^H = Q\Sigma^2Q^H = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 1/9 \end{pmatrix} \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 1 \\ -1 & 3 \end{pmatrix}.$$

[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.

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

$$\begin{pmatrix}1 & 8/3\\ 0 & 1\end{pmatrix} = \frac{1}{\sqrt{10}}\begin{pmatrix}3 & -1\\ 1 & 3\end{pmatrix}\begin{pmatrix}3 & 0\\ 0 & 1/3\end{pmatrix}\frac{1}{\sqrt{10}}\begin{pmatrix}1 & 3\\ -3 & 1\end{pmatrix}.$$

■

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 3.60. A linear transformation $T\colon U \to U$ on an inner product spaces U is called *self-adjoint* or *Hermitian* if $T^* = T$. ■

PROPOSITION 3.61. *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 3.62. *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 3.63. 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 3.64. 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 3.65. *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

$$u = x_1u_1 + \ldots + x_nu_n$$

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

$$Tu = \lambda_1x_1u_1 + \ldots + \lambda_nx_nu_n.$$

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 3.66. *Assume that U and V are finite dimensional inner product spaces and $T\colon U \to V$ is a linear transformation. Then $T^*T\colon U \to U$ and $TT^*\colon V \to V$ 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 3.67. *Assume that U and V are finite dimensional inner product spaces and $T\colon U \to V$ 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 3.68. *Assume that U and V are finite dimensional inner product spaces and that $T\colon U \to V$ is linear. Let $(u_1, \dots, u_n)$ be an orthonormal eigenbasis of T^*T for the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$.*

Denote $p = \operatorname{rank} T$, and define

$$\sigma_i = \sqrt{\lambda_i} \qquad \text{and} \qquad v_i = \frac{Tu_i}{\sigma_i}$$

for $k = 1, \dots, p$. Then

$$Tu = \sum_{i=1}^{p} \sigma_i \langle u, u_i\rangle v_i \tag{22}$$

is a singular value decomposition of T. In particular, the singular values of T are precisely the square roots of 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 $(u_1, \dots, u_p)$ and $(v_1, \dots, v_p)$ are orthonormal, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p > 0$, and that Tu actually has the form given in (22).

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 $\lambda_1 \geq \lambda_2 \geq \dots$, we also have that $\sigma_1 \geq \sigma_2 \geq \dots$.

Since $(u_1, \dots, u_n)$ is an orthonormal basis of U, the family $(u_1, \dots, u_p)$ 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 $i = p+1, \dots, n$, since the corresponding eigenvalue λ_i of T^*T is by assumption equal to 0. Thus $Tu_i = 0$ for $i = p+1, \dots, n$, 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 (22) is valid.

It remains to show that the family $(v_1, \dots, v_p)$ is orthonormal as well. For that, we compute for $1 \leq i, j \leq p$ the inner product

$$\begin{aligned}\langle v_i, v_j\rangle_V = \frac{1}{\sigma_i\sigma_j}\langle Tu_i, Tu_j\rangle_V &= \frac{1}{\sigma_i\sigma_j}\langle u_i, T^*Tu_j\rangle_U \\ &= \frac{1}{\sigma_i\sigma_j}\langle u_i, \sigma_j^2 u_j\rangle_U = \frac{\sigma_j}{\sigma_i}\langle u_i, u_j\rangle_U = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j, \end{cases}\end{aligned}$$

as the family $(u_1, \dots, u_p)$ is orthonormal. This concludes the proof. □

CHAPTER 4

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. Normed Spaces

We start by defining the notion of a *norm* on vector spaces, which is a function that assigns a length to each vector.

Definition 4.1 (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:* $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in U$.

A vector space together with a norm is called a *normed space*. ■

Example 4.2. The space $\mathbb{K}^n$ with the Euclidean norm $\|x\|_2 := \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$ is a normed space. ■

The Euclidean norm is by no means the only norm on $\mathbb{K}^n$. Amongst those that are used regularly in applications are the 1-norm and the ∞-norm defined below.

Example 4.3. On the space $\mathbb{K}^n$ we define

$$\|x\|_1 := |x_1| + \ldots + |x_n|$$

and

$$\|x\|_\infty := \max\bigl\{|x_1|, \ldots, |x_n|\bigr\}$$

for $x \in \mathbb{K}^n$. 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. ■

[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 required a significantly more complicated argumentation, but this was relegated to the optional part of these notes.

[2] Differentiability of mappings would be the next important idea, but, alas, we won't have time in this course to go that far.

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 4.1.

LEMMA 4.4. *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 3.12. □

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 4.5 (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

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

□

EXAMPLE 4.6. The norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{K}^n$ with $n \geq 2$ 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 4.7 (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.) □

Analoga of the 1-norm and ∞-norm can also be defined for function spaces:

EXAMPLE 4.8. The space $C([0,1]; \mathbb{C}^n)$ of continuous functions $f\colon [0,1] \to \mathbb{C}^n$ can be made a normed space with the norm

$$\|f\|_\infty := \max_{x \in [0,1]} |f(x)|_2.$$

Here we denote by $|\cdot|_2$ the Euclidean norm on $\mathbb{C}^n$ (in order to minimise the confusion possibilities with the norm on $C([0,1]; \mathbb{C}^n)$). There are (reasonable) alternatives to this norm, though. For instance, we can define the norm

$$\|f\|_1 := \int_0^1 |f(x)|_2 \, dx.$$

We will see later that these two norms have pretty different properties. ■

We will now state a consequence of the triangle inequality that turns out to be useful in a large number of estimates.

LEMMA 4.9 (Reverse triangle inequality). *Assume that $(U, \|\cdot\|)$ is a normed space and that u, $v \in U$. Then*

$$\big|\|u\| - \|v\|\big| \leq \|u - v\|.$$

PROOF. We have that

$$\|u\| = \|u - v + v\| \leq \|u - v\| + \|v\|$$

and therefore

$$\|u\| - \|v\| \leq \|u - v\|.$$

Similarly,

$$\|v\| = \|v - u + u\| \leq \|v - u\| + \|u\|$$

and thus

$$\|v\| - \|u\| \leq \|u - v\|.$$

Together, these inequalities imply that

$$\big|\|u\| - \|v\|\big| \leq \|u - v\|,$$

which proves the assertion. □

2. Convergence and Continuity

On a normed space, we can say whether two vectors are close to each other, as the norm gives us a method for calculating distances between vectors. This in turn gives a method for defining convergence of sequences:

DEFINITION 4.10 (Convergence and limits). Assume that $(U, \|\cdot\|)$ is a normed space and that $(u_n)_{n\in\mathbb{N}} \subset U$ is a sequence in U. We say that the sequence *converges* to $u \in U$, if there exists for every $\varepsilon > 0$ some index $N \in \mathbb{N}$ such that

$$\|u_n - u\| < \varepsilon \qquad \text{whenever } n \geq N.$$

We call u a *limit* of the sequence $(u_n)_{n\in\mathbb{N}}$ and denote the convergence by

$$u = \lim_{n\to\infty} u_n.$$

■

REMARK 4.11. The statement that for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|u_n - u\| < \varepsilon$ whenever $n \geq \mathbb{N}$ is precisely the same as the convergence of the sequence of *real numbers* $\|u_n - u\|$ to 0. That is, we have that

$$\lim_{n\to\infty} u_n = u \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} \|u_n - u\| = 0.$$

■

EXAMPLE 4.12. Consider the space $C([0,1])$ of continuous real valued functions on the interval $[0,1]$ with the norm

$$\|f\|_1 = \int_0^1 |f(x)|\, dx,$$

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

$$\|f_n - f\|_1 = \int_0^1 |x^n - 0|\, dx = \int_0^1 x^n\, dx = \frac{1}{n+1},$$

which shows that

$$\lim_{n\to\infty} \|f_n - f\|_1 = 0.$$

However, if we consider the same space with the norm

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

then the same sequence does not converge to 0 (in fact, it does not converge at all). In this case, we have that

$$\|f_n - 0\|_\infty = \max_{x\in[0,1]} |x^n - 0| = 1$$

for all $n \in \mathbb{N}$. This shows that the notion of convergence of a sequence can change drastically if we change the norm on the space. ■

LEMMA 4.13 (Uniqueness of the limit). *Assume that $(U, \|\cdot\|)$ is a normed space and that $(u_n)_{n\in\mathbb{N}} \subset X$ is a sequence in U. If the limit $\lim_{n\to\infty} u_n$ exists, it is unique.*

PROOF. Let $u = \lim_{n\to\infty} u_n$ and $v = \lim_{n\to\infty} u_n$. Assume to the contrary that $u \neq v$. Then $\|u - v\| > 0$. Define now $\varepsilon := \|u - v\|/2$. Because of the convergence of the sequence $(u_n)_{n\in\mathbb{N}}$ to u and v, there exist N_1, $N_2 \in \mathbb{N}$ such that $\|u_n - u\| < \varepsilon$

whenever $n \geq N_1$ and $\|u_n - v\| < \varepsilon$ whenever $n \geq N_2$. Let now $n := \max\{N_1, N_2\}$. Then the triangle inequality and positive homogeneity of the norm imply that

$$2\varepsilon = \|u - v\| \leq \|u - u_n\| + \|u_n - v\| = \|u_n - u\| + \|u_n - v\| < \varepsilon + \varepsilon = 2\varepsilon,$$

which is clearly a contradiction. This shows that $u = v$. □

As a consequence of this uniqueness result, we can indeed talk about *the* limit of a convergent sequence $(u_n)_{n\in\mathbb{N}}$ in a normed space.

DEFINITION 4.14 (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 $A \subset B_r(0)$. ■

PROPOSITION 4.15. *Assume that $(U, \|\cdot\|)$ is a normed space and that $(u_k)_{k\in\mathbb{N}} \subset U$ is a convergent sequence in U. Then the set $\{u_k : k \in \mathbb{N}\}$ is bounded.*

PROOF. Denote $u := \lim_{k\to\infty} u_k$. Then there exists $K \in \mathbb{N}$ such that $\|u - u_k\| \leq 1$ whenever $k \geq K$. (Here we have used the definition of continuity with the choice $\varepsilon = 1$.) Now define

$$r := \max\{\|u_1\|, \|u_2\|, \ldots, \|u_{K-1}\|, \|u\| + 1\} + 1.$$

Then we obviously have that $\|u_k\| < r$ whenever $k < K$. Moreover, for $k \geq K$ it follows from the triangle inequality that

$$\|u_k\| \leq \|u_k - u\| + \|u\| \leq 1 + \|u\| < r.$$

Thus $\|u_k\| < r$ for all $k \in \mathbb{N}$, or, put differently, $u_k \in B_r(0)$ for all $k \in \mathbb{N}$. This shows the boundedness of the set $\{u_k : k \in \mathbb{N}\}$. □

Having defined convergence of sequences in normed spaces, we can now continue with the definition of continuity of functions between normed spaces. The definition is, in a sense, a straight-forward generalisation of continuity of functions on $\mathbb{R}$; we simply replace each occurrence of an absolute value by the norm on the respective normed space.

DEFINITION 4.16 (Continuity). Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are normed spaces and that $f\colon U \to V$ is a mapping. We say that f is *continuous at a point* $u \in U$ if for every sequence $(u_n)_{n\in\mathbb{N}} \subset U$ with $\lim_{n\to\infty} u_n = u$ we have that $\lim_{n\to\infty} f(u_n) = f(u)$.

We say that f is *continuous everywhere* or simply *continuous*, if it is continuous at every point $u \in U$. ■

Alternatively to this definition of continuity, we have the ε-δ-characterisation that everyone loves and likes:

THEOREM 4.17. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. The function $f\colon U \to V$ is continuous at $u \in U$ if and only if for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\|f(u) - f(v)\|_V < \varepsilon \qquad \text{whenever} \qquad v \in V \text{ satisfies } \|u - v\|_U < \delta. \tag{23}$$

PROOF. Assume that (23) holds for all $\varepsilon > 0$ and $\delta > 0$. Now let $(u_n)_{n\in\mathbb{N}}$ be a sequence converging to u. We have to show that $(f(u_n))_{n\in\mathbb{N}}$ converges to $f(x)$. That is, we have to show that there exists for every $\varepsilon > 0$ some $N \in \mathbb{N}$ such that $\|f(u_n) - f(u)\|_V < \varepsilon$ whenever $n \geq N$.

Let therefore $\varepsilon > 0$ and let $\delta > 0$ be such that (23) holds. Since $u = \lim_{n\to\infty} u_n$, there exists $N \in \mathbb{N}$ such that $\|u_n - u\|_U < \delta$ for all $n \geq N$. According to (23), this implies that $\|f(u_n) - f(u)\|_V < \varepsilon$ for all $n \geq N$. This proves that f is continuous at u.

Now assume that (23) does not hold. That is, assume that there exists some $\varepsilon > 0$ such that for every $\delta > 0$ there exists $\hat{u} \in U$ with $\|u - \hat{u}\| < \delta$ and $\|f(u) - f(\hat{u})\|_V \geq \varepsilon$. Applying this with $\delta = 1/n$, this means that we can find a sequence $(u_n)_{n\in\mathbb{N}}$ such that $\|u-u_n\|_U < 1/n$ and $\|f(u)-f(u_n)\| \geq \varepsilon$. Since $\|u-u_n\|_U < 1/n$, we then have that $u = \lim_{n\to\infty} u_n$. On the other hand, since $\|f(u) - f(u_n)\|_V \geq \varepsilon$, the sequence $(f(u_n))_{n\in\mathbb{N}}$ does not converge to $f(u)$. This shows that f is not continuous at u. □

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.

THEOREM 4.18. *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

$$\big|\|u\| - \|u_n\|\big| \leq \|u - u_n\| \to 0 \qquad \text{as } n \to \infty,$$

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

$$\|u + v - (u_n + v_n)\| = \|u - u_n + v - v_n\| \leq \|u - u_n\| + \|v - v_n\| \to 0 \qquad \text{as } n \to \infty,$$

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

$$\begin{aligned}\|\lambda u - \lambda_n u_n\| &= \|\lambda u - \lambda u_n + \lambda u_n - \lambda_n u_n\| \leq \|\lambda(u - u_n)\| + \|(\lambda - \lambda_n)u_n\| \\ &= |\lambda|\|u - u_n\| + |\lambda - \lambda_n|\|u_n\| \to 0 \qquad \text{as } n \to \infty,\end{aligned}$$

as the sequence $(\|u_n\|)_{n\in\mathbb{N}}$ converges to $\|u\|$ and therefore is bounded. □

EXAMPLE 4.19. 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 $\|e_1\|_0 = 1$, but also $\|2e_1\|_0 = 1$.

The other properties of a norm are satisfied, though: Non-negativity is clear from the definition, and $\|u\|_0 = 0$ if and only if the vector u has no non-zero components, that is, if $u = 0$. Also, the number of non-zero components of a vector $u + v$ cannot be larger the sum of non-zero components of u and v, 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. ■

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 normed spaces.

3. Open and Closed Sets

DEFINITION 4.20 (Open ball). Assume that $(U, \|\cdot\|)$ is a normed space, let $u \in U$ and $r > 0$. By

$$B_r(u) := \big\{v \in U : \|u - v\| < r\big\}$$

we denote the (open) ball of radius r around u. ■

DEFINITION 4.21 (Open set). A subset $S \subset U$ of a normed space $(U, \|\cdot\|)$ is *open* if there exists for every $u \in S$ some $r > 0$ such that $B_r(u) \subset S$. ■

That is, a set S is open if we can find for each point u in S a small ball around u that is still completely contained in S.

In particular, open balls are actually open sets according to this definition. While this statement might seem trivial at first glance, it actually is not and does deserve a proof: In order to show that $B_r(u)$ is an open set, we have to find for each $v \in B_r(u)$ some ball $B_s(v)$ such that $B_s(v) \subset B_r(u)$.

LEMMA 4.22. *Let $(u, \|\cdot\|)$ be a normed space, let $u \in U$ and $r > 0$. Then the open ball $B_r(u)$ is an open set.*

PROOF. Exercise! □

EXAMPLE 4.23. Consider the space $C([0,1])$ with the norm $\|\cdot\|_\infty$. Then the set

$$U := \big\{f \in C([0,1]) : f(x) > 0 \text{ for all } x \in [0,1]\big\}$$

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 (bounded and closed) 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

$$\begin{aligned} g(x) \geq f(x) - |f(x) - g(x)| \geq f(x) - \max_{y\in[0,1]} |f(y) - g(y)| \\ = f(x) - \|f - g\|_\infty > f(x^*) - r = 0 \end{aligned}$$

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 norm.

EXAMPLE 4.24. Consider the space $C([0,1])$ with the norm $\|\cdot\|_1$. Then the set

$$U := \big\{f \in C([0,1]) : f(x) > 0 \text{ for all } x \in [0,1]\big\}$$

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

$$\|f_r - f\|_1 = \int_0^1 |f_r(x) - f(x)|\, dx = \int_0^{r/2} \frac{4x}{r}\, dx = \frac{r}{2} < r.$$

Thus $f_r \in B_r(f)$ and $f_r \notin U$. Since r was arbitrary, this shows that U is not open. ■

DEFINITION 4.25 (Closed set). A subset $K \subset U$ of a normed space $(U, \|\cdot\|)$ is *closed*, if every sequence in K that converges in U has its limit in K. That is, if $(u_n)_{n\in\mathbb{N}}$ is a sequence with $u_n \in K$ for all $n \in \mathbb{N}$ such that $u = \lim_{n\to\infty} u_n$ exists in U, then $u \in K$. ■

PROPOSITION 4.26 (Relation between open and closed sets). *A set S in a normed space $(U, \|\cdot\|)$ is open, if and only if its complement $S^C := U \setminus S$ is closed.*

PROOF. Assume that $S \subset U$ is open. Let moreover $(u_k)_{k\in\mathbb{N}}$ be a convergent sequence with $u_k \in S^C$ for all $k \in \mathbb{N}$. We have to show that $u := \lim_{k\to\infty} u_k \in S^C$.

Assume to the contrary that $u \in S$. Since S is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(u) \subset S$. Next, since $u = \lim_{k\to\infty} u_k$, there exists $K \in \mathbb{N}$ such that $\|u_k - u\| < \varepsilon$ for all $k \geq K$. In other words, $u_k \in B_\varepsilon(u) \subset S$ for all $k \geq K$. This is clearly a contradiction to the assumption that $u_k \notin S$ for all $k \in \mathbb{N}$. Thus $u \in S^C$, which shows that S^C is closed.

Now assume that S is not open. Then there exists $u \in S$ such that for every $\varepsilon > 0$ we have that $B_\varepsilon(u) \not\subset S$. That is, for every $\varepsilon > 0$ we have that $B_\varepsilon(u) \cap S^C \neq \emptyset$. As a consequence, we can find for each $k \in \mathbb{N}$ some $u_k \in S^C$ with $\|u_k - u\| < 1/k$. The resulting sequence $(u_k)_{k\in\mathbb{N}}$ then by construction converges to u and satisfies that $u_k \in S^C$ for all $k \in \mathbb{N}$. However, $\lim_{k\to\infty} u_k = u \notin S^C$, which shows that S^C is not closed. □

LEMMA 4.27. *Assume that $(U, \|\cdot\|)$ is a normed space, that $u \in U$, and $r > 0$. Denote by*

$$\overline{B}_r(u) := \big\{v \in U : \|u - v\| \leq r\big\}$$

the closed ball *of radius r around u. Then $\overline{B}_r(u)$ is a closed set.*

PROOF. Exercise! □

Now we can formulate an alternative characterisation of continuity:

THEOREM 4.28 (Continuity). *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and let $f\colon U \to V$ be a function. The following are equivalent:*

1. *The function f is continuous.*
2. *For every open set $A \subset V$ the set $f^{-1}(A) = \big\{u \in U : f(u) \in A\big\}$ is open.*
3. *For every closed set $K \subset V$ the set $f^{-1}(K) = \big\{u \in U : f(u) \in K\big\}$ is closed.*

PROOF. For the equivalence of the last two items, we observe that

$$f^{-1}(A)^C = \big\{u \in U : f(u) \notin A\big\} = f^{-1}(A^C).$$

Now the equivalence of these items follows from the fact that a set is open if and only if its complement is closed.

It remains to show the equivalence of the first item with any of the others. For this, we recall that the function f is continuous, if and only if for every $u \in U$ and every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|f(v) - f(u)\|_V < \varepsilon$ whenever $v \in V$ satisfies $\|v - u\|_U < \delta$. Now observe that this is equivalent to stating that for every $u \in U$ and every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$B_\delta(u) \subset f^{-1}(B_\varepsilon(u)).$$

Assume now that Item 2 holds. Let moreover $u \in U$ and let $\varepsilon > 0$. Since the set $B_\varepsilon(f(u)) \subset V$ is open, it follows that also $C := f^{-1}(B_\varepsilon(f(u)))$ is open. Moreover, we obviously have that $u \in C$. Thus, as C is open, there exists $\delta > 0$ such that $B_\delta(u) \subset C = f^{-1}(B_\varepsilon(f(u)))$. Since $u \in U$ and $\varepsilon > 0$ were arbitrary, it follows that f is continuous.

Assume now that f is continuous and let $A \subset V$ be open. We have to show that $f^{-1}(A)$ is open. Assume therefore that $u \in f^{-1}(A)$. Then $f(u) \in A$. Since A is open, there exists $\varepsilon > 0$ such that $B_\varepsilon(f(u)) \subset A$. Now the continuity of f implies the existence of $\delta > 0$ such that $B_\delta(u) \subset f^{-1}(B_\varepsilon(f(u)))$. Now the fact that $B_\varepsilon(f(u)) \subset A$ implies that

$$B_\delta(u) \subset f^{-1}(B_\varepsilon(f(u))) \subset f^{-1}(A).$$

Since $u \in f^{-1}(A)$ was arbitrary, this proves that $f^{-1}(A)$ is open. □

REMARK 4.29. Assume that $(U, \|\cdot\|)$ is a normed space and that $f\colon U \to \mathbb{R}$ is continuous. Then the sets

$$\big\{u \in U : f(u) = 0\big\} = f^{-1}(0)$$

and

$$\big\{u \in U : f(u) \geq 0\big\} = f^{-1}([0, \infty))$$

are closed, and the set

$$\big\{u \in U : f(u) > 0\big\} = f^{-1}((0, \infty))$$

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 4.30. *Assume that $(U, \|\cdot\|)$ is a normed space and that A_i, $i \in I$, are open subsets of U. Then the set $A := \bigcup_{i\in I} A_i$ is open.*

Similarly, if K_i, $i \in I$, are closed subsets of U, then the set $K := \bigcap_{i\in I} K_i$ is closed.

PROOF. Assume that $u \in A = \bigcup_{i\in I} A_i$. Then there exists an index $j \in I$ such that $u \in A_j$. Since A_j is open, there exists $r > 0$ such that $B_r(u) \subset A_j$. This implies that

$$B_r(u) \subset A_j \subset \bigcup_{i\in I} A_i = A,$$

which shows that A 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 4.31. *Assume that $(U, \|\cdot\|)$ is a normed space and that $A_1, \ldots, A_n$ are finitely many open sets. Then the set $A := \bigcap_{k=1}^n A_k$ 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 $u \in \bigcap_{k=1}^n A_k$. Then $u \in A_k$ for each k. Since each set A_k is open, there exist $r_k > 0$, $k = 1, \ldots, n$, such that $B_{r_k}(u) \subset A_k$. Now define $r := \min\{r_1, \ldots, r_n\}$. Then $B_r(u) \subset B_{r_k}(u) \subset A_k$ for each k, and thus $B_r(u) \subset \bigcap_{k=1}^n A_k = A$. 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 4.32 (Interior, closure, and boundary). Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ 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 4.33. 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 4.34. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ be a set. Then $u \in A^\circ$ if and only if there exists $r > 0$ such that $B_r(u) \subset A$.*

PROOF. Assume first that $u \in A^\circ$. By the definition of A°, there exists an open set $D \subset A$ such that $u \in D$. Since D is open, there exists $r > 0$ such that $B_r(u) \subset D$. As a consequence, $B_r(u) \subset D \subset A$.

Conversely, assume that there exists $r > 0$ such that $B_r(u) \subset A$. Since the set $B_r(u)$ is open and contained in A, it follows that $B_r(u) \subset A^\circ$. In particular, this shows that $u \in A^\circ$. □

LEMMA 4.35. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ 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 $D := A^C$. Applying the first result we have shown in this proof, we obtain that $\overline{D^C} \subset (D^\circ)^C$, 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 $D = A^C$, we obtain that $(\overline{D})^C \subset (D^C)^\circ$, or $(\overline{A^C})^C \subset A^\circ$, which shows that $(A^\circ)^C \subset \overline{A^C}$. □

LEMMA 4.36. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ 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 4.35 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 4.37. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ be a set. Then $u \in \overline{A}$, if and only if for all $r > 0$ we have that $B_r(u) \cap A \neq \emptyset$.*

PROOF. By Lemma 4.35, we have that $u \in \overline{A}$, if and only if $u \notin (A^C)^\circ$. By Proposition 4.34, this is the case if and only if for all $r > 0$ we have that $B_r(u) \not\subset A^C$. This, in turn, is the case if and only if for all $r > 0$ we have that $B_r(u) \cap A \neq \emptyset$, which proves the assertion. □

LEMMA 4.38. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ be a set. Then $u \in \partial A$, if and only if for all $r > 0$ we have that $B_r(u) \cap A \neq \emptyset$ and $B_r(u) \cap \overline{A} \neq \emptyset$.*

PROOF. By Lemma 4.36, we have that $u \in \partial A$, if and only if $u \in \overline{A}$ and $u \in \overline{A^C}$. By applying Lemma 4.37 to both $\overline{A}$ and $\overline{A^C}$, we thus obtain the claimed result. □

PROPOSITION 4.39. *Let $(U, \|\cdot\|)$ be a normed space, and let $A \subset U$ be a set. The closure $\overline{A}$ of A consists of all limits of convergent sequences contained in A. That is, $u \in \overline{A}$, if and only if there exists a sequence $(u_k)_{k\in\mathbb{N}} \subset A$ such that $u = \lim_{k\to\infty} u_k$.*

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 $u \in \overline{A}$. By Lemma 4.37, we have for all $k \in \mathbb{N}$ that $B_{1/k}(u) \cap A \neq \emptyset$. In other words, for every $k \in \mathbb{N}$ we can find $u_k \in A$ such that $\|u_k - u\| < 1/k$. Thus $u = \lim_{k\to\infty} u_k$, which proves the assertion. □

DEFINITION 4.40 (Dense set). Assume that $(U, \|\cdot\|)$ is a normed space, that $A \subset B \subset U$ 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 $u \in B$ some sequence $(u_k)_{k\in\mathbb{N}}$ with $u = \lim_{k\to\infty} u_k$ and $u_k \in A$ for all k.

EXAMPLE 4.41. 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 4.42 (Stone–Weierstraß). *The space $\mathcal{P}$ of polynomials is dense in $C([0, 1])$ with respect to the norm $\|\cdot\|_\infty$.*

THEOREM 4.43 (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 norm $\|\cdot\|_\infty$.*

REMARK 4.44. The previous result does *not* state that the Fourier series of an arbitrary continuous function f converges with respect to $\|\cdot\|_\infty$ 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. ■

4. Cauchy–sequences and Completeness

A large portion of linear algebra does not make full use of the real numbers, but can as well be formulated just for vector spaces over the rational numbers $\mathbb{Q}$. For instance, if a system of linear equations with only rational coefficients is solvable within the real numbers, it is already solvable within the rational numbers. The situation changes a bit if we want to talk about eigenvalues and singular values, and the corresponding decompositions, as these necessarily involve solutions of algebraic equations. Still, one can formulate all the theory that we covered in this course up

til now by moving from the rational numbers to the so called algebraic numbers, which include all roots of polynomials with rational coefficients.

However, we are now in the process of introducing analytic tools into linear algebra, and analysis does not work well with rational (or algebraic) numbers, as many of the well-known results from calculus fail to hold in that setting. 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 $f(b)$ does not hold if we only allow for function values that are rational.

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

$$x_k := \sum_{\ell=0}^{k} \frac{1}{\ell!}. \tag{24}$$

This is a sequence of rational numbers that converges to Euler's number e. However, Euler's number is not rational. Thus, if we restrict ourselves only to the rational numbers, the sequence does not converge at all, even though it should (the remainder $\sum_{\ell=k+1}^{\infty} 1/\ell!$ tends very fast to 0).

In the case of the rational numbers, we get around that problem by introducing the real numbers and working with them instead when we want to do calculus. Still, when we start working with (infinite dimensional) normed spaces over the real numbers, we observe similar phenomena to occur. Take for instance the space $C([0, 1])$ of continuous real functions on the interval $[0, 1]$ with the norm $\|f\|_1 = \int_0^1 |f(x)|\,dx$, and consider the sequence of functions

$$f_k(x) = \begin{cases} 0 & \text{if } x < \dfrac{1}{2}, \\ k\Big(x - \dfrac{1}{2}\Big) & \text{if } \dfrac{1}{2} \le x \le \dfrac{1}{2} + \dfrac{1}{k}, \\ 1 & \text{if } x > \dfrac{1}{2} + \dfrac{1}{k}. \end{cases} \tag{25}$$

This sequence appears to converge to the function

$$f(x) = \begin{cases} 0 & \text{if } x \le \dfrac{1}{2}, \\ 1 & \text{if } x > \dfrac{1}{2}. \end{cases}$$

This function, however, is not continuous and thus not contained in the space $C([0, 1])$ we are interested in. Even though the sequence *appears* to converge, it does not actually do so within the space of continuous functions.

On the space $C^1([-1, 1])$ of continuously differentiable real-valued functions on the interval $[-1, 1]$, we can consider the norm

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

Now consider the sequence of functions $(f_k)_{k\in\mathbb{N}} \subset C^1([-1, 1])$, $f_k(x) := \sqrt{x^2 + 1/k}$. Then $|f_k(x) - |x|| \to 0$ for all k, and also $f_k'(x) \to \operatorname{sgn}(x)$ for all $x \neq 0$ (which is the derivative of the function $x \mapsto |x|$ outside of 0), and it therefore appears as if the sequence $(f_k)_{k\in\mathbb{N}}$ would converge to the function $f(x) = |x|$. Again, we have that $f \notin C^1([-1, 1])$, but the situation is even worse than in the previous case: We

cannot even define $\|f\|_{\infty,1}$, because we would need the derivative of f to be able to do so.

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 case of sequences in $\mathbb{Q}$, and some space of piecewise continuous functions or piecewise diffentiable functions in the second case.

Now there are (at least) two questions: First, can we formally characterise the situation where this phenomenon occurs. That is, can we formalise a notion of completeness and incompleteness of normed spaces, and can we do so in an *intrinsic* way without regarding any larger spaces. Second, if we know that a given normed space $(U, \|\cdot\|)$ is incomplete, can we find a larger space $(V, \|\cdot\|)$ containing U, such that every sequence in U that should converge, actually converges in V? 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 normed space: The sequence $(u_n)_{n\in\mathbb{N}}$ converges in the normed space $(U, \|\cdot\|)$ to $u \in U$, if there exists for each $\varepsilon > 0$ some $N \in \mathbb{N}$ such that $\|u_n - u\| < \varepsilon$ for all $n \geq N$. Now assume that this is the case, and take n, $m \geq N$. Then the triangle inequality implies that

$$\|u_n - u_m\| \leq \|u_n - u\| + \|u_m - u\| < 2\varepsilon.$$

Note here that the inequality $\|u_n - u_m\| < 2\varepsilon$ makes no mention of the limit u of the sequence. Thus we can use this in a definition of sequences that should converge.

Definition 4.45 (Cauchy sequence). A sequence $(u_n)_{n\in\mathbb{N}}$ in a normed space $(U, \|\cdot\|)$ is a *Cauchy sequence*, if there exists for each $\varepsilon > 0$ some $N \in \mathbb{N}$ such that

$$\|u_n - u_m\| < \varepsilon \qquad \text{whenever } n, m \geq N.$$

■

From our discussion before the definition, we immediately get the following result:

Lemma 4.46. *Every convergent sequence in a normed space is a Cauchy sequence.*

Conversely, the sequence defined in (24) 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. As a preparation for this result, we show that Cauchy sequences, like convergent sequences, are bounded.

Lemma 4.47. *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, $m \geq N$. Define now

$$r := \max\{\|u_1\|, \|u_2\|, \ldots, \|u_N\|\} + 1.$$

Then by construction $u_n \in B_r(0)$ for all $1 \leq n \leq N$. Also, for $n \geq N$ we have that

$$\|u_n\| \leq \|u_n - u_N\| + \|u_N\| < \|u_N\| + 1 \leq r.$$

Thus we have that $u_n \in B_r(0)$ for all $n \in \mathbb{N}$, which shows that the sequence is bounded. □

Theorem 4.48. *Every Cauchy sequence in the real numbers $\mathbb{R}$ converges.*

PROOF*. Assume that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. Then by Lemma 4.47, the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded. Define now for $N \in \mathbb{N}$

$$a_N := \inf\{x_n : n \geq N\} \qquad \text{and} \qquad b_N := \sup\{x_n : n \geq N\}.$$

Note that the existence of a_N and b_N follows from Axiom 1.36. Also, we have that

$$\inf\{x_n : n \in \mathbb{N}\} = a_1 \leq a_N \leq b_M \leq b_1 = \sup\{x_n : n \in \mathbb{N}\}$$

for every N, $M \in \mathbb{N}$. Define now

$$x := \sup\{a_N : N \in \mathbb{N}\}.$$

Then we have by definition that $x \geq a_N$ for all N. In addition, since $a_N \leq b_M$ for all N and M, it follows that also $x \leq b_M$ for all $M \in \mathbb{N}$. We now claim that $x = \lim_{n\to\infty} x_n$.

Let therefore $\varepsilon > 0$. Since $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, there exists some $N \in \mathbb{N}$ such that $|x_n - x_m| < \varepsilon/2$ for all n, $m \geq N$. Let now $n \geq N$ be fixed.

Since $a_N = \inf\{x_k : k \geq N\}$, there exists some $m \geq N$ such that $x_m < a_N + \varepsilon/2$. Thus

$$(x_n - x) \leq x_n - a_N < x_n - (x_m - \varepsilon/2) \leq |x_n - x_m| + \varepsilon/2 < \varepsilon.$$

Moreover, since $x \leq b_N$, there exists some $\ell \geq N$ such that $x_\ell > b_N - \varepsilon/2$. Thus

$$(x - x_n) \leq b_N - x_n < x_\ell + \varepsilon/2 - x_n \leq |x_\ell - x_n| + \varepsilon/2 < \varepsilon.$$

This shows that $|x - x_n| < \varepsilon$ for all $n \geq N$, which proves that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to x. □

REMARK 4.49. 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 $\|x_n - x_{n+1}\|$, 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 4.50 (Completeness). A normed space $(U, \|\cdot\|)$ is *complete*, if every Cauchy sequence in U converges. ■

Thus our Theorem above can be rephrased as saying that the real numbers are complete.

LEMMA 4.51. *The spaces $\mathbb{C}^d$ and $\mathbb{R}^d$ with the Euclidean norm $\|\cdot\|_2$ are complete.*

PROOF. We only prove the claim in the (slightly more cumbersome) case of $\mathbb{C}^d$.

Assume that $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}^d$. Denote now, for $1 \leq j \leq d$, by $z_k^{(j)} \in \mathbb{C}$ the j-th component of the vector z_k and write $z_k^{(j)} = a_k^{(j)} + ib_k^{(j)}$, where $a_k^{(j)}$ and $b_k^{(j)}$ are the real and imaginary parts of $z_k^{(j)}$, respectively. Now let $\varepsilon > 0$. Since $(z_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{C}^d$, there exists some $N \in \mathbb{N}$ such that $\|z_n - z_m\|_2 < \varepsilon$ whenever n, $m \geq N$. Thus we have for every n, $m \geq N$ that

$$|a_n^{(j)} - a_m^{(j)}| \leq |z_n^{(j)} - z_m^{(j)}| \leq \|z_n - z_m\| < \varepsilon,$$

and similarly

$$|b_n^{(j)} - b_m^{(j)}| \leq |z_n^{(j)} - z_m^{(j)}| \leq \|z_n - z_m\| < \varepsilon.$$

Thus the sequences $(a_n^{(j)})_{n\in\mathbb{N}}$ and $(b_n^{(j)})_{n\in\mathbb{N}}$ are Cauchy sequences in $\mathbb{R}$ and therefore have limits, say

$$a^{(j)} := \lim_{n\to\infty} a_n^{(j)} \qquad \text{and} \qquad b^{(j)} := \lim_{n\to\infty} b_n^{(j)}.$$

Define now $z \in \mathbb{C}^d$ by $z := (z^{(1)}, \ldots, z^{(d)})$ with

$$z^{(j)} := a^{(j)} + ib^{(j)}.$$

We claim that $z = \lim_{n\to\infty} z_n$.

Let therefore $\varepsilon > 0$. Since $a^{(j)} = \lim_{n\to\infty} a_n^{(j)}$ and $b^{(j)} = \lim_{n\to\infty} b_n^{(j)}$, there exist $N^{(j)} \in \mathbb{N}$ such that

$$|a_n^{(j)} - a^{(j)}| < \frac{\varepsilon}{\sqrt{2d}} \qquad \text{and} \qquad |b_n^{(j)} - b^{(j)}| < \frac{\varepsilon}{\sqrt{2d}}$$

for every $n \geq N^{(j)}$. Define now $N := \max\{N^{(1)}, \dots, N^{(d)}\}$ and let $n \geq N$. Then

$$\begin{aligned}\|z_n - z\|_2 = \Big(\sum_{j=1}^{d} |z_n^{(j)} - z^{(j)}|^2\Big)^{1/2} &= \Big(\sum_{j=1}^{d} |a_n^{(j)} - a^{(j)}|^2 + \sum_{j=1}^{d} |b_n^{(j)} - b^{(j)}|^2\Big)^{1/2} \\ &\leq \Big(\sum_{j=1}^{d} \frac{\varepsilon^2}{2d} + \sum_{j=1}^{d} \frac{\varepsilon^2}{2d}\Big)^{1/2} = \varepsilon.\end{aligned}$$

This proves that $z = \lim_{n\to\infty} z_n$, which in turn proves the completeness of $\mathbb{C}^d$. □

LEMMA 4.52. *The space $C([0,1])$ with the norm $\|f\|_1 = \int_0^1 |f(x)|\,dx$ is* not *complete.*

PROOF. Let $f_k \colon [0,1] \to \mathbb{R}$ be as in (25). We will show explicitly that the sequence $(f_k)_{k\in\mathbb{N}}$ is a Cauchy sequence with respect to $\|\cdot\|_1$. 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

$$\|f_n - f_m\|_1 = \int_{1/2}^{1/2+1/N} |f_n(x) - f_m(x)|\,dx \leq \int_{1/2}^{1/2+1/N} 1\,dx = \frac{1}{N} < \varepsilon.$$

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 norm.

THEOREM 4.53. *The space $C([0,1])$ with the norm $\|f\|_\infty = \max_{x\in[0,1]}|f(x)|$ 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

$$|f_n(x) - f_m(x)| \leq \max_{y\in[0,1]} |f_n(y) - f_m(y)| = \|f_n - f_m\|_\infty$$

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 $\|f_n - f\|_\infty \to 0$ 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

$$\|f_n - f\|_\infty = \max_{x\in[0,1]} |f_n(x) - f_m(x)| < \varepsilon/3$$

for all n, $m \in \mathbb{N}$. Since $f(x) = \lim_{m\to\infty} f_m(x)$, we have that

$$|f_n(x) - f(x)| = \lim_{m\to\infty} |f_N(x) - f_m(x)| \leq \varepsilon/3 \qquad \text{for all } x \in [0,1] \text{ and } n \geq N. \tag{26}$$

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 (26) with $n = N$ we thus obtain for every z with $|y - z|$ that

$$|f(y) - f(z)| \leq |f(y) - f_N(y)| + |f_N(y) - f_N(z)| + |f_N(z) - f(z)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.$$

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 (26) again to conclude that

$$\|f_n - f\|_\infty = \max_{x\in[0,1]} |f_n(x) - f(x)| \leq \varepsilon/3 \qquad \text{for all } n \geq N.$$

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 $\|\cdot\|_\infty$. □

Because of their importance, complete normed spaces as well as complete inner product spaces have earned their own names:

DEFINITION 4.54 (Banach space). A *Banach space* is a complete normed space. ■

DEFINITION 4.55 (Hilbert space). A *Hilbert space* is a complete inner product space. ■

Thus we can rephrase the previous results as follows:

EXAMPLE 4.56. The space $(\mathbb{K}^d, \|\cdot\|_2)$ is a Hilbert space. ■

THEOREM 4.57. *The space* $(C([0,1]), \|\cdot\|_\infty)$ *is a Banach space.*

LEMMA 4.58. *The space* $(C[0,1], \|\cdot\|_1)$ *is* not *a Banach space.*

REMARK 4.59. Assume that $(f_n)_{n\in\mathbb{N}}$ is a sequence of functions $f_n\colon [0,1] \to \mathbb{R}$, and assume that $f\colon [0,1] \to \mathbb{R}$.

- 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 norm $\|\cdot\|_\infty$ (though it makes sense to use the expression "uniform convergence" also for discontinuous functions). Thus Theorem 4.53 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 $f\colon [0,1] \to \mathbb{R}$ we cannot conclude that f is continuous as well. An example is the sequence of continuous functions from (25) that converges pointwise to the discontinuous function $f\colon [0,1] \to \mathbb{R}$, $f(x) = 0$ for $0 \le x \le 1/2$ and $f(x) = 1$ for $1/2 < x \le 1$. ■

REMARK 4.60. In Theorem 4.53, 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 norm $\|\cdot\|_\infty$.

Now assume that $(V, \|\cdot\|_V)$ is an arbitrary normed space. Then we can consider the space $C([0,1], V)$ of continuous functions $f\colon [0,1] \to V$ with the norm
$$\|f\|_\infty := \max_{x\in[0,1]} \|f(x)\|_V.$$
With essentially the same proof, we can show that the space $C([0,1], V)$ with this norm is complete as well provided that V is complete. 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]

Since the space $\mathbb{R}^d$ is complete respect to the Euclidean norm, this implies in particular that the space $C(I;\mathbb{R}^d)$ is complete with respect to $\|\cdot\|_\infty$ whenever $I \subset \mathbb{R}$ is a closed and bounded interval. This is something we will make use of in Section 7 below. ■

THEOREM 4.61. *Assume that* $(U, \|\cdot\|)$ *is a Banach space and that* $V \subset U$ *is a subspace. Then* V *is complete (with respect to the restriction of the norm* $\|\cdot\|$ *to* V*) if and only if* V *is closed (as a subspace of* U*).*

PROOF. Assume that V is complete as a normed space. Let moreover $(u_n)_{n\in\mathbb{N}}$ be a sequence with $u_n \in V$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} u_n = u$ for some $u \in U$. We have to show that $u \in V$.

Since the sequence $(u_n)_{n\in\mathbb{N}}$ converges in U, it is a Cauchy sequence in U. This, however, implies that it is a Cauchy sequence in V, which is complete. Thus the sequence $(u_n)_{n\in\mathbb{N}}$ converges to some $w \in V$. Because of the uniqueness of the limit

[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 intervals, this is no longer the case, and the "norm" of a function could become $+\infty$. We could remedy this problem, though, by restricting ourselves to *bounded* continuous functions, but we won't go that far.

of convergent sequences in normed spaces, it follows that $u = w \in V$. Thus V is closed.

Now assume that V is closed, and let $(u_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in V. Then this is a Cauchy sequence in U as well, and thus the completeness of U implies that it converges to some $u \in U$. Because V is closed and $u = \lim_{n\to\infty} u_n$, it follows that $u \in V$. Thus the Cauchy sequence $(u_n)_{n\in\mathbb{N}}$ converges in V, and thus V is complete. □

5. Equivalence of Norms

DEFINITION 4.62. Assume that U is a vector space and that $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on U. We say that the norms are *equivalent*, if there exist constants $0 < c < C < \infty$ such that

$$c\|u\|_a \leq \|u\|_b \leq C\|u\|_a$$

for all $u \in U$. ■

EXAMPLE 4.63. The norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{K}^n$ are equivalent. Indeed, we have for every $u \in \mathbb{K}^n$ the estimate

$$\|u\|_\infty = \max_{1\leq i\leq n} |u_i| \leq \sum_{i=1}^{n} |u_i| = \|u\|_1 \leq n \max_{1\leq i\leq n} |u_i| = n\|u\|_\infty.$$

Thus we can use the constants $c = 1$ and $C = n$. ■

EXAMPLE 4.64. The norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $C([0,1])$ are *not* equivalent. Although we can estimate

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

for all $f \in C([0,1])$, there exists no constant $C > 0$ such that $\|f\|_\infty \leq C\|f\|_1$ for all $f \in C([0,1])$. This can be seen be regarding the functions $f_n(x) := x^n$. We have $\|f_n\|_\infty = 1$ for all $n \in \mathbb{N}$, but $\|f_n\|_1 = 1/(n+1)$. Thus such a constant $C > 0$ would have to satisfy the estimate $C \geq n + 1$ for all $n \in \mathbb{N}$, which is of course impossible. ■

If $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms, then they induce the same convergence of sequences:

PROPOSITION 4.65. *Assume that U is a vector space and $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on U. Assume moreover that $(u_n)_{n\in\mathbb{N}}$ is a sequence in U.*

- *$(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence with respect to $\|\cdot\|_a$ if and only if it is a Cauchy sequence with respect to $\|\cdot\|_b$.*
- *$(u_n)_{n\in\mathbb{N}}$ converges with respect to $\|\cdot\|_a$ if and only if it converges with respect to $\|\cdot\|_b$. Moreover, in this case the limits are the same.*

PROOF. Exercise. □

As a consequence, equivalent norms define the same closed (and thus open) sets on a vector space.

COROLLARY 4.66. *Assume that U is a vector space and $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on U.*

- *A subset $A \subset U$ is open with respect to $\|\cdot\|_a$ if and only if it is open with respect to $\|\cdot\|_b$.*
- *A subset $K \subset U$ is closed with respect to $\|\cdot\|_a$ if and only if it is open with respect to $\|\cdot\|_b$.*

COROLLARY 4.67. *Assume that U is a vector space and $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms on U. Then U is complete with respect to $\|\cdot\|_a$ if and only if it is complete with respect to $\|\cdot\|_b$.*

Of particular importance is the case of a finite dimensional vector space. Here we can show that in fact all norms are equivalent. Thus the choice of the norm has no influence on the convergence of sequences or continuity of functions.

THEOREM 4.68. *Assume that U is a finite dimensional vector space. Then all norms on U are equivalent.*

PROOF. Let $(u_1, \dots, u_n)$ be a basis of U. We define a norm $\|\cdot\|_1$ on U setting

$$\|u\|_1 := \sum_{i=1}^n |x_i| \text{ if } u = \sum_{i=1}^n x_i u_i.$$

That is, $\|u\|_1$ is the 1-norm of the coordinate representation of u with respect to the basis $(u_1, \dots, u_n)$. It is then sufficient to show that all norms on U are equivalent to $\|\cdot\|_1$.

Assume therefore that $\|\cdot\|$ is a norm on U. Define

$$C := \max\{\|u_1\|, \dots, \|u_n\|\}.$$

Assume moreover that

$$u = \sum_{i=1}^n x_i u_i$$

is a vector in U. Then

$$\|u\| = \Big\|\sum_{i=1}^n x_i u_i\Big\| \le \sum_{i=1}^n |x_i| \|u_i\| \le \Big(\sum_{i=1}^n |x_i|\Big) \max_{1\le j\le n} \|u_j\| = C\|u\|_1.$$

Next define

$$c := \inf\big\{\|v\| : \|v\|_1 = 1\big\}.$$

Clearly, $c \ge 0$. In fact, we will show that $c > 0$. For this we note first that we can equivalently write c as

$$c = \inf\Big\{\Big\|\sum_{i=1}^n x_i u_i\Big\| : x \in \mathbb{K}^n,\ \sum_{i=1}^n |x_i| = 1\Big\}.$$

Define now the mapping $f\colon \mathbb{K}^n \to \mathbb{R}$,

$$f(x) := \Big\|\sum_{i=1}^n x_i u_i\Big\|.$$

We claim that f is continuous. By the reverse triangle inequality for the norm $\|\cdot\|$ we have for all x, $y \in \mathbb{K}^n$ that

$$\begin{aligned} |f(x) - f(y)| &= \Big|\Big\|\sum_{i=1}^n x_i u_i\Big\| - \Big\|\sum_{i=1}^n y_i u_i\Big\|\Big| \le \Big\|\sum_{i=1}^n (x_i - y_i) u_i\Big\| \\ &\le \sum_{i=1}^n |x_i - y_i| \|u_i\| \le C \sum_{i=1}^n |x_i - y_i| = C \sum_{i=1}^n 1 \cdot |x_i - y_i| \le C\sqrt{n}\|x - y\|_2. \end{aligned}$$

Here we have used the Cauchy–Schwarz–Bunyakovsky inequality in the last step. This, however, shows that f in fact is Lipschitz continuous on $\mathbb{K}^n$. Now the Extremal Value Theorem implies that the function f attains its minimum (and maximum) on the closed and bounded set $K := \big\{x \in \mathbb{K}^n : \sum_{i=1}^n |x_i| = 1\big\}$. Moreover, by construction of f we have that $f(x) > 0$ for all $x \in K$. Thus

$$c = \inf_{x\in K} f(x) > 0.$$

Now let $u \in U \setminus \{0\}$ and denote $v := u/\|u\|_1$. Then $\|v\|_1 = 1$ and thus $\|v\|_1 \geq c$. Thus

$$\|u\| = \|u\|_1 \left\| \frac{u}{\|u\|_1} \right\| = \|u\|_1 \|v\| \geq c\|u\|_1,$$

which concludes the proof. □

6. Banach's Fixed Point Theorem

DEFINITION 4.69 (Lipschitz continuity). Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and let $f\colon U \to V$ be a function. We say that f is *Lipschitz continuous*, if there exists $L \geq 0$ such that

$$\|f(u) - f(v)\|_V \leq L\|u - v\|_U \tag{27}$$

for all u, $v \in U$. A constant L with this property is called a *Lipschitz constant* of f. ■

REMARK 4.70. Sometimes one defines *the Lipschitz constant of* f as the smallest constant for which (27) 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 4.71 (Continuity of Lipschitz functions). *Every Lipschitz continuous function is continuous.*

PROOF. Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are normed spaces and let $f\colon U \to V$ 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 u, $v \in U$ be such that $\|u - v\|_U < \delta$. Then

$$\|f(u) - f(v)\|_V \leq L\|u - v\|_U < L\delta = \varepsilon.$$

Thus f is continuous. □

REMARK 4.72. 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

$$f(y) = f(x) + \int_0^1 \big\langle \nabla f\big(x + t(y - x)\big), y - x \big\rangle \, dt$$

and therefore

$$\begin{aligned} \|f(y) - f(x)\| &= \left\| \int_0^1 \big\langle \nabla f\big(x + t(y - x)\big), y - x \big\rangle \, dt \right\| \\ &\leq \int_0^1 \big| \big\langle \nabla f\big(x + t(y - x)\big), y - x \big\rangle \big| \, dt \\ &\leq \int_0^1 \|\nabla f\big(x + t(y - x)\big)\| \|y - x\| \, dt \leq L\|y - x\| \end{aligned}$$

for all x, $y \in \mathbb{R}^n$. ■

DEFINITION 4.73 (Contraction). Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces and let $f\colon U \to V$ be a function. We say that f is a contraction, if f is Lipschitz continuous with a Lipschitz constant L satisfying $0 \leq L < 1$. ■

In other words, f is a contraction, if there exists $0 \leq L < 1$ such that

$$\|f(u) - f(v)\|_V \leq L\|u - v\|_U$$

for all u, $v \in U$.

REMARK 4.74. Note that it is not sufficient for a function f to be a contraction that it satisfies

$$\|f(u) - f(v)\|_V < \|u - v\|_U$$

for all $u \neq v$. 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 4.75 (Banach's Fixed Point Theorem). *Assume that $(U, \|\cdot\|)$ is a Banach space and that $f\colon U \to U$ is a contraction. Then there exists a unique element $u^* \in U$ (a* fixed point of f*) such that*

$$f(u^*) = u^*. \tag{28}$$

PROOF. Let $0 \leq L < 1$ be a Lipschitz constant of f.

We start by showing that the solution of (28), *if it exists*, is unique. Assume therefore that u, $v \in U$ satisfy $f(u) = u$ and $f(v) = v$. Then

$$\|u - v\| = \|f(u) - f(v)\| \leq L\|u - v\|.$$

Since $L < 1$, this is only possible if $\|u - v\| = 0$ and thus $u = v$. This shows uniqueness of a fixed point of f. It remains to show that a fixed point actually exists.

Let now $u_0 \in U$ be arbitrary and consider the sequence given by

$$u_{n+1} = f(u_n).$$

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 $n \in \mathbb{N}$,

$$\|u_{n+1} - u_n\| = \|f(u_n) - f(u_{n-1})\| \leq L\|u_n - u_{n-1}\|.$$

By induction over n we thus obtain that

$$\|u_{n+1} - u_n\| \leq L^n\|u_1 - u_0\|.$$

Thus we have for all $n \in \mathbb{N}$ and $k \in \mathbb{N}$ that

$$\begin{aligned} \|u_{n+k} - u_n\| &\leq \sum_{j=1}^{k} \|u_{n+j} - u_{n+j-1}\| \leq \sum_{j=1}^{k} L^{n+j-1}\|u_1 - u_0\| \\ &= L^n\|u_1 - u_0\| \sum_{j=1}^{k} L^{j-1} \leq L^n\|u_1 - u_0\| \sum_{j=0}^{\infty} L^j = \frac{L^n}{1-L}\|u_1 - u_0\|. \end{aligned} \tag{29}$$

Now let $\varepsilon > 0$ and let $N \in \mathbb{N}$ be such that

$$\frac{L^N}{1-L}\|u_1 - u_0\| < \varepsilon.$$

Let moreover n, $m \geq N$. Without loss of generality, assume that $m \geq n$. Then we can apply (29) with $k = m - n$ and obtain that

$$\|u_m - u_n\| \leq \frac{L^n}{1-L}\|u_1 - u_0\| \leq \frac{L^N}{1-L}\|u_1 - u_0\| < \varepsilon.$$

Thus the sequence $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Since $(U, \|\cdot\|)$ is a Banach space, there exists $u^* \in U$ such that $u^* = \lim_{n\to\infty} u_n$.

It remains to show that u^* is a fixed point of f. Since f is a contraction, it is in particular Lipschitz continuous and thus continuous. Since $u^* = \lim_{n\to\infty} u_n$ by definition, we thus have that

$$f(u^*) = \lim_{n\to\infty} f(u_n) = \lim_{n\to\infty} u_{n+1} = u^*,$$

which concludes the proof. □

REMARK 4.76. 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 u^* of f. In fact, we have shown that the sequence

$$u_{n+1} = f(u_n)$$

converges for each starting point $u_0 \in U$ to u^*. 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 u^* after a finite number of steps:

By letting k tend to infinity in (29), we obtain the *a-priori estimate*

$$\|u^* - u_n\| \leq \frac{L^n}{1-L}\|u_1 - u_0\|.$$

This allows us to estimate already after the first step how close we will be to u^* after n steps.

In addition, we can show the *a-posteriori* estimate

$$\|u^* - u_n\| \leq \frac{L}{1-L}\|u_n - u_{n-1}\|.$$

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. ■

Sometimes we are given a mapping f that is not a contraction on the whole space U but only on some subset $K \subset U$. In this case, it is still possible to apply a variation of Banach's fixed point theorem, provided that K is closed and f maps K onto itself:

PROPOSITION 4.77. *Assume that $(U, \|\cdot\|)$ is a Banach space, that $f\colon U \to U$ is a function, and that $K \subset U$ is a closed subset. Assume moreover that*

$$f(u) \in K \qquad \text{whenever } u \in K$$

and that there exists $0 \leq L < 1$ such that

$$\|f(u) - f(v)\| \leq L\|u - v\| \qquad \text{for all } u, v \in K.$$

Then there exists a unique $u^* \in K$ such that $f(u^*) = u^*$.

PROOF. Choose $u_0 \in K$ arbitrary and define again a sequence setting $u_{n+1} := f(u_n)$. Then $u_n \in K$ for all $n \in \mathbb{N}$. Applying the same estimates as in the proof of Theorem 4.75, we then obtain that $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and therefore converges to some $u^* \in U$. Because K is closed and $u_n \in K$ for all n, it follows that $u^* \in K$. Moreover, as in the proof of Theorem 4.75 we obtain that u^* is a fixed point of f. For the uniqueness of that fixed point (within K), we can also apply the same ideas as there. □

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

$$f(x) + \int_0^1 k(x,y)\, f(y)\, dy = h(x). \tag{30}$$

That is, we want to find a function $f \in C([0,1])$ such that (30) 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 define 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 (30) 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 norm $\|\cdot\|_\infty$. For that

$$\begin{aligned}
\|Tf - Tg\|_\infty &= \max_{x\in[0,1]} \big|Tf(x) - Tg(x)\big| \\
&= \max_{x\in[0,1]} \Big| \int_0^1 k(x,y)\big(f(y) - g(y)\big)\, dy \Big| \\
&\le \max_{x\in[0,1]} \int_0^1 \big|k(x,y)\big|\big|f(y) - g(y)\big|\, dy \\
&\le \max_{x\in[0,1]} \int_0^1 c\big|f(y) - g(y)\big|\, dy \\
&= c \int_0^1 \big|f(y) - g(y)\big|\, dy \\
&\le c \max_{y\in[0,1]} \big|f(y) - g(y)\big| \\
&= c\|f - g\|_\infty.
\end{aligned}$$

Thus T is a contraction on the Banach space $\big(C([0,1]), \|\cdot\|_\infty\big)$, and therefore T has a unique fixed point. Put differently, the equation (30) has, under the given assumptions on k and h, a unique continuous solution f.

7. 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

$$\begin{aligned} y' &= f(t,y), \\ y(t_0) &= y_0, \end{aligned} \tag{31}$$

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 4.78 (Picard–Lindelöf). *Let $a > 0$ and $b > 0$, and define the cylinder*

$$Z := \big\{(t,y) \in \mathbb{R}\times\mathbb{R}^d : |t - t_0| \le a \text{ and } |y - y_0|_2 \le b\big\}.$$

Assume that the function $f\colon \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ *is continuous on* Z *and that it satisfies a* Lipschitz condition *with respect to its second component in the sense that there exists* $L \geq 0$ *such that*

$$\big|f(t,y) - f(t,z)\big|_2 \leq L|y-z|_2$$

for all (t,y), $(t,z) \in Z$. *Assume moreover that*

$$M := \max_{(t,y)\in Z} |f(t,y)|_2 \leq \frac{b}{a}.$$

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 (31). 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

$$y(t) = y_0 + \int_{t_0}^t f\big(s, y(s)\big)\,ds, \tag{32}$$

then y is necessarily differentiable (because the right hand side is) and it satisfies (31).

Define therefore the mapping $T\colon C(I;\mathbb{R}^d) \to C(I;\mathbb{R}^d)$,

$$Ty(t) := y_0 + \int_{t_0}^t f\big(s, y(s)\big)\,ds.$$

Then y solves the initial value problem 31 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

$$K := \big\{y \in C(I;\mathbb{R}^d) : |y(t) - y_0|_2 \leq b \text{ for all } t \in I\big\}$$

and assume that $y \in K$. Then

$$\big|f\big(s, y(s)\big)\big|_2 \leq M$$

for all $s \in I$ and thus

$$\big|Ty(t) - y_0\big|_2 = \Big|\int_{t_0}^t f\big(s, y(s)\big)\,ds\Big|_2 \leq \Big|\int_{t_0}^t \big|f\big(s, y(s)\big)\big|_2\,ds\Big| \leq |t-t_0|M \leq a\frac{b}{a} = b$$

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 (32) (or, equivalently, of the ODE (31)). 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 $|y(t) - y_0|_2 > b$. Since $y(t) = y_0$ by assumption, it follows that $t \neq t_0$.

Assume now that $t > t_0$. Since the mapping $t \mapsto |y(t)-y_0|_2$ is continuous, there exists a minimal value $t_0 < t_1 < t_0 + a$ such that $|y(t_1) - y_0|_2 = b$. In particular, $|y(t_1) - y_0|_2 < b$ for all $t_0 < s < t_1$. As a consequence

$$b = |y(t_1) - y_0|_2 = \Big| \int_{t_0}^{t_1} f\big(s, y(s)\big)\, ds \Big|_2 \le M|t_1 - t_0| < Ma = b,$$

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 norm on $C(I;\mathbb{R}^d)$.

We now need to define a norm on $C(I;\mathbb{R}^d)$ with respect to which the mapping T is a contraction on K. For that we define

$$\|y\| := \max_{t\in I} e^{-2L|t-t_0|}|y(t)|_2. \tag{33}$$

Assume now that y, $z \in K$. Then

$$\begin{aligned}
\|Ty - Tz\| &= \max_{t\in I} e^{-2L|t-t_0|}|Ty(t) - Tz(t)|_2 \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big| y_0 + \int_{t_0}^{t} f\big(s, y(s)\big)\, ds - y_0 - \int_{t_0}^{t} f\big(s, z(s)\big)\, ds \Big|_2 \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} f\big(s, y(s)\big) - f\big(s, z(s)\big)\, ds \Big|_2 \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} \big| f\big(s, y(s)\big) - f\big(s, z(s)\big)\big|_2\, ds \Big| \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} L|y(s) - z(s)|_2\, ds \Big| \\
&= \max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} Le^{2L|s-t_0|}e^{-2L|s-t_0|}|y(s) - z(s)|_2\, ds \Big| \\
&\le \max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} Le^{2L|s-t_0|}\, ds \Big| \Big(\max_{s\in I} e^{-2L|s-t_0|}|y(s) - z(s)|_2 \Big) \\
&= \Big(\max_{t\in I} e^{-2L|t-t_0|}\Big| \int_{t_0}^{t} Le^{2L|s-t_0|}\, ds \Big| \Big) \|y - z\| \\
&= \Big(\max_{t\in I} e^{-2L|t-t_0|}\frac{e^{2L|t-t_0|}}{2} \Big) \|y - z\| \\
&= \frac{1}{2}\|y - z\|.
\end{aligned}$$

This shows that T is a contraction with respect to the norm defined in (33).

Step 5: Show that $C(I;\mathbb{R}^d)$ is complete and K closed.

Denote by $\|\cdot\|_\infty$ the standard ∞-norm on $C(I;\mathbb{R}^d)$, that is,

$$\|y\| := \max_{t\in I}|y(t)|_2.$$

Since

$$1 \ge e^{-2L|t-t_0|} \ge e^{-2La}$$

for all $t \in I$, we have that

$$e^{2La}\|y\|_\infty \le \|y\| \le \|y\|_\infty$$

for all $y \in C(I;\mathbb{R}^d)$. Thus $\|\cdot\|$ and $\|\cdot\|_\infty$ are equivalent norms on $C(I;\mathbb{R}^d)$. Since $C(I;\mathbb{R}^d)$ with the norm $\|\cdot\|_\infty$ is complete, this implies that it is also complete with the norm $\|\cdot\|$ and thus a Banach space.

Also, a set is closed with respect to $\|\cdot\|$ if and only if it is closed with respect to $\|\cdot\|_\infty$. Since K is precisely the closed ball of radius b around the constant function y_0 with respect to $\|\cdot\|_\infty$, it is closed for that norm. Thus K is closed for the norm $\|\cdot\|$ as well.

Step 6: Collect everything and call it a day.

We have shown that K is a closed subset of the Banach 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 (31) is equivalent to finding a fixed point of T in $C(I;\mathbb{R}^d)$ (Step 1), we can conclude that the solution (31) 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. □

REMARK 4.79. It is possible to show that the ODE (31) 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

$$y' = y^{1/3}, \qquad y(0) = 0.$$

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 $t \geq 0$ are also solutions of the same ODE. Even worse, for any $T > t_0$ we can define the functions

$$y_{\pm,T}(t) := \begin{cases} 0 & \text{if } t < T, \\ \pm(2(t-T)/3)^{3/2} & \text{if } t \geq T. \end{cases}$$

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 4.80 (Picard iteration). Since we have proved Theorem 4.78 by showing that the Banach fixed point Theorem is applicable, we obtain in addition that the fixed point iteration (the *Picard iteration*)

$$y_{k+1}(t) = Ty_k(t) = y_0 + \int_{t_0}^t f\big(s, y_k(s)\big)\, ds \tag{34}$$

(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 (31).

If we apply this to the ODE

$$y' = y, \qquad y(0) = 1,$$

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 (34) 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. ■

8. 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 4.81 (p-norm on $\mathbb{K}^n$). Let $1 \leq p < \infty$. We define the mapping $\|\cdot\|_p \colon \mathbb{K}^n \to \mathbb{R}$,

$$\|x\|_p := \big(|x_1|^p + \ldots + |x_n|^p\big)^{1/p}$$

for $x \in \mathbb{K}^n$.

Similarly, we define $\|\cdot\|_\infty \colon \mathbb{K}^n \to \mathbb{R}$ by

$$\|x\|_\infty := \max\big\{|x_1|, \ldots, |x_n|\big\}$$

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 4.82 (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 4.83 (Young's inequality). *Let $1 < p < \infty$, and let $1 < p_* < \infty$ be the conjugate of p. Then for all $a, b \geq 0$ the inequality*

$$ab \leq \frac{1}{p} a^p + \frac{1}{p_*} b^{p_*} \tag{35}$$

holds.

PROOF. For $b = 0$, this inequality holds obviously. Now assume that $b > 0$ and consider the function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$,

$$f(a) := ab - \frac{1}{p}a^p.$$

Then (35) is equivalent to the statement that $f(a) \leq \frac{1}{p_*}b^{p_*}$ for all $a \geq 0$. In order to show this, we compute the maximum of f. Since $\lim_{a\to\infty} f(a) = -\infty$ (because the term $-\frac{1}{p}a^p$ decreases superlinearly towards $-\infty$, whereas the term ab only increases linearly), this maximum actually exists in $\mathbb{R}_{\geq 0}$. 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 $f(a) \leq f(b^{1/(p-1)})$ for all $a \geq 0$. Thus

$$\begin{aligned} ab - \frac{1}{p}a^p = f(a) \leq f(b^{1/(p-1)}) &= bb^{1/(p-1)} - \frac{1}{p}b^{p/(p-1)} \\ &= b^{p/(p-1)} - \frac{1}{p}b^{p/(p-1)} = \frac{p-1}{p}b^{p/(p-1)} = \frac{1}{p_*}b^{p_*}, \end{aligned}$$

which proves (35). □

LEMMA 4.84 (Hölder's inequality). *Assume that $1 < p < \infty$. Then we have for all x, $y \in \mathbb{K}^n$ that*

$$\Big|\sum_{k=1}^n x_k y_k\Big| \leq \sum_{k=1}^n |x_k||y_k| \leq \Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p}\Big(\sum_{k=1}^n |y_k|^{p_*}\Big)^{1/p_*} = \|x\|_p\|y\|_{p_*}. \tag{36}$$

PROOF. Assume without loss of generality that x, $y \neq 0$, else the claim is trivial.

The estimate

$$\Big|\sum_{k=1}^n x_k y_k\Big| \leq \sum_{k=1}^n |x_k||y_k|$$

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

$$\begin{aligned} \frac{1}{\|x\|_p\|y\|_{p_*}}\sum_{k=1}^n |x_k||y_k| = \sum_{k=1}^n a_k b_k &\leq \frac{1}{p}\sum_{k=1}^n a_k^p + \frac{1}{p_*}\sum_{k=1}^n b_k^p \\ &= \frac{1}{p}\sum_{k=1}^n \frac{|x_k|^p}{\|x\|_p^p} + \frac{1}{p_*}\sum_{k=1}^n \frac{|y_k|^{p_*}}{\|y\|_{p_*}^{p_*}} = \frac{1}{p} + \frac{1}{p_*} = 1. \end{aligned}$$

□

LEMMA 4.85 (Minkowski's inequality). *Assume that $1 < p < \infty$. Then*

$$\|x+y\|_p = \Big(\sum_{k=1}^n |x_k+y_k|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^n |y_k|^p\Big)^{1/p} = \|x\|_p + \|y\|_p. \tag{37}$$

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

$$\sum_{k=1}^{n}|x_k + y_k|^p \leq \sum_{k=1}^{n}|x_k + y_k|^{p-1}\big(|x_k| + |y_k|\big) = \sum_{k=1}^{n}|x_k + y_k|^{p-1}|x_k| + \sum_{k=1}^{n}|x_k + y_k|^{p-1}|y_k|. \tag{38}$$

Now we apply Hölder's inequality and obtain that

$$\sum_{k=1}^{n}|x_k + y_k|^{p-1}|x_k| \leq \Big(\sum_{k=1}^{n}|x_k + y_k|^{(p-1)p_*}\Big)^{1/p_*}\Big(\sum_{k=1}^{n}|x_k|^p\Big)^{1/p} = \Big(\sum_{k=1}^{n}|x_k + y_k|^p\Big)^{(p-1)/p}\Big(\sum_{k=1}^{n}|x_k|^p\Big)^{1/p},$$

where we have used that the conjugate p_* of p satisfies

$$p_* = \frac{p}{p-1}.$$

Similarly, we get that

$$\sum_{k=1}^{n}|x_k + y_k|^{p-1}|y_k| \leq \Big(\sum_{k=1}^{n}|x_k + y_k|^p\Big)^{(p-1)/p}\Big(\sum_{k=1}^{n}|y_k|^p\Big)^{1/p}.$$

Inserting these estimates in (38), we obtain

$$\sum_{k=1}^{n}|x_k + y_k|^p \leq \Big(\sum_{k=1}^{n}|x_k + y_k|^p\Big)^{(p-1)/p}\Big(\Big(\sum_{k=1}^{n}|x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{n}|y_k|^p\Big)^{1/p}\Big).$$

Now dividing both sides of this inequality by $\Big(\sum_{k=1}^{n}|x_k + y_k|^p\Big)^{(p-1)/p}$ results in (37). □

THEOREM 4.86. *For all $1 \leq p \leq \infty$, 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 4.85. □

Similarly, we can define the p-norm of a continuous function:

DEFINITION 4.87. Let $1 \leq p < \infty$. 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 4.88 (Hölder's inequality). *Assume that $1 < p < \infty$. Then we have for all f, $g \in C([0,1])$ that*

$$\Big|\int_0^1 f(x)g(x)\,dx\Big| \le \int_0^1 |f(x)g(x)\,dx| \le \Big(\int_0^1 |f(x)|^p\,dx\Big)^{1/p}\Big(\int_0^1 |g(x)|^{p_*}\,dx\Big)^{1/p_*} = \|f\|_p\|g\|_{p_*}. \tag{39}$$

PROOF. Without loss of generality assume that f, $g \neq 0$, 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

$$\tilde f(x) := \frac{f(x)}{\|f\|_p} \qquad \text{and} \qquad \tilde g(x) := \frac{g(x)}{\|g\|_{p_*}}.$$

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 4.89 (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 (39) 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 4.90. *For all $1 \le p \le \infty$, the function $\|\cdot\|_p \colon C([0,1]) \to \mathbb{R}$ is a norm on $C([0,1])$.*

[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.

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 4.89. □

9. 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 4.91 (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 4.92. 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 4.93. One or both sides of the inequality (40) may in principle by infinite. However, the inequality in this case still is satisfied. That is, if the right hand side of (40) 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 4.94 (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 2.20 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 $\|x\|_p \geq 0$ 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 4.85) we have that

$$\Big(\sum_{k=1}^{n} |x_k + y_k|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^{n} |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{n} |y_k|^p\Big)^{1/p}$$

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

$$\|x+y\|_p = \Big(\sum_{k=1}^{\infty} |x_k + y_k|^p\Big)^{1/p} \leq \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} + \Big(\sum_{k=1}^{\infty} |y_k|^p\Big)^{1/p} = \|x\|_p + \|y\|_p < \infty.$$

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 4.95 (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*

$$\Big|\sum_{k=1}^{\infty} x_k y_k\Big| \leq \sum_{k=1}^{\infty} |x_k y_k| \leq \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{\infty} |y_k|^{p_*}\Big)^{1/p_*}. \tag{40}$$

PROOF. The first inequality in (40) is obvious.

Moreover, we obtain from Hölder's inequality in $\mathbb{K}^n$ (see Lemma 4.84) that

$$\sum_{k=1}^{n} |x_k y_k| \leq \Big(\sum_{k=1}^{n} |x_k|^p\Big)^{1/p} \Big(\sum_{k=1}^{n} |y_k|^{p_*}\Big)^{1/p_*}$$

for all $n \in \mathbb{N}$. Taking the limit $n \to \infty$ now results in (40). □

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 4.96. *Assume that $1 < p < q < \infty$. Then*

$$\|x\|_1 \geq \|x\|_p \geq \|x\|_q \geq \|x\|_\infty$$

for all sequences $x = (x_k)_{k\in\mathbb{N}}$. In particular,

$$\ell^1 \subset \ell^p \subset \ell^q \subset \ell^\infty.$$

Moreover, all inclusions are proper, that is, in fact

$$\ell^1 \subsetneq \ell^p \subsetneq \ell^q \subsetneq \ell^\infty.$$

PROOF. It is obvious from the definition that

$$\|x\|_r^r = \sum_{k=1}^{\infty} |x_k|^r \geq \sup_{k\in\mathbb{N}} |x_k|^r = \|x\|_\infty^r$$

for all sequences $x = (x_k)_{k\in\mathbb{N}}$ and all $1 \leq r < \infty$. Thus the last inequality holds. Moreover, if $x \in \ell^q$, then $\|x\|_q < \infty$, and thus also $\|x\|_\infty \leq \|x\|_q < \infty$, which implies that $x \in \ell^\infty$. This proves the inclusion $\ell^q \subset \ell^\infty$.

Next we want to show that $\|x\|_p \geq \|x\|_q$ for all sequences $x = (x_k)_{k\in\mathbb{N}}$. To that end, we note that this inequality trivially holds $\|x\|_p = \infty$. Assume therefore that $x \in \ell^p$. Then we can estimate

$$\begin{aligned}\|x\|_q^q = \sum_{k=1}^{\infty} |x_k|^q = \sum_{k=1}^{\infty} |x_k|^{q-p}|x_k|^p &\leq \Big(\sup_{k\in\mathbb{N}} |x_k|^{q-p}\Big) \sum_{k=1}^{\infty} |x_k|^p \\ &= \Big(\sup_{k\in\mathbb{N}} |x_k|\Big)^{q-p} \|x\|_p^p = \|x\|_\infty^{q-p}\|x\|_p^p \leq \|x\|_q^{q-p}\|x\|_p^p.\end{aligned}$$

This shows that $\|x\|_q \leq \|x\|_p < \infty$ and thus also $x \in \ell^q$. Moreover, the inequality $\|x\|_1 \leq \|x\|_p$ and 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 4.97 (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. □

As a final result, we show that all the ℓ^p-spaces are actually complete.

THEOREM 4.98. *The space $(\ell^p, \|\cdot\|_p)$ is a Banach space for all $1 \leq p \leq \infty$, 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, $m \geq N$. In particular, this shows that

$$|x_k^{(n)} - x_k^{(m)}| < \varepsilon$$

for all $k \in \mathbb{N}$ and all n, $m \geq N$, 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 4.47. 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

$$\begin{aligned} \sum_{k=1}^{K} |x_k| = \sum_{k=1}^{K} \Big| \lim_{n\to\infty} x_k^{(n)} \Big| = \sum_{k=1}^{K} \lim_{n\to\infty} |x_k^{(n)}| = \lim_{n\to\infty} \sum_{k=1}^{K} |x_k^{(n)}| \\ \leq \sup_{n\in\mathbb{N}} \sum_{k=1}^{K} |x_k^{(n)}| \leq \sup_{n\in\mathbb{N}} \sum_{k=1}^{\infty} |x_k^{(n)}| = \sup_{n\in\mathbb{N}} \|x^{(n)}\|_1 \leq R. \end{aligned} \tag{41}$$

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 (41) 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. □

10. 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 4.99 (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 4.100. 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 4.101. 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 4.102. 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 4.103 (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 4.104. 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 4.105. 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 4.106 (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.*

[5] There are some exceptions to this, though. Try to find them!

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 4.107. The space $(V, \|\cdot\|_V)$ from Theorem 4.106 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 4.108. Denote by

$$c_{\text{fin}} := \big\{x = (x_k)_{k\in\mathbb{N}} : \text{ there exists } K \in \mathbb{N} \text{ such that } x_k = 0 \text{ for all } k > K\big\}$$

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

$$x^{(n)} := (x_1, \dots, x_n, 0, 0, \dots) \in c_{\text{fin}}.$$

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 $n \ge N$. Then

$$\|x - x^{(n)}\|_p = \Big(\sum_{k=n+1}^{\infty} |x_k|^p\Big)^{1/p} \le \Big(\sum_{k=N+1}^{\infty} |x_k|^p\Big)^{1/p} < \varepsilon.$$

Since this holds for all $n \ge N$ 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 $1 \le p < \infty$.

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 $x = (1, 1, \dots) \in \ell^\infty$. Then, if $y = (y_1, \dots, y_K, 0, \dots) \in c_{\text{fin}}$ is any finite sequence, it follows that $\|x - y\|_\infty \ge 1$. 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 4.109. 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 4.110. 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 5

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 always 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

$$T\big(\lim_{n\to\infty} f_n\big) = T(0) = 0 \neq 1 = \lim_{n\to\infty} Tf_n.$$

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, these conditions can be simplified significantly.

Definition 5.1 (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

$$\|Tu\|_V \leq C\|u\|_U \qquad \text{for all } u \in U. \tag{42}$$

■

Example 5.2. 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

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m \Big|\sum_{j=1}^n a_{ij}x_j\Big| \leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}||x_j| \\ &= \sum_{j=1}^n \Big(\sum_{i=1}^m |a_{ij}|\Big)|x_j| \leq \Big(\max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|\Big)\Big(\sum_{j=1}^n |x_j|\Big) = \Big(\max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|\Big)\|x\|_1.\end{aligned}$$

Thus (42) holds with

$$C = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|$$

and thus the mapping A is bounded. ∎

EXAMPLE 5.3. 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

$$\begin{aligned}\|Tf\|_\infty &= \max_{x\in[0,1]}|Tf(x)| = \max_{x\in[0,1]}\Big|\int_0^1 k(x,y)f(y)\,dy\Big| \\ &\le \max_{x\in[0,1]}\Big(\max_{y\in[0,1]}|k(x,y)|\int_0^1|f(y)|\,dy\Big) = \Big(\max_{(x,y)\in[0,1]^2}|k(x,y)|\Big)\Big(\int_0^1|f(y)|\,dy\Big) \\ &\le \Big(\max_{(x,y)\in[0,1]^2}|k(x,y)|\Big)\|f\|_\infty.\end{aligned}$$

Thus (42) holds with

$$C = \max_{(x,y)\in[0,1]^2}|k(x,y)|$$

and T is bounded. ∎

EXAMPLE 5.4. 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 (42) 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 5.5 (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) $\implies$ (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

$$\|Tw\|_V \leq \frac{2}{\delta} \|w\|_U.$$

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 5.6. 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 5.7. 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.
- 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 5.8. Let $1 \leq p \leq \infty$. 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 $\|Lx\|_p \leq \|x\|_p$ for all $x \in \ell^p$. ∎

EXAMPLE 5.9. Assume that $y = (y_1, y_2, \ldots) \in \ell^\infty$ 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

$$\Big|\sum_{k=1}^{\infty} x_k y_k\Big| \leq \big(\sup_k |y_k|\big) \sum_k |x_k| = \|y\|_\infty \|x\|_1.$$

In fact, this last inequality actually shows that T is bounded, the constant C in (42) being $\|y\|_\infty$.

More general, let $1 \leq p \leq \infty$ 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

$$|Tx| = \Big|\sum_{k=1}^{\infty} x_k y_k\Big| \leq \|x\|_p \|y\|_{p_*},$$

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 5.10. Let U and V be vector spaces. By $L(U, V)$ we denote the vector space of linear operators $T\colon U \to V$. ∎

Assume now that U and V are normed spaces. Our next goal is to show that the set of *bounded* linear operators $T\colon U \to V$ is a subspace of $L(U, V)$. Moreover, we will be able to define a natural norm on that space.

DEFINITION 5.11. Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. By $B(U, V)$ we denote the set of bounded linear operators $T\colon U \to V$. ∎

LEMMA 5.12. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. The set $B(U, V)$ is a vector space.*

PROOF. We have to show that $B(U,V)$ is a subspace of $L(U,V)$. 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 $0 \in B(U,V)$.

Next, assume that T, $S \in B(U,V)$ 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 $T \in B(U,V)$ 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 $B(U,V)$. By definition, we can find for every $T \in B(U,V)$ 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 5.13. Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. We define the mapping $\|\cdot\|_{B(U,V)} \colon B(U,V) \to \mathbb{R}$ by

$$\|T\|_{B(U,V)} := \inf\big\{C > 0 : \|Tu\|_V \leq C\|u\|_U \text{ for all } u \in U\big\}. \tag{43}$$

■

The following result provides an alternative characterisation of $\|\cdot\|_{B(U,V)}$ that is more useful in practice.

LEMMA 5.14. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces with $U \neq \{0\}$. Then*

$$\|T\|_{B(U,V)} = \sup_{\substack{u \in U,\\ u \neq 0}} \frac{\|Tu\|_V}{\|u\|_U} = \sup_{\substack{u \in U,\\ \|u\|_U = 1}} \|Tu\|_V. \tag{44}$$

PROOF. We note first that, for all $u \in U$ with $u \neq 0$, we have

$$\frac{\|Tu\|_V}{\|u\|_U} = \Big\|T\Big(\frac{u}{\|u\|_U}\Big)\Big\|_V \qquad \text{and} \qquad \Big\|\frac{u}{\|u\|_U}\Big\|_U = 1.$$

This shows the second equality in (44).

Now denote

$$K := \sup_{\substack{u \in U,\\ u \neq 0}} \frac{\|Tu\|_V}{\|u\|_U}.$$

We have to show that $K = \|Tu\|_{B(U,V)}$.

To that end, 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

$$C \geq \frac{\|Tu\|_V}{\|u\|_U}.$$

Since this holds for all $u \neq 0$, we obtain that

$$C \geq \sup_{\substack{u \in U, \\ u \neq 0}} \frac{\|Tu\|_V}{\|u\|_U} = K.$$

Taking the infimum over all such $C \geq 0$, we then conclude that

$$K \leq \inf\big\{C > 0 : \|Tu\|_V \leq C\|u\|_U \text{ for all } u \in U\big\} = \|T\|_{B(U,V)}.$$

Conversely, we have for all $u \neq 0$ that

$$\|Tu\|_V = \frac{\|Tu\|_V}{\|u\|_U}\|u\|_U \leq \Big(\sup_{\substack{v \in U, \\ v \neq 0}} \frac{\|Tv\|_V}{\|v\|_U}\Big)\|u\|_U = K\|u\|_U.$$

Thus also

$$\|T\|_{B(U,V)} = \inf\big\{C > 0 : \|Tu\|_V \leq C\|u\|_U \text{ for all } u \in U\big\} \leq K,$$

as K can be used as a constant within the infimum. This proves the assertion. □

REMARK 5.15. An immediate consequence of Lemma 5.14 is the inequality

$$\|Tu\|_V \leq \|T\|_{B(U,V)}\|u\|_U$$

for all $u \in U$ and all $T \in B(U,V)$. ■

THEOREM 5.16. *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed spaces. Then $\|\cdot\|_{B(U,V)}$ as defined in (43) defines a norm on $B(U,V)$.*

PROOF. Exercise. □

LEMMA 5.17. *Assume that $(U, \|\cdot\|_U)$, $(V, \|\cdot\|_V)$, and $(W, \|\cdot\|_W)$ are normed spaces and that $T\colon U \to V$ and $S\colon V \to W$ are bounded linear. The the composition $S \circ T\colon U \to W$ is bounded linear and*

$$\|S \circ T\|_{B(U,W)} \leq \|S\|_{B(V,W)}\|T\|_{B(U,V)}. \tag{45}$$

PROOF. For every $u \in U$ we have

$$\|(S \circ T)u\|_W = \|S(Tu)\|_W \leq \|S\|_{B(V,W)}\|Tu\|_V \leq \|S\|_{B(V,W)}\|T\|_{B(U,V)}\|u\|_U.$$

This shows that $S \circ T$ is bounded. Moreover, we can use $C = \|S\|_{B(V,W)}\|T\|_{B(U,V)}$ as a constant in the definition (42). Now (45) follows from the definition of $\|S \circ T\|_{B(U,W)}$ as the smallest such constant. □

EXAMPLE 5.18. We now consider specifically the case where $A \in \mathbb{K}^{m \times n}$ is a matrix defining a linear operator $A\colon \mathbb{K}^n \to \mathbb{K}^m$. Moreover, we consider on $\mathbb{K}^n$ and $\mathbb{K}^m$ the p-norm for some $1 \leq p \leq \infty$. 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 3.36 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 5.2. 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 5.19. 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

$$\|A\|_{1\mapsto q} = \max_{1\le j\le n} \Big(\sum_{i=1}^{m} |a_{ij}|^q\Big)^{1/q} \tag{46}$$

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 (46) 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 (46) 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 5.20. Let $1 \leq p \leq \infty$ 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 $1 \leq p < \infty$, 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

$$\|Lx\|_p = \Big(\sum_{k=1}^{\infty} |(Lx)_k|^p\Big)^{1/p} = \Big(\sum_{k=1}^{\infty} |x_{k+1}|^p\Big)^{1/p} = \Big(\sum_{k=2}^{\infty} |x_k|^p\Big)^{1/p} \leq \|x\|_p,$$

and thus

$$\|L\| = \sup_{x \neq 0} \frac{\|Lx\|_p}{\|x\|_p} \leq 1.$$

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

$$\|L\| \geq \frac{\|(1, 0, \ldots)\|_p}{\|(0, 1, 0, \ldots)\|_p} = 1.$$

Together, we obtain again that $\|L\| = 1$. ■

EXAMPLE 5.21. 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

$$\|Tf\|_\infty \leq \Big(\max_{(x,y)\in[0,1]^2} |k(x,y)|\Big) \|f\|_\infty$$

for all $f \in C([0,1])$. This shows that

$$\|T\| \leq \max_{(x,y)\in[0,1]^2} |k(x,y)|.$$

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:

$$\begin{aligned}(47)\quad \|Tf\|_\infty &= \max_{x\in[0,1]} |Tf(x)| = \max_{x\in[0,1]} \Big| \int_0^1 k(x,y) f(y)\, dy \Big| \\ &\le \max_{x\in[0,1]} \Big(\Big(\int_0^1 |k(x,y)|\, dy \Big) \Big(\max_{y\in[0,1]} |f(y)| \Big) \Big) = \Big(\max_{x\in[0,1]} \int_0^1 |k(x,y)|\, dy \Big) \|f\|_\infty .\end{aligned}$$

We therefore see that

$$(48)\qquad \|T\| \le \max_{x\in[0,1]} \int_0^1 |k(x,y)|\, dy.$$

In fact, we have equality in (48).

If $k(x,y) \ge 0$ for all $(x,y) \in [0,1]^2$, this is easy to see: We simply can use the function $f(x) = 1$ in (47) 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 $T\colon U \to V$ is a linear mapping. Then T is uniquely determined by the action of T on a basis of U. That is, if $(u_1, \dots, u_n)$ is a basis of U, then we know T as soon as we know the values of $Tu_1, \dots, Tu_n \in V$: Because $(u_1, \dots, u_n)$ 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 $v_1, \dots, v_n \in V$ and a basis $(u_1, \dots, u_n)$, then there exists a unique linear mapping $T\colon U \to V$ with $Tu_i = v_i$ for all $1 \le i \le n$, 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 Hamel 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 $i_1, \dots, i_N \in I$, and coefficients $\alpha_1, \dots, \alpha_N \in \mathbb{K}$. 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"

$$(49)\qquad e_i = (0, \dots, 0, \underbrace{1}_{i\text{-th position}}, 0, \dots)$$

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

$$T\colon c_{\text{fin}} \to V,$$
$$(x_1, \ldots, x_N, 0, \ldots) \mapsto x_1 v_1 + \ldots + x_N v_N.$$

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 5.22 (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 5.23. *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*

$$\|\bar{T}\|_{B(U,V)} = \|T\|_{B(M,V)}.$$

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

$$\|Tu_n - Tu_m\|_V = \|T(u_n - u_m)\|_V \leq \|T\|_{B(M,V)} \|u_n - u_m\|_U.$$

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

$$\|u_n - v_n\|_U \leq \|u_n - u\|_U + \|u - v_n\|_U \to_{n\to\infty} 0.$$

Thus it follows that

$$\begin{aligned}\|\bar{T}u - Tv_n\|_V &\leq \|\bar{T}u - Tu_n\|_V + \|Tu_n - Tv_n\|_V \\ &\leq \|\bar{T}u - Tu_n\|_V + \|T\|\|u_n - v_n\|_U \to_{n\to\infty} 0,\end{aligned}$$

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 $\|\bar{T}\|_{B(U,V)} = \|T\|_{B(M,V)}$. 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

$$\|\bar{T}u\|_V = \lim_{n\to\infty} \|Tu_n\|_M \le \|T\|_{B(M,V)} \lim_{n\to\infty} \|u_n\|_U = \|T\|_{B(M,V)} \|u\|_U.$$

This shows that $\bar{T}$ is bounded linear and $\|\bar{T}\|_{B(U,V)} \le \|T\|_{B(M,V)}$. On the other hand, we have that

$$\|\bar{T}\|_{B(U,V)} = \sup_{\substack{u\in U,\\ \|u\|_U=1}} \|\bar{T}u\|_U \ge \sup_{\substack{u\in M,\\ \|u\|_U=1}} \|Tu\|_U = \|T\|_{B(M,V)},$$

which shows that, in fact, $\|\bar{T}\|_{B(U,V)} = \|T\|_{B(M,V)}$.

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 5.24. *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 5.23. □

REMARK 5.25. The construction in the proof of Theorem 5.23 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 5.26. Let $1 \leq p < \infty$. Then c_{fin} is dense in ℓ^p and the unit sequences e_i, $i \in \mathbb{N}$, as defined in (49) 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 $x = (x_1, \ldots, x_N, 0, \ldots) \in c_{\text{fin}}$ that

$$\|Tx\|_V = \|x_1 v_1 + \ldots + x_N v_N\|_V \leq |x_1| \|v_1\|_V + \ldots + |x_N| \|v_N\|_V.$$

Now we can use Hölder's inequality and obtain that

$$\begin{aligned} |x_1| \|v_1\|_V + \ldots + |x_N| \|v_N\|_V &\leq \Big(\sum_{n=1}^N |x_1|^p \Big)^{1/p} \Big(\sum_{n=1}^N \|v_n\|_V^{p_*} \Big)^{1/p_*} \\ &\leq \|x\|_p \, \|(\|v_1\|_V, \|v_2\|_V, \ldots)\|_{p_*}. \end{aligned}$$

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 $x = (x_1, x_2, \ldots) \in \ell^p$. ■

EXAMPLE 5.27. 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 $|x| \geq R$. 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

$$\|f\|_1 = \int_{-\infty}^\infty |f(x)| \, dx = \int_{-R}^R |f(x)| \, dx \leq 2R \max_{x \in [-R,R]} |f(x)| < \infty,$$

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

$$\|\hat{f}\|_\infty = \sup_{\omega\in\mathbb{R}}|\hat{f}(\omega)| \le \frac{1}{\sqrt{2\pi}}\sup_{\omega\in\mathbb{R}}\int_{-\infty}^{\infty}|f(x)e^{-i\omega x}|\,dx$$
$$= \frac{1}{\sqrt{2\pi}}\sup_{\omega\in\mathbb{R}}\int_{-\infty}^{\infty}|f(x)|\,dx = \frac{1}{\sqrt{2\pi}}\|f\|_1.$$

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 2.81 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

$$\ker T = \big\{u \in U : Tu = 0\big\},$$
$$\operatorname{ran} T = \big\{v \in V : \text{ there exists } u \in U \text{ with } Tu = v\big\}.$$

LEMMA 5.28. *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

[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 $u \in \ker T$.

$v \notin \operatorname{ran} T$. 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 5.29. 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 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

$$\|Tf\|_\infty = \max_{x\in[0,1]}\left|\int_0^x f(y)\,dy\right| \le \int_0^1 |f(y)|\,dy \le \max_{x\in[0,1]}|f(x)| = \|f\|_\infty,$$

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 5.30. *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*

$$\|Tu\|_V \ge C\|u\|_U \tag{50}$$

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

$$\|T^{-1}\|_{B(\operatorname{ran} T, U)} \le \frac{1}{C},$$

where $C > 0$ is the constant from (50).

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 (50), 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

$$\|u_n - u_m\|_U \leq \frac{1}{C}\|Tu_n - Tu_m\|_V = \frac{1}{C}\|v_n - v_m\|. \tag{51}$$

By assumption, the sequence $(v_n)_{n\in\mathbb{N}}$ converges, which implies that it is a Cauchy sequence. Now (51) 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 (50). 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 $\|T^{-1}\|_{B(\operatorname{ran} T,U)} \leq 1/C$. □

We now show that the inequality (50) is also a necessary condition for T to have a bounded inverse.

LEMMA 5.31. *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*

$$\|Tu\|_V \geq C\|u\|_U \qquad \text{for all } u \in U. \tag{52}$$

PROOF. Without loss of generality assume that U, $V \neq \{0\}$. Then $T^{-1} \neq 0$ and thus $\|T^{-1}\|_{B(V,U)} > 0$. Now let $u \in U$ be arbitrary. Then $u = T^{-1}Tu$, and thus we can estimate

$$\|u\|_U = \|T^{-1}Tu\|_U \leq \|T^{-1}\|_{B(V,U)}\|Tu\|_V.$$

This shows that (52) holds with $C = 1/\|T^{-1}\|_{B(U,V)}$. □

REMARK 5.32. Assume that $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ are Banach spaces and that $T\colon U \to V$ is bounded linear. Then Theorem 5.30 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 (50) 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 5.30 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 (50) 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] ■

[3] I refer to the class TMA4230 – Functional Analysis for more details.

5. Spaces of Bounded Linear Mappings

We have seen previously that the space $B(U,V)$ is a normed space whenever U and V are normed spaces. The next question would then be whether $B(U,V)$ 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 5.33. *Assume that $(U,\|\cdot\|_U)$ is a normed space and that $(V,\|\cdot\|_V)$ is a Banach space. Then $B(U,V)$ is a Banach space with the norm $\|\cdot\|_{B(U,V)}$.*

PROOF. We have to show that the normed space $B(U,V)$ is complete, that is, that every Cauchy sequence in $B(U,V)$ converges. Assume thus that $(T_n)_{n\in\mathbb{N}} \subset B(U,V)$ is a Cauchy sequence. We have to show that there exists $T \in B(U,V)$ such that $\|T_n - T\|_{B(U,V)} \to 0$ 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

$$\|T_n - T_m\|_{B(U,V)} < \varepsilon \qquad \text{whenever } n,\, m \geq N. \tag{53}$$

Let now $u \in U$. Then

$$\|T_n u - T_m u\|_U \leq \|T_n - T_m\|_{B(U,V)}\|u\|_U < \varepsilon\|u\|_U \text{ whenever } n,\, m \geq N.$$

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 $\|T_n - T\|_{B(U,V)} \to 0$ 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 $\|T_n\|_{B(U,V)} \leq C$ 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 $\|T_n - T\|_{B(U,V)} \to 0$ as $n \to \infty$. Let again $\varepsilon > 0$, and let $N \in \mathbb{N}$ be such that (53) holds. Then we have for all $u \in U$ and all $n \geq N$ that

$$\begin{aligned}\|(T_n - T)u\|_V = \|T_n u - Tu\|_V &= \lim_{m\to\infty} \|T_n u - T_m u\|_V \\ &\leq \lim_{m\to\infty} \|T_n - T_m\|_{B(U,V)}\|u\|_U \leq \varepsilon\|u\|_U.\end{aligned}$$

Thus

$$\|T_n - T\|_{B(U,V)} \leq \varepsilon \qquad \text{whenever } n \geq N,$$

which shows that $\|T_n - T\|_{B(U,V)} \to 0$ as $n \to \infty$. □

In particular, we may apply the previous result to the case $V = \mathbb{K}$.

DEFINITION 5.34 (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 5.35. 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 sheets.
- Consider again the Banach space $(C([0,1]), \|\cdot\|_\infty)$, and let $g \in C([0,1])$ be a fixed function. Then the mapping $T_g \colon C([0,1]) \to \mathbb{K}$,
$$T_g(f) := \int_0^1 f(x)g(x)\,dx$$
is a bounded linear functional on $C([0,1])$, as
$$|T_g(f)| \leq \int_0^1 |f(x)||g(x)|\,dx \leq \|f\|_\infty \|g\|_\infty$$
for all $f \in C([0,1])$.

■

Definition 5.36 (dual space). Assume that U is a normed apce. The *dual space* of U is defined as
$$U^* := B(U, \mathbb{K}) = \big\{ f \colon U \to \mathbb{K} : f \text{ is bounded linear } \big\}.$$

■

Theorem 5.37. *Let $(U, \|\cdot\|_U)$ be a normed space. Then U^* is a Banach space with the norm*
$$\|f\|_{U^*} = \sup_{\substack{u\in U,\\ \|u\|_U \leq 1}} |f(u)|.$$

Proof. This immediately follows from Theorem 5.33 using the completeness of $\mathbb{K}$. □

Example 5.38. We consider the case $U = \mathbb{K}^n$ with the norm $\|\cdot\|_p$ for some $1 \leq p \leq \infty$. Then every linear mapping $T \colon \mathbb{K}^n \to \mathbb{K}$ 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 $T \colon \mathbb{K}^n \to \mathbb{K}$ 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
$$\|A\|_{(K^n)^*} = \sup_{\|x\|_p=1} |Ax| = \sup_{\|x\|_p=1} \Big|\sum_{i=1}^n a_i x_i\Big|.$$
By Hölder's inequality, we can now estimate
$$\Big|\sum_{i=1}^n a_i x_i\Big| \leq \|A\|_{p_*} \|x\|_p,$$

[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.

and thus

$$\|A\|_{(K^n)^*} \leq \|A\|_{p_*}.$$

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 $x \neq 0$ 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

$$x_i := \begin{cases} \overline{a_i}|a_i|^{p_*-2} & \text{if } a_i \neq 0, \\ 0 & \text{if } a_i = 0. \end{cases}$$

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 5.39. *Let $1 \leq p < \infty$ 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

$$|f(x)| = \Big|\sum_{k=1}^\infty x_k y_k\Big| \leq \|x\|_p \|y\|_{p_*},$$

which shows that f is bounded linear and that

$$\|f\|_{(\ell^p)^*} \leq \|y\|_{p_*}. \tag{54}$$

(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 5.38 we define a sequence $x = (x_1, x_2, \ldots)$ setting

$$x_k = \begin{cases} \overline{y_k}|y_k|^{p_*-2} & \text{if } x_k \neq 0, \\ 0 & \text{if } x_k = 0. \end{cases}$$

Then $x^{(N)} := (x_1, \ldots, x_N, 0, \ldots) \in c_{\text{fin}} \subset \ell^p$ for all $N \in \mathbb{N}$. Define now $y^{(N)} := (y_1, \ldots, y_N, 0, \ldots)$. Similarly as in Example 5.38 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

$$\|y\|_{p_*} \le \|f\|_{(\ell^p)^*} \tag{55}$$

and thus $y \in \ell^{p_*}$.

Since $c_{\text{fin}} \subset \ell^p$ is dense, it follows from Theorem 5.23 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 5.23 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 (54) and (55). □

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 5.40. 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 5.23. ■

CHAPTER 6

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 6.1 (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 6.2. 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 6.3. Assume $(U, \|\cdot\|_U)$ is a normed space and let $R > 0$. Then the ball

$$B_R(0) = \big\{u \in U : \|u\|_U < R\big\}$$

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

$$\|\lambda u+(1-\lambda)v\|_U \leq |\lambda|\|u\|_U+|1-\lambda|\|v\|_U = \lambda\|u\|_U+(1-\lambda)\|v\|_U < \lambda R+(1-\lambda)R = R,$$

which shows that $\lambda u + (1-\lambda)v \in B_R(0)$. ■

DEFINITION 6.4 (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 6.5 (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 4.5) to $z_n - z_m$ and $2u - z_n - z_m$ and obtain that

$$\begin{aligned} 2\|z_n - z_m\|_U^2 + 2\|2u - z_n - z_m\|_U^2 \\ = \|z_n - z_m + 2u - z_n - z_m\|_U^2 + \|z_n - z_m - (2u - z_n - z_m)\|_U^2 \\ = 4\|u - z_m\|_U^2 + 4\|u - z_n\|_U^2. \end{aligned}$$

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 6.6 (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

$$\Re(\langle u - z, v - z \rangle) \leq 0 \qquad \text{for all } v \in K \tag{56}$$

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

$$\|u - z\|_U^2 \leq \|u - \lambda v - (1 - \lambda) z\|_U^2 \text{ for all } v \in K \text{ and } 0 \leq \lambda \leq 1. \tag{57}$$

Now we can write

$$\|u - z\|_U^2 = \|u\|_U^2 - 2\Re(\langle u, z\rangle) + \|z\|_U^2$$

and

$$\|u - \lambda v - (1-\lambda)z\|_U^2 = \|u\|_U^2 + \lambda^2\|v\|_U^2 + (1-\lambda)^2\|z\|_U^2 \\ - 2\lambda\Re(\langle u, v\rangle) - 2(1-\lambda)\Re(\langle u, z\rangle) + 2\lambda(1-\lambda)\Re(\langle v, z\rangle).$$

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 (57) can be reformulated as

$$f_v(\lambda) \geq 0 \qquad \text{for all } v \in K \text{ and } 0 \leq \lambda \leq 1. \tag{58}$$

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 (58) holds if and only if $f_v'(0) \geq 0$ 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 (58) and therefore (57) is equivalent to the variational inequality

$$-f_v'(0) = \Re(\langle z-u, z-v\rangle) \leq 0 \qquad \text{for all } v \in K.$$

This shows that the assumption $z = \pi_K(u)$ implies (56).

Conversely, assume that $z \in K$ and that (56) holds. As seen above, this is equivalent to z satisfying (57). 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 6.7. *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

$$W^\perp = \big\{v \in U : \langle v, w\rangle = 0 \text{ for all } w \in W\big\}$$

is the orthogonal complement of W in U.

Proof. The existence and uniqueness follow from Theorem 6.5. Moreover, Theorem 6.6 implies that $z = \pi_W(u)$ is the unique point satisfying $z \in W$ and

$$\Re(\langle u - z, v - z\rangle) \leq 0 \qquad \text{for all } v \in W. \tag{59}$$

We thus have to show that (59) holds if and only if $u - z \in W^\perp$.

Assume therefore that (59) 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 (59) implies that

$$0 \geq \Re(\langle u - z, \lambda w + z - z\rangle) = \Re(\langle u - z, \lambda w\rangle) \qquad \text{for all } \lambda \in \mathbb{K}.$$

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 (59) holds. □

COROLLARY 6.8. *Assume that U is a Hilbert space, that $W \subset U$ is a closed subspace, and that $w_0 \in U$. Define $\hat{W} := w_0 + W$. Then there exists for each $u \in U$ a unique point $z = \pi_{\hat{W}}(u)$ such that $\pi_{\hat{W}}(u) \in \hat{W}$ and*

$$\|z - u\| = \operatorname{dist}(u, \hat{W}) = \inf_{w \in \hat{W}} \|u - w\|.$$

Moreover, z is the unique point satisfying

$$z \in \hat{W} \qquad \text{and} \qquad u - z \in W^\perp.$$

PROOF. The existence and uniqueness of $\pi_{\hat{W}}(u)$ follow from Theorem 6.5. Denote now $\hat{u} := u - w_0$. Then it is easy to see that $\pi_{\hat{W}}(u) = \pi_W(\hat{u}) + w_0$. Now by Corollary 6.7, $\hat{z} := \pi_W(\hat{u})$ is characterised by the conditions $\hat{z} \in W$ and $\hat{z} - \hat{u} \in W^\perp$. Since $\hat{z} = z - w_0$ and $\hat{u} = u - w_0$, these conditions are the same as the conditions $z \in w_0 + W$ and $z - u \in W^\perp$, which proves the assertion. □

REMARK 6.9. We have stated Theorem 6.6 and Corollary 6.7 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 6.10. 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

$$\pi_K(u) = \begin{cases} u & \text{if } \|u\| \le R \\ R\dfrac{u}{\|u\|} & \text{if } \|u\| > R \end{cases} = \frac{Ru}{\max\{\|u\|, R\}}. \tag{60}$$

First assume that $u \in K$, that is, $\|u\| \le R$. Then obviously $\pi_K(u) = u$. Thus (60) 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 6.6 we thus have to show that

$$\Re(\langle u - z, v - z\rangle) \leq 0 \qquad \text{for all } v \in K. \tag{61}$$

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

$$\Re(\langle u, v\rangle) \leq |\langle u, v\rangle| \leq \|u\|\|v\|.$$

Since $\|v\| \leq R$ it follows that

$$\Re(\langle u - z, v - z\rangle) = \big(\|u\| - R\big)\Big(\frac{\Re(\langle u, v\rangle)}{\|u\|} - R\Big) \leq \big(\|u\| - R\big)\big(\|v\| - R\big) \leq 0.$$

Thus (61) holds, which in turn shows that $\pi_K(u) = z = Ru/\|u\|$. ∎

Remark 6.11. 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 $|a| \leq 1$ satisfies $\|(a, 0) - u\|_\infty = 1 = \operatorname{dist}(u, K)$. That is, every point $(a, 0) \in K$ with $|a| \leq 1$ can be considered a projection of $(0, 1)$ onto K. In particular, the projection is not unique.

Now we consider the case $U = \ell^1$ the space of all real-valued summable sequences, and we define

$$K = \Big\{x \in \ell^1 : \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)x_n \geq 1\Big\}.$$

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

$$1 \leq \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)x_n \leq \sum_{n=1}^{\infty}\Big(1 - \frac{1}{n}\Big)|x_n| < \sum_{n=1}^{\infty}|x_n| = \|x\|_1.$$

That is, $\|x\|_1 > 1$ for all $x \in \ell^1$, which also implies that $\operatorname{dist}(0, K) \geq 1$. On the other hand, for all $n \geq 2$ we have that

$$\frac{n}{n-1}e_n \in K,$$

which shows that

$$\operatorname{dist}(0, K) \leq \Big\|\frac{n}{n-1}e_n\Big\|_1 = \frac{n}{n-1}$$

for all $n \geq 2$. Taking the infimum over all $n \geq 2$, it follows that $\operatorname{dist}(0, K) \leq 1$. 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 6.12. *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*

$$\|\pi_K(u) - \pi_K(v)\| \leq \|u - v\|.$$

PROOF. If $\pi_K(u) = \pi_K(v)$, the claim is trivial, as the left hand side is equal to zero. Assume therefore that $\pi_K(u) \neq \pi_K(v)$. 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 (56), which shows that

$$\begin{aligned}\Re(\langle \pi_K(v) - \pi_K(u), u - \pi_K(u)\rangle) \leq 0, \\ \Re(\langle \pi_K(u) - \pi_K(v), v - \pi_K(v)\rangle) \leq 0.\end{aligned}$$

Moreover, using the Cauchy–Schwarz–Bunyakovsky inequality, we can estimate

$$\begin{aligned}\Re(\langle \pi_K(v) - \pi_K(u), u - v\rangle) &\leq |\langle \pi_K(v) - \pi_K(u), u - v\rangle| \\ &\leq \|\pi_K(v) - \pi_K(u)\|\|u - v\|.\end{aligned}$$

Thus

$$\begin{aligned}\|\pi_K(u) - \pi_K(v)\|^2 &= \Re\big(\|\pi_K(u) - \pi_K(v)\|\big)^2 \\ &= \Re\big(\langle \pi_K(v) - \pi_K(u), u - \pi_K(u)\rangle \\ &\qquad + \langle \pi_K(u) - \pi_K(v), u - v\rangle \\ &\qquad + \langle \pi_K(u) - \pi_K(v), v - \pi_K(v)\rangle\big) \\ &\leq \Re(\langle \pi_K(v) - \pi_K(u), v - u\rangle) \\ &\leq \|\pi_K(v) - \pi_K(u)\|\|v - u\|.\end{aligned}$$

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 6.7.

The results to be derived will be a generalisations of our discussion in Section 3 of the finite dimensional case.

PROPOSITION 6.13. *Assume that U is a Hilbert space and that $W \subset U$ is a closed subspace. Then the projection $\pi_W \colon U \to U$ 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 6.7 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 6.12.

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 6.14. *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 6.15. *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

$$|f(u)| = |\langle u, v\rangle| \leq \|u\|\|v\|.$$

□

LEMMA 6.16. *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 6.17. *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 6.14 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 6.18. *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 6.19. *Assume that U is a Hilbert space and W is a subspace. Then*

$$U = \overline{W} \oplus W^\perp.$$

PROOF. Since $W^\perp$ is closed, we can write

$$U = W^\perp \oplus (W^\perp)^\perp = W^\perp \oplus \overline{W}.$$

□

COROLLARY 6.20. *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 6.18 that

$$(\overline{\operatorname{span} S})^\perp = (((\operatorname{span} S)^\perp)^\perp)^\perp = (\operatorname{span} S)^\perp.$$

□

COROLLARY 6.21. *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 $U^* = B(U, \mathbb{K})$ 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 6.22 (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 6.23. *Assume that U is an inner product space and that $v \in U$ is some vector. Then $v = 0$ if and only if*

$$\langle u, v\rangle = 0 \qquad \text{for all } u \in U.$$

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 $v = 0$. □

THEOREM 6.24 (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

$$|f_v(u)| = |\langle Tu, v\rangle_V| \le \|Tu\|_V \|v\|_V \le \|T\|_{B(U,V)} \|v\|_V \|u\|_U,$$

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 $T^*(v+w) = T^*v + T^*w$.

[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.

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

$$\|T^*v\|_U^2 = \langle T^*v, T^*v\rangle_U = \langle TT^*v, v\rangle_V \leq \|TT^*v\|_V\|v\|_V \leq \|T\|_{B(U,V)}\|T^*v\|_U\|v\|_V.$$

This shows that

$$\|T^*v\|_U \leq \|T\|_{B(U,V)}\|v\|_V, \tag{62}$$

which proves the boundedness of T^*.

Uniqueness: Assume now that $S \in B(V,U)$ 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 6.25. 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 6.26 (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:*

- $(T^*)^* = T$.
- $\|T^*\|_{B(V,U)} = \|T\|_{B(U,V)}$.
- $\|T^*T\|_{B(U,U)} = \|TT^*\|_{B(V,V)} = \|T\|^2_{B(U,V)}$.

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 $\|T^*\|_{B(V,U)} = \|T\|_{B(U,V)}$. For that, we note first that the estimate $\|T^*\|_{B(V,U)} \leq \|T\|_{B(U,V)}$ follows from (62). Applying the same estimate to T^*, we obtain moreover that

$$\|T\|_{B(U,V)} = \|(T^*)^*\|_{B(U,V)} \leq \|T^*\|_{B(V,U)},$$

which proves the assertion.

For showing that $\|T^*T\|_{B(U,U)} = \|T\|^2_{B(U,V)}$, we note first that

$$\|T^*T\|_{B(U,U)} \leq \|T^*\|_{B(V,U)}\|T\|_{B(U,V)} = \|T\|^2_{B(U,V)},$$

as $\|T^*\|_{B(V,U)} = \|T\|_{B(U,V)}$. For the opposite inequality we note that, for all $u \in U$,

$$\|Tu\|_V^2 = \langle Tu, Tu\rangle_V = \langle u, T^*Tu\rangle_U \leq \|u\|_U\|T^*Tu\|_U \leq \|u\|_U^2\|T^*T\|_{B(U,U)}.$$

Thus

$$\|T^*Tu\|_U \geq \sup_{\substack{u\in U\\ u\neq 0}} \frac{\|Tu\|_V^2}{\|u\|_U^2} = \|T\|^2_{B(U,V)}.$$

This shows that $\|T^*T\|_{B(U,U)} = \|T\|^2_{B(U,V)}$.

Finally, we have that

$$\|TT^*\|_{B(V,V)} = \|(T^*)^*T^*\|_{B(V,V)} = \|T^*\|^2_{B(V,U)} = \|T\|^2_{B(U,V)},$$

which concludes the proof. □

LEMMA 6.27. *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 6.28 (Right and left shift). Let $U = \ell^2$ and let $T = R\colon U \to U$ be the right shift operator, that is,

$$Rx = (0, x_1, x_2, \ldots)$$

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 6.29. 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

$$\begin{aligned}\|Tf\|_2^2 = \int_0^1 |Tf(x)|^2\,dx &= \int_0^1 \Big|\int_0^1 k(x,y)f(y)\,dy\Big|^2\,dx \\ &\le \int_0^1 \Big(\int_0^1 |k(x,y)|^2\,dy\Big)\Big(\int_0^1 |f(y)|^2\,dy\Big)\,dx \\ &= \Big(\int_0^1\int_0^1 |k(x,y)|^2\,dx\,dy\Big)\|f\|_2^2,\end{aligned}$$

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 6.30. Consider the linear initial value problem with varying coefficients

$$\begin{aligned} y'(t) &= A(t)y(t), \\ y(0) &= x, \end{aligned} \tag{63}$$

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 (63). 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 (63) as the *final value problem*

$$\begin{aligned} v'(t) &= -A(t)^* v(t), \\ v(1) &= w, \end{aligned} \tag{64}$$

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 (64).

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

$$\begin{aligned} 0 = \int_0^1 \langle y(t;x), v(t;w)\rangle'\, dt &= \langle y(1;x), v(1;w)\rangle - \langle y(0;x), v(0;w)\rangle \\ &= \langle Tx, w\rangle - \langle x, v(0;w)\rangle \end{aligned}$$

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 (64). ∎

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 3.47). The next result provides a generalisation to the infinite dimensional case.

PROPOSITION 6.31. *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 $T^*v = 0$ 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 6.32. *Assume that U and V are Hilbert spaces and that $T\colon U \to V$ 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 family $(u_i)_{i\in I} \subset U$ 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 6.33 (Countable set). An infinite set S is called *countable*, if there exists a surjective mapping $f\colon \mathbb{N} \to S$. 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 6.34. 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 6.35. If S is a countable set, then so is the Cartesian product S^n for every $n \geq 1$. 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 $n \geq 3$.

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 6.36. 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 6.37. 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

$$s_n = \pm a^{(n)}_{-N_n} a^{(n)}_{-(N_n-1)} \ldots a^{(n)}_0, a^{(n)}_1 a^{(n)}_2 a^{(n)}_3 \ldots$$

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 6.38. *Let U be an infinite dimensional Banach space. Then every Hamel basis of U is uncountable.*

REMARK 6.39. For incomplete normed spaces, the situation is different as we have already seen. For instance, the family $(e_1, e_2, \ldots, \ldots)$ 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.

[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 never setting $b_j = 0$ or $b_j = 9$.

DEFINITION 6.40 (Schauder basis). Let $(U, \|\cdot\|_U)$ be an infinite dimensional normed space. A countable family $(u_1, u_2, u_3, \ldots) \subset U$ 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 6.41. Let $1 \leq p < \infty$ and consider the family 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. ■

EXAMPLE 6.42. The family of unit sequences $(e_1, e_2, \ldots)$ is *not* a Schauder basis of ℓ^∞. Take for instance $x = (1, 1, \ldots)$. Then

$$\Big\| x - \sum_{n=1}^{N} \lambda_n e_n \Big\| \geq 1$$

for all sequences $\lambda_1, \lambda_2, \ldots$, and thus we cannot represent x in the desired form. ■

DEFINITION 6.43 (Separability). A normed space U is called *separable*, if there exists a countable dense subset of U. ■

EXAMPLE 6.44. 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 6.45. *Assume that the normed space U has a Schauder basis. Then U is separable.*

PROOF. Exercise! □

REMARK 6.46. 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 6.47. 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 $\{I : I \subset \mathbb{N}\}$ (and thus also the set $\{e^{(I)} : I \subset \mathbb{N}\}$) 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 $\{s_1, s_2, \ldots\} \subset \ell^\infty$ 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 $\{s_1, s_2, \ldots\}$ is countable. ∎

We will now discuss in more detail the case where U is a Hilbert space. Recall to that end that a sequence $(u_1, u_2, \ldots)$ 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.

THEOREM 6.48. *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 6.49 (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 $u \in U$.

[3] Try it yourself! This is really only Cantor's diagonal argument again.

PROOF. Applying the result of Theorem 6.48 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 6.50. *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 6.51. *Assume that U is an infinite dimensional Hilbert space and that $(u_1, u_2, \ldots) \subset U$ is an orthonormal system. Denote moreover*

$$W = \overline{\operatorname{span}\{u_1, u_2, \ldots\}}.$$

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 $\operatorname{span}\{u_1, u_2, \ldots\}$, 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

$$\langle z, u_k\rangle = \lim_{n\to\infty} \sum_{j=1}^{n} \big\langle \langle u, u_j\rangle u_j, u_k\big\rangle = \lim_{n\to\infty} \sum_{j=1}^{n} \langle u, u_j\rangle\langle u_j, u_k\rangle = \langle u, u_k\rangle,$$

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

$$u - z \in \{u_1, u_2, \ldots\}^\perp = \overline{\operatorname{span}\{u_1, u_2, \ldots\}}^\perp = W^\perp.$$

Thus we have shown that $z \in W$ and $u - z \in W^\perp$, which proves that $z = \pi_W(u)$. □

REMARK 6.52. Assume that U is a Hilbert space and that $(u_1, u_2, \ldots)$ 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 $(u_1, u_2, \ldots)$. ■

DEFINITION 6.53 (total subset). An orthonormal system $(u_1, u_2, \ldots)$ in a Hilbert space U is *total* or *maximal*, if

$$\overline{\operatorname{span}(\{u_1, u_2, \ldots\})} = U,$$

or, equivalently, if

$$\{u_1, u_2, \ldots\}^\perp = \{0\}.$$

■

DEFINITION 6.54 (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 6.55. *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

$$\sum_{j=1}^{\infty} |\langle u, u_j\rangle|^2 = \|u\|^2 \tag{65}$$

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 6.56. *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 6.57. *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 6.58 (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 6.59. In the space ℓ^2 the family $(e_j)_{j\in\mathbb{N}}$ of unit sequences is a Hilbert basis. ■

EXAMPLE 6.60. 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.

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 again to least squares solutions of linear systems.

THEOREM 6.61. *Assume that U and V are Hilbert spaces, that $T\colon U \to V$ is bounded linear, and that $v \in V$ is given. Then $u \in U$ solves the* least squares problem

$$\min_{u\in U} \|Tu - v\|_V^2 \tag{66}$$

if and only if u solves the normal equation

$$T^*Tu = T^*v. \tag{67}$$

PROOF. The least squares problem 66 is equivalent to the problem

$$\min_{w\in \operatorname{ran}(T)} \|w - v\|_V^2 \tag{68}$$

in the sense that w solves (68) if and only if $w = Tu$ for some solution u of (66).

Since (68) is just a projection problem, we know that $w \in V$ is a solution of (68) 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 $T^*(Tu - v) = 0$. □

REMARK 6.62. Note that it can happen that neither of the problems (66) or (67) has a solution. A part of the claim of the theorem is, however, that if one of the problems has a solution, then so has the other. ■

Assume now that U and V are infinite dimensional Hilbert spaces, that $T\colon U \to V$ 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 $S := T|_W\colon W \to V$ 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

$$S^*Su = S^*v, \tag{69}$$

where $S^*\colon V \to W$ 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 $(u_1, u_2, \dots, u_N)$ is a basis of W. Then $u \in W$ solves (69), 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

$$\langle Tu, Tu_i\rangle = \langle v, Tu_i\rangle \qquad \text{for all } 1 \le i \le N. \tag{70}$$

Finally, we will rewrite this system in matrix-vector form. Since $(u_1, \dots, u_N)$ is a basis of W, we can write every solution u of (70) as

$$u = \sum_{j=1}^{N} x_j u_j$$

with unique coordinates $x_1, \dots, x_N \in \mathbb{K}$. Inserting this form of u in (70), we obtain the system of equations

$$\sum_{j=1}^{N} x_j \langle Tu_j, Tu_i\rangle = \langle v, Tu_i\rangle \qquad \text{for all } 1 \le i \le N \tag{71}$$

for the coordinates $x_1, \dots, x_N$ of u. Now define the matrix

$$A = \big(\langle Tu_j, Tu_i\rangle\big)_{i,j=1,\dots,N}$$

and the vector

$$b = \big(\langle v, Tu_i\rangle\big)_{i=1,\dots,N}.$$

Then the system (71) 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 6.63 (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 $(u_1, u_2, \dots)$ is a Hilbert basis of U. For $N \in \mathbb{N}$ denote

$$W_N := \operatorname{span}\{u_1, u_2, \dots, u_N\} \subset U$$

and denote by $u^{(N)} \in W_N$ the (unique) solution of the problem

$$\min_{u \in W_N} \|Tu - v\|_V^2. \tag{72}$$

Then

$$u^\dagger = \lim_{N \to \infty} u^{(N)}.$$

PROOF. Since $T^{-1}\colon V \to U$ is bounded linear, we can estimate

$$\|u^{(N)} - u^\dagger\|_U \leq \|T^{-1}\|_{B(V,U)} \|Tu^{(N)} - Tu^\dagger\|_V = \|T^{-1}\|_{B(V,U)} \|Tu^{(N)} - v\|_V.$$

Thus it is sufficient to show that $\|Tu^{(N)} - v\|_V \to 0$. Since $u^{(N)}$ solves the least squares problem (72) we have that

$$\|Tu^{(N)} - v\|_V \leq \|T\pi_{W_N}(u^\dagger) - v\|_V$$

for all $N \in \mathbb{N}$.

Since $(u_1, u_2, \ldots)$ 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 6.64. In the proof of the previous result, we actually obtain the concrete estimate

$$\|u^{(N)} - u^\dagger\|_U \leq \|T\|_{B(U,V)} \|T^{-1}\|_{B(V,U)} \|u^\dagger - \pi_{W_N}(u^\dagger)\|,$$

which can be rewritten as

$$\|u^{(N)} - u^\dagger\|_U \leq \|T\|_{B(U,V)} \|T^{-1}\|_{B(V,U)} \sum_{n=N+1}^{\infty} |\langle u^\dagger, u_i \rangle|^2.$$

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. ■

7. Reproducing Kernel Hilbert Spaces

Assume that $\Omega \subset \mathbb{R}^d$ is some subset and that U is a Hilbert space consisting of functions defined on Ω. That is, every $u \in U$ is a function $u\colon \Omega \to \mathbb{K}$. We now assume that for each $x \in \Omega$, the point evaluation

$$\delta_x\colon U \to \mathbb{K}, \qquad u \mapsto \delta_x(u) := u(x),$$

is a bounded linear functional. That is, there exists for each $x \in \Omega$ some constant $c_x > 0$ such that

$$|u(x)| \leq c_x \|u\|_U \qquad \text{for all } u \in U.$$

Then the Riesz Representation Theorem implies that there exists for each $x \in \Omega$ some element $k_x \in U$ such that

$$u(x) = \delta_x(u) = \langle u, k_x \rangle_U \qquad \text{for all } u \in U.$$

Now $k_x \in U$, and U consists of functions on Ω. Thus the element k_x is itself a function $k_x \colon \Omega \to \mathbb{K}$. We can therefore define a function $k \colon \Omega \times \Omega \to \mathbb{K}$,

$$k(y, x) := k_x(y).$$

Then we have that

$$u(x) = \langle u, k(\cdot, x) \rangle_U \qquad \text{for all } u \in U.$$

These considerations give rise to the following definitions:

DEFINITION 6.65 (RKHS). A *reproducing kernel Hilbert space* (RKHS) is a Hilbert space U consisting of $\mathbb{K}$-valued functions on some set Ω such that for each $x \in \Omega$ the point evaluation $\delta_x \colon U \to \mathbb{K}$ is a bounded linear functional. The function $k \colon \Omega \times \Omega \to \mathbb{K}$ satisfying

$$u(x) = \langle u, k(\cdot, x) \rangle_U \qquad \text{for all } u \in U$$

is called the *kernel* of the RKHS U. ■

EXAMPLE 6.66. The *Sobolev space* $H_0^1([0,1])$ is the completion of the space of all functions $u \in C^1([0,1])$ satisfying $u(0) = u(1) = 0$ with respect to the inner product

$$\langle u, v \rangle_1 := \int_0^1 u'(x) \overline{v}'(x)\, dx.$$

One can show that this space is indeed a space of functions on the interval $[0,1]$ and that the point evaluations δ_x are continuous.[4] Thus $H_0^1([0,1])$ is an RKHS. Moreover, it is possible to show that the kernel $k \colon [0,1] \times [0,1] \to \mathbb{K}$ is the function

$$k(y, x) = \begin{cases} (1-x)y & \text{if } x \geq y, \\ x(1-y) & \text{if } x \leq y. \end{cases}$$

Indeed, if $u \colon [0,1] \to \mathbb{K}$ is (for simplicity) continuously differentiable and satisfies $u(0) = u(1) = 0$, then

$$\begin{aligned} \langle u, k(\cdot, x) \rangle_U = \int_0^1 u'(y)\, \partial_y k(y, x)\, dy &= \int_0^x u'(y)(1-x)\, dy - \int_x^1 u'(y) x\, dy \\ &= u(x)(1-x) - u(0) - u(1) + u(x)x = u(x). \end{aligned}$$

■

LEMMA 6.67. *Assume that U is an RKHS with kernel $k \colon \Omega \times \Omega \to \mathbb{K}$. Then k is Hermitian symmetric in the sense that*

$$k(y, x) = \overline{k(x, y)} \qquad \textit{for all } x,\, y \in \Omega.$$

PROOF. By definition, the function $y \mapsto k_x(y) := k(y, x)$ is an element of U. Thus we have that

$$k(y, x) = k_x(y) = \langle k_x, k(\cdot, y) \rangle_U = \langle k(\cdot, x), k(\cdot, y) \rangle_U.$$

Similarly, we can write

$$k(x, y) = k_y(x) = \langle k_y, k(\cdot, x) \rangle_U = \langle k(\cdot, y), k(\cdot, x) \rangle_U.$$

Now the claim follows from the fact that $\langle f, g \rangle_U = \overline{\langle g, f \rangle}_U$ for all $f,\, g \in U$. □

[4] More details will e.g. be discussed in the course TMA4212 Numerical Solution of Differential Equations by Difference Methods, as Sobolev spaces are a central tool in the modern theory of partial differential equations and their analytic and numerical solution.

LEMMA 6.68. *Assume that U is an RKHS with kernel $k\colon \Omega\times\Omega \to \mathbb{K}$. Then k is positive semi-definite in the following sense: For all distinct points $x_1, \dots, x_N \in \Omega$ and all $\alpha_1, \dots, \alpha_N \in \mathbb{K}$ we have that*

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j k(x_j, x_i) \geq 0.$$

PROOF. Define the function $f \in U$,

$$f(y) = \sum_{i=1}^{N} \alpha_i k(y, x_i).$$

Then

$$\begin{aligned} 0 \leq \|f\|_U^2 = \langle f, f\rangle_U &= \Big\langle \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \sum_{j=1}^{N} \alpha_j k(\cdot, x_j)\Big\rangle_U \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j \big\langle k(\cdot, x_i), k(\cdot, x_j)\big\rangle_U = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \overline{\alpha}_j k(x_j, x_i). \end{aligned}$$

□

An alternative way of interpreting these last results is the following: Given distinct points $x_1, \dots, x_N \in \Omega$ we define a matrix $K \in \mathbb{K}^{N\times N}$ by setting

$$K_{ij} := k(x_i, x_j).$$

Then Lemma 6.67 states that K is a Hermitian matrix, and Lemma 6.68 states that K is positive semi-definite.

REMARK 6.69. Assume again that $x_1, \dots, x_N \in \Omega$ are distinct points, and define the space

$$V := \operatorname{span}\{k_{x_1}, \dots, k_{x_N}\} \subset U.$$

Then V is a Hilbert space itself (being a finite dimensional subspace of U) and the family $(k_{x_1}, \dots, k_{x_N})$ by construction contained in V. Moreover, we have that

$$\langle k_{x_j}, k_{x_i}\rangle_U = \langle k(\cdot, x_j), k(\cdot, x_i)\rangle_U = k(x_i, x_j).$$

Thus the matrix $K \in \mathbb{K}^{N\times N}$ with $K_{ij} = k(x_i, x_j)$ is in fact the Gram matrix of the family $(k_{x_1}, \dots, k_{x_N})$ in V.

In particular, we obtain by Lemma 3.24 that the family $(k_{x_1}, \dots, k_{n_N})$ is linearly independent (and thus a basis of V), if and only if the matrix K is positive definite. ■

Conversely, it is possible to show that every function k that satisfies the conditions of Lemmas 6.67 and 6.68 defines a unique RKHS:

THEOREM 6.70 (Moore–Aronszajn). *Let Ω be a set and let $k\colon \Omega\times\Omega \to \mathbb{K}$ be Hermitian symmetric and positive semi-definite. Then there exists a unique Hilbert space U of functions on Ω with reproducing kernel k. Moreover, the space*

$$U_0 := \operatorname{span}\big\{k(\cdot, x) : x \in \Omega\big\}$$

is a dense subspace of U.

PROOF. See [**BTA04**, Thm. 3]. □

REMARK 6.71. The statement in Theorem 6.70 that the space U_0 is dense in U is the same as saying that functions in U can be arbitrarily well approximated by finite linear combinations of functions $k(\cdot, x_j)$ for some $x_j \in \Omega$. That is, for

each $u \in U$ and each $\varepsilon > 0$ there exist points $x_1, \dots, x_N \in \Omega$ and coefficients $c_1, \dots, c_N \in \mathbb{K}$ such that

$$\Big\| u - \sum_{j=1}^{N} c_j k(\cdot, x_j) \Big\|_U < \varepsilon.$$

■

EXAMPLE 6.72. In the case $\Omega = \mathbb{R}^d$, so called *radial kernels* of the form $k(y,x) = \kappa(\|x\|_2^2)$ for some function $\kappa\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are of specific interest in view of their symmetry properties. Particular examples are the following (cf. [**BTA04**, p. 43]):

- The *Gaussian kernel* with *bandwidth* $\sigma > 0$ (or *shape parameter* $1/\sigma$) defined by
$$k(y,x) = e^{-\frac{\|y-x\|_2^2}{2\sigma^2}}.$$
- The *Laplacian kernel* with bandwidth $\sigma > 0$ (or *shape parameter* $1/\sigma$) defined by
$$k(y,x) = e^{-\frac{\|y-x\|_2}{\sigma}}.$$
- The *inverse multiquadric* kernel with bandwidth $\sigma > 0$ and parameter $s > 0$ defined by
$$k(y,x) = \frac{1}{(\sigma^2 + \|y-x\|_2^2)^s}.$$

■

Interpolation and Regression.

THEOREM 6.73. *Let U be an RKHS over the set Ω and let $k\colon \Omega \times \Omega \to \mathbb{K}$ be the reproducing kernel of U. Let $x_1, \dots, x_N$ be distinct points in Ω and let $z = (z_1, \dots, z_N) \in \mathbb{K}^N$ be given. Assume that the matrix $K \in \mathbb{K}^{N\times N}$, $K_{ij} := k(x_i, x_j)$, is positive definite. Then the problem*

$$\min_{u \in U} \|u\|_U^2 \qquad \text{s.t. } u(x_i) = z_i \text{ for all } i = 1, \dots, N, \tag{73}$$

has a unique solution $u^\dagger \in U$ given by

$$u^\dagger = \sum_{i=1}^{N} c_i k(\cdot, x_i),$$

where the vector $c = (c_1, \dots, c_N)$ is defined as

$$c = K^{-1} z.$$

PROOF. Since the matrix K is positive definite, it follows from the discussion in Remark 6.69 that the family $(k_{x_1}, \dots, k_{x_N})$ in U is linearly independent. In particular, there exists some $\hat{c} \in \mathbb{K}^N$ such that

$$\sum_{j=1}^{N} \hat{c}_j k(x_i, x_j) = z_i$$

for all $i = 1, \dots, N$. Denote now

$$\hat{w} := \sum_{j=1}^{N} \hat{c}_j k(\cdot, x_j) \in U,$$

let

$$W := \big\{ u \in U : u(x_i) = 0 \text{ for all } i = 1, \dots, N \big\}$$

and denote

$$\hat{W} := \hat{w} + W = \big\{u \in U : u(x_i) = z_i \text{ for all } i = 1, \dots, N\big\}.$$

Then we can rewrite the problem (73) as

$$\min_{u \in U} \|u\|_U^2 \qquad \text{s.t. } u \in \hat{W},$$

which is the same as the problem of computing the projection of the point $0 \in U$ onto the affine space $\hat{W}$. Note moreover that W is finite dimensional and therefore a closed subspace of U. Thus Corollary 6.8 implies that (73) has a unique solution $u^\dagger$. Moreover, $u^\dagger$ is uniquely characterised by the conditions

$$u^\dagger \in \hat{W} \qquad \text{and} \qquad u^\dagger \in W^\perp.$$

Now define the mapping $T\colon U \to \mathbb{K}^n$,

$$Tu := \begin{pmatrix} \langle u, k_{x_1}\rangle_U \\ \vdots \\ \langle u, k_{x_N}\rangle_U \end{pmatrix} = \begin{pmatrix} u(x_1) \\ \vdots \\ u(x_N) \end{pmatrix}.$$

Then $W = \ker T$. As a consequence, since W is closed, we have that

$$W^\perp = (\ker T)^\perp = \operatorname{ran} T^*.$$

We now determine the mapping $T^*\colon \mathbb{K}^N \to U$. Here we regard $\mathbb{K}^N$ as a Hilbert space with the standard Euclidean inner product. Let $c \in \mathbb{K}^N$ and $u \in U$. Then

$$\langle u, T^*c\rangle_U = \langle Tu, c\rangle_{\mathbb{K}^N} = \sum_{i=1}^N \langle u, k_{x_i}\rangle \overline{c}_i = \Big\langle u, \sum_{i=1}^N c_i k_{x_i} \Big\rangle.$$

Thus

$$T^*c = \sum_{i=1}^N c_i k_{x_i}.$$

As a consequence,

$$\operatorname{ran} T^* = \operatorname{span}\{k_{x_1}, \dots, k_{x_N}\}.$$

This further implies that $u^\dagger \in U$ is the unique function of the form

$$u^\dagger = \sum_{j=1}^N c_j k_{x_j}$$

such that

$$u^\dagger(x_i) = z_i \qquad \text{for all } i = 1, \dots, N.$$

Now we note that (cf. Proposition 3.26)

$$z_i = u^\dagger(x_i) = \langle u^\dagger, k_{x_i}\rangle_U = \Big\langle \sum_{j=1}^N c_j k_{x_j}, k_{x_i} \Big\rangle_U = \sum_{j=1}^N c_j k(x_i, k_j) = \sum_{j=1}^N K_{ij} c_j.$$

In other words, the coefficients $c \in \mathbb{K}^N$ satisfy the equation $Kc = z$, which implies that $c = K^{-1}z$. □

THEOREM 6.74. *Let U be an RKHS over the set Ω and let $k\colon \Omega \times \Omega \to \mathbb{K}$ be the reproducing kernel of U. Let $x_1, \dots, x_N$ be (not necessarily distinct) points in Ω and let $z = (z_1, \dots, z_N) \in \mathbb{K}^N$ be given. Let moreover $\lambda > 0$. Then the problem*

$$\min_{u \in U} \lambda \|u\|_U^2 + \sum_{i=1}^N |u(x_i) - z_i|^2 \tag{74}$$

has a unique solution $u_\lambda \in U$ given by

$$u_\lambda = \sum_{i=1}^{N} c_i k(\cdot, x_i),$$

where the vector $c = (c_1, \dots, c_N) \in \mathbb{K}^N$ is defined as

$$c = (\lambda \operatorname{Id} + K)^{-1} z.$$

Proof. We can write

$$\lambda \|u\|_U^2 + \sum_{i=1}^{N} |u(x_i) - z_i|^2 = \lambda \Big(\|u\|_U^2 + \sum_{i=1}^{N} \Big| \frac{u(x_i)}{\sqrt{\lambda}} - \frac{z_i}{\sqrt{\lambda}} \Big|^2 \Big).$$

Thus we can rewrite (74) as the problem

$$\min_{(u,y) \in U \times \mathbb{K}^N} \Big(\|u\|_U^2 + \sum_{i=1}^{N} \Big| y_i - \frac{z_i}{\sqrt{\lambda}} \Big|^2 \Big) \qquad \text{s.t. } y_i = \frac{u(x_i)}{\sqrt{\lambda}} \text{ for all } i = 1, \dots, N.$$

Define now the mapping $T \colon U \times \mathbb{K}^N \to \mathbb{K}^N$,

$$T(u, y) := \Big(\frac{u(x_i)}{\sqrt{\lambda}} - y_i \Big)_{i=1}^{N}.$$

Define moreover on $U \times \mathbb{K}^N$ the inner product

$$\langle (u, y), (v, w) \rangle_{U \times \mathbb{K}^N} := \langle u, v \rangle_U + \langle y, w \rangle_{\mathbb{K}^N}$$

with corresponding norm $\|\cdot\|_{U \times \mathbb{K}^N}$. Then this the problem is the same as the problem

$$\min_{(u,y) \in U \times \mathbb{K}^N} \Big\| (u, y) - \Big(0, \frac{z}{\sqrt{\lambda}} \Big) \Big\|_{U \times \mathbb{K}^N} \qquad \text{s.t. } (u, y) \in \ker T.$$

Since $\ker T$ is a closed linear subspace of the Hilbert space $U \times \mathbb{K}^N$, this problem has a unique solution (u_λ, y_λ), which is the projection of $(0, \frac{z}{\sqrt{\lambda}})$ on the space $\ker T$. This projection is uniquely characterised by the conditions

$$(u_\lambda, y_\lambda) \in \ker T \qquad \text{and} \qquad (u_\lambda, y_\lambda - z/\sqrt{\lambda}) \in (\ker T)^\perp = \overline{\operatorname{ran} T^*}. \tag{75}$$

We now compute the mapping T^*. Let therefore $(u, y) \in U \times \mathbb{K}^N$ and $w \in \mathbb{K}^N$. Then

$$\begin{aligned} \langle (u, y), T^* w \rangle_{U \times \mathbb{K}^N} = \langle T(u, y), w \rangle_{\mathbb{K}^N} &= \sum_{i=1}^{N} \Big(\frac{u(x_i)}{\sqrt{\lambda}} - y_i \Big) \overline{w_i} \\ &= \sum_{i=1}^{N} \langle u, k_{x_i} \rangle_U \frac{\overline{w_i}}{\sqrt{\lambda}} - \langle y, w \rangle_{\mathbb{K}^N} = \Big\langle u, \sum_{i=1}^{N} \frac{w_i k_{x_i}}{\sqrt{\lambda}} \Big\rangle_U + \langle y, -w \rangle_{\mathbb{K}^N}. \end{aligned}$$

This shows that

$$T^* w = \Big(\sum_{i=1}^{N} \frac{w_i k_{x_i}}{\sqrt{\lambda}}, -w \Big) \in U \times \mathbb{K}^N.$$

We note in particular that $\operatorname{ran} T^*$ is a finite dimensional subspace of $U \times \mathbb{K}^N$ and therefore closed. Thus $(\ker T)^\perp = \overline{\operatorname{ran} T^*} = \operatorname{ran} T^*$. As a consequence, the

characterisation (75) of the solution of (74) is equivalent to the system of equations

$$\begin{aligned} u_\lambda &= \sum_{j=1}^{N} \frac{w_j}{\sqrt{\lambda}} k_{x_j}, \\ y_\lambda - \frac{z}{\sqrt{\lambda}} &= -w, \\ \frac{u_\lambda(x_i)}{\sqrt{\lambda}} &= y_{\lambda,i} \qquad\qquad \text{for all } i = 1, \dots, N. \end{aligned}$$

Denote now

$$c_j := \frac{w_j}{\sqrt{\lambda}}.$$

Then we can rewrite this system as

$$\begin{aligned} u_\lambda &= \sum_{j=1}^{N} c_j k_{x_j}, \\ u_\lambda(x_i) &= z_i - \lambda c_i \qquad \text{for all } i = 1, \dots, N. \end{aligned}$$

In particular,

$$z_i - \lambda c_i = u_\lambda(x_i) = \sum_{j=1}^{N} c_j k_{x_j}(x_i) = \sum_{j=1}^{N} c_j k(x_i, x_j) = \sum_{j=1}^{N} K_{ij} c_j.$$

Thus the vector $c = (c_1, \dots, c_N)$ solves the equation

$$z = Kc + \lambda c.$$

Since the matrix K is positive semi-definite, the matrix $K + \lambda \operatorname{Id}$ is positive definite. Thus this equation has a unique solution $c = (K + \lambda \operatorname{Id})^{-1} z$, which concludes the proof. □

Bibliography

[Axl24] Sheldon Axler. *Linear algebra done right*. Undergraduate Texts in Mathematics. Springer, Cham, fourth edition, 2024.

[BTA04] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004. With a preface by Persi Diaconis.

[Dev12] Keith J. Devlin. *Introduction to mathematical thinking*. Keith Devlin, Palo Alto, CA, USA, 2012.

[Hal74] Paul R. Halmos. *Naive set theory*. Undergraduate Texts in Mathematics. Springer-Verlag, New York-Heidelberg, 1974. Reprint of the 1960 edition.

[HS75] Edwin Hewitt and Karl Stromberg. *Real and abstract analysis*, volume No. 25 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Heidelberg, 1975. A modern treatment of the theory of functions of a real variable, Third printing.