Linear Algebra

Lecture notes by Jochen Kuttler

October 17, 2019

University of Alberta

Contents

Remarks

These notes are based on the two honours linear algebra courses taught at the University of Alberta over several years. The material (and also the exposition) is taken mostly from the following textbooks which these notes follow quite closely in parts:

– Charles Curtis, *Linear Algebra: An Introductory Approach*, Springer, 4th edition (1984), ISBN-13 978-0387909929

– Michael Artin, *Algebra*, Pearson, 2nd edition (2010), ISBN-13 978-0132413770

– Michael Artin, *Algebra*, (German translation), Birkhäuser, 1993, ISBN 3-7643-2927-0

These notes are not for redistribution.

Part I.

MATH 127: Honours Linear Algebra I

Introduction

1. Prerequisites

At the centre of Linear Algebra is the question of how to solve linear equations. We all recall the simplest such equation: given real numbers a, b consider the equation

$$ax = b. \tag{1.1}$$

That is, we are looking for real numbers that, when substituted for x, make this equation a true statement. Such a value is called a *solution*. Depending on the values of a and b there are no, one, or infinitely many such solutions. However, in many applications (inside and outside of mathematics) we need to allow more general coefficients. Clearly, (1.1) makes sense if a, b are integers, or rational numbers, and we may look for solutions in these (smaller) number systems. The general methods we will develop to solve an equation like this won't depend on whether we work over rationals or reals. However, it is crucial that we can divide: for instance the equation $2x = 1$ has no integer solutions. As a first step we will now fix the precise rules we need in order to be able to talk meaningfully about linear equations.

1.1. Number systems

We will base our discussion on the real numbers. That is, we treat the set[1] $\mathbb{R}$ together with its addition and multiplication as a given[2]. It is assumed that you are familiar with manipulating real numbers how to add, multiply them etc. The purpose of this section is to describe the important properties of these operations in mathematically precise terms.

Arithmetics. We have two operations[3] on the set $\mathbb{R}$ of real numbers, called *addition*, denoted by the symbol $+$, and *multiplication*, denoted by $\cdot$; that is, given two elements $a, b \in \mathbb{R}$ we may form their sum $a + b$ and their product $a \cdot b$ to obtain again elements of $\mathbb{R}$.

In the following we write $\mathbb{F}$ instead of $\mathbb{R}$ for reasons that will be clear soon. Also, we write $a \in \mathbb{F}$ to mean "a is an element of $\mathbb{F}$" – if $\mathbb{F} = \mathbb{R}$ this just says "a is a real number." We usually omit the $\cdot$ in the multiplication and simply write ab instead of $a \cdot b$.

1.1 Axioms. The two operations are subject to the following rules (also "axioms"):

[1] A *set* is a collection of objects; see the appendix.

[2] In a more foundational class, one would *construct* the real numbers step by step out of more basic sets, such as the set of natural numbers.

[3] Technically, an (binary) operation on a set X is simply a map $f\colon X \times X \to X$, that assigns to a pair (x, y) of elements of X a new element, denoted $f(x, y)$. Usually, we pick a symbol (e.g. $+$, $\times$, $*$) and write $x + y$ (resp. $x \times y$ or $x * y$) instead of $f(x, y)$. See the appendix.

a. *Associative Law.* For $a, b, c \in \mathbb{F}$ we have

$$a + (b + c) = (a + b) + c \qquad \text{and} \qquad a(bc) = (ab)c.$$

b. *Commutative Law.* For $a, b \in \mathbb{F}$ we have

$$a + b = b + a \qquad \text{and} \qquad ab = ba.$$

c. *Additive Identity.* There is an element, denoted 0, such that $a + 0 = 0 + a = a$ for all $a \in \mathbb{F}$.

d. *Additive Inverse.* For each $a \in \mathbb{F}$ there is an element $-a$ for which $a + (-a) = 0$.

e. *Multiplicative Identity.* There is an element 1, not equal to 0, such that $1a = a1 = a$ for all $a \in \mathbb{F}$.

f. *Multiplicative Inverse.* For each $a \in \mathbb{F}$ which is not equal to 0, there is an element a^{-1} for which $aa^{-1} = a^{-1}a = 1$.

g. *Distributive Law.* For $a, b, c \in \mathbb{F}$ we have[4]

$$a(b + c) = ab + ac.$$

These are the main rules from which all other arithmetic rules (see e.g. the Proposition 1.7 below) follow.

For quite a while the only properties of $\mathbb{R}$ that we will use are only those in this list of seven rules. In particular, it will always be completely irrelevant how addition and multiplication of real numbers are defined. All we need to know is our seven rules.

With this in mind, it makes sense to ask whether there are other mathematical structures that obey them.

Here is a very elementary one, which nevertheless appears in important applications (Computer Science comes to mind): Consider any set $S = \{a, b\}$ with exactly two elements. Define two operations on S by

$$a + b = b + a = b$$
$$a + a = b + b = a$$

[4] When dealing with two operations written as addition and multiplication, it is usually understood that any multiplication is carried out first (unless there are brackets). Thus an equation of the form $ab + c$ means "first compute ab, then add the result to c."

and

$$ab = ba = aa = a$$

and

$$bb = b.$$

Then S together with these two operations satisfies all seven rules. Indeed, a plays the role of 0 and b plays the role of 1. It makes therefore sense to relabel them and then $\mathbb{F}_2 = \{0, 1\}$ is called *the field with two elements* . In $\mathbb{F}_2$ we have the interesting relation that $-1 = 1$ (or $1 + 1 = 0$).

1.1.1 Problem. Show that there are only two ways to define operations on a set with two elements such that the result satisfies all seven properties. The two ways differ only by which element plays the role of 0 and which element serves as 1.

Here comes our first definition:

1.2 Definition. A *field* is a set $\mathbb{F}$ together with two operations, called addition $+$ and multiplication $\cdot$, such that all seven properties in 1.1 are satisfied.

We sometimes refer to the elements of a field $\mathbb{F}$ as *scalars* .

We usually write ab instead of $a \cdot b$. Again, just to emphasize the point, it is not relevant how these operations are defined. The elements of a field $\mathbb{F}$ need not be numbers, and the addition and multiplication need not have anything to do with addition or multiplication of actual numbers.

What is the purpose of this definition in Linear Algebra? It turns out that in many areas of mathematics or science in general one needs to solve linear equations where the variables and coefficients take values in some arbitrary field. If we understand the basic properties of a field well, then we can develop methods for solving these equations; methods that work in any context.

Since we make such a big deal out of it, it should be no surprise that there are many more fields than just the real numbers $\mathbb{R}$, or $\mathbb{F}_2$. For us, the most important one, the field of complex numbers, will be introduced in the next section.

An important subset of the real numbers $\mathbb{R}$ is of course the set $\mathbb{Z}$ of integers:

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\} \tag{1.2}$$

$\mathbb{Z}$ is *closed under addition and multiplication* in $\mathbb{R}$, that is, for all integers a, b both $a + b$ and ab are again integers. Also, $0, 1$ are elements of $\mathbb{Z}$ by definition, so there is an additive and multiplicative identity. Finally, since the two operations $+$ and $\cdot$ on $\mathbb{R}$ are associative, commutative, and obey the distributive law, they still do so when we restrict their arguments to integers. The only axiom in 1.1 that is violated is therefore f.: most integers don't have a multiplicative inverse that is also an integer: indeed, for any nonzero integer a, unless $a = \pm 1$, there is no integer b such that $ab = 1$.

To fix this problem, we introduce the set $\mathbb{Q}$ of *rational numbers*:

$$\mathbb{Q} = \{ab^{-1} \in \mathbb{R} \mid a, b \in \mathbb{Z}, b \neq 0\} \tag{1.3}$$

It is elementary (but non-trivial if one is only to use the Axioms 1.1 for $\mathbb{R}$) to verify that $\mathbb{Q}$ is again closed under $+$ and $\cdot$, and now satisfies all Axioms 1.1. Indeed, a multiplicative inverse for ab^{-1} is ba^{-1} which is again an element of $\mathbb{Q}$.

1.3 Definition. Let $\mathbb{F}$ be a field. A *subfield* of $\mathbb{F}$ is a subset S of $\mathbb{F}$ that is closed under addition and multiplication (ie. $a+b$ and ab are elements of S for all $a, b \in S$) such that all Axioms 1.1 hold for S with respect to the addition and multiplication inherited from $\mathbb{F}$. In particular, S is nonempty.

1.1.2 Problem. Show that a subset $S \subset \mathbb{F}$ is a subfield if

a. For all $a, b \in S$, $a + b$ and ab are in S.

b. $1 \in S$.

c. For all $a \in S$, $-a \in S$.

d. For all $a \neq 0 \in S$, $a^{-1} \in S$.

Show that b. can be replaced by the requirement that S is nonempty and contains an element other than 0. Show that the requirements $a + b \in S$ and $-a \in S$ may be replaced by: for all $a, b \in S$, also $a - b = a + (-b) \in S$. Observe that $\{0\}$ and $\emptyset$ satisfy all requirements except b.

Remark. $\mathbb{Q}$ is a subfield of $\mathbb{R}$. $\mathbb{Z}$ is not a subfield of $\mathbb{R}$. Any subfield (together with the addition and multiplication) is again a field.

For now we have three fields: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{F}_2$. Notice that the fundamental difference between these three is that one is finite whereas the others are infinite[5].

Convention. Throughout this class, unless otherwise specified, $\mathbb{F}$ will always denote an arbitrary field.

For the time being, you are welcome to think of $\mathbb{F}$ simply as the field of real (or rational) numbers, unless explicitly stated otherwise.

We won't dwell on the axioms of a field for long, but as one example, how they are used (in fact how we often use them subconsciously) let us prove the following

1.4 Lemma. *Let $a, b \in \mathbb{F}$. Then there is a unique $c \in \mathbb{F}$ such that $a + c = b$. In particular, if $b = 0$ and $a + c = 0$, then it follows $c = -a$.*

A similar statements holds for the multiplication: If $a \neq 0$, there is a unique $d \in \mathbb{F}$ such that $ad = b$. In particular, if $b = 1$, and $ad = 1$ then it follows that $d = a^{-1}$.

[5] One could argue that the real dividing line between $\mathbb{Q}$ and $\mathbb{R}$ on the one and $\mathbb{F}_2$ on the other hand is the fact that in $\mathbb{F}_2$, "$2 = 0$," that is $1+1 = 0$, whereas no finite sum $1+1+\cdots+1 = 0$ in $\mathbb{Q}$ or $\mathbb{R}$: indeed, there are infinite fields $\mathbb{F}$ where $1+1 = 0$, and in certain aspects these fields behave more closely like $\mathbb{F}_2$ than $\mathbb{R}$ or $\mathbb{Q}$.

(It may strike you as odd that one needs to prove a seemingly trivial fact that $-a$ is the only solution of $a+x=0$. But the reason this may seem odd is mostly that we are very much used to this fact. Anyway, since it is not specifically stated in our list of seven rules, and since we claim it holds for any field, we must prove it regardless.)

Proof. We first treat the case of the addition. There are two (distinct) assertions. First, the equation $a+x=b$ has a solution, meaning there is at least one solution. Second, the solution is unique, that is, there is at most one.

Put $c=(-a)+b$. Then

$$a+c=a+((-a)+b)=(a+(-a))+b=0+b=b \tag{1.4}$$

and hence c is a solution.

We now show that there is at most one solution: Suppose c, c' both are solutions of $a+x=b$. Then

$$a+c=a+c'$$

and after adding $-a$ to both sides of this equation we conlcude that $(-a)+(a+c)=(-a)+(a+c')$. Observe that

$$(-a)+(a+c)=((-a)+a))+c=0+c=c \tag{1.5}$$

whereas $(-a)+(a+c')=c'$. This shows that $c=c'$.

Since $-a$ is a solution for $a+x=0$ it follows it is the only such solution.

The statement about the multiplication is very similar: $d=a^{-1}b$ is a solution:

$$ad=a(a^{-1}b)=(aa^{-1})b=1b=b.$$

Any such solution is uniquely determined: If $ad=ad'$ then $a^{-1}(ad)=a^{-1}(ad')$ which by associativity yields

$$(a^{-1}a)d=(a^{-1}a)d'$$

and so $1d=1d'$. But this means $d=d'$.

Again, an immediate consequence is the fact that for any nonzero a, a^{-1} is the only solution of $ax=1$. □

As a corollary we obtain the fact that the two identities in a field are unique:

1.5 Corollary. *The additive and multiplicative identities in a field $\mathbb{F}$ are uniquely determined: whenever $ea=a$ for all a, then $e=1$. Similarly, whenever $a+n=a$ for all a, then $n=0$.*

Proof. 0 is the (unique) solution of $0+x=0$. n is also one. Hence $n=0$. 1 is the (unique) solution of $1x=1$. e is another one, hence $e=1$. □

1.1.3 Problem. Let $S\subseteq\mathbb{F}$ be a subfield. Show that the identities of S and $\mathbb{F}$ coincide. Also, show that for an element $a\in S$ its additive (or multiplicative, if $a\neq 0$) inverse does not depend on whether we think of a as an element of $\mathbb{F}$ or S.

1.1.4 Problem. In each of the Equations (1.4) and (1.5), check where we used which of the Axioms 1.1.

1.6 Remarks.

a. Our Axioms 1.1 make no mention of subtraction. Rather than being an entirely new operation, it is determined by the addition: if $a, b \in \mathbb{F}$, then $b - a$ is the unique element c of $\mathbb{F}$ for which $a + c = b$. In other words,

$$b - a = (-a) + b = b + (-a).$$

It is important to understand that this is a definition of the symbol $b - a$.

b. Similarly, if $a, b \in \mathbb{F}$ and $a \neq 0$, we sometimes write b/a to denote the unique element d of $\mathbb{F}$ for which $ad = b$. Thus $b/a = ba^{-1} = a^{-1}b$.

Our final observation in this context is that the usual rules of arithmetic apply in any field:

1.7 Proposition. *Let a, b be elements of a field $\mathbb{F}$:*

a. $0 \cdot a = a \cdot 0 = 0.$

b. $(-1) \cdot a = -a.$ *More generally,* $(-b)a = -(ba) = b(-a).$

c. $-(-a) = a$ *and, if* $a \neq 0$, $(a^{-1})^{-1} = a.$

d. $(-a)(-b) = ab.$

e. $-(a + b) = -a + (-b).$

Proof.

a. $0a + 0a = (0 + 0)a = 0a$. Hence $0a$ solves the equation $0a + x = 0a$. As 0 is the only such solution, $0a = 0$. Of course, $0a = a0$, which finishes the first part.

b. To verify that $(-1)a = -a$ it is enough to check that $a + (-1)a = 0$. But this follows from a.: $a + (-1)a = (1 + (-1))a = 0a = 0$. Using this[6] (twice), we find that $(-b)a = ((-1)b)a = (-1)(ba) = -(ba)$; similarly, $b(-a) = b((-1)a) = (-1)(ba) = -(ba)$.

c. $a + (-a) = 0$, which makes a the (unique) solution of $(-a) + x = 0$. Hence $a = -(-a)$. Similarly, $aa^{-1} = 1$ forces $a = (a^{-1})^{-1}$.

d. $(-a)(-b) = -(a(-b)) = -(-(ab)) = ab$. (Here we applied a. and c.).

e. Left as an exercise.

□

[6] Try to prove $(-b)a = -(ba)$ directly without using the case $b = 1$.

1.1.5 Problem. Let $S = \{a, b, c\}$ be a set with three elements. Show that there are operations $+$ and $\cdot$ that follow the seven rules.

(As a starting point, let a play the role of 0, b the one of 1. As a first step show that necessarily then $c = b + b$.)

1.8 Remark. We adopt the usual abbreviations commonly used in the arithmetic of real numbers: We usually write $a - b$ instead of $a + (-b)$. Similarly, we sometimes write a/b instead of ab^{-1}. Similarly, if $k \in \mathbb{Z}$ and $a \in \mathbb{F}$ is nonzero, we define

$$a^k = \begin{cases} \underbrace{a \cdot a \cdots a}_{k \text{ factors}} & k > 0 \\ 1 & k = 0 \\ (a^{-1})^{-k} & k < 0 \end{cases} \tag{1.6}$$

(Before we go on, let us briefly mention, that the associative laws extend to sums and products of more than just two elements of a field: it is possible to show that if $a_1, a_2, \ldots, a_n \in \mathbb{F}$, then the sum $a_1 + a_2 + \cdots + a_n$ is well defined (regardless of the order in which we perform the additions). Likewise, $a_1 a_2 \cdots a_n$ is independent of the order of multiplication. We will return to this question in a more general context later.)

1.2. The complex numbers

Our first project is to extend the real number system to the most important field of all, the field of complex numbers.

Real numbers have an important property (which is not shared by all fields), namely they can be *ordered*: for any two real numbers a, b we always have one and only one of the three possibilities $a < b$, $a = b$, or $b < a$. Note that these three possibilities correspond to the three statements $0 < b - a$, $0 = a - b$, or $0 < a - b$. The order is therefore determined by the collection of *positive* real numbers.

1.9 Definition. A field $\mathbb{F}$, together with a subset $P \subseteq \mathbb{F}$ of so called *positive elements*, is an *ordered* field, if the following holds:

a. $a + b \in P$ for all $a, b \in P$.

b. $ab \in P$ for all $a, b \in P$.

c. For each $a \in \mathbb{F}$ one and only one of the following three statements is true: $a \in P$, $-a \in P$, $a = 0$.

The first two properties simply mean that "positive + positive = positive" and "positive · positive = positive." What about "negative · negative?" Let $x, y \in \mathbb{F}$ be negative, that is, $x, y \neq 0$ and $x, y \notin P$. Then $-x, -y \in P$, and so $(-x)(-y) = xy \in P$ as well. It follows that the product of two negatives is positive. In particular, P can never be empty. In fact, 1 is alsways positive: $1 = (-1) \cdot (-1)$ is a nonzero square and hence positive.

1.10 Example. $\mathbb{R}$ is ordered with $P = \{x \in \mathbb{R} \mid x > 0\}$. Similarly, $\mathbb{Q}$ is ordered (with set of positive elements given by $P \cap \mathbb{Q}$).

1.2.1 Problem. Let $\mathbb{F}$ be an ordered field (with positive elements P). For $a, b \in \mathbb{F}$ define $a < b$ if $b - a \in P$. Show that this defines a total order on $\mathbb{F}$: for each $a, b \in \mathbb{F}$ exactly one of the three statements is true: $a < b$, $a = b$, $b < a$.

Also, show that if $a < b$ and $b < c$, then $a < c$.

In an ordered field $\mathbb{F}$, the equation $x^2 = -1$ has no solution: Indeed, we have seen that 1 is positive, so -1 is negative. We have also seen that squares are always positive. So let a be a solution. Then a cannot be positive, because if so then so is a^2 which is -1. Thus, $a \notin P$. But then $-a \in P$, and hence so is $(-a)(-a) = a^2 = -1$, a contradiction. As we all know, this applies in particular to the field $\mathbb{R}$ of real numbers, where we cannot take square roots of negative numbers: The equation

$$x^2 = -1 \tag{1.7}$$

has no solution.

Our plan is to extend the set of real numbers by adding solutions to this equation. To motivate the definitions that are about to come, let us work backwards: let us assume that there exists a field C, say, that contains the real numbers as a subfield such that in C we can solve the above equation: there is an element $\mathbf{i}$ for which $\mathbf{i}^2 = -1$.Since C contains $\mathbb{R}$ we can form the set $C' = \{a + b\mathbf{i} \mid a, b \in \mathbb{R}\}$. Notice the rules of a field imply that indeed if a, b are real numbers then there must be an element $b\mathbf{i} \in C$ and hence also an element $a + b\mathbf{i}$.

What can we say about C'? Well, if $a + b\mathbf{i}$ and $c + d\mathbf{i}$ are both in C', then so is their sum: $(a + b\mathbf{i}) + (c + d\mathbf{i}) = (a + c) + (b + d)\mathbf{i}$ because $a + c \in \mathbb{R}$ and $b + d \in \mathbb{R}$. What about their product?

$$(a + b\mathbf{i})(c + d\mathbf{i}) = ac + (b\mathbf{i})c + a(d\mathbf{i}) + (b\mathbf{i})(d\mathbf{i}) = ac + (bc + ad)\mathbf{i} + bd(\mathbf{i})^2 \tag{1.8}$$

where we used several of the field axioms which hold in C. But $\mathbf{i}^2 = -1$, so this product is equal to

$$(a + b\mathbf{i})(c + d\mathbf{i}) = (ac - bd) + (ad + bc)\mathbf{i} \tag{1.9}$$

and is again an element of C'. Notice that because $\mathbb{R}$ is supposed to be a subfield of C, the result of a computation involving only real numbers is independent of whether we view the real numbers as real numbers or as elements of C. Thus, we can compute in C' without even knowing how C' is defined. This is ample motivation to simply try and use (1.9) as a definition of a new multiplication. But multiplication of what? An element $a + b\mathbf{i}$ is determined by the (ordered) pair (a, b) of real numbers. Multiplying $(a + b\mathbf{i})(c + d\mathbf{i})$ and writing it as $e + f\mathbf{i}$ produces a new pair of real numbers (namely, (e, f)), defined by $e = ac - bd$ and $f = ad + bc$.

Recall that a (ordered) *tuple* of real numbers numbers is an ordered pair[7] (a, b) where a, b are real numbers. Two such pairs are equal if and only if both entries coincide; that is, $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$. As an example, this means that $(1, 2) \neq (2, 1)$.

Let

$$\mathbb{C} = \{(a, b) \mid a, b \in \mathbb{R}\}$$

be the collection of all tuples of real numbers[8]. On this set, we introduce two operations. Let $(a, b), (c, d) \in \mathbb{C}$. The addition is defined componentwise as

$$(a, b) + (c, d) = (a + c, b + d)$$

and it should be clear that this is associative and commutative:

$$(a, b) + (c, d) = (a + c, b + d) = (c + a, d + b) = (c, d) + (a, b).$$

(We leave the associativity as an exercise.)

In the end we will show that $\mathbb{C}$ satisfies all field axioms, so we check some of them right away: the addition has an identity element, which we will denote by $0_{\mathbb{C}}$ for the time being:

$$0_{\mathbb{C}} = (0, 0).$$

Then $0_{\mathbb{C}} + z = z$ for all $z \in \mathbb{C}$. Indeed, if $z = (a, b)$, then $z + 0_{\mathbb{C}} = (a + 0, b + 0) = (a, b) = z$. Also

$$(a, b) + (-a, -b) = (a + (-a), b + (-b)) = (0, 0) = 0_{\mathbb{C}}.$$

So every element has an additive inverse. Next, we introduce the following multiplication:

$$(a, b) \cdot (c, d) = (ac - bd, ad + bc).$$

It should be clear from the definition that this is a commutative operation because the multiplication and addition of real numbers is commutative. It is also associative. For this let $(a, b), (c, d), (e, f) \in \mathbb{C}$:

$$\begin{aligned} ((a, b)(c, d))(e, f) &= (ac - bd, ad + bc)(e, f) \\ &= (ace - bde - adf - bcf, acf - bdf + ade + bce) \\ &= (a(ce - df) - b(de + cf), a(cf + de) + b(ce - df)) \\ &= (a, b)(ce - df, cf + de) = (a, b)((c, d)(e, f)) \end{aligned} \tag{1.10}$$

A similar computation shows that $(a, b)((c, d) + (e, f)) = (a, b)(c, d) + (a, b)(e, f)$. We even have an identity element for the multiplication: if we put $1_{\mathbb{C}} = (1, 0)$, then

$$1_{\mathbb{C}}(a, b) = (1, 0)(a, b) = (1a - 0b, 1b + 0a) = (a, b)$$

[7] The ordered pair (a, b) is not to be confused with the *set* $\{a, b\}$ containing a, b.

[8] You may have seen (and will see again) this set labeled as $\mathbb{R}^2$ or $\mathbb{R} \times \mathbb{R}$.

> The set $\mathbb{C}$ together with this addition and multiplication is called the *field of complex numbers* (and its elements accordingly are referred to as complex numbers).

We mentioned earlier that we want to construct an extension of the real number system. So how does $\mathbb{C}$ contain $\mathbb{R}$? It turns out that $\mathbb{C}$ contains a set that is sufficiently "similar" to the set of real numbers so that it makes sense to treat them as equal:

Let

$$R = \{(a, 0) \mid a \in \mathbb{R}\} \subseteq \mathbb{C}.$$

Notice that for $(a, 0), (b, 0) \in R$, by definition, we have $(a, 0)(b, 0) = (ab, 0) \in R$ and $(a, 0) + (b, 0) = (a + b, 0)$. The natural one-to-one correspondence $\mathbb{R} \leftrightarrow R$ that pairs $a \in \mathbb{R}$ with $(a, 0) \in R$ therefore preserves the addition and multiplication. There is no material difference whether we write a or $(a, 0)$ for the real number a.

> We may therefore *identify* the set $\mathbb{R}$ of real numbers with the set R and think of $\mathbb{R}$ as a subset of the set of complex numbers, which is closed under addition and multiplication.

From now on, if $a \in \mathbb{R}$ we will write a also for the complex number $(a, 0)$. Notice that keeping this in mind, we can write the complex number $z = (a, b)$ also as $z = a + b(0, 1)$. Also, in this sense, the real number 1 is identified with $1_{\mathbb{C}}$, and 0 is identified with $0_{\mathbb{C}}$. We therefore drop these subscripts. Finally, put $i = (0, 1)$. Then every complex number $z = (a, b)$ can be written *uniquely* as

$$z = a + bi$$

with $a, b \in \mathbb{R}$.

The most important computation we need to do is figure out whether we can now solve the equation $x^2 + 1 = 0$: indeed, put $x = i = (0, 1)$:

$$i^2 = (0 \cdot 1 - 1 \cdot 1, 0 \cdot 1 + 1 \cdot 0) = (-1, 0) = -1.$$

Recalling that in $\mathbb{C}$, $1_{\mathbb{C}}$ is actually equal to the real number 1 we conclude that we indeed constructed a new "number" system where we can take square roots of -1.

What is still missing is to observe that $\mathbb{C}$ really is a field. For this, the only fact we still need is that every nonzero complex number z must have a multiplicative inverse. One can find z^{-1} explicitly by interpreting the equation $zz^{-1} = 1$ as a system of equations on the two real coefficients of z^{-1}.

However, we can also proceed as follows: recall that for real numbers a, b we have $(a + b)(a - b) = a^2 - b^2$. Similarly, a direct computation shows that

$$(a + bi)(a - bi) = a^2 + b^2$$

which is a *real number*. In particular, the real number $(a+bi)(a-bi)$ is nonzero (and positive), if $a + bi \neq 0$. Because of this interesting property we call $a - bi$ the *complex conjugate* of

$z = a + bi$ and denote it by $\overline{z}$. We know how to to invert real numbers, so if $z = a + bi \neq 0$, we conclude that

$$z^{-1} = (a+bi)^{-1} = (a^2+b^2)^{-1}\overline{z} = \frac{a}{a^2+b^2} - \frac{b}{a^2+b^2}i$$

is a multiplicative inverse for z.

Summarizing we obtain

1.11 Theorem. *The complex numbers $\mathbb{C}$ form a field.*

You may have observed that for solving for square roots of negative real numbers, it is enough to have a square root of -1: indeed, if $a < 0$ is a negative real number, then $\sqrt{-a} \cdot i$ will be a square root of a. But note that since in $\mathbb{C}$, $x^2 = -1$ has a solution, $\mathbb{C}$ cannot be turned into an ordered field. Therefore, it makes no sense to talk about a "positive" square root of a number a.

More generally, the usual quadratic formula now yields a solution of *every* quadratic equation:

> Let a, b, c be real numbers with $a \neq 0$. Then the solutions of $ax^2 + bx + c$ are
>
> $$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1.11}$$
>
> where we write $\sqrt{r}$ for any particular (possibly complex) solution of $x^2 = r$.

1.12 Example. The solutions of $x^2 + 2x + 5 = 0$ are $x_1 = 1 + 2i$ and $x_2 = 1 - 2i$.

One might think that, since we extended the number system, we may have obtained new quadratic equations that have no solution by allowing complex coefficients. However, this is not the case: every quadratic equation over the complex numbers has a solution. Even better, *every polynomial equation* (of arbitrary positive degree) with complex coefficients has solutions in $\mathbb{C}$. This is the very reason why the complex numbers are so important. There is no need for further extending the number system: we can solve any (algebraic) equation, at least in principle. We will return to this later.

Before we move on, let us introduce the following commonly used notation: if $z = a + bi$ is a complex number, a is called the *real part* of z, often denoted $\mathrm{Re}(z)$, and b is called the *imaginary part*, denoted $\mathrm{Im}(z)$.

1.2.2 Problem. Let z, w be arbitrary complex numbers. Show:

a. $z = \overline{z}$ if and only if $z \in \mathbb{R}$.

b. $\overline{z} = -z$ if and only if $z \in \mathbb{R}i := \{bi \mid b \in \mathbb{R}\}$.

c. $z\overline{z} = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2 \in \mathbb{R}_{\geq 0}$.

d. $\overline{z+w} = \overline{z} + \overline{w}$

e. $\overline{zw} = \overline{z} \cdot \overline{w}$

f. $\overline{z^{-1}} = \overline{z}^{-1}$ if $z \neq 0$.

1.3. Mathematical induction

In these notes we treat the set $\mathbb{R}$ of real numbers as the fundamental set out of which many other sets are constructed. We introduced the set $\mathbb{Z} \subseteq \mathbb{R}$ in Chapter 1. Of course, we are all familiar with the set $\mathbb{N} = \{1, 2, \dots\}$ of positive integers, the set of *natural numbers*. The fundamental "philosophical" property of the natural numbers is that they are used for counting. Intuitively, if we "count" or enumerate the numbers in $\mathbb{N}$, we will never have counted them all (after all, there are infinitely many of them), however eventually, *every* natural number will be counted. A mathmatical precise formulation of this fact is the following

The Principle of Induction. Let S be a set of natural numbers, such that the following holds:

a. $1 \in S$;

b. whenever $n \in S$ then also $n + 1 \in S$.

Then $S = \mathbb{N}$.

Here is a little experiment: try to define the set of natural numbers as a subset of $\mathbb{R}$. Of course, it is the set of positive integers. Well, what is an integer? Its decimal expansion is "simple." What is its decimal expansion? In particular, what is a digit in there? It turns out, one way to define the (positive) integers as a subset of $\mathbb{R}$ is precisely by the Principle of Induction: $\mathbb{N}$ is the *smallest* subset of $\mathbb{R}$ satisfying a. and b.; that is, any other subset of $\mathbb{R}$ that satisfies a. and b. actually contains $\mathbb{N}$ as a subset. In other words, it is the intersection of all such subsets of $\mathbb{R}$.

The POI is fundamental for proving statements that depend on natural numbers n: Suppose we have a statement $A(n)$ for each natural number n. For instance, $A(n)$ could be "A set with n elements has 2^n subsets." How can we prove such a statement? There is no way we can prove it for each natural number individually, since this involves infinitely many individual proofs.

However, we can proceed as follows: Let $S \subseteq \mathbb{N}$ be the subset of all natural numbers for which $A(n)$ holds: that is,

$$S = \{n \in \mathbb{N} \mid A(n)\}.$$

If we want to prove that $A(n)$ holds for all $n \in \mathbb{N}$, we need to prove that $S = \mathbb{N}$. By the Principle of Induction, all that is needed is to show that $1 \in S$ and that whenever a given natural number n is a member of S, then so is $n + 1$. Thus we have to show two statements

a. $A(1)$ holds.

b. For each $n \in \mathbb{N}$, if $A(n)$ holds, then $A(n + 1)$ holds.

We usually call the case $n = 1$ the *base case*, whereas we refer to the second part as the *induction step*. When showing the second step it is enough to assume that $A(n)$ is true and then show that this implies $A(n+1)$ is true. We usually refer to this as the *induction assumption* or *induction hypothesis*. However it is important to keep in mind that the n in the induction hypothesis is some arbitrary but specific integer.

There are two variants of the POI. Both have their advantages in certain circumstances but it is important to keep in mind that they are logically equivalent.

Proof by complete induction: Suppose for each $n \in \mathbb{N}$, $A(n)$ is a statement that is either true or false. To prove that $A(n)$ is true for all $n \in \mathbb{N}$ it is enough to prove:

a. $A(1)$ is true.

b. For each natural number n, the following is true: If $A(r)$ is true for all natural numbers $r < n$, then $A(n)$ is true.

The POI is equivalent to the following fact:

Well-ordering Principle Let S be a *nonempty* set of natural numbers. Then S has a smallest element. That is, there is a natural number $n_0 \in S$ such that for all $n \in S$, $n_0 \leq n$.

Here is a simple example:

1.13 Proposition. *a be a real number, different from 1. Then*

$$1 + a + a^2 + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}$$

Proof. We proceed by induction. Here the statement $A(n)$ is of course that the above equation is true for n.

Base case: The left hand side for $n = 0$ evaluates to 1. The right hand side is $(1-a)/(1-a) = 1$. Thus $A(0)$ is true. Next, for $n = 1$, the left hand side is $1 + a$, and the right hand side is $(1-a^2)/(1-a) = 1+a$, provided $a \neq 1$. Thus $A(1)$ is true as well. But see also Problem 1.3.1 below.

Induction step: Let n be given. Suppose $A(n)$ is true. We need to show that then also $A(n + 1)$ is true. Now

$$1 + a + a^2 + \cdots + a^{n+1} = (1 + a + \cdots + a^n) + a^{n+1}$$

Since $A(n)$ is supposed to be true, we may replace the first part on the right hand side by $(1 - a^{n+1})/(1 - a)$, to get

$$1 + a + \cdots + a^{n+1} = \frac{1 - a^{n+1}}{1 - a} + a^{n+1}.$$

But

$$\frac{1 - a^{n+1}}{1 - a} + a^{n+1} = \frac{1 - a^{n+1} + (1 - a)a^{n+1}}{1 - a} = \frac{1 - a^{n+2}}{1 - a}.$$

But this is precisely the right hand side that is needed for $A(n+1)$ to be true. The set of all natural numbers n for which $A(n)$ holds is therefore the set of all natural numbers. And the above formula is true. □

To conclude this section let us look at the statement "A set with n elements has 2^n subsets." used as an example abobe:

1.14 Proposition. *Let $n \geq 0$ be an integer. A set with n elements has 2^n subsets.*

Proof. The case $n = 0$ covers the empty set: the only subset of the empty set is the empty set itself. Luckily $2^0 = 1$, so the proposition holds for $n = 0$. We will now prove the proposition for sets of sizes $n > 0$ by induction on n:

Base case: If S is a set with one element, it has precisely two subsets: the empty set and S. Hence the number of subsets is $2 = 2^1$ as claimed.

Induction step: Let $n > 0$ be an integer. Suppose we know the proposition is true for sets with n elements (Induction Assumption; IA).

Let S be a set with $n+1$ elements. In particular, S is not empty. Let $s \in S$ be a fixed element and define $S' = S - \{s\}$. Then S' has n elements. Hence the IA applies to S'. Any subset of S' is also a subset of S because $S' \subseteq S$. Thus S has exactly 2^n subsets that are also subsets of S'. These are precisely the subsets of S that do not contain s as an element. For each such subset T, say, we create a new subset of S by adding s: $T' = T \cup \{s\}$. Notice that if $T_1, T_2 \subseteq S'$, then $T_1' \neq T_2'$ unless $T_1 = T_2$. We obtain another 2^n substes of S, this time those that do contain s as an element. Moreover, all such subsets U are of the form T' for $T = U \cap S' = U - \{s\}$. Since each subset of S either contains s or doesn't contain S, we find that S has a total of $2^n + 2^n$ subsets, which is equal to $2 \cdot 2^n = 2^{n+1}$ as needed. □

1.3.1 Problem. In the previous proof, we could have used $n = 0$ as the base case. Indeed, prove the following version of the POI:

Let $S \subseteq \mathbb{Z}$ be a set of integers and let $k \in \mathbb{Z}$ be any integer such that

a. $k \in S$

b. whenever $n \in S$ then also $n+1 \in S$.

Then S contains $\mathbb{Z}_{\geq k} = \{n \in \mathbb{Z} \mid n \geq k\}$.

(*Hint:* Consider the set $S' = \{n' \in \mathbb{N} \mid (k-1) + n' \in S\}$ and apply the regular POI to S'.)

1.3.1. *Recursive definitions

Here is an important consequence or application: Often we want to define a mathematical object A_n depending on the natural number n in a recursive fashion. That is, there is an explicit rule how to obtain A_{n+1} if A_n has been constructed, and there is an object A_1. The question is, does A_n exist for all natural numbers n?

For instance, we could ask whether there exists a sequence $\{a_n\}_n$ of natural numbers such that $a_1 = 1$ and $a_{n+1} = (n+1)a_n$. The answer is yes, and the resulting sequence is $a_n = n!$ (n factorial). Colloquially we write $n! = n\cdot(n-1)\cdot\dots\cdot 2\cdot 1$, which, in pure logical terms, is nothing but a recursive definition. Usually, whenever we write $\dots$ in a mathematical construction, we implicitly use a recursive definition.

We don't want do stress this point at this stage but nevertheless, we will prove the following abstract version of a recursive definition to gain some experience in applying mathematical reasoning.

1.15 Theorem. *Let X be a nonempty set and suppose $f\colon X \to X$ is a function. For each $x \in X$, there exists a unique sequence $x_1, x_2, \dots \in X$ such that for each $n \in \mathbb{N}$ we have $x_{n+1} = f(x_n)$ and $x_1 = x$.*

Since all sequences of mathematical objects that we usually encounter are elements of some (sometimes very large) set, the theorem solves our problem completely.

Proof. Let $S \subseteq \mathbb{N}$ be the subset of natural numbers n for which there is a unique sequence $x_1, x_2, \dots, x_n$ satisfying $x_1 = x$ and $x_{k+1} = f(x_k)$ whenever $k = 1, 2, \dots, n-1$. We will first show that $S = \mathbb{N}$ using the principle of induction.

First, let us show the base case: $1 \in S$. This is evident, as $x_1 = 1$ is a sequence satisfying all requirements, but obviously is also the only one doing so.

Next, given an integer n, suppose $n \in S$. We have to show that this implies that also $n+1 \in S$. For this, let $x_1, x_2, \dots, x_n$ be the unique sequence that exists because $n \in S$. If we put $x_{n+1} = f(x_n)$ we obtain a sequence $x_1, x_2, \dots, x_{n+1}$ as needed. It remains to show this sequence is unique. Suppose $y_1, y_2, \dots, y_{n+1}$ is also a sequence such that $y_1 = x$ and $f(y_k) = y_{k+1}$. Then $y_1, y_2, \dots, y_n$ must coincide with $x_1, x_2, \dots, x_n$ because $n \in S$. In particular, $y_n = x_n$ and hence $y_{n+1} = f(y_n) = f(x_n) = x_{n+1}$. Thus the sequence $x_1, x_2, \dots, x_{n+1}$ is unique as well and so $n+1 \in S$. By the principle of induction we conclude that $S = \mathbb{N}$.

Hence for each integer n there exists a unique sequence $x_1, x_2, \dots, x_n$. A priori this sequence might depend on n, so we should write $x_{n1}, x_{n2}, \dots, x_{nn}$. However, our proof of the uniqueness part showed that indeed, all sequences coincide where they overlap (i.e. $x_{ni} = x_{mi}$ whenever $n \geq m$ and $i \leq m$). Thus, if we put $x_n = x_{nn}$, we obtain a sequence that satisfies $x_{n+1} = f(x_n)$ and $x_1 = x$.

Is it unique? Yes: Let $y_1, y_2, \dots$ be another such sequence. We need to conclude that $x_i = y_i$. For this, we may argue that for each n, $y_1, y_2, \dots, y_n$ is the sequence constructed above which we showed to be unique. Hence in particular $y_n = x_n$. □

Often, in the recursive construction, x_{n+1} does not just depend on x_n but also on the integer n in some way. The following reformulation is therefore sometimes more convenient:

1.16 Corollary. *Let X be a nonempty set and $f\colon \mathbb{N}\times X \to X$ be a function. For each $x \in X$ there exists a unique sequence $x_1, x_2, \ldots$ such that $x_1 = x$ and $x_{n+1} = f(n, x_n)$.*

Proof. Let $Y = \mathbb{N}\times X$ and define $F\colon Y \to Y$ as $F(n,x) = (n+1, f(n,x))$. For $x \in X$ put $y = (1,x)$. By the theorem, there exists a unique sequence $y_1, y_2, \ldots$ such that $y_1 = y$ and $y_{n+1} = F(y_n)$. If we write $y_i = (i, x_i)$, then the sequence x_n is the one we are looking for: $x_{n+1} = f(n, x_n)$ and $x_1 = x$.

The uniqueness part follows because the sequence y_n is unique. □

Normally to define a sequence of objects recursively, we usually simply specify the first object and how to construct the object A_n out of the objects constructed earlier.

We have used this construction implicitly before: given $a \in \mathbb{F}$, and a natural number n we defined a^n to be $a \cdot a \cdots a$ (n factors). If you think about it, you will realize that this is nothing but a recursive definition: we specify how to compute a^{n+1} once we know how to compute a^n, namely, $a^{n+1} = a \cdot a^n$. Thuse, in this example, $f\colon \mathbb{F} \to \mathbb{F}$ would be given by $f(x) = ax$. Then the theorem asserts there is a unique sequence $a_1, a_2, \ldots,$ such that $a_1 = a$, and $a_{n+1} = af(a_n)$, which is precisely what we wanted.

In practice, we often define a sequence A_n of mathematical object as follows: we specify how A_1 is defined. We then explain how A_{n+1} is constructed out of A_n (and sometimes also out of all previous objects $A_1, A_2, \ldots, A_n$.

As a final example consider the Fibonacci sequence $f_0, f_1, \ldots$: it is a sequence of nonnegative integers subject to the following recurrence relation: $f_0 = 0$, $f_1 = 1$, and $f_{n+2} = f_n + f_{n+1}$.

1.3.2 Problem. Show that the Fibonacci sequence exists (using the theorem or its corollary).

2. Matrices

2.1. Intersecting two lines

In the following let E be the Euclidean plane. By choosing a coordinate system, we may identify E with the set $\mathbb{R}^2$ of all ordered pairs of real numbers. If $L \subseteq E$ is a *line*, then the points $p = (x, y)$ in L satisfy a linear equation of the form $ax + by = c$. Thus,

$$L = \{(x, y) \in E \mid ax + by + c = 0\}$$

If L' is another line, with equation $a'x + b'y = c'$, then the points in $L \cap L'$ are precisely the the points (x, y) that satisfy the two equations

$$\begin{aligned} ax + by &= c \\ a'x + b'y &= c' \end{aligned} \tag{2.1}$$

simultaneously.

Assuming L (resp. L') is a proper line, a and b (resp. a', b') cannot both be zero at the same time. We all know that if L and L' are neither equal nor parallel, then $L \cap L'$ contains a single point. If both $b, b' \neq 0$ this means that $a/b \neq a'/b'$, or $ab' \neq a'b$. This last form also covers the case where b or b' is zero (why?). We conclude that $L \cap L'$ consist of one point if and only if

$$ab' - a'b \neq 0.$$

Let us give an algebraic proof of this. Suppose $ab' - a'b \neq 0$. Then multiplying the first equation by b' and the second by b we obtain the new system

$$\begin{aligned} ab'x + bb'y &= b'c \\ a'bx + bb'y &= bc' \end{aligned}$$

Now we can subtract the second equation from the first and get

$$(ab' - a'b)x = b'c - bc'.$$

(Note that the coefficients of y cancel.)

This gives us

$$x = \frac{b'c - bc'}{ab' - a'b}. \tag{2.2}$$

Similarly, multiplying the first equation by a' and the second by a we get

$$\begin{aligned} aa'x + a'by &= a'c \\ aa'x + ab'y &= ac' \end{aligned}$$

which yields

$$(a'b - ab')y = a'c - ac'$$

or

$$y = \frac{a'c - ac'}{a'b - ab'}. \tag{2.3}$$

The upshot is we have a unique solution of the System 2.1 as long as $ab' - a'b \neq 0$. (If $ab' - a'b = 0$ an elementary compuation shows that $L \cap L' = L$ if $L = L'$, or $L \cap L' = \emptyset$ otherwise.

2.2. Linear equations

In large parts, Linear Algebra is about keeping track of information in a clever (and efficient) way. The prime example of this are matrices. Their use will be apparent in due course, a few examples will appear at the end of this section. In what follows, $\mathbb{F}$ will be a fixed (but arbitrary) field, but for the time being you may always think of $\mathbb{F}$ as either $\mathbb{Q}$ or $\mathbb{R}$ (or $\mathbb{C}$, for that matter). We will put restrictions on $\mathbb{F}$ once necessary.

Most of the time, in terms of applications, Linear Algebra is used in the context of systems of linear equations. What is the problem? The formal definition is as follows: A *system of linear equations in n variables* is a collection of m equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{2.4}$$

Here the x_i are thought of as variables and the a_{ij} are elements of a (fixed) field $\mathbb{F}$. A *solution* of such a system is an n-tuple $S = (s_1, s_2, \ldots, s_n)$ with $s_i \in \mathbb{F}$ such that all equations in (2.4) are satisfied when s_i is substituted for x_i.

Remarks.

a. With the advent of serious computer power, linear approximations of many problems become feasible and consequently more and more important. Typical systems contain thousands of equations and variables and you might imagine that the methods to solve these must be quite sophisticated, computers notwithstanding. Nevertheless in principle most of these algorithms all go back to Gaussian elemination. While this is an interesting topic in its own right, we are more interested in the principal questions: can we solve it, if so, what can we say about the solutions etc. rather than actually develop algorithms for efficiently solving huge systems.

b. A typical example for linear approximation of a problem is the derivative of a function. Suppose $f\colon I \to \mathbb{R}$ is a differentiable function (where $I \subseteq \mathbb{R}$ is some interval). If f is differentiable at $p \in I$ then for each $X \in \mathbb{R}$ we have $f(p+X) = f(p) + f'(p)X + r(X)$ with $r(X)$ "very small", that is, $r(X)/X$ is small for small nonzero X. Thus, for sufficiently small X, around p, f can be approximated by the linear function $g(q) = f(p) + f'(p)(q-p)$. The situation is analogous for functions in more than one variable.

c. A geometric example of linear approximation is the following: The surface of a sphere can be approximated locally by a plane. Indeed, if we pick a small segment of the surface (small meaning small compared to say the radius of the sphere), then it "looks just like" a small segment of the plane. An everyday example is the very fact that for thousands of years, we believed that Earth is flat. The algebraic/analytic expression of this is that the equation of a two dimensional sphere is $x^2 + y^2 + z^2 = 1$. You will learn in calculus that the derivative of $f(x,y,z) = x^2 + y^2 + z^2$ at the point $p = (p_x, p_y, p_z)$ is $(2p_x, 2p_y, 2p_z)$. So if $f(p_x, p_y, p_z) = 1$, then it turns out that the solutions (X, Y, Z) of

$$2p_xX + 2p_yY + 2p_zZ = 0$$

for "small" X, Y, Z are an approximation of the sphere in the sense that $(p_x + X, p_y + Y, p_z + Z)$ is "close" to a point of the sphere. Indeed, the plane[1] in $\mathbb{R}^2$ defined by the equation $2p_xX + 2p_xY + 2p_zZ = 0$ is the plane perpendicular to (p_x, p_y, p_z) and hence tangent to the sphere at (p_x, p_y, p_z).

2.1 Example. Consider the following system of equations over $\mathbb{R}$:

$$\begin{array}{ccccc} x_1 & + & 2x_2 & +x_4 & = 5 \\ x_1 & +x_2 & +5x_3 & +2x_4 & = 12 \end{array}$$

There are various ways to solve this system. One would be to solve the first equation for x_1: $x_1 = 5 - 2x_2 - x_4$ and than substitute this into the second equation:

$$5 - 2x_2 - x_4 + x_2 + 5x_3 + 2x_4 = 12$$

which of course is equivalent to

$$-x_2 + 5x_3 + x_4 = 7.$$

We obtain $x_2 = 5x_3 + x_4 - 7$. Resubstituting this into the expression for x_1 we get $x_1 = 5 - 10x_3 - 2x_4 + 14 - x_4$ and end up with the two equations

$$\begin{aligned} x_1 &= 19 - 10x_3 - 3x_4 \\ x_2 &= -7 + 5x_3 + x_4 \end{aligned}$$

[1] We will define these terms later. For now you should be content with an intuitive notion.

Thus, for each choice of values for x_3 and x_4 we obtain a unique solution. Thus, this system has infinitely many solutions.

For systems of more than two equations this ad hoc method becomes more than tedious. Also, it is very hard to find general principles that way (eg. to find criteria whether a system has a solution). We will describe below the method using elimination of variables which is essentially the same algorithm but differs in the way we write things down. Computationally both are equally fast (or slow).

The first step in solving a system is a clever method of book keeping.

2.2 Definition. Let $n, m > 0$ be integers. A $m \times n$*-matrix with entries in* $\mathbb{F}$ or simply $m \times n$-matrix if $\mathbb{F}$ is understood, is a rectangular array[2] with m rows and n columns of elements in $\mathbb{F}$. We write

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{21} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

for a matrix with entries $a_{ij} \in \mathbb{F}$ (where a_{ij} denotes the entry in row i and column j). The set of all $m \times n$-matrices with entries in $\mathbb{F}$ will be denoted[3] $M_{m\times n}(\mathbb{F})$. In the special case that $m = n$, we simply write $M_n(\mathbb{F})$ instead of $M_{n\times n}(\mathbb{F})$. A $1 \times n$-matrix is often called a *row vector* and a $n \times 1$-matrix is a *column vector*. We write $\mathbb{F}^n$ for the set of all $n \times 1$-column vectors.

Of course, two matrices are equal if and only if all their entries at corresponding positions are equal.

By abuse of language we often call a matrix "real" (resp. "complex") to indicate that its entries are real (resp. complex) numbers.

Remark. The elements of $\mathbb{F}^n$ are often simply called vectors. Here we made a choice, we could have called the row vectors vectors and denoted the set of all row vectors by $\mathbb{F}^n$. This would have been in line with the usual convention that $\mathbb{F}^n$ is the set of all n-tuples of elements in $\mathbb{F}$, that is $\mathbb{F}^n = \{(a_1, a_2, \dots, a_n) \mid a_i \in \mathbb{F}\}$.

It will become clear why we "prefer" the columns over the rows. However, this adds some notational inconvenience, so we will often write $(a_1, a_2, \dots, a_n)$ also for the column vector

$$(*) \qquad \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

[2] To be abolutely precise, such an array is not a matrix but a presentation of a matrix ["rectangular array" is not a very precisely defined notion; how we write the elements of a set should be irrelevant.]. In set theoretic terms a matrix is a function $f \colon D \to \mathbb{F}$ where $D = \{(i, j) \mid 1 \le i \le m, 1 \le j \le n\}$ (so that the element a_{ij} in the presentation is $f(i, j)$.

[3] Another common notation is $\mathbb{F}^{m\times n}$.

We will always expressedly say so, if we think of an n-tuple as a $1 \times n$ matrix.

Sometimes we write something like "Let $A = [a_{ij}]$ be an $m \times n$-matrix." to simultaneously indicate that A is an $m \times n$ matrix and its entries are denoted by a_{ij}.

There are many "special" matrices.

a. For example, a $m \times n$-matrix $D = [d_{ij}]$ is called *diagonal* if its only nonzero entries lie on the *main diagonal*, i.e. the only nonzero entries are the d_{ij} for which $i = j$.

b. An element of $M_n(\mathbb{F})$ for some n (ie. a matrix where the number of rows equals the number of columns) is called a *square matrix*.

c. A square matrix $A = [a_{ij}]$ is called *symmetric* if $a_{ij} = a_{ji}$. For instance any square diagonal matrix is symmetric, as is the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

d. Some matrices have their own name: the $m \times n$ matrix with all zero entries is called the $(m \times n)$ *zero matrix* and we often denote it by $0_{m,n}$ or simply 0_n (if $m = n$), or even just 0.

e. For reasons that will become clear soon the $n \times n$ diagonal matrix with all diagonal entries equal to 1 is called the $n \times n$*-identity matrix* and denoted I_n or simply I.

$$I_n = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Example. The matrices

$$[1] \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix} \quad \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

all are real diagonal matrices.

Notice that as the individual entries of a matrix are field elements, we can add them componentwise: If A, B are matrices of the *same* dimension, $m \times n$, say, we define their sum to be the $m \times n$ matric C defined as

$$C = A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \dots & \vdots & \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{bmatrix} \tag{2.5}$$

Of course, here the entries of A are the a_{ij} and the entries of B are the b_{ij}.

Observation I: The addition hereby defined is *associative* and *commutative*, that is, for all $m \times n$-matrices A, B, C we have

$$(A+B)+C = A+(B+C) \qquad \text{(associative law)} \tag{2.6}$$

and

$$A+B = B+A \qquad \text{(commutative law).} \tag{2.7}$$

This follows immediately from the definition by the same properties + of the field $\mathbb{F}$. We can now add more than two matrices: $A_1 + A_2 + \cdots + A_n$ is a well defined matrix: we just compute it in any order we like. Thanks to the associative law, the result will always be the same[4].

The addition is the reason why $0_{m,n}$ is special:

Observation II: For $A \in M_{m\times n}(\mathbb{F})$, $A + 0_{m,n} = 0_{m,n} + A = A$. So $0_{m,n}$ "behaves" in $M_{m\times n}(\mathbb{F})$ like the zero element of the field $\mathbb{F}$.

For $A = [a_{ij}] \in M_{m\times n}(\mathbb{F})$, we put $-A = [-a_{ij}]$, that is $-A$ is the $m \times n$ matrix whose entry at position (i, j) is $-a_{ij}$. Then by definition $A + (-A) = (-A) + A = 0_{m\times n}$. So we do have additive inverses as well.

If $A = [a_{ij}] \in M_{m\times n}(\mathbb{F})$ and $c \in \mathbb{F}$ is any element (we usually refer to them as *scalars*) we can define

$$cA = \begin{bmatrix} ca_{11} & ca_{12} & \dots & ca_{1n} \\ ca_{21} & ca_{22} & \dots & ca_{2n} \\ \vdots & \vdots & \dots & \vdots \\ ca_{m1} & ca_{m2} & \dots & ca_{mn} \end{bmatrix}.$$

We refer to this as *scaling* A by c. Notice that $(-1)A = -A$, since if $a \in \mathbb{F}$, $-a = (-1)a$. As usual we write $A - B$ instead of $A + (-B)$.

Observation III: For $A, B \in M_{m\times n}(\mathbb{F})$, and $a, b \in \mathbb{F}$ we have

$$a(A+B) = aA + aB \quad \text{and} \quad (a+b)A = aA + bB \quad \text{and} \quad (ab)A = a(bA).$$

We could now move on and similarly define a multiplication by componentwise multiplying entries. While this is possible, it turns out a much more useful definition is the following: Instead of two $m \times n$ matrices we can only multiply matrices with compatible dimensions as follows: let A be an $\ell \times m$-matrix and B be an $m \times n$-matrix. Then AB is defined as the $\ell \times n$-matrix $C = [c_{ij}]$ with entries

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj} \tag{2.8}$$

[4] See Proposition 3.1.1 and the remark thereafter for a more thorough discussion of this fact

for $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, n$. Notice that on the right hand side the column index of the appearing a_{ik} is the row index of the corresponding b_{kj}. We will motivate this definition below.

2.3 Example.

$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} \tag{2.9}$$

The Formula 2.8 shows that we can view this product as A acting on the columns of B. What we mean by this is that we can multiply A with elements of $\mathbb{F}^m$ and the result will be column vectors in $\mathbb{F}^\ell$. If we write B as $B = [\, B_1\, B_2\, \ldots\, B_n \,]$ with columns $B_i \in \mathbb{F}^m$, then the n columns of AB are exactly given by

$$AB = [\, AB_1\, AB_2\, \ldots\, AB_n \,].$$

> The i-th column of AB is A *times the i-th column of* B.

This can be pushed one step further: the product of a $1 \times m$-row vector with a $m \times 1$-column vector is a 1×1-matrix. We simply identify[5] a 1×1-matrix with its single entry. Then for the product of a matrix A with a vector X we have: the ith entry of AX is the product of the ith row of A with X.

A good way to remember the rules of matrix multiplication is the following:

> The entry of AB at *row* i and *column* j is the product of the ith row of A with the jth column of B.

2.4 Example. (Here all matrices are considered to be real.)

$$\begin{bmatrix} 1 & 3 & 0 & 9 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 2 \\ -2 \end{bmatrix} = 1 \cdot 2 + 3 \cdot 1 + 0 \cdot 2 + 9 \cdot (-2) = -13. \tag{2.10}$$

Also, to keep track of the dimension requirements, we can formally write $(\ell \times m) \cdot (m \times n) = \ell \times n$ (so the "meeting" integers "cancel").

[5] "Identify" means that, even though two objects may not be logically identical, they are sufficiently similar as to be treated as equal. A typical example is the common fact in calculus to identify the real numbers with the constant functions.

Observation IV: The single most important two products one has to compute once are the following: if A is a $m \times n$ matrix then $I_mA = A = AI_n$. Thus, the various identity matrices I_n behave like the multiplicative identities in $\mathbb{F}$.

Combining this with the remarks on matrix multiplication above, we get:

> Let e_i be the i-th column of I_n. Then $A = AI_n$ and hence **The ith column of A is equal to Ae_i.**

This will play an important role later on.

Observation V: If A, B, C are matrices such that $(AB)C$ is defined, then also $A(BC)$ is defined and $(AB)C = A(BC)$ (and similarly vice versa). In this sense, the matrix multiplication is associative.

This actually requires a proof. The assertion that $A(BC)$ if and only if $(AB)C$ is defined follows immediately by comparing dimensions. The associativity claim is a (nasty) straight forward computation which we omit. We will see another much more elegant proof when we talk about linear transformation. □

Because of the associative law, we simply write ABC instead of $A(BC)$, or $(AB)C$. In fact, as in the case of the addition, arbitrary products now make sense without any parentheses. Also, for a positive integer n we define A^n in the same way as we defined it for elements of a field: $A^n = \underbrace{A \cdot A \cdots A}_{n \text{ times}}$.

Our next observation connects addition and multiplication by means of the *distributive laws*:

Observation VI: Suppose A, B, C are matrices such that B and C have the same dimensions and AB (and therefore also AC and $A(B + C)$) is defined. Then $A(B + C) = AB + AC$. Similarly, $(E + F)G = EG + FG$ whenever E, F, G are matrices such that this makes sense.

Let A be an $m \times n$ and B, C be an $n \times p$-matrices. Then AB, AC, and $A(B+C)$ are $m \times p$ matrices.

This is easy to verify directly. We do as an example the case $A(B + C) = AB + AC$. Let $M = A(B + C)$ with entries m_{ij} then

$$m_{ij} = \sum_{k=1}^{n} a_{ik}(b_{kj} + c_{kj}) = \sum_{k=1}^{n} a_{ik}b_{kj} + \sum_{k=1}^{n} a_{ik}c_{kj}.$$

The right hand side is the entry at position (i, j) of $AB + AC$. □

Observation VII: We can always pull out scalars: If A, B are matrices such that AB is defined, and if $a \in \mathbb{F}$ is any scalar, then $A(aB) = (aA)B = a(AB)$.

It is important to note that the commutative law does not hold for the matrix multiplication. For instance, if AB is defined, there is no reason to assume that also BA is defined. Indeed, if AB and BA are both defined, then necessarily the number of columns of A and the number of rows of B are the same and vice versa. However, if AB and BA are defined and *equal*, this forces A and B to be square matrices of the same dimension. But even in that case "most" pairs of $n \times n$ matrices do not commute. For instance

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ whereas } \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.11}$$

Remark. Observations I through V gain particular weight in the case of square matrices: in this case all restrictions on dimensions are void: if A, B are $n \times n$ matrices then both AB and BA are defined (and are again $n \times n$ matrices). Thus, $M_{n\times n}(\mathbb{F})$ has two operations, an addition and a multiplication. These two operations are associative and satisfy Observations I, II, VI. Such a mathematical structure is called a *ring*. Moreover, the multiplication has an identity element (namely, I_n by Observation V). Such rings are called *rings with identity* or *unital rings*. To mention another mathematical structure Observation VII (together with the other observations) guarantees that $M_n(\mathbb{F})$ is what is called an $\mathbb{F}$-*algebra*.

For the moment however, we are happy to know that $M_n(\mathbb{F})$ is a ring with identity and we may forget it for the time being. But before that, quickly note that a ring is nothing but something that satisfies the axioms of a field with the exceptions that there need not be a multiplicative identity (and by extension, no multiplicative inverses). Also, the multiplication need not be commutative.

Examples.

a. Becoming forgetful, Gandalf has the recipes of his magical potions written down in a matrix. Each row represents the amount of different ingredients needed for a particular type of brew. For instance, assuming 7 ingredients (all top secret of course), he writes $[\, a_{i1}\, a_{i2}\, \ldots\, a_{i7}\,]$ into the row corresponding to the ith potion where a_{ij} is the amount of ingredient j needed to produce that particular potion.

   Assuming three different recipes, the result will be a 3×7 matrix A. This matrix can serve several purposes: for instance if Gandalf wants to figure out how much each potions costs him, he simply multiplies A with the vector

$$P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_7 \end{bmatrix}$$

   where p_i is the number of gold pieces for each unit of ingeredient i. The result AP will be a vector whose ith row will be the price per potion of each given type.

   On the other hand, he might want to figure out how many ingredients of each type are needed to make n_i potions of type i. For this, he simply computes

$$[\, n_1\, n_2\, n_3]A$$

which will result in a row vector with 7 entries $a_1, a_2, \ldots, a_7$. The entry a_i is the number of units of ingredient i.

b. Matrix multiplication is very useful in describing certain dynamical systems: For instance, suppose we have a system whose state is determined by two real numbers say: A very simple model to describe the dynamics of a bunny populations is as follows: The state of the population in year n of our study is determined by the number A_n of adult bunnies and the number C_n of child bunnies. Between one year and the next, the following happens: some proportion d of adult bunnies will die. Also, coyotes will reduce the total population by a factor c (we assume here that the coyotes like young bunnies as well as old ones). The surviving young bunnies will become adults. Finally, these are bunnies after all, so there will be new bunnies born, at a rate proportional to the adult population. We obtain the following recurrence relation:

$$\begin{aligned} A_{n+1} &= A_n - dA_n - cA_n + (1-c)C_n \\ C_{n+1} &= bA_n \end{aligned}$$

We can rewrite this as follows:

$$\begin{bmatrix} A_{n+1} \\ C_{n+1} \end{bmatrix} = P \begin{bmatrix} A_n \\ C_n \end{bmatrix} \tag{2.12}$$

where

$$P = \begin{bmatrix} 1-d-c & 1-c \\ b & 0 \end{bmatrix} \tag{2.13}$$

We we start with an initial population of

$$X_0 = \begin{bmatrix} A_0 \\ C_0 \end{bmatrix}$$

then after n years the bunny population will have developed to

$$X_n = P^n X_0.$$

We find that understanding this process for various initial conditions is more or less the same as understanding the behaviour of P^n as n grows.

Observation IV asserts that I_n has the same role on $M_n(\mathbb{F})$ as does 1 in $\mathbb{F}$ itself: it is an identity element for the matrix multiplication. It may therefore make sense to consider those matrices in $M_n(\mathbb{F})$ that have an *inverse*: Analogous to the equation $ab = 1$ in $\mathbb{F}$ we could look at equations of the form $AB = I_n$. One additional complication is here introduced due to the fact that the multiplication is not commutative, so we need to consider both products AB and BA.

2.5 Definition. Let A be an $n \times n$-matrix. We say A is *invertible* if there is an $n \times n$ matrix B such that $AB = BA = I_n$.

If we say A is invertible we usually implicitly mean that A is also a square matrix (otherwise invertible makes no sense).

2.6 Example.

a. The following are examples of invertible matrices: I_n, cI_n where $c \in \mathbb{F}$ is nonzero.

b. The matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

is invertible. Indeed, $AB = BA = I_2$ for

$$B = \begin{bmatrix} \frac{-1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{-1}{3} \end{bmatrix}$$

c. Not every matrix is invertible. For instance

$$N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

is not invertible. For various reasons. One is that $N^2 = 0$. Hence, if there was B such that $NB = BN = I$ then $(NB)^2 = NBNB = I$. But also $NBNB = BN^2B = B0B = 0$. Since obviously $0 \neq I_n$, no such B exists.

Right now we haven't yet a good method of a) determining whether a given square matrix is invertible, and b), if so, how to compute its inverse. Both questions will be addressed shortly.

Returning to our system of equations (2.4) the astute reader may have noticed that the *coefficients* of the system, the a_{ij} form an $m \times n$ matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{n1} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Collecting the b_i into a column vector B we see that the system (2.4) is equivalent to the *matrix equation*

$$AX = B \tag{2.14}$$

where we look for elements $X \in \mathbb{F}^n$ making $AX = B$ a correct statement. A is called the *coefficient matrix* of the system (2.4). Together with B it contains all the information about the original system. This language emphasizes the viewpoint that the multiplication by a matrix A transforms the vector X into a new vector $Y = AX$. Allowing different vectors X we can look at this as a transformation from the set of all $n \times 1$-column vectors to the set of all $m \times 1$-column vectors much like we study functions $y = f(x)$ in calculus. We will look at this transformation later in great detail.

An important linear combination. Note that we can express a matrix equation in terms of the columns of A: Let $A = [\,A_1\, A_2\, \ldots\, A_n\,]$ where $A_i \in \mathbb{F}^m$. Then $AX = B$ is the same statement as

$$x_1A_1 + x_2A_2 + \cdots + x_nA_n = B. \tag{2.15}$$

where the x_i are the entries of X. If B is equal to the sum on the lefthand side, we also say B is a *linear combination* of the A_i.

Thus, solving the matrix equation $AX = B$ is equivalent to expressing B as a linear combination of the columns of A.

2.7 Example. We could ask for which $B \in \mathbb{R}^2 = M_{2\times 1}(\mathbb{R})$, the equation $AX = B$ has a solution where

$$A = \begin{bmatrix} 2 & 1 & 7 \\ 1 & 3 & 4 \end{bmatrix}$$

Now

$$AX = A\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 + 7x_3 \\ x_1 + 3x_2 + 4x_3 \end{bmatrix} = x_1\begin{bmatrix} 2 \\ 1 \end{bmatrix} + x_2\begin{bmatrix} 1 \\ 3 \end{bmatrix} + x_3\begin{bmatrix} 7 \\ 4 \end{bmatrix}$$

Consider the first two columns A_1, A_2 of A: we can view them as points in the plane $\mathbb{R}^2$. Let L_1 be the line through the origin $(0,0)$ containing the first one, $A_1 = (2,1)$, and L_2 the one through the second one, $A_2 = (1,3)$ (we write them as tuples to save space). Notice that in matrix language, L_1 is the set of all $Y \in M_{2\times 1}(\mathbb{R})$ such that $Y = cA_1$ for some real number c. It is clear that L_1 and L_2 are not parallel or equal[6]: $L_1 \cap L_2 = (0,0)$. Now let $B \in \mathbb{R}^2$ be an arbitrary point. Euclidean geometry asserts that there is a unique line L_2' through B parallel to L_2. Since L_1 and L_2 are not parallel, neither are L_1 and L_2'. Hence there is a unique point P in the intersection $L_1 \cap L_2'$ (cf. the beginning of the chapter). Then $P = x_1A_1$ for some (unique!) $x_1 \in \mathbb{R}$. The same reasoning suggests that the unique parallel L_1' to L_1 through B intersects L_2 at a point $Q = x_2A_2$. It is very useful to convince yourself that then $B = x_1A_1 + x_2A_2$ (indeed, the points $(0,0), P, B, Q$ are the vertices of a parallelogram). Thus, $AX = B$ has a solution for all $B \in \mathbb{R}^2$: x_1, x_2 are obtained as outlined above, and we put $x_3 = 0$. Then $B = x_1A_1 + x_2A_2 + x_3A_3$. Notice that in this example, we could have chosen A_2, A_3, or A_1, A_3 instead of A_1, A_2 as none of the columns of A are on the same line through the origin.

Remark. Note that strictly speaking the reasoning in the previous Example does not constititute a proof that $AX = B$ always has a solution, unless we verified that the the usual rules of Euclidean geometry actually apply to lineas and points in $\mathbb{R}^2$ as defined by us. This is good exercise. To be precise, you should convince yourself of the following:

If we define a *point* as an element of $\mathbb{R}^2$ and a *line* as a subset $L \subseteq \mathbb{R}^2$ that has the form $L = \{u + tv \mid t \in \mathbb{R}\}$. Then the following holds:

[6] Prove this algebraically!

Figure 2.1.: Graphic solution of $AX = B$ in Example 2.7

a. If P, Q are distinct points, there is a unique line, denoted PQ, containing P and Q.

b. If L is a line and P a point, then there is a unique line L' containing P such that $L = L'$ or $L \cap L' = \emptyset$.

c. If L, L' are lines then $L \cap L'$ is either empty, a single point, or equal to both L and L'.

Together with these facts, the reasoning above is rigorous and has nothing to do with any picture we draw. For instance "parallelogram" just means a set of our distinct points A, B, C, D such that the lines (unique by a.) AB and CD (as well as AD and BC) are parallel (ie. don't intersect or are equal).

2.3. Row operations

From now on we will discuss mostly systems of the form $AX = B$. Our goal is to "simplify" A and B, resulting in a matrix A' and a vector B', in such a way that

$$A'X = B'$$

has the same solutions as the original system, but whis is simple to solve. This process is known as *Gaussian elimination*. We will write $[A \mid B]$ for the matrix obtained from A by appending B as a column. This larger matrix is usually called the *augmented matrix*.

When defining matrix multiplication (and in Example 2.7) it turned out to be useful to think of a matrix as an ordered collection of column vectors. Of course, one can equally well think of a matrix as a list of row vectors. This is what we will do now: we will define three types[7] of so called *elementary row operations* on any matrix which when applied to $[A \mid B]$ will turn out to leave the solutions unaffected.

Row opererations.

Type I: Exchanging two rows.

Type II: Adding to a row a multiple of another row.

Type III: Scaling a row by a nonzero scalar $c \in \mathbb{F}$.

[7] Note that the order of operations varies depending on the author; there is no universally recognized "Type I" operation.

(Here addition of rows means addition of the corresponding row vectors. Similarly, scaling a row by c refers to multiplying each entry of the row by c.)

It should be clear that Type I and Type III operations applied to $[A \mid B]$ won't change the solutions: the first corresponds to simply swapping two equations in (2.4), and the third to multiplying all coefficients (including the corresponding b_i) in one equation by c. Also a Type II operation will not affect the solutions; we will prove it below.

The goal is to use row operations to obtain a simpler augmented matrix, for which we can read off the solutions immediately. Before we actually discuss a very convenient form to bring a matrix into by means of Type I-III operations, we will first realize each of the operations as the multiplication by a matrix from the left.

For example, the formulae for matrix multiplication show that applying a Type III operation to an $m \times n$ matrix A is the same as multiplying A with the $m \times m$ matrix matrix

$$E = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \tag{2.16}$$

where c is the ith entry on the diagonal. We follow our convention that any matrix entry that is not explicitly written down is equal to 0. Thus, if we want to scale the third row in a $4 \times m$ matrix A by c we have to compute

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & c & \\ & & & 1 \end{bmatrix} A.$$

For the other two types we can find matrices as well. For example, if we want to exchange rows $i < j$, we multiply A with the $m \times m$-matrix

$$E = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & 1 & & & \\ & & & & I & & & & \\ & & & 1 & & 0 & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{bmatrix} \tag{2.17}$$

Here the central I is meant to represent just diagonal 1's. Also, the two 0s on the diagonal are precisely at positions (i,i) and (j,j), and the two off diagonal 1s are at (i,j) and (j,i).

Finally, a Type III operations corresponds to a matrix of the following form: if we want to add k times row j to row i (where $i<j$), then the result is EA where E is the $m\times m$ matrix

$$E=\begin{bmatrix}1 & & & \\ & \ddots & k & \\ & & \ddots & \\ & & & 1\end{bmatrix} \tag{2.18}$$

where the solitary k is at position (i,j). The matrix looks similar, if $i>j$ but then k is below the diagonal.

> The matrices of the form (2.17), (2.16), (2.18) are called *elementary matrices*. The *type* of an elementary matrix is the type of the associated row operation.

2.8 Examples. Let

$$A=\begin{bmatrix}3 & 2 & 3 & -2\\ 1 & 1 & 1 & 0\\ 1 & 2 & 1 & -1\end{bmatrix}$$

and we now apply a row operation of each type to A.

a. Subtracting 3 times row 2 from row 1:

$$\begin{bmatrix}1 & -3 & \\ & 1 & \\ & & 1\end{bmatrix}A=\begin{bmatrix}0 & -10 & -2 & \\ 1 & 1 & 1 & 0\\ 1 & 2 & 1 & -1\end{bmatrix}$$

b. Exchanging rows 1 and 2:

$$\begin{bmatrix}0 & 1 & \\ 1 & 0 & \\ & & 1\end{bmatrix}A=\begin{bmatrix}1 & 1 & 1 & 0\\ 3 & 2 & 3 & -2\\ 1 & 2 & 1 & -1\end{bmatrix}$$

c. Scaling the first row by $\frac{1}{3}$:

$$\begin{bmatrix}\frac{1}{3} & & \\ & 1 & \\ & & 1\end{bmatrix}A=\begin{bmatrix}1 & \frac{2}{3} & 1 & \frac{-2}{3}\\ 1 & 1 & 1 & 0\\ 1 & 2 & 1 & -1\end{bmatrix}$$

2.3.1 Problem. Show that row operations can indeed be expressed by left multiplication with elementary matrices.

2.3.2 Problem. We define two $m \times n$ matrices A and B as *row equivalent*, written $A \sim B$, if B is obtained from A by a sequence of elementary row operations. Show that $\sim$ is an *equivalence relation*, that is, show that

a. $\sim$ is *reflexive*: $A \sim A$ for all matrices $A \in M_{m \times n}(\mathbb{F})$.

b. $\sim$ is *symmetric*: $A \sim B$ if and only if $B \sim A$.

c. $\sim$ is *transitive*: If $A \sim B$ and $B \sim C$, then also $A \sim C$.

(*Caution*: Often $A \sim B$ is also used to indicate that the matrices A and B are *similar*, which is an entirely different concept.)

Sometimes it is more convenient to write these matrices in a different way. For this we introduce the *matrix units*: for a given integer n, let e_{ij} be the $n \times n$-matrix with a single nonzero entry, equal to 1, at position i, j. So if $n = 3$, then e_{13} is

$$e_{13} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Using $m \times m$-matrix units we can express the three different matrices above in a very compact form: The elementary matrix E corresponding to switching rows i and j can be written as

$$E = I - e_{ii} - e_{jj} + e_{ij} + e_{ji}.$$

The Type II matrix (2.18) can be written as

$$E = I + ke_{ij}.$$

and finally, the Type III matrix (2.16) is equal to

$$E = I + (c - 1)e_{ii}.$$

As we will see soon, most of the time, it is completely irrelevant, how the actual form of an elementary matrix is. What will mostly be important is the fact that they exist.

2.3.3 Problem. Let δ_{ij} be the KRONECKER delta, that is

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

Prove that if e_{ij} and $e_{k\ell}$ are matrix units (of the same size) then

$$e_{ij}e_{k\ell} = \delta_{jk}e_{i\ell}.$$

Any elementary row operation of any type is reversible, and the inverse operation is again of the same type. It follows easily:

2.9 Lemma. *Elementary matrices are invertible.*

Proof. Let E be an $n \times n$ elementary matrix of any type. Let E' be the elementary matrix corresponding to the reverse procedure: that is, if E corresponds to

a. exchanging rows i and j, then $E' = E$;

b. adding k times row j to row i, then E' corresponds to adding $-k$ times row j to row i.

c. scaling row i by $c \neq 0$, then E' corresponds to scaling row i by c^{-1}.

By construction, then, we have for each $n \times n$ matrix A that $E(E'A) = A = E'(EA)$. This applies in particular to the case $A = I = I_n$ and we conclude that $EE' = E'E = I$. □

Notice that if we apply several row operations to a matrix A, then the result is obtained as

$$A' = E_p E_{p-1} \cdots E_1 A$$

where E_i is the elementary matrix corresponding to the ith operation.

(Here we use the associative law: indeed, a priori A' is equal to $E_p(E_{p-1}(\cdots(E_1 A))\cdots)$. However, we can leave the parentheses away.)

Hence $A' = PA$ where $P = E_p E_{p-1} \cdots E_1$ is a product of elementary matrices. Thus the result of applying a sequence of row operations to a matrixcan A can be expressed as multiplying A with a matrix P from the left.

2.10 Proposition. *Let $[A \mid B]$ be the augmented matrix of the system of equations $AX = B$. Suppose $[A' \mid B']$ is obtained from $[A \mid B]$ by elementary row operations. Then the solutions of $AX = B$ and $A'X = B'$ are the same.*

Proof. Let $M = [A \mid B]$ and $M' = [A' \mid B']$. Then $M' = PM$ where P is the product of elementary matrices corresponding to the operations that transform M into M'. If $P = E_p E_{p-1} \cdots E_1$, then we may put $Q = E_1' E_2' \cdots E_p'$ where E_i' is defined as in the proof of Lemma 2.9 as the matrix for the reverse procedure corresponding to E_i. Then it is easy to verify that $QP = PQ = I$. And so $M = IM = QPM = QM'$.

Let X be a solution of $AX = B$. Then also $P(AX) = PB$, and so $(PA)X = PB$. But $PA = A'$ and $PB = B'$ and so $A'X = B'$, making X a solution of $A'X = B'$ as well. Similarly, if X is a solution of $A'X = B'$, then X is also a solution of $QA'X = QB'$ and hence of $AX = B$ because $A = QA'$ and $B = QB'$. □

Let us see how this helps.

Consider the equation

$$AX = B$$

where

$$A = \begin{bmatrix} 3 & 2 & 3 & -2 \\ 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & -1 \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

The augmented matrix is then

$$M = \begin{bmatrix} 3 & 2 & 3 & -2 & 1 \\ 1 & 1 & 1 & 0 & 3 \\ 1 & 2 & 1 & -1 & 2 \end{bmatrix} \tag{2.19}$$

(Even though we don't always separate the last column it is important to remember that it corresponds to B here.)

We first erase all nonzero entries below the top position in the first column. For convenience, we first arrange for a 1 in the top left spot (to avoid fractions).

$$M \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 3 & 2 & 3 & -2 & 1 \\ 1 & 2 & 1 & -1 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -2 & -8 \\ 0 & 1 & 0 & -1 & -1 \end{bmatrix} \tag{2.20}$$

We multiply the second row by -1 and then erase the leading 1 in the last row.

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & -1 & 0 & -2 & -8 \\ 0 & 1 & 0 & -1 & -1 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 & 8 \\ 0 & 1 & 0 & -1 & -1 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 & 8 \\ 0 & 0 & 0 & -3 & -9 \end{bmatrix} \tag{2.21}$$

Finally, we divide the last row by -3 and then use the first nonzer entry in each row (working backwards) to erase all the entries above it.

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 & 8 \\ 0 & 0 & 0 & -3 & -9 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 2 & 8 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \tag{2.22}$$

So here

$$[A' \mid B'] = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \tag{2.23}$$

The system $A'X = B'$ is easily solved: indeed it corresponds to the three equations

$$\begin{aligned} x_1 + x_3 &= 1 \\ x_2 &= 2 \\ x_4 &= 3 \end{aligned}$$

The only choice here is therefore x_3 (or x_1; but not both). We find that a general solution X looks like

$$X = \begin{bmatrix} 1-t \\ 2 \\ t \\ 3 \end{bmatrix}$$

where t is some arbitrary real number. So the set S of solutions can be written as

$$S = \left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \\ 3 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \middle| t \in \mathbb{R} \right\}$$

The reason why we can solve this system is that each equation expresses one variable in terms of other variables which are not restricted in any way.

We will now describe the general result. A nonzero row is one where at least one entry is nonzero (in line with our convention to denote the elements in $M_{1,n}(\mathbb{F})$ of all zeros by 0). If a row is nonzero, its "first" – that is, leftmost – nonzero entry is called the *leading entry*.

Row Echelon Form. An $m \times n$ matrix A is in row echelon form if it satisfies the following properties:

a. Every nonzero row is above any zero row.

b. Any leading entry is always (strictly) to the right of any leading entry of a row above.

If in addition

c. Every leading entry is 1.

d. The entries (in the same column) above any leading entry are all zero.

the matrix A is said to be in *reduced* row echelon form. A leading entry in a row echelon form is sometimes called *pivot*, and a column containing a pivot is also called *pivot column*.

A typical example is the the $n \times n$ identity matrix I_n. Here is another typical example:

$$\begin{bmatrix}
0 & \dots & 0 & 1 & * & \dots & * & 0 & * & \dots & * & 0 & * & \dots \\
\vdots & \dots & \dots & 0 & 0 & \dots & 0 & 1 & * & \dots & * & 0 & * & \dots \\
\vdots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 0 & \dots & 0 & 1 & * & \dots \\
\vdots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \dots & \dots \\
\vdots & & & & & & & & & & & & & \\
0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & \dots
\end{bmatrix}$$

where $*$ denotes an arbitrary (possibly zero) entry.

It should be clear from the example above that every matrix can be transformed into reduced row echelon form. The procedure used is usually called Gaussian elimination. To be absolutely precise, let us prove that it works.

2.11 Theorem (Gaussian elimination)**.** *Let A be an $m \times n$ matrix with entries in a field $\mathbb{F}$. Then there are finitely many row operations that transform A into reduced row echelon form.*

Proof. We describe an algorithm that produces the row echelon form in finitely many steps. To make our live simple, we will use induction on the number m of rows of A.

Base case: If A has exactly one row, then A is almost in reduced row echelon form: the only possible problem is that its leading entry is not 1; but after scaling by a nonzero scalar (the inverse of the leading entry) we have a leading entry of 1 and hence a reduced row echelon form.

Induction step: So suppose $m > 0$ is an integer and suppose the theorem holds for all matrices with m rows. Let A be a matrix with $m+1$ rows. If A is the zero matrix, it is in reduced row echelon form. Otherwise, A has a left most nonzero entry (ie. a nonzero entry whose column index is minimal among the column indices of all other nonzero entries). After a row exchange if necessary, we may assume this entry is in the first row. To the east and southeast of our entry, all entries are zero. Using row operations of Type II we can erase all entries below our entry. By a Type III operation, we can make our entry equal to 1. Our matrix has now the form

$$(*) \qquad \left[\begin{array}{cccc|c} 0 & \dots & 0 & 1 & * \\ 0 & \dots & 0 & 0 & \\ \vdots & \dots & \vdots & \vdots & B \\ 0 & \dots & 0 & 0 & \end{array}\right]$$

Where B is a smaller matrix with m rows. (B is only present if our entry is not in the last column.) By the induction assumption, we can reduce B into reduced row echelon form. Applying the same row operations to our matrix, we obtain a matrix as in (*) where B is in reduced row echelon form. Using the pivots in B and row operations of Type II, we then erase all the entries in the first row above the pivots. The result is the reduced row echelon form. □

Remark. Note that the algorithm sketched in the proof of the previous theorem does erase the entries *above* the pivots only in the end. That way, considerable time is saved: if we erased all nonzero entries in a pivot column, we do too much arithmetic in the pivot columns to the right, as there are still nonzero entries there.

To be precise, the algorithm works as follows:

Pick a nonzero entry in the left most column that contains a nonzero entry. Move it to the top position by a row exchange. Erase all nonzero entries below it by Type II operations.

Next, move to the left most column that has a nonzero entry below the first row. Apply the same procedure (ignoring the first row, howver).

Repeat until the matrix is in echelon form. Now, working backwards, erase all entries above the pivots.

Why are matrices in row echelon form helpful? The matrix in (2.23), for instance, is in reduced row echelon form; we have seen above how this helped when solving the corresponding system.

The general principle is the following: Suppose R is an $m \times (n+1)$ matrix in reduced row echelon form.

If we interpret R as the augmented matrix of a system of linear equations, that is if $R = [A \mid B]$ then we can immediately read of the solutions of $AX = B$.

- First of all, if the last column of R is a pivot column, then the system is *inconsistent*, ie. has no solution: indeed, the last nonzero row of R corresponds to an equation $0x_n = 1$, which cannot be satisfied no matter what x_n.

- If on the other hand, the last column of R is *not* a pivot column, then to obtain a solution, we may choose an arbitrary value for each variable corresponding to one of the first n columns which is not a pivot column. Substituting these in the equations we obtain a unique value for each variable corresponding to a pivot column. For this reason, the variables corresponding to non-pivot columns are called *free variables* the others are sometimes called *basic variables*.

  Sometimes there are no free variables, which means that the solution is uniquely determined: in this case all variables are basic variables, and their values are uniquely determined (if a solution exists).

Thus, solving equations for matrices in reduced row echelon form is a simple exercise.

2.12 Example. Suppose the last row in our echelon form in Example (2.23) had been $[\,0\,0\,0\,0\,1\,]$ So that

$$[A' \mid B'] = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then the system has clearly no solution, because the last row corresponds to $0x_4 = 1$.

On the other hand, had the last row been a row of all zeros, such that

$$[A' \mid B'] = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

we would have gained an additional free parameter (x_4 in this case). No matter what happened, the solution would never be unique since always at least one column corresponding to a variable is not a pivot column, since there are four variables but at most three pivots.

To summarize the above discussion let us note:

2.13 Theorem. *Let $A \in M_{m\times n}(\mathbb{F})$ be a matrix and $B \in \mathbb{F}^m$. Let $U = [A' \mid B']$ be a reduced row echelon form for $[A \mid B]$.*

Then $AX = B$ has a solution if and only if the last column of U does not contain a leading entry of some row. If a solution exists, it is unique, if and only if every column of A' is a pivot column of U. Otherwise, and one can prescribe an arbitrary value for each of the free variables, and the solution is not unique. In particular, if $m < n$ it is never unique.

Proof. This follows immediately form the fact that a) the solutions of $AX = B$ are the same as the solutions of $A'X = B'$, b) our discussion of equations corresponding to a reduced row echelon form matrix, and c), that if $m < n$ there are always free variables (thus, if a solution exists in this case, then we can choose a second one by assigning one of the free variables a different value[8]). □

The discussion above applies in particular to *homogeneous equations*

$$AX = 0.$$

Of course, this equation has always the *trivial* solution $X = 0 = 0_{n,1}$.

2.14 Corollary. *Let A be an $m \times n$ matrix with $m < n$. Then the homogeneous equation*

$$AX = 0$$

has a non-trivial solution.

Proof. This is exactly Theorem 2.13 applied in the base $B = 0$: since we do have a solution and since $m < n$ not every column is a pivot column and we can prescribe any of the free variables arbitrarily. In particular, we can pick a nonzero value for some of them. □

The best way to understand this algorithm and its conclusions is to apply it.

We conclude this discussion with the remark that all we said applies to *column operations* in complete analogy. This should be clear; however, here is a "quick and dirty" proof:

2.15 Definition. Let $A = [a_{ij}]$ be an $m \times n$ matrix. Its *transpose*, denoted A^T, is the $n \times m$ matrix with entry a_{ji} at position (i, j).

2.16 Example.

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

The main thing to remember about the transpose is that the rows of A are the columns of A^T and vice versa.

[8] Notice however, that this is a close call: if $\mathbb{F} = \mathbb{F}_2$, we may end up with exactly two solutions.

2.3.4 Problem. Prove the following facts of "transpose arithmetic": Let A, B be matrices such that the left hand sides below make sense (so they have the same size in the first, and compatible dimensions in the second part):

a. $(A+B)^T = A^T + B^T$.

b. For any $c \in \mathbb{F}$ we have $(cA)^T = cA^T$.

c. $(AB)^T = B^T A^T$.

d. $(A^T)^T = A$.

Now let A be an $m \times n$ matrix and let E be an $n \times n$ elementary matrix. Then

$$AE = ((AE)^T)^T = (E^T A^T)^T.$$

For each elementary matrix E, E^T is again an elementary matrix of the same type. The operation of E^T on the rows of A^T translates into an operation of E on the columns of A. From this equation it is immediate that

- if E corresponds to a Type I operation switching rows i and j, then AE is obtained from A by switching columns i and j;
- if E is a Type II operation, adding k times row j to row i, say, then AE is obtained from A by adding k times column i to column j (**note that the roles of i and j are exchanged here**);
- if E is a Type III operation, scaling row i by $c \in \mathbb{F}$, say, then AE corresponds to scaling column i by c.

Example. Adding the k times the first column to the third column of a 3×3 matrix is obtained as

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & \ell \end{bmatrix} \begin{bmatrix} 1 & & k \\ & 1 & \\ & & 1 \end{bmatrix} = \begin{bmatrix} a & b & ka+c \\ d & e & kd+f \\ g & h & kg+\ell \end{bmatrix}$$

We could now discuss a column echelon form, but that would not yield any new insights (the result would be just the transpose of a row echelon form). We will discuss later what happens if we perform row and column operations simultaneously to a matrix A.

2.4. Invertible matrices

The main point in the last section was that if we have a matrix A, then we can apply a sequence of row operations and obtain a matrix A' in reduced row echelon form. We will now show how we can use this to characterize invertible matrices.

NB that a priori to check that a matrix A is invertible one has to test both products AB and BA. It is conceivable that $AB = I$ but BA isn't. We will see later however that *for matrices*[9] this is not the case.

Recall that we have shown that for a nonzero scalar $a \in \mathbb{F}$, its inverse a^{-1} is uniquely determined by the fact that $aa^{-1} = 1$. If we copy these arguments we find:

2.17 Proposition-Definition. *Let A be an $n \times n$ matrix. If A is invertible, the matrix B such that $AB = BA = I_n$ is uniquely determined. It is called the* inverse *of A and denoted by A^{-1}.*

In fact, if $AB = I$ or *$BA = I$ it readily follows that $B = A^{-1}$.*

Proof. We copy the arguments of the proof of Lemma 2.19.

Suppose A is invertible and let $AB = BA = I$. Suppose B' is another matrix such that $AB' = I$.

Then in particular $I = AB = AB'$ and hence $B = B(AB) = B(AB')$ from which we deduce that $B = (BA)B' = IB' = B'$.

The case $B'A = I$ follows similarly (by right multiplying with B). □

2.4.1 Problem. If A is invertible show that A^T and A^{-1} are invertible and that $(A^{-1})^{-1} = A$ and $(A^T)^{-1} = (A^{-1})^T$.

2.18 Remark. One important fact about invertible matrices is the following: if A and B are invertible (of the same size) then so is AB and its inverse is $B^{-1}A^{-1}$ and by an easy induction it follows that any product $A_1A_2 \cdots A_n$ of invertible matrices of the same size is invertible. We have seen an example of this already in the proof of Lemma 2.9 where we showed that a certain product P of elementary matrices was invertible.

Invertible matrices have a very nice property. As in Lemma 1.4, we can solve all equations and we can solve them uniquely!

2.19 Lemma. *Let A be an invertible $n \times n$ matrix. Then for each B in $\mathbb{F}^n$ the equation*

$$AX = B$$

has the unique solution $A^{-1}B$.

Proof. The proof is essentially a copy of the proof of Lemma 1.4: First we show that $A^{-1}B$ is a solution: indeed, $A(A^{-1}B) = (AA^{-1})B = IB = B$.

To see uniqueness, suppose $AX = B$. Then $A^{-1}(AX) = A^{-1}B$. But $A^{-1}(AX) = (A^{-1}A)X = IX = X$, showing that $X = A^{-1}B$. □

Remark. In the proof of the lemma it is instructive to observe that the *existence* of a solution used the fact that $AA^{-1} = I_n$ whereas the *uniqueness* required $A^{-1}A = I_n$.

[9] To be absolutely precise, for finite matrices this is not the case. There are example of infinitely sized matrices where this fails.

Instead of using Gaussian elimination to solve an equation $AX = B$ where A is invertible, we might be tempted to use A^{-1} instead. However, as it turns out, to out compute A^{-1} we need to perform Gaussian elimination as well – in fact for n matrix equations of the form $AX = B_i$ $(i = 1, 2, \dots, n)$.

Summarizing all we know about matrices we obtain the following result:

2.20 Theorem. *Let A be an $n \times n$-matrix over $\mathbb{F}$. Then the following statements are equivalent:*

a. *A is invertible.*

b. *The homogeneous equation $AX = 0$ has only the trivial solution $X = 0$.*

c. *A can be reduced to the identity matrix by elementary row operations.*

d. *A is a product of elementary matrices.*

Proof. We will prove the following: $a. \implies b. \implies c. \implies d. \implies a.$

Suppose a. holds. We have to show that $X = 0$ is the only solution of $AX = 0$. But since $X = 0$ is a solution and by Lemma 2.19 any solution is unique, $X = 0$ is the only one.

Suppose $AX = 0$ has only the solution $X = 0$. Let $[A' \mid B']$ be a reduced echelon form for $[A \mid 0]$. Notice that also $B' = 0$. Since the solution $X = 0$ is unique, every column of A' in $[A' \mid 0]$ must be a pivot column. Since A' is a square matrix this means that $A' = I_n$. The row operations that turn $[A \mid 0]$ into $[I_n \mid 0]$ obviously turn A into I_n proving c.

If c. holds, d. is merely a reformulation: there are elementary matrices $E_1, E_2, \dots, E_p$ such that $(E_1 E_2 \cdots E_p)A = I_n$. Each of the E_i is invertible, so $A = E_p^{-1} E_{p-1}^{-1} \cdots E_1^{-1} I_n = E_p^{-1} E_{p-1}^{-1} \cdots E_1^{-1}$. Now E_i^{-1} is again an elementary matrix so d. follows.

Finally, suppose $A = F_1 F_2 \cdots F_p$ is a product of elementary matrices F_i. Then A is invertible with inverse $A^{-1} = F_p^{-1} F_2^{-1} \cdots F_1^{-1}$ proving a.. □

2.21 Corollary. *An $n \times n$ matrix where a row or column is identically zero is not invertible.*

Proof. This is a good exercise and is left to the reader. □

2.22 Corollary. *Let A, B be $n \times n$ matrices. If $AB = I$ or $BA = I$ then it readily follows that A is invertible and $B = A^{-1}$.*

Proof. First suppose $BA = I$. According to Theorem 2.20 it is enough to show that $AX = 0$ has only the trivial solution. So suppose $AX = 0$. Then $B(AX) = B0 = 0$ and hence $(BA)X = 0$. But this means $IX = X = 0$. Hence A is invertible.

If $AB = I$, then the same reasoning shows that B is invertible. By Lemma 2.19 it follows that $A = B^{-1}$ and so $AB = BA = I$. But then A is invertible as well. □

2.4.2 Problem. Show: An $n \times n$ upper triangular matrix is invertible if and only if all entries on the diagonal are nonzero.

2.4.3 Problem. In the following let A be an $n \times n$ matrix. For each of the following cases decide whether it is possible in principle that A is invertible.

a. $A = 0$.

b. A row of A is identically zero.

c. Two rows of A are identical.

d. Two columns of A are identical.

e. All diagonal entries of A are zero.

f. $\operatorname{trace}(A) = 0$.

How can we decide wether a matrix is invertible and if so how can we compute its inverse?

Suppose A is invertible and suppose that $B = A^{-1}$ is its inverse. Let $B_1, B_2, \ldots, B_n$ be the columns of B. Then

$$AB = [AB_1\, AB_2\, \ldots AB_n\,] = I_n.$$

It follows that B_i is a solution of the equation

$$AX = e_i$$

where e_i is the ith column of the identity matrix. Thus,

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where the single nonzero entry 1 is in the ith row. To compute the columns of B (and hence B) we simply have to solve the matrix equations $AX = e_i$ for $i = 1, 2, \ldots, n$. But it gets better, we can solve these equations simultaneously.

Notice that in the augmented matrix $[A \mid e_i]$ the final form will always be the reduced row echelon form of A plus some column. Hence, it will be $[I_n \mid f_i]$. But what is f_i? It is exactly B_i. Thus, we can shorten the proceedings by simultaneously solving the systems and instead of row reducing $[A \mid e_i]$ we row reduce $[A \mid I]$ to reduced echelon form and the result will be $[I_n \mid B]$: the columns of B then are exactly the solutions of $AX = e_i$. More formally, if

$$\underbrace{E_1 E_2 \cdots E_p}_{A^{-1}} A = I_n,$$

then applying the same procedures to I_n we obtain $A^{-1} = E_1 E_2 \cdots E_p I_n = E_1 E_2 \cdots E_p$ (here $E_1, E_2, \ldots, E_p$ are the elementary matrices corresponding to the row operation sthat transform A to I).

2.23 Example. Let us compute the inverse of

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{array}\right] \to \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & -3 & -2 & 1 \end{array}\right]$$

$$\to \left[\begin{array}{cc|cc} 1 & 2 & 1 & 0 \\ 0 & 1 & \frac{2}{3} & \frac{-1}{3} \end{array}\right] \to \left[\begin{array}{cc|cc} 1 & 0 & \frac{-1}{3} & \frac{2}{3} \\ 0 & 1 & \frac{2}{3} & \frac{-1}{3} \end{array}\right]$$

which coincides with the inverse of Example 2.6.

Invertible matrices are fundamentally important: we observed above that if A, B are invertible, then also AB is invertible. This endows the collection of all invertible $n \times n$ matrices with an important structure:

2.24 Definition. A *group* is a pair $(G, \bullet)$ where G is a nonempty set and $\bullet$ is an operation on G such that

a. $\bullet$ is associative: for all $g, h, k \in G$ we have

$$g \bullet (h \bullet k) = (g \bullet h) \bullet k.$$

b. There is a distinguished element $e \in G$, called the *identity*, such that $e \bullet g = g \bullet e = g$ for all $g \in G$.

c. For each $g \in G$, there exists an element $g^{-1} \in G$ such that $g \bullet g^{-1} = g^{-1} \bullet g = e$.

We usually just write G instead of $(G, \bullet)$.

If in additon $g \bullet h = h \bullet g$ for all $g, h \in G$, then G is called an *abelian* or *commutative* group.

We usually simply write gh instead of $g \bullet h$, or even $g \cdot h$ if we really need a symbol.

We see that a group satisfies some of the field axioms. But we have only one operation, which need not be commutative. But we do have an identity and inverses.

Remark. Our list of axioms for a field could have been shortened considerably: A field is a set $\mathbb{F}$ together with operations $+$, $\cdot$ such that

a. $(\mathbb{F}, +)$ is an abelian group with identity element 0.

b. $\mathbb{F} - \{0\}$ is closed under the $\cdot$-operation and $(\mathbb{F} - \{0\}, \cdot)$ is an abelian group with identity element 1.

c. For all $a, b, c \in \mathbb{F}$ we have $a \cdot (b + c) = a \cdot b + a \cdot c$.

The concept of a group is the right notion to discuss invertible matrices:

2.25 Definition. The *general linear group* $\mathrm{GL}_n(\mathbb{F})$ is the set of all $n \times n$ invertible matrices with entries in $\mathbb{F}$, together with matrix multiplication.

As the name suggests, the general linear group together with matrix multiplication is a group, this follows from Observation V in Section 2.2, Definition 2.5, and Remark 2.18.

Much like in the case of fields, there is an abundance of different groups. Some are groups consisting of matrices, some are groups consisting of completely different objects; but all share many properties. The concept of a group allows to study these common properties independent of a particular case, all at once.

The various general linear groups vary greatly in type if you change the field in question. For instance, $\mathrm{GL}_n(\mathbb{F}_2)$ is a finite group for each n. On the other hand, $\mathrm{GL}_n(\mathbb{R})$ is infinite. Both groups are complicated in their own ways. There is a whole "industry" of mathematicians working at understanding such groups.

2.26 Example. $\mathrm{GL}_2(\mathbb{F}_2)$ is the following set of 6 matrices:

$$\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ & 1 \end{bmatrix}, \begin{bmatrix} 1 & \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & \end{bmatrix}, \begin{bmatrix} & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}$$

There are many examples of groups. Groups are interesting in their own right, because they are somehow the "simplest" interesting mathematical objects one step above sets: they are sets with one additional structure (an operation). They are also interesting because they arise as building blocks of more complex structures (like rings or fields, or, as we will see, vector spaces). One of the most interesting aspects of groups however is the fact that they can be used to describe symmetries (see wallpaper or friese symmetries).

To illustrate the latter concept, consider a regular triangle (all sides even length). It has six symmetries: doing nothing (denoted by $\mathbf{1}$), a rotation by 120 degrees, and a rotation by 240 degrees (counter-clockwise), denoted by r_{120} and r_{240}, respectively. Moreover, there are 3 reflexions, in the median lines (the lines passing through a vertex and the midpoint of the opposing edge), denoted by s_1, s_2, s_3, say. Let D_6 be the set of these six symmetries of some fixed triangle. It is easy to see, that performing two symmetries in a row is again a symmetry of the triangle. So we obtain an operation on D_6: if x, y are two elements of D_6, we define xy as the symmetry of the triangle obtained by first applying y, and then applying x. One can show that D_6 together with this composition law is a group with identity elements 1. For example, we then have $r_{120}r_{120} = r_{240}$ and $r_{240}r_{240} = r_{120}$, and $r_{240}r_{120} = \mathbf{1}$. Similarly, $s_is_i = \mathbf{1}$. If labelled appropriately, then $s_1s_2 = r_{120}$, $s_1s_3 = r_{240}$. (This is best seen by labelling the vertices counterclockwise as $1, 2, 3$. If s_i is the reflexion fixing vertex i, then s_1s_2 is r_{120}, and $s_1s_3 = r_{240}$. We leave the remaining products to the reader.

An abstract concept of "symmetry" is embodied by the *permutations* of a finite set, say, $\{1, 2, \ldots, n\}$. These are functions (reorderings) $\sigma\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ that are bijective (injective and surjective, see the appendix). In other words, if we write σ in the following form as

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(1) & \sigma(2) & \ldots & \sigma(n) \end{pmatrix}$$

then *every* integer between 1 and n is listed in the second row exactly once; so the second row is just a permutation of the first row. Let S_n be the set of all such permutations. It is called the *symmetric group* in n letters. And indeed, it is a group with the operation given by composition: we write $\sigma\mu$ for the permutation obtained by first applying μ and then applying σ. For example if $n = 3$, then there are precisely 6 permutations:

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$$

And we have for example

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

Note if we associate to each element x in D_6 the permutation π_x it induces on the three vertices labelled, $1, 2, 3$, then

$$\pi_{s_1} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \pi_{s_2} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \pi_{r_{120}} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$$

and now a miracle happens:

$$\pi_{xy} = \pi_x \pi_y,$$

so it doesn't matter whether we first compute xy and then figure out the permutation, or whether we multiply the permutations!

2.27 Definition. Let G be a group. A *subgroup* is a subset $H \subseteq G$ such that

a. H is nonempty;

b. whenever $g, h \in H$ then $gh \in H$;

c. if $g \in H$ then also $g^{-1} \in H$.

Remark. If $H \subseteq G$ is a subgroup of a group G, then $e \in H$: indeed, H is nonempty, so there is some $h \in H$; then also $h^{-1} \in H$ and hence $hh^{-1} = e \in H$.

This shows that H together with the operation inherited from G is again a group.

It can be shown, that *every* finite group can be realized as a subgroup of S_n for some n. Similarly, many (virtually all important) groups can be realized as subgroups of $\mathrm{GL}_n(\mathbb{F})$ for some n and some $\mathbb{F}$.

2.4.4 Problem. Let G be any group and let $a, b \in G$. Then the equations $ax = b$ and $ya = b$ have the unique solution $x = a^{-1}b$ and $y = ba^{-1}$ in G. But note that these solutions differ in general, that is, $x \neq y$ in general.

It is a good exercise in matrix arithmetic to check that the following subsets are all matrix goups.

a. The *special linear group*

$$\mathrm{SL}_2(\mathbb{F}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \middle| \, ad - bc = 1 \right\} \subseteq \mathrm{GL}_2(\mathbb{F})$$

There is an analogue for other values of n which we will define later.

b. The *orthogonal group*

$$\mathrm{O}_n(\mathbb{F}) = \{A \in M_n(\mathbb{F}) \mid AA^T = I\}.$$

c. Important in Physics is the *unitary group*: Let $A = [a_{ij}] \in M_n(\mathbb{C})$. We define $\overline{A}$ to be the matrix whose entry at (i, j) is $\overline{a}_{ij}$. So $A \in M_n(\mathbb{R}) \subseteq M_n(\mathbb{C})$ if and only if $A = \overline{A}$. Let us define $A^* = \overline{A}^T$. Then the unitary group

$$\mathrm{U}_n(\mathbb{C}) = \{A \in M_n(\mathbb{C}) \mid AA^* = I\}.$$

d. Notice that since $\mathbb{R} \subseteq \mathbb{C}$ we also have

$$\mathrm{GL}_n(\mathbb{R}) \subseteq \mathrm{GL}_n(\mathbb{C}).$$

It is instructive to verify that $\mathrm{O}_n(\mathbb{R}) = \mathrm{GL}_n(\mathbb{R}) \cap \mathrm{U}_n(\mathbb{C})$.

e. The set $B_n(\mathbb{F}) = \{A \in \mathrm{GL}_n(\mathbb{F}) \mid A$ is upper triangular$\}$.

f. The set $T_n(\mathbb{F}) = \{A \in \mathrm{GL}_n(\mathbb{F}) \mid A$ is diagonal$\}$.

With the exception of the last one, all groups in this list are not abelian (for $n > 1$).

3. Vector spaces

In many areas of science objects are described by various quantities associated to them. Here "quantity" is meant in a loose sense, it doesn't have to be a number. For instance, in classical mechanics, the state of a particle (usually thought of as a "point mass" with no physical size) is described by its position and its dynamic properties: its speed and its direction of travel. Similarly, a force (acting on the particle, for instance) is described by its magnitude and its direction. In this context many physical quantities are described by a magnitude and a direction. The common feature of these "things" is that they can be scaled (that is, their magnitude can be multiplied by a real number) and added (using what is called the parallelogram rule). We won't go into this any deeper because sometimes this picture of little arrows in the plane is misleading. For instance in quantum mechanics the state of a particle still is described by what we call a vector, but it does not have a natural "direction" in a concrete sense (in fact, depending on the concrete description used, this vector is a certain complex valued function). In what follows it is best not to attribute any concrete meaning to the word "vector." It will turn out that many structures behave very much in the same way: they form sets where two elements can be added in a more or less natural way, and where elements can be multiplied by an element in a given (fixed) field $\mathbb{F}$. Linear algebra can be thought of as the study of the common properties of such structures. We will apply this later on to a thorough understanding of linear equations, for instance, whose solutions serve as both, motivational examples, and main applications of the theory.

3.1. Vector spaces

We will follow a similar approach to vector spaces as we did for general fields: we list the properties we require vector spaces to share and then deduce some consequences. This has two advantages. First, we treat many mathematical objects simultaneously without having to do the same work over and over again. But second, even when we are interested only in $\mathbb{R}^2$, say, the general treatment may help by allowing us to focus only on the essential properties.

3.1 Definition. Let $\mathbb{F}$ be a field. An $\mathbb{F}$-*vector space* or simply *vector space* if $\mathbb{F}$ is understood is a triple $(V, +, \cdot)$ where V is a nonempty set and $+$ is an *associative* operation on V, called the *addition* of V, and $\cdot$ is a map $\mathbb{F} \times V \to V$ called the *scalar multiplication* (which associates to each $c \in \mathbb{F}$ and each $v \in V$ an element $cv = c \cdot v \in V$), such that the following properties hold:

a. The addition is *commutative*: $v + w = w + v$ for all $v, w \in V$.

b. There is an identity element for the addition: There is an element 0 called the *zero vector* or simply *zero* of V such that $0 + v = v + 0 = v$ for all $v \in V$.

c. Each element of V has an additive inverse: for each $v \in V$ there is an element $-v$ of V such that $v + (-v) = 0$.

d. The scalar multiplication is associative: for each $a, b \in \mathbb{F}$ and each $v \in V$, we have $a(bv) = (ab)v$.

e. $1 \in \mathbb{F}$ is an identity element for the scalar multiplication: $1v = v$ for all $v \in V$.

f. The scalar multiplication is distributive in the following two senses: for each $a, b \in \mathbb{F}$ and each $v \in V$ we have
$$(a + b) \cdot v = av + bv;$$
and for each $c \in \mathbb{F}$ and $v, w \in V$ also
$$c(v + w) = cv + cw.$$

We usually write just V to denote the vector space $(V, +, \cdot)$ with the operations understood.

A vector space over the field $\mathbb{R}$ of real numbers is often called *real*, and a vector space over the field $\mathbb{C}$ of complex numbers is often called *complex*.

Remark. The elements of V are sometimes called *vectors*. One should not attribute any intrinsic meaning to the word "vector," other than being an element of a vector space.

Also, even though we denote the zero vector by 0, it has usually *nothing* to do with the additive identity of $\mathbb{F}$ (unless, of course, $V = \mathbb{F}$ in which case both zeros coincide). If this confuses you, you could write (for a while) 0_V and $0_{\mathbb{F}}$ for the zero elements of V and $\mathbb{F}$, respectively. However, this rarely causes any concern.

Remark. As in the case of fields, the definition of a vector space can be streamlined a little bit, by replacing the first three axioms in Definition 3.1 by a single one: $(V, +)$ is an abelian group (with identity element denoted by 0).

3.2 Examples. We begin with two "obvious" examples.

a. If $\mathbb{F}$ is any field, there is always the[1]*trivial* vector space: if V is any set with one element a, say, then we can turn V into an $\mathbb{F}$-vector space by simply relabelling a by 0 and defining $0 + 0 = 0$ and $c0 = 0$ for all $c \in \mathbb{F}$.

b. $V = \mathbb{F}$ with the addition and multiplication of the field $\mathbb{F}$.

3.3 Example. In Section 2.2 we defined an addition and scalar multiplication on the set $M_{m\times n}(\mathbb{F})$ of all $m \times n$ matrices over $\mathbb{F}$. Observations I through III guaranteed that this turns $M_{m\times n}(\mathbb{F})$ into a vector space.

[1] Technically, there is not *one* trivial vector space: there are as many as there are sets with one element. However all of them are of course similar enough to warrant to identify them.

3.4 Example. One of the most fundamental examples is the vector space $\mathbb{F}^n$ of n-tuples with entries in $\mathbb{F}$ (where $n \geq 1$ is an integer). By convention we define $\mathbb{F}^0 = \{0\}$. By identifying $\mathbb{F}^n$ with the set of column vectors $M_{n\times 1}(\mathbb{F})$ we obtain a vector space structure on $\mathbb{F}^n$ which of course in the case of $n = 1$ coincides with the one defined above.

If $v = (a_1, a_2, \ldots, a_n), w = (b_1, b_2, \ldots, b_n)$ are elements of $\mathbb{F}^n$, then

$$v + w = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n). \tag{3.1}$$

Similarly, if $c \in \mathbb{F}$ and v is as above, then

$$cv = (ca_1, ca_2, \ldots, ca_n). \tag{3.2}$$

The next example is fundamental. In a sense it generalizes the cases of $\mathbb{F}^n$ and $M_{m\times n}(\mathbb{F})$ introduced above.

3.5 Example. Let X be any (nonempty) set and $\mathbb{F}$ a field. On the set $\mathcal{F}(X, \mathbb{F})$ of $\mathbb{F}$-valued functions on $\mathbb{F}$ we define an addition and scalar multiplication pointwise. That is, if $f, g \in \mathcal{F}(X, \mathbb{F})$, then $f + g \colon X \to \mathbb{F}$ is defined by

$$(f + g)(x) = f(x) + g(x). \tag{3.3}$$

If $c \in \mathbb{F}$, then $cf \colon X \to \mathbb{F}$ is defined as

$$(cf)(x) = c(f(x)). \tag{3.4}$$

These two operations turn $\mathcal{F}(X, \mathbb{F})$ into a vector space. We often write $\mathcal{F}(X)$ instead of $\mathcal{F}(X, \mathbb{F})$ if the field is understood.

Let us verify some of the axioms: First of all, the addition is associative: indeed, if $f, g, h \in \mathcal{F}(X)$, then we have to verify that $f + (g + h) = (f + g) + h$. Two functions are equal if and only they have the same value for each $x \in X$. So let $x \in X$ then

$$\begin{aligned}(f + (g + h))(x) &= f(x) + (g + h)(x) = f(x) + (g(x) + h(x)) \\ &= (f(x) + g(x)) + h(x) = (f + g)(x) + h(x) = ((f + g) + h)(x)\end{aligned}$$

Since $x \in X$ was arbititrary, the functions $f + (g + h)$ and $(f + g) + h$ are equal.

Now as for the axioms in 3.1:

a. $f+g = g+f$: indeed, for each $x \in X$ we have $(f+g)(x) = f(x)+g(x) = g(x)+f(x) = (g + f)(x)$.

b. There is an additive identity: we define $0 \colon X \to \mathbb{F}$ as the constant function that maps everything to $0 \in \mathbb{F}$: $0(x) = 0$ for all $x \in X$. (It is common to write 0 for this function.)

   Then $f + 0 = f$ because $(f + 0)(x) = f(x) + 0(x) = f(x)$ for all $x \in X$.

c. For any $f \in \mathcal{F}(X)$, define $-f$ to be the function that maps x to $-f(x)$.

   Then $(f + (-f))(x) = f(x) + (-f(x)) = 0$ in $\mathbb{F}$, hence $f + (-f) = 0$. (By the way, in $\mathbb{F}$ we know that $-a = (-1)a$ for all $a \in \mathbb{F}$, so it follows that $-f = (-1)f$ again by pointwise comparison.)

d. $(a(bf) = (ab)f$: again by pointwise comparison, this follows from the associative law in $\mathbb{F}$: $(a(bf))(x) = a((bf)(x)) = a(bf(x)) = (ab)f(x) = ((ab)f)(x)$.

e. $1f = f$ is clear: $(1f)(x) = 1f(x) = f(x)$ for all $x \in X$.

f. The distributive laws are also easily verified by pointwise considerations. We do one of them: $a(f + g) = af + ag$ for all $a \in \mathbb{F}$ and $f, g \in \mathcal{F}(X)$.

$$\begin{aligned}(a(f+g))(x) = a \cdot (f+g)(x) &= a \cdot (f(x) + g(x)) \\ &= af(x) + ag(x)(af)(x) + (ag)(x) = (af + ag)(x)\end{aligned}$$

For convenience, we define[2] $\mathcal{F}(\emptyset, \mathbb{F}) = \{0\}$.

3.6 Example. Here is an example of a "weird" vector space: it is a vector space over the field $\mathbb{F}_2$. Let X be a fixed set. Define $V = P(X) = \{S \subseteq X\}$, the power set[3] of X.

We define the addition as follows: $S + T := S \cup T - S \cap T$: Thus $S + T$ is the subset of X defined by $x \in S + T$ if and only if $x \in S$ or $x \in T$ but $x \notin S \cap T$. If $X = \{1, 2, \ldots, n\}$, say, computer scientists will recognize this as the XOR operation (exclusive or) on strings of n bits.

It is a good exercise to check that $+$ satisfies all that is required for the addition of a vector space: indeed, the empty set $\emptyset$ is an identity element. Also for each subset S we have $S + S = \emptyset$. Hence $S = -S$ is an additive inverse. Also, associativity is not hard to verify.

Since $\mathbb{F}_2$ has only two elements, the scalar multiplication alone carries not much information: $1v = v$ is required from the vector space axioms and we define $0v = 0$ (we actually don't have a choice here: if V is to be a vector space, this is forced; see Proposition 3.10 below). So what does it actually mean to have a vector space over $\mathbb{F}_2$? The crucial part is contained in the distributive law: we must have that $0 = 0v = (1 + 1)v = 1v + 1v = v + v$ for each $v \in V$. Hence $v = -v$ and this is precisely what we obtained above.

3.7 Example. Let $V = \mathbb{C}$. We know that this is a complex vector space. However, we can view it as a *real* vector space as well (by simply restricting the scalar multiplication to scalars in $\mathbb{R}$). In field theory this is a very important concept: if K is a subfield of a field L, then the structure of L as a vector space over K is often used.

3.8 Example. In Quantum Mechanics, the state of a system is described by a vector, often denoted $|\psi\rangle$, in a complex vector space $\mathcal{H}$. The point here is that $\mathcal{H}$ is typically an "abstract"

[2] Depending on your conventions there is precisely one map from the empty set into any other set. We just call this map 0 and use the trivial vector space structure.

[3] The set whose elements are the subsets of X.

vector space (that is, just an object satisfying the axioms of the vector space; its elements are not typically *equal* to n-tuples of complex numbers, but representatives of "states").

In algebraic statistics, probability distributions are often elements in vector spaces that are "tensor products" of other vector spaces.

3.9 Example. Sometimes it is useful to be able to create new vector spaces out of old ones: Let V, W be vector spaces over $\mathbb{F}$. Then

$$V \times W = \{(v, w) \mid v \in V, w \in W\}$$

becomes a vector space if we define

$$(v, w) + (x, y) = (v + x, w + y)$$

($v, x \in V$, $w, y \in W$) and

$$c(v, w) = (cv, cw).$$

($c \in \mathbb{F}$). We refer to $V \times W$ as the (*direct*) *product*[4] of V and W.

Remark. A purist (or set theorist) might object to our definition of a matrix as a "rectangular array" because what is that? Set theory allows only for certain constructions out of known sets and forming a "rectangular array" is not one of them. As mentioned in a footnote when we defined matrices, we could be more precise and view a matrix with m rows and n columns and entries in $\mathbb{F}$ as nothing but a function $f\colon D \to \mathbb{F}$ where

$$D = \{(i, j) \mid 1 \le i \le m, 1 \le j \le n\}$$

is a subset of $\mathbb{N}^2$. Indeed, if $A = [a_{ij}]$ is a matrix such a function is specified by putting $f(i, j) = a_{ij}$. Thus, one may think of a matrix as a function, and of our rectangle as simply a convenient notation.

It is a good exercise to check that if we identify matrices and functions in this manner that indeed sums are identified with sums and scalar multiples are identified with scalar multiples. In other words, $M_{m \times n}(\mathbb{F})$ may be identified with the vector space $\mathcal{F}(D, \mathbb{F})$. An even simpler exercise shows that $\mathbb{F}^n$ may be identified with $\mathcal{F}(\{1, 2, \ldots, n\}, \mathbb{F})$.

Before we expand our universe of vector spaces, let's briefly turn to some of the basic properties that hold in every vector space.

First, as for fields, a priori there might be several elements in V that could serve as an additive identity element. However this is not the case: Suppose $0'$ is any identity, then

$$0 = 0 + 0' = 0' \tag{3.5}$$

where the first equality uses that $0'$ and the second that 0 is an additive identity.

The following list of properties summarizes the basic arithmetic in a vector space.

[4] Not to be confused with the *tensor product* of V and W.

3.10 Proposition. *Let V be a vector space. Then the following holds:*

a. *For all $u, v, w \in V$, if $u + v = u + w$, then $v = w$.*

b. *An equation of the form $u + x = v$ has the unique solution $v + (-u)$, denoted by $v - u$.*

c. *$-(-v) = v$ for all $v \in V$.*

d. *$c0 = 0$ for all $c \in \mathbb{F}$. (Here 0 is the zero element of V on both sides.)*

e. *$0v = 0$ for all $v \in V$.*

f. *$(-c)v = c(-v) = -(cv)$ for each $v \in V$ and each $c \in \mathbb{F}$. In particular, $-v = (-1)v$.*

g. *If $c \in \mathbb{F}$ is nonzero and if for some $v, w \in V$, $cv = cw$ then $v = w$.*

Proof. The proofs of most of these statements are identical (word by word) to corresponding results about a field $\mathbb{F}$: cf. Lemma 1.4 and Proposition 1.7. We will prove e. to g..

e. $0_{\mathbb{F}} = 0_{\mathbb{F}} + 0_{\mathbb{F}}$ so $0_{\mathbb{F}}v = (0_{\mathbb{F}} + 0_{\mathbb{F}})v = 0v + 0v$. By b. it follows $0v = 0$.

f. $cv + (-c)v = (c + (-c))v = 0v = 0$ by e. This shows $(-c)v = -(cv)$ by b.. Similarly, if $cv + c(-v) = c(v + (-v)) = c0 = 0$ which implies that $c(-v) = -(cv)$ as well.

   The remaining assertion is obtained by applying this in case $c = 1$.

g. Suppose $cv = cw$. Then also $c^{-1}(cv) = c^{-1}(cw)$. Using Axioms d. and e. we get $v = w$.

□

3.1.1. *Generalized associativity

One question we have been ducking so far is the following: we are very familiar with the notion of adding many numbers. For instance, we have no difficulty interpreting

$$3 + 8 + 2 + (-1) + 0 + 123 + 23. \tag{3.6}$$

Why is that? Consider the problem of adding four real numbers a_1, a_2, a_3, a_4: There are several ways to give meaning to the expression $a_1 + a_2 + a_3 + a_4$: it can be any of the following

$$\begin{aligned} &a_1 + (a_2 + a_3 + a_4)),\ a_1 + ((a_2 + a_3) + a_4),\ (a_1 + a_2) + (a_3 + a_4), \\ &\qquad (a_1 + (a_2 + a_3)) + a_4,\ ((a_1 + a_2) + a_3) + a_4. \end{aligned} \tag{3.7}$$

As you may have guessed, because of the associative law of the addition of real numbers, all these expressions evaluate to the same real number. This is the very reason why we can give meaning to an expression like

$$a_1 + a_2 + \cdots + a_n \tag{3.8}$$

where the a_i are n real numbers. We will prove this now once and for all for an associative binary operation. So let X be a (nonempty) set and let $\bullet$ be any associative binary operation on X. By associativity, for all $x, y, z \in X$ we have $x \bullet (y \bullet z) = (x \bullet y) \bullet z$. You may think of X as a field and $\bullet$ as the multiplication; or, you may think of $\bullet$ as the addition in a vector space. Given n elements $x_1, x_2, \ldots, x_n$, we want to prove that there is one and only one way to define $x_1 \bullet x_2 \bullet \cdots \bullet x_n$ (that is, computing a "product" like this does not depend on the order the pairings are evaluated). To be precise, let $X^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in X\}$. A *product* for $\bullet$ is a collection of functions

$$P_n \colon X^n \to X \tag{3.9}$$

(one for each $n \in \mathbb{N}$) such that

a. $P_1(x) = x$ for all $x \in X$.

b. If $n > 1$, then for each $i = 1, 2, \ldots, n-1$ and all $(x_1, x_2, \ldots, x_n) \in X^n$, we have $P_n(x_1, x_2, \ldots, x_n) = P_i(x_1, x_2, \ldots, x_i) \bullet P_{n-i}(x_{i+1}, x_{i+2}, \ldots, x_n)$.

For example, for $n = 2$, the second condition b. just means that $P_2(x_1, x_2) = P_1(x_1) \bullet P_1(x_2)$ which by a. means that $P_2(x_1, x_2) = x_1 \bullet x_2$. For $n = 3$, this just is the associative law: we can define $P_3(x_1, x_2, x_3)$ as $P_2(x_1, x_2) \bullet x_3 = (x_1 \bullet x_2) \bullet x_3$. The associative law will then make sure that a. and b. are satisfied.

3.1.1 Proposition. *Let $\bullet$ be an associative operation on a nonempty set X. Then there is a unique product satisfying* a. *and* b. *above.*

Proof. We will prove the following result: for each natural number n, there exists a unique collection of functions $P_1, P_2, \ldots, P_n$ such that $P_i \colon X^i \to X$ and P_i satisfies a. and b.. We will proceed by induction on n.

Base case: If $n = 1$, we define $P_1(x) = x$ and nothing is to do.

Induction step: Now let n be a fixed positive integer, for which we assume to know that there is a unique collection $P_1, P_2, \ldots, P_n$ satisfying a. and b.

We define a collection $P'_1, P'_2, \ldots, P'_{n+1}$ be defining $P'_i = P_i$ for $i \leq n$ and

$$P_{n+1}(x_1, x_2, \ldots, x_{n+1}) = P_n(x_1, x_2, \ldots, x_n) \bullet x_{n+1}.$$

Then a. and b. are still true for P_i $(i < n+1)$. So we only need to worry about P_{n+1}. Let $1 \leq i \leq n$. If $i = n$, then a. is precisely the definition of P_{n+1}. Otherwise, $i < n$ and we have

$$\begin{aligned} P_{n+1}(x_1, x_2, \ldots, x_{n+1}) &= P_n(x_1, x_2, \ldots, x_n) \bullet x_{n+1} \\ &= (P_i(x_1, x_2, \ldots, x_i) \bullet P_{n-i}(x_{i+1}, x_{i+2}, \ldots, x_n)) \bullet x_{n+1} \\ &= P_i(x_1, \ldots, x_i) \bullet (P_{n-i}(x_{i+1}, \ldots, x_n) \bullet x_{n+1}) \\ &= P_i(x_1, x_2, \ldots, x_i) \bullet P_{n+1-i}(x_{i+1}, x_{i+2}, \ldots, x_{n+1}). \end{aligned} \tag{3.10}$$

(Note where we used that $\bullet$ is associative.)

Thus $P'_1, \dots, P'_{n+1}$ is again such a sequence. As for uniqueness, suppose $Q_1, Q_2, \dots, Q_{n+1}$ is also a sequence satisfying a. and b., then $Q_i = P_i$ for $i = 1, 2, \dots, n$ because of the uniqueness in the case n. But then

$$\begin{aligned} Q_{n+1}(x_1, x_2, \dots, x_{n+1}) &= Q_n(x_1, \dots, x_n) \bullet Q_1(x_{n+1}) \\ &= P_n(x_1, \dots, x_n) \bullet x_{n+1} = P_{n+1}(x_1, \dots, x_{n+1}). \end{aligned} \tag{3.11}$$

□

(Note that one could substantially reduce this proof by means of the Recursive Definition Theorem 1.15.)

As a corollary, we now can form sums of arbitrarily many vectors without specifiying any brackets. So if $v_1, v_2, \dots, v_n$ are elements of a vector space V, we define

$$v_1 + v_2 + \cdots + v_n = \sum_{i=1}^{n} v_i := P_n(v_1, v_2, \dots, v_n)$$

where P_n is the unique product function associated to $\bullet = +$ (if $\bullet$ is a commutative operation and written as $+$, we usually call a product a sum).

The same applies to sums and products of elements in a field $\mathbb{F}$ (or a ring, like $M_n(\mathbb{F})$).

One can use similar reasoning as above to deduce that arbitrary products of matrices are defined (as long as they make sense, o.e.).

3.2. Subspaces

Often a subset W of a vector space V, together with the addition function of V and the scalar multiplication of V (both restricted to W of course) is again a vector space. For this it is clearly necessary that W is *closed* under the two operations $+$ and $\cdot$. In fact this is almost enough:

3.11 Definition. Let V be an $\mathbb{F}$-vector space. A subset $W \subseteq V$ is called a *subspace* of V if it satisfies the following three properties:

a. W is not empty.

b. If $v, w \in W$ then also $v + w \in W$.

c. If $w \in W$ and $r \in \mathbb{F}$ then $rv \in W$.

Before we list important examples let us first note:

3.12 Lemma. *Let $W \subseteq V$ be a subspace of a vector space V. Then W together with $+$ and $\cdot$ is a vector space.*

Moreover, the zero vector of V is contained in W and if $w \in W$ then $-w \in W$.

Proof. The two operations are defined on W by the properties a. and b.. They clearly remain associative and the addition is still commutative (we only restricted the arguments of the operations). Similarly, the distributive laws still hold.

Thus, all is left is to show that there exists an identity element and additive inverses. The final statement of the lemma gives a hint: they are the same as in V: indeed, if $w \in W$ then also $-w = (-1)w \in W$ because of c.. Finally, because of a., there is at least one $w \in W$. By Proposition 3.10 e., $-w = (-1)w$ and so $-w \in W$ by c. Hence $0 = w + (-w) \in W$ because of b.. □

This opens a wealth of new examples:

3.13 Examples.

a. If V is any vector space then $\{0\}$ and V are subspaces. $\{0\}$ is often referred to as the *trivial subspace*; sometimes it is also denoted simply by 0.

b. Let X be any set and let $V = \mathcal{F}(X) = \mathcal{F}(X, \mathbb{F})$. For a function $f\colon X \to \mathbb{F}$ its *support* is defined as the set where it is nonzero: $\operatorname{supp}(f) = \{x \in X \mid f(x) \neq 0\}$. Let $\mathbb{F}X$ be the subset of V of functions with finite support:

$$\mathbb{F}X = \{f \in \mathcal{F}(X) \mid \operatorname{supp}(f) \text{ finite}\}$$

Then $\mathbb{F}X$ is a subspace[5] of V.

c. If $\mathbb{F} = \mathbb{R}$, and $X \subseteq \mathbb{R}$ is an interval, then the set $\mathcal{C}(X)$ of continuous real valued functions, and the set $\mathcal{C}^p(X)$ of p-times continuously differentiable functions are subspaces of $\mathcal{F}(X)$.

d. Again, if $\mathbb{F} = \mathbb{R}$, then, $L^p(\mathbb{R}) = \{f \in \mathcal{F}(\mathbb{R}) \mid \int_{-\infty}^{\infty} |f(x)|^p dx < \infty\}$ is a subspace of $\mathcal{F}(\mathbb{R})$ (where one usually uses Lebesgue-integration). (Also, to be precise, one identifies functions in $L^p(\mathbb{R})$ if their difference is a function g for which $\int_{-\infty}^{\infty} |g(x)|^p dx = 0$. Of course, the only *continuous* such g is $g = 0$; but some elements of $L^p(\mathbb{R})$ are not continuous.) Often, one considers functions with values in $\mathbb{C}$, rather than $\mathbb{R}$. Let

$$L^2(\mathbb{R}, \mathbb{C}) = \{f \in \mathcal{F}(\mathbb{R}, \mathbb{C}) \mid \int_{-\infty}^{\infty} f(x)\overline{f(x)}dx < \infty\}.$$

Note, that $L^2(\mathbb{R}, \mathbb{C})$ is a *complex* vector space.

In fact, in Quantum Mechanics, one model for a single particle on a line is $\mathcal{H} = L^2(\mathbb{R}, \mathbb{C})$, the space of square-integrable complex valued function. The state $|\psi\rangle \in \mathcal{H}$ is then a function $f\colon \mathbb{R} \to \mathbb{C}$, for which $p(x) := \overline{f(x)}f(x)$ is the probability density of finding the particle at x: thus, the probability of finding the particle in the interval $[a, b]$ is $\int_a^b p(x)dx$.

[5] Of course, if X itself is finite, then $\mathbb{F}X = V$

e. A function $f\colon \mathbb{R} \to \mathbb{R}$ is called a *polynomial function* if there exists a (fixed) list of real numbers $a_0, a_1, \dots, a_n$ such that for each $x \in \mathbb{R}$

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Let $\mathcal{P}(\mathbb{R})$ be the set of all polynomial functions on $\mathbb{R}$. Then $\mathcal{P}(\mathbb{R}) \subseteq \mathcal{F}(\mathbb{R})$ is a subspace.

3.14 Example. In many real world situations (e.g. the harmonic oscillator in classical mechanics) one comes across a differential equation of the form

$$y'' + my = 0$$

where m is some fixed constant and y is a function of a variable t, say, and y'' indicates the second derivative of y. For instance if $m = 1$, then both $y(t) = \sin t$ and $y(t) = \cos t$ are solutions of this equation. In fact, any linear combination $\lambda \sin t + \mu \cos t$ is a solution for $\lambda, \mu \in \mathbb{R}$. More generally, it is very easy to check that whenever f, g are solutions then also $f + g$ is a solution and also λf is a solution. It follows that the set of all solutions is a subspace of the space of all functions on $\mathbb{R}$. Many problems in calculus (or analysis) are of this form: we are looking for specific elements of a certain subspace of a space of functions.

The following example is crucial.

3.15 Proposition. *Let $A \in M_{m\times n}(\mathbb{F})$. Then the set of solutions of the homogeneous matrix equation*

$$AX = 0$$

is a subspace of $\mathbb{F}^n$.

Proof. Indeed, let $S \subseteq \mathbb{F}^n$ be the set of all X for which $AX = 0$. Then $0 \in S$ and so S is nonempty. Also if $X, Y \in S$ then

$$A(X + Y) = AX + AY = 0 + 0 = 0$$

so $X + Y \in S$ and finally if $X \in S$ and $c \in \mathbb{F}$ then also $A(cX) = c(AX) = c0 = 0$ and so $cX \in S$. (Compare with the Observations in Section 2.2.) □

The space S is called the *null space* of A and denoted by $\mathcal{N}(A)$.

*Coding Theory

What is Coding Theory? The basic idea is simple: suppose you have a channel over which you can send messages in the form of strings of symbols. In practice this could be a cell-phone, space communication, or as mundane as reading a DVD.

The problem is that most real life channels are error prone: This is clear for outer space communications where the transmission signal is subject to a lot of noise in the form of radiation, planets, etc. But also your DVD could be scratched, the signal from your cell phone tower could be weak, or your computer's memory could be faulty due to faulty contacts or similar.

In all these cases the problem is that the message received at the other end of the channel might not be the one originally sent.

Coding Theory basically asks two questions:

- how to send messages such that it is very likely that the other party can *detect* whether an error occurred.
- related but not the same, how to send a message that the other party can *correct* an error that occurred.

The second is what your DVD player does with a scratched DVD. As an instance for the first question: usually the other party performs an elementary check when you enter a credit card number whether the number is potentially a valid number.

The basic idea for binary channels is as follows: a single symbol is either 0 or 1. We think of them as elements of $\mathbb{F}_2$. The admissible words that we can send over a channel are sequences of fixed length, n, say. Thus, we can represent them as *vectors* in $\mathbb{F}^n$. Now the *codewords*, that is, the messages that actually mean something, form a subset $C \subseteq \mathbb{F}^n$, usually called the *code*. n is called the *length* of the code.

The idea is now to "spread out" the words in C so that they are very "far apart." For instance, it is a very bad idea to use $C = \mathbb{F}^n$ because then no matter what word is received, the recipient will always think it is a valid codeword.

3.2.1 Example. A very simple code is $C = \{(0, 0, 0, \ldots, 0), (1, 1, \ldots, 1)\}$.

To detect an error proceed as follows: if we receive any word where not all symbols are either equal to 0 or equal to 1, then we know an error occurred.

But notice that this is not perfect: if n errors occur, so that all symbols change their value, then we would still interpret the result as a valid codeword. No code that can actually be used to transmit any information can be perfect in that sense. There is always a positive probability that the wrong codeword comes out at the other end (assuming the channel is not perfect).

How could we correct an error? We could use the following procedure: any sequence w of length n we receive is interpreted as its *nearest neighbour* in C. We pick the element in C that is closest to w in the sense that it differs in the least number of positions from w. For instance, if $n = 3$ we would interpret 101 as 111 and 011 as 111 whereas 010 would be interpreted as 000 (it is customary to write sequences just without any brackets and commata).

A careful analysis of the example leads to the following definition: if $v, w \in \mathbb{F}_2^n$, the *Hamming distance* of v, w is

$$d(v, w) = |\{i \mid v_i \neq w_i\}|$$

where $v_1, v_2, \ldots, v_n$ (resp. $w_1, w_2, \ldots, w_n$) are the entries of v (resp. w).

What Coding Theory tries to do is find subsets $C \subseteq \mathbb{F}_2^n$ such that $v, w \in C$ and $v \neq w$ means $d(v, w)$ is "large." Distinct codewords should be far apart. More formally, if C has more than one element we define its *minimum distance* $d(C)$ as

$$d(C) = \min\{d(v,w) \mid v,w \in C, v \neq w\}.$$

Thus Coding Theory is looking for subsets C with $d(C)$ as large as possible (usually given other restrictions).

A very fruitful concept here is to look for subsets C that are actually *subspaces*. These codes are called *linear* and they are used in many applications.

3.2.1 Problem. For $x \in \mathbb{F}_2^n$ define its *weight* as $w(x) = |\{i \mid x_i \neq 0\}| = d(0,x)$.

Let $C \neq \{0\}$ be a linear code. Show that

$$d(C) = \min\{w(x) \mid x \in C - \{0\}\}.$$

3.2.2 Problem. Let C be a code with odd minimum distance $d = 2t+1$. Prove that if t or less errors occurred during a transmission, and if x is received, there is a unique codeword $v \in C$ such that $d(x,v) \leq t$. v is called the *nearest neighbour* of x.

Explain how C could be used to correct up to t errors.

Explain how C could be used to detect up to $d-1$ errors.

Remark. The (linear) *binary repetition code* C of length n is the code of the example above. It is not hard to show that C is optimal in the sense that it is the only code such that $d(C) = n$ (up to equivalence; roughly speaking equivalent codes have the same number of elements and are equally good as codes for all purposes).

But it is a highly inefficient code: to transmit information worth a single bit, it needs n bits. So the bandwith requirements of such a code would be n times the information content of the messages. For many practical purposes this is far from good enough.

Think about outer space communications: historically, during the Mariner expeditions (one of the early uses of Coding Theory), the data transmission rate from the spacecraft to Earth was several bits per second (in the beginning). To transmit a single picture took several hours. Using n times as many hours instead is impractical if n is large.

The code used in the later Mariner expeditions is known as the *Reed-Muller Code*. It was a code $C \subseteq \mathbb{F}^{32}$ of length 32 with $64 = 2^6$ elements. Thus, it used 32 bits to transmit 6 bits worth of data. However, it could correct up to 7 errors. 8 or more errors (per 32 sent message bits) need to happen for a faulty transmission. Using 32 bits you could repeat a 6 bit word roughly 5 times. This would allow only for the guaranteed correction of 2 errors (depending on where they happen).

3.3. Generators and bases

In general, a vector space is just a set of some objects which we may not have a lot of knowledge about, so it might appear that it is very hard to compute with them. We will now learn why this is not so.

3.16 Definition. Let V be a vector space and $v_1, v_2, \dots, v_n \in V$. We say an element $v \in V$ is a *linear combination* of $v_1, v_2, \dots, v_n$ if there exist scalars $c_1, c_2, \dots, c_n \in \mathbb{F}$ for which

$$v = c_1 v_1 + c_2 v_2 + \cdots + c_n v_n.$$

Compare this definition to (2.15) in Chapter 2.

Why are linear combinations interesting? A first hint is the following simple observation:

3.17 Lemma. *Let $W \subseteq V$ be a subspace. If $w_1, w_2, \dots, w_m \in W$, then W contains every linear combination of $w_1, w_2, \dots, w_m$.*

Proof. This is a good opportunity to practice induction proofs.

The base case is obvious: if $m = 1$, i.e. if we are given a single element $w \in W$, then $\operatorname{Span}(w) = \{cw \mid c \in \mathbb{F}\}$ is contained in W because W is a subspace.

Suppose now that for a given natural number m, all linear combinations of m arbitrary elements of W are contained in W. Let $w_1, w_2, \dots, w_{m+1} \in W$ and $c_1, c_2, \dots, c_{m+1} \in \mathbb{F}$ be given. We need to show that $c_1 w_1 + c_2 w_2 + \cdots + c_{m+1} w_{m+1} \in W$. By the induction hypothesis, we know that

$$w = c_1 w_1 + c_2 w_2 + \cdots + c_m w_m \in W \tag{3.12}$$

By the case $m = 1$ we also know that $c_{m+1} w_{m+1} \in W$. Combined, it follows that also $w + c_{m+1} w_{m+1} \in W$ (once again, because of b. in the definition of a subspace). □

Let $v_1, v_2, \dots, v_n$ be a subset of V. Let $\operatorname{Span}(v_1, v_2, \dots, v_n)$ denote the set of all vectors in V that may be expressed as a linear combination of $v_1, v_2, \dots, v_n$.

3.18 Lemma. *Let $v_1, v_2, \dots, v_n \in V$ $(n > 0)$. Then $\operatorname{Span}(v_1, \dots, v_n)$ is a subspace of V.*

In fact it is the minimal subspace containing $v_1, v_2, \dots, v_n$ in the following sense: if W is any subspace of V containing $v_1, v_2, \dots, v_n$ as elements, then $\operatorname{Span}(v_1, v_2, \dots, v_n) \subseteq W$. Thus,

$$\operatorname{Span}(v_1, v_2, \dots, v_n) = \bigcap_{\substack{W \subseteq V \\ v_1, v_2, \dots, v_n \in W}} W$$

where the intersection ranges over all subspaces of V that contain $v_1, v_2, \dots, v_n$.

Proof. We first show that $\operatorname{Span}(v_1, v_2, \dots, v_n)$ is a subspace of V. First of all, it is nonempty: $v_1 = 1v_1 + 0v_2 + \cdots + 0v_n \in \operatorname{Span}(v_1, v_2, \dots, v_n)$ by construction; similarly, $v_i \in \operatorname{Span}(v_1, \dots, v_n)$ for $i = 2, 3, \dots, n$.

Let $\sum_{i=1}^n c_i v_i$ and $\sum_{i=1}^n d_i v_i$ be elements of $\operatorname{Span}(v_1, v_2, \dots, v_n)$ (where of course $c_i, d_i \in \mathbb{F}$). Then their sum

$$c_1 v_1 + \cdots + c_n s_n + d_1 v_1 + d_2 v_2 + \cdots + d_n v_n = \sum_{i=1}^n (c_i + d_i) v_i \tag{3.13}$$

is again a linear combination of $v_1, v_2, \ldots, v_n$ (using the commutative and distributive laws). Similarly, if $\sum_{i=1}^n c_i v_i \in \mathrm{Span}(v_1, v_2, \ldots, v_n)$, and $c \in \mathbb{F}$ then $c(\sum_{i=1}^n c_i v_i) = \sum_{i=1}^n (cc_i)v_i$ is contained in $\mathrm{Span}(v_1, v_2, \ldots, v_n)$.

If W is a subspace containing $v_1, v_2, \ldots, v_n$, then also every linear combination of the v_i is contained in W by Lemma 3.17. Hence $\mathrm{Span}(v_1, v_2, \ldots, v_n) \subseteq W$. Finally, since $\mathrm{Span}(v_1, v_2, \ldots, v_n)$ is a subspace itself, it is then clear that it is equal to the intersection of all subspaces containing $v_1, v_2, \ldots, v_n$. □

3.19 Definition. Let $v_1, v_2, \ldots, v_n \in V$. The subspace $W = \mathrm{Span}(v_1, v_2, \ldots, v_n)$ is called the *subspace generated* or *spanned by* $v_1, v_2, \ldots, v_n$, and $v_1, v_2, \ldots, v_n$ are called *generators* for W. If $L = (v_1, v_2, \ldots, v_n)$ is an ordered list of elements in V we also simply write $\mathrm{Span}(L)$ instead of $\mathrm{Span}(v_1, v_2, \ldots, v_n)$.

A subspace S of V is *finitely generated* if there exist $v_1, v_2, \ldots, v_n \in S$ such that $S = \mathrm{Span}(v_1, v_2, \ldots, v_n)$.

By convention we agree that $\{0\}$ is also generated by the empty list of generators. So we write $\{0\} = \mathrm{Span}(\emptyset) \subset V$.

3.20 Examples.

a. If $v \in V$ is nonzero, then $L = \mathrm{Span}(v) = \{cv \mid c \in \mathbb{F}\}$ is what we call a *line through the origin*. If $\mathbb{F} = \mathbb{R}$ and $V = \mathbb{R}^2$ this is a line on the ordinary sense.

b. If $A_1, A_2, \ldots, A_n \in \mathbb{F}^m$ are the columns of the matrix $A \in M_{m\times n}(\mathbb{F})$, then we call

$$\mathrm{Col}(A) = \mathrm{Span}(A_1, A_2, \ldots, A_n)$$

the *column space* of A. It is a subspace of $M_{m\times 1}(\mathbb{F})$ which as usual we identify with $\mathbb{F}^m$.

It is the set of all $B \in \mathbb{F}^m$ for which the matrix equation $AX = B$ has a solution: indeed, $AX = B$ has a solution if and only if B can be expressed as a linear combination of the columns $A_1, A_2, \ldots, A_n$ of A.

c. Another subset associated to an $m \times n$ matrix is its null space $\mathcal{N}(A) = \{X \in \mathbb{F}^n \mid AX = 0\}$ (see Proposition 3.15). When we say, "Solve the matrix equation $AX = 0$!" what we really mean is "Find generators for $\mathcal{N}(A)$!".

Indeed, when we solve the equation $AX = 0$ what we do is find an expression for X of the form

$$X = x_{f_1}X_1 + x_{f_2}X_2 + \cdots + x_{f_p}X_p$$

where the x_{f_i} are the "free" variables, and the X_i are the coefficient vectors of the free variable x_{f_i}. We will return to this in greater detail shortly.

3.3.1 Problem. Show that indeed the set $L = \{rv \mid r \in \mathbb{R}\}$ is a line in the earlier sense (ie. the set of solutions of $ax + bx = c$).

Notice that if we are given a subspace $W \subseteq V$, then having a list $w_1, w_2, \dots, w_m \in W$ of generators such that $\mathrm{Span}(w_1, w_2, \dots, w_n) = W$ completely describes W: To specify an element of W, we simply need to pick a list $c_1, c_2, \dots, c_m \in \mathbb{F}$ as the coefficients in a linear combination; and every element of W is obtained in that way.

Remark. Many problems arising in this context are of the following form:

a. Show that a certain subspace W of a vector space V is finitely generated.

b. Find an explicit list of generators.

If you think about it, when we say, "Solve the matrix equation $AX = 0$!" what we actually mean is "Find generators for $\mathcal{N}(A)$." Many problems in Linear Algebra are reformulations of this task.

It may seem that a. is redundant because of course if we have solved b. then a. is clear. However, often it is true by some abstract reasoning that a subspace is finitely generated but it is not at all obvious how to find an explicit set of generators. (This shows both, the power and limitations of abstraction.)

For instance, we will prove a theorem that states as long as V itself is finitely generated, *every* subspace of V is also finitely generated. But this does not tell us anything about how we could find explicit generators in a particular case.

Similarly, one can show that the solution set W of a differential equation of the form $y'' + by' + cy = 0$ is always finitely generated, *even if b and c itself are continuous functions!* There is no easy way of determining any nonzero solution of such an equation.

Clearly, we would like to use an *optimal* spanning set, i.e. we don't want a set that is too large. For example, consider $V = \mathbb{F}^2$. It is immediate that $V = \mathrm{Span}(e_1, e_2)$; but also $V = \mathrm{Span}(e_1, e_2, e_1 + e_2)$. What distinguishes the two cases? In the first case, the set is obviously minimal: we cannot remove any element and still have a spanning set. In the second case however, we can leave off any single element. How can we precisely describe this situation?

Let $(v_1, v_2, \dots, v_p)$ be an ordered list of vectors $v_i \in V$. Suppose we can express v_i as a linear combination of the remaining elements in the list (or $v_i = 0$ if this list is empty). Let us write $L' = (v_1, v_2, \dots, \hat{v}_i, \dots, v_p)$ for the list with v_i removed. Then $v_i \in \mathrm{Span}(L')$ and hence $\mathrm{Span}(L) = \mathrm{Span}(L')$: indeed, $\mathrm{Span}(L')$ contains $v_1, v_2, \dots, v_p$, and is a subspace, so by Lemma 3.18, $\mathrm{Span}(L) \subseteq \mathrm{Span}(L')$. But obviously also $\mathrm{Span}(L') \subseteq \mathrm{Span}(L)$, and hence $\mathrm{Span}(L) = \mathrm{Span}(L')$.

Conversely, if $\mathrm{Span}(L') = \mathrm{Span}(L)$, then $v_i \in \mathrm{Span}(L')$ and so v_i is a linear combination of the elements in L' (or, $v_i \in \mathrm{Span}(\emptyset) = \{0\}$ if L' is empty; then $v_i = 0$).

So v_i may be dropped from the list of generators for $\mathrm{Span}(L)$ without changing the subspace, if and only if it is a linear combination of the elements in L'.

Now, v_i is a linear combination of the elements of L' means that there are scalars

$$c_1, c_2, \dots, c_{i-1}, c_{i+1}, \dots, c_p \in \mathbb{F}$$

such that

$$v_i = c_1v_1 + c_2v_2 + \cdots + c_{i-1}v_{i-1} + c_{i+1}v_{i+1} + \cdots + c_pv_p.$$

Here we interpret the right hand side simply as 0 if $p = 1$. Adding on both sides $-v_i = (-1)v_i$ and using the commutative law in V we obtain

$$0 = c_1v_1 + c_2v_2 + \cdots + c_iv_i + \cdots + c_pv_p$$

where $c_i = (-1)$. In particular, we have shown that there exist $c_1, c_2, \dots, c_p \in \mathbb{F}$, not all equal to zero since at least $c_i \neq 0$, such that $0 = c_1v_1 + \cdots + c_pv_p$.

Conversely, suppose that there are some $c_1, c_2, \dots, c_p$ such that not all c_i are equal to zero but still

$$c_1v_1 + c_2v_2 + \cdots + c_pv_p = 0.$$

Let us assume that $c_i \neq 0$ for some i. Adding $-(c_iv_i) = (-c_i)v_i$ to both sides of this equation we get

$$c_1v_1 + \cdots + c_{i-1}v_{i-1} + c_{i+1}v_{i+1} + \cdots + c_pv_p = (-c_i)v_i.$$

(Again, we interpret the left hand side as zero if $p = 1$.) Multiplying both sides by $-c_i^{-1}$ which is possible because $c_i \neq 0$ we get

$$(-c_i^{-1})(-c_i)v_i = (-c_i^{-1})c_1v_1 + \cdots + (-c_i^{-1})c_{i-1}v_{i-1} + (-c_i^{-1})c_{i+1}v_{i+1} + \cdots + (-c_i^{-1})c_pv_p$$

and observing that $(-c_i^{-1})(-c_i) = 1$ (cf. Prop 1.7) we have expressed $1v_i = v_i$ as a linear combination of the elements in L'.

Summarizing, we have shown:

3.21 Lemma. *Let $L = (v_1, v_2, \dots, v_p)$ be a list of $p > 0$ vectors in V. Then $\mathrm{Span}(L) = \mathrm{Span}(L')$ where L' is the list obtained from L by removing v_i, if and only if there are elements $c_1, c_2, \dots, c_p \in \mathbb{F}$ with $c_i \neq 0$ such that*

$$c_1v_1 + c_2v_2 + \cdots + c_pv_p = 0.$$

To formalize this situation we make the following

3.22 Definition. An ordered list $(v_1, v_2, \dots, v_p)$ of vectors $v_1, v_2, \dots, v_p \in V$ is called *linearly dependent* if there are scalars $c_1, c_2, \dots, c_p \in \mathbb{F}$ *not all zero* such that

$$c_1v_1 + c_2v_2 + \cdots + c_pv_p = 0.$$

Such a formula is called a (*linear*) *dependence relation*. We also call the vectors $v_1, v_2, \dots, v_p$ linearly dependent if the list $(v_1, v_2, \dots, v_p)$ is.

The list $(v_1, v_2, \dots, v_p)$ is called *linearly independent* if it is not linearly dependent. In other words, it is linearly independent if

$$c_1v_1 + c_2v_2 + \cdots + c_pv_p = 0$$

implies that $c_1 = c_2 = \cdots = c_p = 0$. Thus, $(v_1, v_2, \dots, v_p)$ is linearly independent if and only if there is one and only one way to write 0 as a linear combination of the v_i: $0 = 0v_1 + 0v_2 + \cdots + 0v_p$. If this is the case we also say the vectors $v_1, v_2, \dots, v_p$ are linearly independent.

A set[6] $S \subseteq V$ is linearly dependent, if there exist distinct elements $v_1, v_2, \dots, v_p \in S$ such that $(v_1, v_2, \dots, v_p)$ is linearly dependent.

A set $S \subseteq V$ that is not linearly dependent is linearly independent.

Note that it follows from the definition that the empty set is always a linearly independent subset of a vector space V: it is not linearly dependent because there exist *no* elements let alone linearly dependent ones.

3.23 Examples. The following lists or sets of vectors are always linearly dependent regardless of the vector space V in question.

a. $\{0\}$ is always linearly dependent.

   Indeed, $1 \cdot 0 = 0$ is a dependency relation because in the field $\mathbb{F}$, $1 \neq 0$.

b. Any set or list containing 0 is linearly dependent.

   For a set this is clear because by definition every set containing a linearly dependent subset is linearly dependent. Suppose $(v_1, v_2, \dots, v_n)$ is a list such that $v_i = 0$. Then putting $c_i = 1$ and $c_j = 0$ for all $j \neq i$ defines a relation of linear dependence.

   In particular, V itself is a linearly dependent set.

c. Any list of vectors containing duplicates is linearly dependent.

   Let $(v_1, v_2, \dots, v_n)$ be a list of vectors and suppose $v_i = v_j$ (where $i \neq j$. Put $c_i = 1$ and $c_j = -1$ and $c_k = 0$ for $k \neq i, j$. Then this gives a dependence relation.

When is a single vector $v \in V$ linearly dependent or independent? If v is linearly dependent, then there is a nonzero $c \in \mathbb{F}$ such that $cv = 0$. We already observed that this means $v = 0$ and hence $v = 0$, which we know is linearly dependent. It follows that if $v \neq 0$, then v is linearly independent.

> A single vector $v \in V$ is linearly dependent if and only if it is the zero vector.

What about two vectors? Let $v, w \in V$ be two vectors. Then v, w linearly dependent means that there are c_1, c_2 not both zero such that

$$c_1 v + c_2 w = 0 \quad \text{or, equivalently,} \quad c_1 v = -c_2 w \tag{3.14}$$

[6] One might wonder why we defined linear dependence for ordered sets. The reason is the following: we want to have a way of expressing that if we have two vectors v, w that they are linearly dependent if they are equal. However, if $v = w$, then the set $\{v, w\}$ is equal to the set $\{v\}$ and hence linearly independent if $v \neq 0$. But the ordered list (v, v) is different from the list (v). This is not a terribly important distinction, but it simplifies the language occasionally.

We may assume that c_1 is not zero (otherwise, we exchange the order of v and w). Then $c_1^{-1}(c_1v) = c_1^{-1}(-c_2w) = (-c_2/c_1)w$, hence $v = (-c_2/c_1)w$. In other words $v = cw$ for a suitable $c \in \mathbb{F}$. Consversely, if $v = cw$ then clearly $1v + (-c)w = 0$ and hence v, w are linearly dependent. (Of course, if $c_2 \neq 0$ above, we may conclude that $w = cv$ for suitable $c \in \mathbb{F}$.)

> Two vectors $v, w \in V$ are linearly dependent if and only if one of them is a multiple of the other.

If both are nonzero, an equivalent statement is to say that they both span the same line through the origin.

3.24 Example. If V is a vector space, then a *plane* in V is a subspace $E \subseteq V$ that is the span of two linearly independent vectors.

Even if V is a (possibly very large) vector space, for instance $\mathbb{R}^n$, many questions regarding two vectors v, w may be dealt with by simply focussing the attention onto the plane that is spanned by them (if they are linearly independent).

3.25 Example. Here is a more involved example: In $\mathbb{F}^n$, the vectors $e_1, e_2, \dots, e_n$ are linearly independent. Indeed, suppose $c_1e_1 + c_2e_2 + \cdots + c_ne_n = 0$. Then observe that

$$c_1e_1 + \cdots + c_ne_n = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = 0$$

if and only if all $c_i = 0$.

3.26 Example. In $\mathbb{R}^3$, the three vectors

$$\begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$$

are linearly independent.

Based on the previous example, here is a simple but crucial observation:

3.27 Fact. Let $v_1, v_2, \dots, v_k$ be vectors in $\mathbb{F}^n$. Then $(v_1, v_2, \dots, v_k)$ is linearly independent if and only if the equation $AX = 0$ has only the trivial solution $X = 0$ where $A = [\, v_1 \, v_2 \, \dots \, v_k \,]$.

In particular, if $k > n$ the vectors $v_1, v_2, \dots, v_k$ are always linearly dependent.

We need only to remark that the nonzero solutions of $AX = 0$ are precisely the vectors

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

for which $x_1 v_1 + \cdots + x_k v_k = 0$ is a dependence relation.

If $k > n$ we know by Theorem 2.13, that the homogeneous equation $AX = 0$ always has a non-trivial solution $X \neq 0$ because A has less rows than columns. □

3.3.2 Problem. Show that if an ordered list L is linearly independent, then so is every reordering of L.

3.3.1. Subspaces of $\mathbb{R}^n$ where $n = 1, 2, 3$

This subsection should be viewed as an extended example. We now have all the tools necessary to talk about different types of subspaces of $\mathbb{R}^n$ for small n.

Let us first consider $n = 1$. This is the least interesting case: indeed, $\{0\}$ and $\mathbb{R}$ are subspaces of $\mathbb{R}$. And that's it: if $v \in \mathbb{R}$ is any nonzero element, than we know that we can write every other element of $\mathbb{R}$ as cv for some $c \in \mathbb{R}$ and hence $\operatorname{Span}(v) = \mathbb{R}$.

A more interesting case is $n = 2$. Here we have three classes of subspaces: Again, $\{0\}$ and $\mathbb{R}^2$ are subspaces. Also, if $v \in \mathbb{R}^2$ is nonzero, then $L = \operatorname{Span}(v)$ is a subspace, called a *line* (through the origin and v). Note that $L \neq \mathbb{R}^2$. Indeed, we have seen that any two elements of L are linearly dependent. However, $\mathbb{R}^2$ contains linearly independent pairs, namely e_1, e_2, for instance. Now let $S \subseteq \mathbb{R}^2$ be any subspace that contains two linearly independent elements. Then $S = \mathbb{R}^2$. One can do this similarly as in Example 2.7. Indeed, let v_1, v_2 be linearly independent. $\operatorname{Span}(v_1)$ and $\operatorname{Span}(v_2)$ are lines that are not parallel (and intersect in 0 only). As in 2.7 we can show that any $v \in \mathbb{R}^2$ is the vertex of a parallelogram whose other vertices are 0 and a point in $\operatorname{Span}(v_1)$ and $\operatorname{Span}(v_2)$ each (except for the cases where $v \in \operatorname{Span}(v_1) \cup \operatorname{Span}(v_2)$). It follows that $v \in \operatorname{Span}(v_1, v_2)$.

An algebraic proof would be to observe that the matrix $A = [\, v_1\, v_2\, v\,]$ has more columns than rows. Therefore $AX = 0$ has a non-trivial solution and consequently (v_1, v_2, v) is linearly dependent. Pick any dependence relation $c_1 v_1 + c_2 v_2 + cv = 0$. Since (v_1, v_2) is linearly independent, we must have $c \neq 0$. Hence $\operatorname{Span}(v_1, v_2)$ contains v.

The subspaces of $\mathbb{R}^2$ are hence

a. $\{0\}$

b. The lines through the origin.

c. $\mathbb{R}^2$

Note that the three cases are separated by the number of generators minimally needed.

Finally, what are the subspaces of $\mathbb{R}^3$? Again, we have $\{0\}$, $\mathbb{R}^3$, and lines $L = \operatorname{Span}(v)$ for nonzero $v \in \mathbb{R}^3$. We also have a new class of subspaces, the *planes*: A subspace $E \subseteq \mathbb{R}^2$ is called a plane, if it is generated by a linearly independent list (v_1, v_2). Notice that a plane is never a line: indeed, any two elements in a line are linearly dependent. Furthermore, if W is a subspace that contains three linearly independent elements w_1, w_2, w_3, then $W = \mathbb{R}^3$. This can be shown as above, by showing that any four elements in $\mathbb{R}^3$ are linearly dependent. We obtain the following classificatoin of subspaces of $\mathbb{R}^n$:

a. $\{0\}$

b. Lines of the form $\mathrm{Span}(v)$ $(v \neq 0)$

c. Planes of the form $\mathrm{Span}(v_1, v_2)$ ($(v_1, v_2$ linearly independent).

d. $\mathbb{R}^3$.

We have cheated a little: a priori, there could be a sort of pathological subspace that does not fit into any of the three categories. However, this is not the case: Let $W \subseteq \mathbb{R}^3$ be any subspace. If it does not contain any three linearly independent elements (that is, if it isn't $\mathbb{R}^3$), then any three elements are linearly dependent. If it contains two linearly independent elements, then, it must be a plane (spanned by these elements). Otherwise, any two elements are linearly dependent, and it follows it is either zero or a line.

Again, we observe that the subspaces are somewhat separated by the number of generators they minimally require. However, we run into a small problem: why is $\mathbb{R}^3$ not a plane? That is, why can't $\mathbb{R}^3$ be generated by two elements? Geometrically this should be clear, and with some effort we could show that indeed this is not the case. We will make precise the ideas needed in the next section.

It should also be noted that nothing here is special about $\mathbb{R}$. The exact same reasoning applies to any vector space of the form $\mathbb{F}^n$.

3.3.3 Problem. This is a longer problem. The problem appeared in similar form on the homepage of the German federal intelligence service (as a test for prospective applicants for their mathematical positions).

The rough structure is as follows: Suppose you are the assistant to the CEO of a large company. The board of directors has 16 members and your task is as follows: organize meetings of the directors in groups of four such that

- every director meets with every other director in a meeting;
- no two directors meet twice;
- at every given time, everybody attends a meeting.

We refer to this below as the "meeting problem."

a. A *line* in a vector space V is a subset L of the form $L = \{u + tv \mid t \in \mathbb{F}\}$ where u, v are some elements of V (depending only on L) and $v \neq 0$. The vector v is also called a direction vector for L (any nonzero multiple of v is also a direction).

   Let $L, M \subseteq V$ be two lines. Show that one and only one of the three statements is true: $L \cap M$ consists of a single element; $L \cap M = \emptyset$; $L = M$.

   In the second (and third) case we say L and M are parallel.

b. Show that two lines L, M are parallel if and only if they have the same direction vector (up to multiplication by a nonzero scalar).

c. Let $\mathbb{F}_q$ be a finite field with q elements. Show that any line has q elements. Fix any nonzero vector $v \in \mathbb{F}^2$. How many parallel lines are there in $\mathbb{F}^2$ with direction vector v?

(*Hint:* Pick a line L that is not parallel to v, i.e. such that v is not a direction vector. If M is a line with direction vector v, M is determined by $M \cap L$.)

d. How many lines are there in $\mathbb{F}_q^2$?

e. Suppose there exists a field with 4 elements. Can you solve the meeting problem?

f. Show that there exists a field with 4 elements. Proceed as follows: Let

$$J = \begin{bmatrix} & 1 \\ 1 & 1 \end{bmatrix} \in M_2(\mathbb{F}_2).$$

Show that $J^2 + J + I = 0$ and show that the set

$$\mathbb{F}_4 = \{aI + bJ \mid a, b \in \mathbb{F}_2\}$$

together with matrix multiplication and addition is a field with 4 elements.

3.3.2. Bases

Lemma 3.21 shows that if a vector space V is generated by a linearly independent list $L = (v_1, v_2, \ldots, v_n)$, then no element of L can be omitted from the generating system: if $M \subsetneq L$ is a sublist then $\operatorname{Span}(M) \subsetneq \operatorname{Span}(L)$: suppose M does not contain v_i, say; then after relabelling, we may assume that $i = n$ and so $M \subseteq (v_1, v_2, \ldots, v_{n-1})$ which implies $\operatorname{Span}(M) \subseteq \operatorname{Span}(v_1, v_2, \ldots, v_{n-1})$. The latter, however, is strictly smaller than $\operatorname{Span}(L)$ by the lemma.

This observation is at the heart of the following procedure to arrive at an *optimal* generating set: Suppose we have a finitely generated vector space V, that is, $V = \operatorname{Span}(L)$ for a finite list L. Let $v_1, v_2, \ldots, v_n$ be the elements of L. We can now arrive at a *minimal* generating set by checking for $i = 1, 2, \ldots, n$ whether $v_i \in \operatorname{Span}(v_1, v_2, \ldots, \hat{v}_i, \ldots, v_n)$ (the $\hat{}$ indicating that v_i is to be omitted in that list). If this is the case, we may delete v_i from L and still have a generating set. We can then repeat the whole process with the remaining $n-1$ vectors. In the end, we will end up in one of two possible cases: We could arrive at a list $L' = (w_1, w_2, \ldots, w_k)$ of generators that cannot be shrunk further: If we delete any of the w_i from L' the resulting set will no longer be a generating set. According to Lemma 3.18 this means L' is linearly independent.

The other possible outcome is that $V = \{0\}$ and hence all elements of L were zero in the first place. We will then arrive at the set $L' = \{0\}$ but cannot delete the last remaining L *unless* we have the convention that the empty set is a generating set for $V = \{0\}$. This is the reason why we adopted the convention that $\operatorname{Span}(\emptyset) = \{0\}$. That way we won't have to distinguish the cases that vector spaces are zero or non-zero.

This whole discussion motivates the following definition:

3.28 Definition. Let V be a vector space. A *basis* is a linearly independent ordered list of generators. Thus, $\mathcal{B} \subseteq V$ is a basis if and only if $\mathcal{B}$ is linearly independent and $\mathrm{Span}(\mathcal{B}) = V$. We write

$$\mathcal{B} = (v_1, v_2, \dots, v_n)$$

if $v_1, v_2, \dots, v_n$ are the elements of $\mathcal{B}$ (in order).

By convention, the empty set is a basis for $V = \{0\}$.

What is a basis good for?

3.29 Example. Let us look at the null space of 1×3 matrix $A = [2\,3\ -1]$. Let us define $f\colon \mathbb{R}^3 \to \mathbb{R}$ as $f(X) = AX$. Then $f(x, y, z) = 2x + 3y - z$. We also write $\mathcal{N}(f)$ for $\mathcal{N}(A)$. Now $(x, y, z) \in \mathcal{N}(f)$ if and only if $z = 2x + 3y$. So for $x, y \in \mathbb{R}$, the vector $(x, y, 2x + 3y)$ is an element of $\mathcal{N}(f)$ and all elements are of this form. Following our convention to write vectors usually as column vectors we find that the solutions are of the form

$$c_1 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} \tag{3.15}$$

and for each solutions the coefficients c_1, c_2 are uniquely determined. Also, the two vectors are clearly linearly independent and it follows that

$$\mathcal{B} = \left(\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} \right) \tag{3.16}$$

is a basis for $\mathcal{N}(f)$. If we want to write down an arbitrary solution, we just pick coefficients c_1, c_2. In the language of analytic geometry $\mathcal{N}(f)$ is a plane in three space. We have two "degrees of freedom" corresponding to two independent solutions. The basis $\mathcal{B}$ allows us to associate an element of $\mathbb{R}^2$ to each point of the plane $\mathcal{N}(f)$. We can think of $\mathcal{B}$ as choosing coordinates on $\mathcal{N}(f)$.

3.30 Theorem. *Let V be a vector space with basis $\mathcal{B} = (v_1, v_2, \dots, v_n)$. If $V \neq 0$, then for each $v \in V$ there is one and only one way to write v as a linear combination*

$$v = x_1 v_1 + x_2 v_2 + \cdots + x_n v_n$$

with $v_i \in \mathcal{B}$ and $x_i \in \mathbb{F}$. That means, if also $v = y_1 v_1 + \cdots + y_n v_n$, then $x_i = y_i$ for all i.

Proof. If $\mathcal{B}$ is the empty set there is nothing to show (as then $V = \{0\}$). Otherwise, since $\mathcal{B}$ is a spanning set, every $v \in V$ is a linear combination of elements in $\mathcal{B}$. So suppose v can be written as

$$x_1 v_1 + x_2 v_2 + \cdots + x_n v_n = y_1 v_1 + y_2 v_2 + \cdots + y_n v_n.$$

An easy algebraic manipulation of this equation shows that

$$(x_1 - y_1)v_1 + (x_2 - y_2)v_2 + \cdots + (x_n - y_n)v_n = 0. \tag{3.17}$$

Since $\mathcal{B}$ is a basis, this forces all coefficients in (3.17) to be zero:$x_1 - y_1 = x_2 - y_2 = \cdots = x_n - y_n = 0$. Hence the claim. □

In particular, for a vector space V with a finite basis $\mathcal{B} = (v_1, v_2, \dots, v_n)$ the significance of Theorem 3.30 rests on the following consequence:

> The map $\mathbb{F}^n \to V$ that associates to $X \in \mathbb{F}^n$ the vector
>
> $$\mathcal{B}X := x_1v_1 + x_2v_2 + \cdots + x_nv_n \tag{3.18}$$
>
> is a one to one correspondence. For $v \in V$, the unique vector X such that $\mathcal{B}X = v$ is often denoted $[v]_{\mathcal{B}}$ and called the *coordinate vector of v with respect to $\mathcal{B}$*.

As soon as we choose a basis, we can describe the elements of V by the vectors of $\mathbb{F}^n$ (with which we are probably more familiar with) very explicitly. Also, this is the very reason why we want $\mathcal{B}$ to be *ordered*. To associate a column vector X to a vector in V by means of $\mathcal{B}$, we have to associate the ith position in X to a certain vector in $\mathcal{B}$. But this is nothing but an ordering.

3.31 Example. Example 3.25 provides a sort of tautological example: Suppose $V = \mathbb{F}^n$. Then $\mathcal{E} = (e_1, e_2, \dots, e_n)$ is a basis. (Example 3.25 showed actually both, that $\mathcal{E}$ is linearly independent and that $\operatorname{Span}(\mathcal{E}) = \mathbb{F}^n$.) For $v \in \mathbb{F}^n$ we have $\mathcal{E}v = v$, so $[v]_{\mathcal{E}} = v$. This makes this particular basis a little special; it is therefore often referred to as the *standard basis* of $\mathbb{F}^n$.

The ordering of the basis is crucial. For instance, suppose instead of $\mathcal{E}$ we chose $\mathcal{B} = (e_2, e_2, \dots, e_n, e_1)$. Then $\mathcal{B}$ is still a basis, but now if

$$v = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}, \quad [v]_{\mathcal{B}} = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_n \\ c_1 \end{bmatrix} \neq v.$$

3.32 Example. Let $V = \mathcal{P}(\mathbb{R})$. For $f \in V$ such that $f(x) = a_0 + a_1x + \cdots + a_nx^n$ with $a_n \neq 0$ we define $\deg f = n$. The degree of the 0-function (defined as $0(x) = 0$ for all $x \in \mathbb{R}$) is $-\infty$ by convention[7].

[7] This is to interpreted completely formally; no intrinsic meaning is to be given to $-\infty$ other than that $-\infty < n$ for all integers n; and that $-\infty + n = -\infty$ for all $n \in \mathbb{Z}$. This formal rule makes sure that the usual rules like $\deg(fg) = \deg f + \deg g$ and $\deg(f + g) \leq \max\{\deg f, \deg g\}$ remain valid also if f or g (or both) are equal to 0.

Let $W = V_n := \{f \in V \mid \deg f \leq n\}$. Then W is a subspace. Moreover, the monomials form a basis: $\mathcal{B} = (1, x, x^2, \ldots, x^{n+1})$. So for

$$X = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

we have

$$\mathcal{B}X = a_0 + a_1 x + \cdots + a_n x^n$$

and so

$$[f]_{\mathcal{B}} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

3.3.4 Problem. Let us keep the notation introduced in the previous example. Let $d_0, d_1, \ldots, d_n \in \mathbb{R}$. Find $f \in W$ such that $f(0) = d_0, f(1) = d_1, \ldots, f(n) = d_n$.

This is a linear algebra problem: indeed, the conditions $f(i) = d_i$ are linear equations in the coordinate vector of f with respect to $\mathcal{B}$. This gives a system with augmented matrix

$$\begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & d_0 \\ 1 & 1 & 1^2 & \ldots & 1^n & d_1 \\ 1 & 2 & 2^2 & \ldots & 2^n & d_2 \\ \vdots & \vdots & \ldots & \vdots & \vdots & \\ 1 & n & n^2 & \ldots & n^n & d_n \end{bmatrix}$$

where the ith row represents the equation $f(i) = d_i$. One can show that the coefficient matrix of this system is invertible. Hence for each choice of d_i there is a *unique* $f \in W$ such that $f(i) = d_i$.

An easier way to solve the problem is to choose a clever basis for W. Let $g_i \in W$ be functions that satisfy $g_i(j) = \delta_{ij}$ $(j = 0, 1, \ldots, n)$. Indeed, g_i exists, for instance we could pick

$$g_i = \prod_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{x-j}{i-j} = \frac{x-0}{i-0}\frac{x-1}{i-1}\cdots\frac{x-i-1}{i-i-1}\frac{x-(i+1)}{x-i+2}\cdots\frac{x-n}{i-n}.$$

$\mathcal{B}' = (g_0, g_1, \ldots, g_n)$ is a basis for W. Moreover, for each f we have

$$[f]_{\mathcal{B}'} = \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(n) \end{bmatrix}$$

Let us first check that $\mathcal{B}'$ is linearly independent. Suppose $0 = c_0g_0 + c_1g_1 + \cdots + c_ng_n$. Then for each i,

$$0 = c_0g_0(i) + c_1g_1(i) + \cdots + c_ng_n(i) = c_ig_i(i) = c_i.$$

Unfortunately, to show that $\mathcal{B}'$ actually spans W, one has to work a little harder. While it is possible to show directly, that $\operatorname{Span}(\mathcal{B}') = W$, there is a very helpful indirect route: we will show below, that in V *any* linearly independent list of $n + 1$ vectors is a basis.

Here we use a direct approach: Let $f \in W$. Then f is uniquely determined by its values at $0, 1, 2, \ldots, n$. Indeed, suppose $g \in W$ and $g(i) = f(i)$ for $i = 0, 1, \ldots, n$. Then also $f - g \in W$ is a polynomial function of degree at most n, and therefore, if $f - g \neq 0$, it has at most n zeros[8]. Since by assumption $f - g$ has $n + 1$ zeros it follows that $f - g = 0$ and hence $f = g$.

Now, to see that $\operatorname{Span}(\mathcal{B}') = W$, let $f \in W$ be arbitrary. Let $g = f(0)g_0 + f(1)g_1 + \cdots + f(n)g_n$. Then $g \in W$ and $g(i) = f(i)$ $(i = 0, 1, \ldots, n)$. By what we just said, it follows that $g = f$. Hence $f \in \operatorname{Span}(\mathcal{B}')$.

It follows that $\mathcal{B}'$ is a basis. To solve the original problem, to find f for given $d_0, d_1, \ldots, d_n$, pick $f = d_0g_0 + d_1g_1 + \cdots + d_ng_n$.

In general, the important observation here is that for any $f \in W$, its coordinate vector with respect to $\mathcal{B}'$ is simply,

$$[f]_{\mathcal{B}'} = \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(n) \end{bmatrix}$$

The previous example indicates that sometimes it is very convenient to be able to choose a basis at will. Even in the case $V = \mathbb{F}^n$ there are many instances where we want to choose a basis other than the standard basis.

3.33 Example. Example 3.29 generalizes as follows: Let A be any matrix in reduced row echelon form. To find a basis for $\mathcal{N}(A)$ proceed as follows: Let $f_1, f_2, \ldots, f_k$ be the indices of the free variables. Then define $v_i \in \mathcal{N}(A)$ to be the column vector obtained by putting $x_{f_i} = 1$ and all other free variables equal to zero and computing the basic variables accordingly. Thus, every element of $\mathcal{N}(A)$ then has the form

$$v = x_{f_1}v_1 + x_{f_2}v_2 + \cdots + x_{f_k}v_k. \tag{3.19}$$

Moreover, $\mathcal{B} = (v_1, v_2, \ldots, v_k)$ is linearly independent: indeed, since v_i has a 1 at position f_i and since v_j has a zero there if $i \neq j$, we conclude that in (3.19), the entry at position f_i is precisely x_{f_i}. From this it follows easily that $v = 0$ if and only if all $x_{f_i} = 0$.

It follows that $\mathcal{B}$ is a basis for $\mathcal{N}(A)$.

[8] One can prove this using calculus, or the fact that if $f(a) = 0$, then $f(x) = (x - a)h(x)$ where h is a polynomial function of smaller degree. By induction, it then follows that h has at most $n - 1$ zeros, and so f has at most n zeros.

3.34 Example. The previous example is instructive also for finding a basis for the column space of A: Let $A_1, A_2, \dots, A_n$ be the columns of A.

By Lemma 3.21, we can eliminate A_i from this list, if there is a dependence relation that has a nonzero coefficient at A_i. What are the dependence relations? They are precisely the nonzero solutions of $AX = 0$.

By the previous example, for every free variable there is an element v_i in $\mathcal{N}(A)$, that has a nonzero coefficient at precisely one position of a free variable and (possibly) only basic variables.

Thus $Av_i = 0$ is a relation among the A_i that expresses A_{f_i} in terms of the columns that correspond to basic (pivot) variables. It follows that if $p_1, p_2, \dots, p_r$ are the column indices of the pivots in a row echelon form for A, then $\mathcal{B} = (A_{p_1}, A_{p_2}, \dots, A_{p_r})$ spans $\mathrm{Col}(A)$: indeed $\mathrm{Span}(\mathcal{B})$ contains every A_i and hence also $\mathrm{Span}(A_1, A_2, \dots, A_n) = \mathrm{Col}(A)$. So $\mathrm{Span}(\mathcal{B}) = \mathrm{Col}(A)$.

Convince yourself that this is true and prove, that $\mathcal{B}$ is a basis (if there are no pivots, then $A = 0$, and the empty set is a basis).

Warning: Even though the pivot indices are determined by the reduced row echelon form of A, the actual columns are the ones of A: the column space of the row echelon form differs in general from $\mathrm{Col}(A)$.

3.3.5 Problem. Let A be an $m \times n$ matrix with entries in $\mathbb{F}$. Define $\mathrm{Row}(A) \subseteq M_{1\times n}(\mathbb{F})$ as the span of the rows of A. For this problem, we may identify row vectors with elements of $\mathbb{F}^n$, so you could think of $\mathrm{Row}(A)$ as a subspace of $\mathbb{F}^n$.

Let A' be a reduced row echelon form of A. Prove:

a. Show that $\mathrm{Row}(A) = \mathrm{Row}(A')$.

b. Show that the nonzero rows of A' form a basis for $\mathrm{Row}(A') = \mathrm{Row}(A)$.

Using these observations, we can find a basis for $\mathrm{Col}(A)$ as well: since $\mathrm{Col}(A)$ corresponds naturally to $\mathrm{Row}(A^T)$, we use the above procedure to find a basis $b_1, b_2, \dots, b_k$ for $\mathrm{Row}(A^T)$. Then the b_i are row vectors, and $b_1^T, b_2^T, \dots, b_k^T$ will be a basis for $\mathrm{Col}(A)$.

3.3.6 Problem. The previous two problems give two methods of finding a basis for $\mathrm{Span}(v_1, v_2, \dots, v_p)$ where $v_i \in \mathbb{F}^n$: we simply compute a basis for $\mathrm{Col}(A)$ where A is the matrix with columns $v_1, v_2, \dots, v_p$. Which one is to prefer?

Convince yourself that the two bases we construct using the above problems have the following properties:

- If we want a basis for $\mathrm{Col}(A)$ such that the basis vectors are among the original vectors $v_1, v_2, \dots, v_p$, we pick the first method.
- The second method has the following advantage: let $p_1, p_2, \dots, p_k$ be the indices pivot columns of the reduced row echelon form of A^T. Then if $v \in \mathrm{Span}(v_1, v_2, \dots, v_p)$, the coordinate vector of v with respect to the basis produced by the second method is exactly the vector formed by the entries of v at positions $p_1, p_2, \dots, p_k$. Thus, we can read it off immediately!

3.35 Example. As an illustration, let us consider $\mathrm{Col}(A)$ where $A \in M_{3\times 5}(\mathbb{R})$ is defined as

$$A = \begin{bmatrix} 1 & 2 & 2 & 4 & 2 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 1 & 1 \end{bmatrix}$$

Now, the first method, yields the following: the reduced row echelon form of A is

$$\begin{bmatrix} 1 & 2 & 2 & 4 & 2 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 2 & 2 & 4 & 2 \\ 0 & 0 & 0 & -2 & -1 \\ 0 & 0 & 1 & -3 & -1 \end{bmatrix} \\ \to \begin{bmatrix} 1 & 2 & 2 & 4 & 2 \\ 0 & 0 & 1 & -3 & -1 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \end{bmatrix} \to \begin{bmatrix} 1 & 2 & 0 & 0 & 5 \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & \frac{1}{2} \end{bmatrix} \tag{3.20}$$

The pivot columns are therefore columns 1, 3, 4. Thus, a basis for $\mathrm{Col}(A)$ is given by the columns 1, 3, 4, of the original matrix A:

$$\mathcal{B} = \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix} \right)$$

Now let us consider the second method.

$$A^T = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 2 & 2 & 3 \\ 4 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

Its row echelon form is

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 2 & 2 & 3 \\ 4 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -2 & -3 \\ 0 & -1 & -1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.21}$$

The basis provided by this method is given by the transposes of the nonzero rows of this row echelon form, which means the basis is

$$\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right)$$

which is the standard basis of $\mathbb{R}^3$. Thus, here we learn immediately that $\mathrm{Col}(A) = \mathbb{R}^3$ (we will show below that $\mathbb{R}^3$ itself is the only subspace of $\mathbb{R}^3$ that has a basis consisting of three elements; so this is no surprise).

It follows that the two methods give radically different results; both have their advantages.

3.3.3. The dimension of a vector space

Intuitively speaking if V is a vector space with a finite basis $\mathcal{B} = (v_1, v_2, \dots, v_n)$ we would like to think of n as a property of V: it should somehow describe the degrees of freedom we have in choosing an element of V. We have seen some indications why the number of elements in a basis might be an interesting quantity in 3.3.1.

Remark. When solving a system $AX = 0$, our method using the reduced row echelon form produces a basis for the solution space: the coefficient vectors of the free variables are linearly independent (they have 1 as at mutually exclusive positions).

We would like to think of the number of free variables as an *invariant* of the solution space; no matter which method for solving a system of linear equations we apply, we should always have the same number of free parameters. However, this is not an obvious fact.

3.36 Example. As a further example, why one should believe that any two bases of a vector space have the same number of elements consider the case of a finite field $\mathbb{F}$ with q elements. By Theorem 3.30, a vector space V with a basis $(v_1, v_2, \dots, v_n)$ is in bijection to $\mathbb{F}^n$ which is a set with q^n elements (q choices for each position). Thus, if $(w_1, w_2, \dots, w_m)$ is another basis for V, then $q^m = q^n$ which forces $m = n$ since $q > 1$. Thus, for a finite field, the number of elements in a basis is independent of the basis.

We begin our discussion with the following reformulation of Lemma 3.21:

3.37 Lemma. *Let $L = (v_1, v_2, \dots, v_p) \subseteq V$ be a linearly independent list of vectors. Then $v \in \operatorname{Span}(L)$ if and only if $(v_1, v_2, \dots, v_p, v)$ is linearly dependent.*

Proof. The only if part is a direct consequence of Lemma 3.21.

So let us consider the if part: Let $L' = (v_1, v_2, \dots, v_p, v)$. Then $\operatorname{Span}(L) = \operatorname{Span}(L')$ if $(v_1, \dots, v_p, v)$ is linearly dependent (and hence $v \in \operatorname{Span}(L)$). Indeed, in any dependence relation

$$c_1 v_1 + c_2 v_2 + \cdots + c_p v_p + cv = 0$$

we must have $c \neq 0$. Otherwise L were linearly dependent. But if $c \neq 0$, then we can solve for v and express v as a linear combination of L as in the proof of Lemma 3.21. □

We will now prove that any two (finite) bases of a vector space V have the same number of elements. Our main tool is the following crucial result:

3.38 Lemma (Exchange Lemma)**.** *Let V be a vector space spanned by elements $v_1, v_2, \dots, v_n$, say.*

Let $v = c_1 v_1 + \cdots + c_n v_n \in V$ be a vector.

If $c_i \neq 0$, then

$$V = \operatorname{Span}(v_1, v_2, \dots, v_{i-1}, v, v_{i+1}, \dots, v_n).$$

Proof. The most important observation is that if $c_i \neq 0$, then

$$v_i = -c_i^{-1}(c_1 v_1 + \cdots + c_{i-1}v_{i-1} - v + c_{i+1}v_{i+1} + \cdots + c_n v_n). \tag{3.22}$$

This shows that $v_i \in \operatorname{Span}(L)$ where $L = (v_1, v_2, \ldots, v_{i-1}, v, v_{i+1}, \ldots, v_n)$. But also $v_j \in \operatorname{Span}(L)$ for all $j \neq i$ and hence $\operatorname{Span}(v_1, v_2, \ldots, v_n) = V \subseteq \operatorname{Span}(L)$ by Lemma 3.18. It follows that L is a generating set for V. □

This is all that is needed in order to show that the number of elements in a basis of a vector space V is a property of V rather than an individual basis.

Recall from 3.27 that in $\mathbb{F}^n$ any list of $n+1$ or more elements is linearly independent. A more general version of this fact is the following:

3.39 Theorem. *Let V be a vector space generated by finitely many elements $(v_1, v_2, \ldots, v_n)$, say.*

If $(w_1, w_2, \ldots, w_k)$ is a linearly independent list of elements of V, then $k \leq n$.

Proof. Let $L = (v_1, v_2, \ldots, v_n)$ and $M = (w_1, w_2, \ldots, w_k)$. If $n = 0$ (that is, if L is empty), then $V = \{0\}$, so any number of elements of V are linearly dependent. Hence $k = 0$ as well.

We may therefore assume that $n > 0$. Suppose precisely $m \geq 0$ of the elements of M are also elements of L. By reordering if necessary, we may assume that $w_1 = v_1, w_2 = v_2, \ldots, w_m = v_m$. We will now show how to increase m by 1 if $k - m > 0$. In this case, $w_{m+1} \notin L$. We may write $w_{m+1} = c_1 v_1 + \cdots + c_m v_m$ for suitable $c_i \in \mathbb{F}$.

Claim: At least one c_i with $i > m$ must be nonzero.

Indeed, otherwise $c_{m+1} = c_{m+2} = \cdots = c_n = 0$ and

$$w_{m+1} = c_1 v_1 + \cdots + c_m v_m = c_1 w_1 + \cdots + c_m w_m,$$

contradicting the fact that M is linearly independent (cf. Lemma 3.21). This proves the claim.

So pick one such i (ie. $i > m$ and $c_i \neq 0$). By the Exchange Lemma, we can replace v_i by w_i in L, obtaining a new list of generators L' which has $m+1$ elements in common with M and still satisfies $V = \operatorname{Span}(L')$.

This process can be repeated as long as $k - m > 0$. Thus eventually, all elements of M must be elements of the newly created list L'. In particular, $n \geq k$. □

Remark. First of all note that strictly speaking we should have done an induction proof, namely, induction on the number $d = k - m$. The assertion then is, if L if has n elements and spans V and M is a linearly independent list of k elements that has d additional elements (not in L), then $|M| \leq n$.

If $d = 1$ we showed this in the above proof. Also, if the assertion is true for a particular value for d, then we showed above, that given a list M with $d+1$ elements not contained in L, we can construct a list L' that has n elements and spans V such that M has d elements in common with L' so the induction assumption applies to the pair M, L', and $|M| \leq n$.

This would be a formal induction proof of the theorem.

3.40 Corollary. *If $\mathcal{B}$ and $\mathcal{C}$ are bases for a vector space V, then $|\mathcal{B}| = |\mathcal{C}|$.*

Proof. Both $\mathcal{B}$ and $\mathcal{C}$ generate V and are linearly independent. The previous theorem therefore asserts that $|\mathcal{B}| \leq |\mathcal{C}|$ and $|\mathcal{C}| \leq |\mathcal{B}|$. □

This motivates the following definition.

3.41 Definition. Let V be a vector space with basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$. The uniquely determined integer n is called the *dimension* of V and denoted $\dim V$.

The empty set by convention is a basis for $V = \{0\}$ (it is after all a linearly independent set that spans V). So $\dim\{0\} = 0$.

If V does not have a (finite) basis, then we say $\dim V = \infty$.

3.42 Example. As expected $\dim \mathbb{R} = 1$ (the list with one element $(1_{\mathbb{R}})$ is a basis), $\dim \mathbb{R}^2 = 2$ and $\dim \mathbb{R}^3 = 3$. More generally,

$$\dim \mathbb{F}^n = n. \tag{3.23}$$

Indeed, the standard basis (Example 3.25) $\mathcal{E} = (e_1, e_2, \ldots, e_n)$ of $\mathbb{F}^n$ has exactly n elements.

$\dim M_{m\times n}(\mathbb{F}) = mn$. Here we may choose as a basis a list whose elements are precisely the mn matrix units e_{ij} (in any ordering).

An extremely important definition is now the following:

3.43 Definition. Let $A \in M_{m\times n}(\mathbb{F})$. Then the *rank* of A is defined as

$$\operatorname{rank} A = \dim \operatorname{Col}(A).$$

3.3.7 Problem. Show that the rank of A is exactly the number of pivots in a row echelon form. (*Hint:* Problem 3.34)

Deduce that $\operatorname{rank} A = \dim \operatorname{Row}(A)$ as well.

Remark. Recall from Example 3.33 that there is a one to one correspondence between a basis and the free variables. Thus the *nullity* of A, defined as $\operatorname{nullity}(A) = \dim \mathcal{N}(A)$, is exactly the number of free variables.

Combining this with our observation above, we find the following important result, sometimes referred to as the *Rank-Nullity Theorem*:

> Let $A \in M_{m\times n}(\mathbb{F})$, then
>
> $$\operatorname{rank}(A) + \operatorname{nullity}(A) = n.$$

Indeed, the number of pivot columns plus the number of free variables is exactly the number of all columns.

3.44 Example. The rank if the identity matrix I_n is n. In fact, the rank of any invertible $n \times n$ matrix is n (indeed, $\operatorname{nullity}(A) = 0$ if A is invertible).

Also,

$$\operatorname{rank} \begin{bmatrix} I_k & 0_{k,n-k} \\ 0_{m-k,k} & 0_{m-k,n-k} \end{bmatrix} = k$$

3.3.4. Finitely generated vector spaces

An obvious question that we will address now is whether every vector space has a basis. The answer is yes (if we allow for infinite bases), but that is far from being obvious. We will only address the case when V is finitely generated: Recall that a vector space V is *finitely generated* if V is generated by a finite list $S \subseteq V$.

3.45 Theorem. *Let V be a finitely generated vector space. Then every finite list of vectors that generates V contains a basis.*

In particular, every finitely generated vector space V has a basis, and $\dim V \neq \infty$.

Proof. The proof is essentially given by the informal algorithm in Section 3.3.2. To be precise, let L be a finite list of generators of V, which exists by assumption.

Let $L' \subseteq L$ be a maximal list of linearly independent elements. Maximal means the following: if M is any list contained in L that contains L', that is, $L' \subseteq M \subseteq L$ then $M = L'$ if M is linearly independent. In other words, no properly larger list than L' can be linearly independent.

If all elements of L are zero, this means L' is empty. Otherwise, L' is not empty. Indeed, if $M = (v)$ and $v \in L$ is nonzero, then M is linearly independent and $\emptyset \subsetneq M$ which shows that $\emptyset$ is not maximal.

Let $v \in L$ be arbitrary. If $v \in L'$ then $v \in \operatorname{Span}(L')$. Otherwise, (L', v) (meant as the list obtained by appending v to the list L') is a sublist of L, properly containing L'. Hence (L', v) is linearly dependent and consequently $v \in \operatorname{Span}(L')$ by Lemma 3.37. This shows that $L \subseteq \operatorname{Span}(L')$. By Lemma 3.18, also $\operatorname{Span}(L) = V \subseteq \operatorname{Span}(L')$. Hence L' is a linearly independent spanning set, which is a basis. □

From now on we usually say a vector space is *finite dimensional* instead of finitely generated to indicate it has a finite generating set.

Remark. The proof of Theorem 3.45 does not indicate how we actually can find the basis L'. As such it is a completely *non-constructive* proof. But we know how to find L': we successively eliminate elements from L that are not needed to span V (always using Lemma 3.21) until we arrive at a list that is linearly independent.

We now know that any generating set contains a basis. We can turn around this question and ask, if we start with linearly independent vectors $v_1, v_2, \dots, v_k$, can we find a basis containing them? The answer is yes. This is a very important result:

3.46 Proposition. *Let V be a finitely generated vector space. Suppose $(v_1, v_2, \dots, v_k)$ is a linearly independent list of vectors. This list can be extended to a basis of V.*

Proof. Let $\mathcal{B} = (w_1, w_2, \dots, w_n)$ be a basis for V which exists since V is finitely generated. We know that $n \geq k$ by Theorem 3.39. We will do induction on $d = n - k$.

Let $L = (v_1, v_2, \dots, v_k)$. If $d = 0$, then $k = n$. By Theorem 3.39, we know that (L, w_i), as a list with $n + 1$ elements, is linearly dependent. Thus, by Lemma 3.37, $w_i \in \operatorname{Span}(L)$ for $i = 1, 2, \dots, n$. By Lemma 3.18, then, $\operatorname{Span}(w_1, w_2, \dots, w_n) = \operatorname{Span}(\mathcal{B}) \subseteq \operatorname{Span}(L)$. By $\mathcal{B}$ is a basis, so this means that $V = \mathrm{Span}(\mathcal{B}) = \mathrm{Span}(L)$. Hence L is a linearly independent list of generators and thus a basis.

Now suppose the assertion is true for a particular value of d. We will now prove it also for $d+1$.

Thus, let L have k elements where $k = n-(d+1)$. Then $k < n$. If $k < n$, we may increase L by adding one of the w_i as follows: $\mathrm{Span}(L) \subsetneq V$ because L is not a basis by Corollary 3.40. Hence there must be i such that $w_i \notin \mathrm{Span}(L)$ (a subspace containing a list of generators for V is equal to V by Lemma 3.18).

By Lemma 3.37, then, the list $L' = (L, w_i)$ obtained by appending w_i is linearly independent.

Now L' has $k+1$ elements, and hence $d = n-(k+1)$. By the induction assumption, L' can be extended to a basis of V and thus also L can be extended to a basis. □

Let us now summarize the facts about generating sets and bases:

3.47 Theorem. *Let V be a vector space of dimension n.*

a. *Any generating set has at least n elements.*

b. *Any linearly independent set has at most n elements.*

c. *Any list of n linearly independent elements forms a basis.*

d. *Any generating set with n elements is a basis.*

Proof. The proof is now straight forward.

a. Any generating set contains a basis by Theorem 3.45. Any basis has n elements by Corollary 3.40.

b. This is an immediate consequence of Theorem 3.39.

c. If L is a list of n linearly independent elements, then it may be extended to a basis, which has n elements. Hence L is already a basis itself (we proved this in the proof of Proposition 3.46.

d. If L is any list of n generators, we have to show that L is linearly independent. But L contains a basis, and any basis has n elements, so L is a basis.

□

Remark. One can formalize the concept of dimension also for vector spaces that are not finitely generated. They still have a basis, and any two bases can be shown to be in bijection with each other. Thus, one can define the dimension of an arbitrary vector space to be the *cardinality* of a basis. However, this requires too much set theory so we won't discuss this any further.

3.48 Corollary. *Let V be a vector space of dimension n. Then any subspace W of V is again finite dimensional of dimension at most n.*

Proof. If W is finitely generated, then W has a finite basis. This basis has at most n elements by Theorem 3.39. It remains to observe that W is indeed finite dimensional.

We now successively construct a basis for W: if $W = \{0\}$ it is clearly finite dimensional and we are done. Otherwise, suppose we have a partial basis $\mathcal{B} = (w_1, w_2, \ldots, w_k)$ constructed. If $\mathcal{B}$ spans W we are done. Otherwise, there is some $w \in W$ that is not contained in $\mathrm{Span}(\mathcal{B})$. We add w as w_{k+1} and continue. Since more than n elements in V are linearly dependent, after at most n steps, this process must stop: we cannot find w in W that is not contained in $\mathrm{Span}(\mathcal{B})$ anymore, and hence $\mathrm{Span}(\mathcal{B}) = W$. □

We conclude this chapter by several problems to test the concepts. All vector spaces in the following are vector spaces over a given field $\mathbb{F}$.

3.3.8 Problem. Let V, W be vector spaces with bases $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ and $\mathbb{C} = (w_1, w_2, \ldots, w_m)$ respectively.

Recall the product space $V \times W$ (cf Example 3.9).

Prove that $\mathcal{D} = ((v_1, 0), (v_2, 0), \ldots, (v_n, 0), (0, w_1), (0, w_2), \ldots, (0, w_m))$ is a basis for $V \times W$. Conclude that $\dim(V \times W) = \dim V + \dim W$.

3.3.9 Problem. Let $U, W \subseteq V$ be subspaces of a vector space V. We define *the sum* $U + W$ of U and W as follows

$$U + W := \{v \in V \mid v = u + w \text{ for some } u \in U, w \in W\}.$$

a. Prove that $U + W$ is a subspace of V.

b. We say $U + W$ is a *direct sum* and write $U \oplus W$ for $U + W$ if $U \cap W = \{0\}$.

   Show that $U + W$ is direct if and only if for each $v \in U + W$ there is a unique $u \in U$ and a unique $w \in W$ such that $v = u + w$.

c. Conclude from b. that if $U + W$ is direct then the map

$$S\colon U \times W \to U \oplus W$$

   that sends $(u, w) \to u + w$ is a bijective map, that maps sums to sums and scalar products to scalar products, ie for all $x, y \in U \times W$, $S(x + y) = S(x) + S(y)$ and for all $c \in \mathbb{F}$ and $x \in U \times W$, $S(cx) = cS(x)$.

3.3.10 Problem. Let V, W be two vector spaces.

a. Let $V' \subseteq V \times W$ be the set $\{(v, 0) \mid v \in V\}$ and $W' \subseteq V \times W$ the set $\{(0, w) \mid w \in W\}$.

   Show that V', W' are subspaces of $V \times W$.

b. Show that $V \times W = V' \oplus W'$.

3.3.11 Problem. Generalize the Problem 3.3.9 as follows: Given subspaces $V_1, V_2, \dots, V_n \subseteq V$ of a vector space V, define their sum

$$V_1 + V_2 + \cdots + V_n = \{v \in V \mid v = v_1 + v_2 + \cdots + v_n \text{ for some } v_i \in V_i\}.$$

We say the V_i are *independent* if the following holds: whenever

$$v_1 + v_2 + \cdots + v_n = 0 \qquad \text{with } v_i \in V_i$$

then $v_i = 0$ for all i.

We say the sum $V_1 + V_2 + \cdots + V_n$ is direct, denoted, $V_1 \oplus V_2 \oplus V_2 \oplus \cdots \oplus V_n$ if the V_i are independent.

a. Show that $V = V_1 \oplus V_2 \oplus \cdots \oplus V_n$ is the direct sum of the subspaces V_i if and only if every $v \in V$ can be written uniquely as

$$v = v_1 + v_2 + \cdots + v_n$$

with $v_i \in V_i$.

b. Show that $V = V_1 \oplus V_2 \oplus \cdots \oplus V_n$ if and only if $V = V_1 + V_2 + \cdots + V_n$ and for each $i = 1, \dots, n-1$,

$$V_{i+1} \cap (V_1 + V_2 + \cdots + V_i) = \{0\}.$$

c. Suppose $V = V_1 \oplus V_2 \oplus \cdots \oplus V_n$. Show that $\dim V = \dim V_1 + \dim V_2 + \cdots + \dim V_n$.

3.4. *Infinite dimensional vector spaces

In important applications the vector spaces that are appearing are not finitely generated. To simplify the discussion let us extend our definition of "Span" to arbitrary subsets of a vector space: if $S \subseteq V$ is a subset of a vector space, let $\mathrm{Span}(S)$ be the set of all vectors that can be written as linear combinations of some (finitely many!) elements of V: $v \in \mathrm{Span}(S)$ if and only if there are $s_1, s_2, \dots, s_t \in S$ and $c_1, c_2, \dots, c_t \in \mathbb{F}$ such that

$$v = c_1 s_1 + \cdots + c_t s_t.$$

For instance, for every vector space V we have $\mathrm{Span}(V) = V$.

A set S for which $V = \mathrm{Span}(S)$ is called a *generating set* or *spanning set*. If S is also linearly independent, we call S a (*unordered*) basis for V. A vector space that does not admit a finite basis is called *infinite dimensional* and we sometimes write $\dim V = \infty$.

As a consequence of what is called Zorn's Lemma and by extension the Axiom of Choice, one can show that *every* vector space has a basis. However, in many cases, this basis will be uncountable and hence too huge to write down very explicitly.

A question that we have dodged so far is the following: if V is a finite dimensional vector space, then no infinite generating set can be a basis (because no infinite subset can be linearly independent). However, a priori, it might happen that V has a infinite list of generators that does not contain a basis. However, modifying our proof of Theorem 3.45 we can actually show:

3.49 Lemma. *Let V be a finite dimensional vector space. Suppose $S \subseteq V$ is a generating set. Then S contains a finite generating set.*

Proof. Let $\mathcal{B} = (v_1, v_2, \dots, v_n)$ be an (ordered) basis for V. Since S generates V, we may write v_i as a linear combination in elements of S. $v_i = \sum_{j=1}^{n_i} c_{ij} s_{ij}$, where the s_{ij} are suitable elements of S depending on v_i. However, the point is that only finitely many s_{ij} are needed to express all basis vectors. So let $T \subseteq S$ be the collection of all the s_{ij}. Then T is finite and $v_i \in \mathrm{Span}(T)$ for all i. But this implies $V = \mathrm{Span}(T)$. □

Therefore it can never happen that a finite dimensional vector space has a generating set that does not contain a basis.

3.4.1 Examples. Here are some important examples of infinite dimensional vector spaces:

a. The space of sequences: $\mathbb{F}^\infty = \{(a_1, a_2, a_3, \dots) \mid a_i \in \mathbb{F}\}$. Addition and scalar multiplication are done coefficientwise. Notice that $\mathbb{F}^\infty$ as a vector space is essentially the same as $\mathcal{F}(\mathbb{N}, \mathbb{F})$. (Can you prove that they are isomorphic?)

b. $\mathbb{R}^\infty$ has a few important subspaces: For $p \geq 1$ put

$$\ell^p(\mathbb{R}) := \{a \in \mathbb{R}^\infty \mid \sum_{i=1}^{\infty} |a_i|^p < \infty\}.$$

c. For $p \geq 0$,

$$\mathcal{C}^p(\mathbb{R}) = \{f \in \mathcal{F}(\mathbb{R}, \mathbb{R}) \mid f \text{ is } p \text{ times continuously differentiable}\}$$

(with $\mathcal{C}^0(\mathbb{R})$ just being the continous functions.) Similar definitions make sense for functions on intervals $I \subseteq \mathbb{R}$ of course.

d. $\mathcal{P}(\mathbb{R})$ is an infinite dimensional vector space.

e. We can view $\mathbb{R}$ as a vector space over the field $\mathbb{Q}$ of rational numbers: the addition is the usual addition of real numbers, and the scalar multiplication is the restriction of the usual multiplication to $\mathbb{Q} \times \mathbb{R}$. As a $\mathbb{Q}$-vector space, $\mathbb{R}$ is infinite dimensional. In fact it does not have a countable basis: any $\mathbb{Q}$-vector space with countable basis is still a countable set since it is a countable union of countable subsets (can you prove this?).

f. Let X be any (nonempty) set. Recall $\mathbb{F}X = \{f \in \mathcal{F}(X, \mathbb{F}) \mid \operatorname{supp} f \text{ finite}\}$. For $x \in X$ let $\delta_x \in \mathbb{F}X$ be defined as

$$\delta_x(y) = \begin{cases} 1 & y = x \\ 0 & y \neq x \end{cases}$$

Then the collection of δ_x forms a basis of $\mathbb{F}X$. This shows that $\mathbb{F}X$ is infinite dimensional if and only if X is infinite. (Of course, if X is finite we would have to order the δ_x to obtain a basis.) $\mathbb{F}X$ is important for theoretical considerations. It allows to think of *any set* as a basis of a vector space. Indeed, the function $X \to \mathbb{F}X$ that maps x to δ_x is a bijection[9] of X with a basis of $\mathbb{F}X$.

Here is a piece for the logically inclined among you: If X is empty, then $\mathbb{F}X = \{f\}$ where f is the the unique function[10] $f\colon \emptyset \to \mathbb{F}$. It is common to define the trivial vector space structure on $\mathbb{F}\emptyset$ and denote f by 0.

Finally, here is a proof that every vector space has a basis. It is based on Zorn's Lemma, which we state first.

A *partially ordered set* (or *poset*) is a set X together with a binary relation $\leq$ (so that for each $x, y \in X$, $x \leq y$ is either true or not true), such that the following statements hold:

a. $x \leq x$ for all x.

b. if $x \leq y$ and $y \leq x$, then $x = y$.

c. if $x \leq y$ and $y \leq z$, then $x \leq z$.

Note that it is *not* required that either $x \leq y$ or $y \leq x$. Elements x, y for which $x \leq y$ or $y \leq x$ are called *comparable*. If $x \leq y$ and $x \neq y$, we write $x < y$. (It should be clear what the statements $x \geq y$ and $x > y$ mean.)

A partial order that has this additional requirement is callled a *total order*, and a partially ordered set where the partial order is a total order is called a *totally ordered set*. If X is any set, the *power set* $P(X)$ of X has a natural partial order: $A \leq B$ if $A \subseteq B$. It is not a total order, because in general two subsets need not be contained in each other.

If X is a poset, then so is any subset. A *maximal* element in a poset X is an element $x \in X$ for which $x \leq y$ implies $x = y$. (So there are no elements for which $x \leq y$ and $x \neq y$.) It does *not* mean that $y \leq x$ for all $y \in X$, but it does mean that if x, y are comparable then $y \leq x$.

An *upper bound* for a subset $Y \subseteq X$ is an element $x \in X$ such that $y \leq x$ for all $y \in Y$.

3.50 Theorem (Zorn's Lemma)**.** *Let X be a poset. If every totally ordered subset of X has an upper bound in X, then X contains a maximal element.*

Zorn's Lemma is equivalent to the Axiom of Choice.

Let now V be a vector space, and $X \subset P(V)$ be the set of *linearly independent subsets*. Note that X is not empty because X always contains the empty set. A basis of V is a maximal element of X: if $\mathcal{B}$ is a maximal element of X, then for every $v \in V$, $v \in \mathcal{B}$, or $\mathcal{B} \cup \{v\}$ is linearly dependent; in either case $v \in \operatorname{Span}(\mathcal{B})$ (this uses that $\mathcal{B}$ is linearly independent in a similar way as does Lemma 3.37 for lists). Hence $\mathcal{B}$ is a linearly independent spanning set, which is a basis (for general vector spaces that are not necessarily finite, we do not require a basis to be an ordered list).

[9] See the appendix for a definition of the term "bijective."

[10] This uses the set theoretic definition of a function $f\colon X \to Y$ as a subset of $X \times Y$ with certain properties. If X is emtpy, $\emptyset \times Y$ (which is just the empty set) is a subset that satisfies all requirements of a function.

3.51 Theorem. *Every vector space has a basis.*

Proof. The set X is partially ordered by inclusion. So we need to show that X has a maximal element. By Zorn's Lemma it is enough to show that if $Y \subseteq X$ is a nonempty totally ordered subset, then Y has an upper bound in X.

So let Y be totally ordered. Let us define

$$M := \bigcup_{S \in Y} S$$

the union of all elements in Y. This is a subset of V. It is also in X: indeed, let $m_1, m_2, \ldots, m_n \in M$ be distinct elements. Then $m_i \in S_i$ say for some $S_i \in Y$. The set $\{S_1, S_2, \ldots, S_n\}$ is totally ordered (since Y is), and so has a maximal element, S_i, say. So $S_j \subseteq S_i$ for all j, and $m_1, m_2, \ldots, m_n \in S_i$. Since S_i is linearly independent, $m_1, m_2, \ldots, m_n$ are linearly independent, and hence M is. M is clearly an upper bound in X for all $S \in Y$.

By Zorn's Lemma, X contains a maximal element, which, as we have seen, is a basis. □

Zorn's Lemma and the Axiom of Choice

The following is slightly technical and meant only for the interested. Since most Linear Algebra texts avoid a proof of Zorn's Lemma, and since one should have seen one at least once in one's mathematical existence, it is included for reference. The proof and constructions here closely follow Kneser's proof from 1950.

Zorn's Lemma and the Axiom of Choice are logically equivalent (assuming the other axioms of the Zermelo-Fraenkel set theory). In set theory elements of sets are again sets (so all mathematical objects are sets).

Axiom of Choice Let X be a set of nonempty sets. Then there exists a choice function.

A choice function in this context is a function $f\colon X \to \bigcup_{S \in X} S$ such that $f(x) \in x$. For example, the Axiom of Choice states that if $\{X_i\}_{i \in I}$ is a family of nonempty sets indexed by a set I, then the direct product $\prod_{i \in I} X_i = \{(x_i)_{i \in I} \mid x_i \in X_i\}$ is non-empty. Here $(x_i)_{i \in I}$ is short-hand for a function $I \to \bigcup_i X_i$.

A subset S of a poset X is called *well-ordered*, if it is totally ordered and every nonempty subset of S has a *minimal* element, that is if $T \subset S$ is nonempty, then there is $t \in T$ such that $t \leq s$ for all $s \in T$. (For example, the natural numbers with the usual ordering are well-ordered.) The Axiom of Choice is also equivalent to the statement that every set admits a well-order, but we won't go into details.

Suppose now that X is a poset, and suppose on the set of *well-ordered subsets* of X we have a function u such that for each such subset C, $u(C) \in X$ is an upper bound with the following property: if not all upper bounds of C are elements of C, then $u(C) \notin C$.

For any well-ordered set $C \subset X$, a *beginning segment* (BS) is is a subset $A \subset C$ such that whenever $a \in A$ and $y \leq a$ in C, then $y \in A$. An example of a beginning segment is $C_x = \{y \in C \mid y < x\}$.

Lemma. *C_x is a beginning segment. Moreover, every proper subset of C that is a beginning segment is of the form C_x for some $x \in C$.*

Proof. To see that C_x is a BS, let $z \in C_x$ and $y \leq z$ in C. Then $y < x$ and so $y \in C_x$.

Note if $A \subset C$ is a proper subset and a BS, and $x \in C$ a minimal element of $C \setminus A$, then $A = C_x$. It is clear that $C_x \subset A$: if $y < x$, then $y \in A$ because x is minimal. Let $y \in A$ and suppose $y \geq x$. Then $x \leq y$ so $x \in A$ because A is a BS. This is a contradiction, so we must have $y < x$ and hence $A \subset C_x$. □

For the purpose of this section, a *chain* in X is a well-ordered subset C with the following property: For every $x \in C$, $x = u(C_x)$. For example the empty set is a chain. Also $\{u(\emptyset)\}$ is a chain. If C is a chain, then $C \cup \{u(C)\}$ is a chain.

It is not so easy to get an intuitive understanding of what a chain is.

Example. Let $\mathbb{N}^* = \mathbb{N} \cup \{\infty\}$, with the unerstanding that $\infty > n$ for all $n \in \mathbb{N}$. Then $\mathbb{N}^*$ is still well-ordered, and every nonempty subset has an upper bound (for instance ∞).

If $C \subset \mathbb{N}^*$ is bounded (that is for all $x \in C$, $x \leq n$ for some $n \in \mathbb{N}$), we define $u(C) = \min\{n \mid n \notin C\}$. If C is not bounded, we define $u(C) = \infty$.

A nonempty subset C of $\mathbb{N}^*$ is a chain, if it is of the form $[1, n] = \{1, 2, \ldots, n\} = \{x \in \mathbb{N}^* \mid x \leq n\}$, where n is allowed to be ∞, or if it is equal to $\mathbb{N}$. Indeed, C contains 1: let $n \in C$ be the minimal element. Then $C_n = \emptyset$ and $n = u(\emptyset) = 1$. Let now $n = \min\{m \in \mathbb{N}* \mid [1, m] \not\subset C\}$. If n is not defined (that is $[1, m] \subset C$ for all $m \in \mathbb{N}$, then $C = \mathbb{N}^*$.

So suppose n exists. Then $n > 1$ and so $[1, n-1] \subset C$. If $C = [1, n-1]$, we are done. Otherwise, let m be the minimum of elements in C not in $[1, m-1]$. Then $[1, n-1] = C_m$. As C is a chain, this means $m = u(C_m) = n$. But we assumed $[1, n] \not\subset C$ – a contradiction. So we must have $C = [1, n-1]$.

Lemma (BS-Lemma)**.** *Let C, D be two chains. Suppose every proper beginning segment of C is a subset of D. If C has no maximal element, then $C \subset D$. If C has a maximal element, m, say, then $C \subset D$, or $D = C_m$.*

Proof. The proper BS of C have the form C_x for $x \in C$. If C has no maximal element, then for every $y \in C$, there is $x > y$ in C, and hence $y \in C_x$ for some $x \in C$. But $C_x \subset D$ for all x, so $y \in D$ for all $y \in C$, ie. $C \subset D$.

So suppose m is a maximal element of C. Then $C_m \subset D$, and $C = C_m \cup \{m\}$. If D is not equal to C_m, there is a minimal $d \in D$ such that $d \notin C_m$, and it follows that $D_d = C_m$. As D and C are chains, we have $d = u(D_m) = u(C_m) = m$, and hence $C \subset D$. □

The BS-Lemma allows us to conclude immediately that out of any two chains, one is always a subset of the other.

Lemma. *Let C, D be two chains in X. Then $C \subset D$ or $D \subset C$.*

Proof. Let $F = \{x \in C \mid C_x \not\subset D\}$. If F is empty, then every proper BS of C is a subset of D, and hence the BS-Lemma shows that $C \subset D$ or $D \subset C$.

Otherwise, F has a minimal element x, say. So for all $y \in C_x$, we have $C_y \subset D$. Note that $C_y = (C_x)_y$, as C_x is a BS of C. So every proper BS of C_x is a subset of D. Applying the BS-Lemma to the chains C_x and D, we conclude that $D \subset C_x$, for we cannot have $C_x \subset D$ by construction. □

Corollary. *If C, D are chains, then one is a BS of the other.*

Proof. We may assume that $C \subset D$. If $C = D$ nothing is to show. So let $x \in C$. Then $D_x \subset C$: for D_x is a chain, and $C \not\subset D_x$, so we must have $D_x \subset C$. □

Proof of Zorn's Lemma. Let X be a poset which satisfies the hypothesis. Then X is not empty, because the empty set is a totally ordered subset of X and hence has an upper bound.

We will now construct a function $m\colon X \to X$ such that $m(x) = x$ if x is maximal, and $m(x) > x$, if x is not maximal. For $x \in X$ let $M_x = \{x\}$ if x is maximal and $M_x = \{y \in X \mid y > x\}$ if x is not maximal. Then $x \mapsto M_x$ defines a function $M\colon X \to P(X)$ ($P(X)$ is the power set of X). Let $T = M(X)$ be the range of M. Then T is a set of nonempty subsets by construction. By the axiom of choice there is a function $f\colon T \to X$ such that $f(S) \in S$ for all $S \in T$. Then let $m = f \circ M$ has the desired property.

If now $S \subset X$ is any well-ordered subset, then S has an upper bound (since it is totally ordered). Similar to the construction above, we can define a function, defined on the set of totally ordered subsets of X, which assigns to each such set an upper bound: let $Y \subset P(X)$ be the subset of the well-ordered subsets of X. Note that Y is not empty because Y contains for instance the empty set.

Then we define $B\colon Y \to P(X)$ by $B(S) = \{y \in X \mid x \leq y \text{ for all } x \in S\}$. In other words $B(S)$ is the set of upper bounds for S. Then the image of B admits a choice function, which composed with B gives a function $b\colon Y \to X$ such that $b(S)$ is an upper bound for S.

Combining m and b we get for every well-ordered set S an element $u(S) = m(b(S))$. Note that if $b(S)$ is not a maximal element of X, then $u(S)$ is not contained in S (indeed, $u(b(S)) > b(S)$ in this case); so if $u(S) \in S$, then $u(S)$ is a maximal element.

Now we are almost done: Let U be the union of all chains in X. Then U is again a chain. For this, note U is totally ordered: if $x, y \in U$, then $x \in C$ and $y \in D$ say where C, D are chains. But then $C \subset D$ or $D \subset C$, so one chain contains both, x and y. Thus x, y are comparable in C or D and hence in U.

Next, suppose $T \subset U$ is a nonempty subset. Let C be any chain such that $T \cap C$ is not empty. Then $T \cap C$ has a minimal element $t_0 \in T \cap C$ as C is well-ordered. Let now $t \in T$. Then if $t \leq t_0$, and D is any chain containing t, then $D \subset C$ (so $t = t_0$), or $C \subset D$. In the latter case, C is a BS of D, and $t_0 \in C$, $t \leq t_0$, so $t \in C$. As $t \in T$, this means $t = t_0$. So t_0 is a minimal element of T.

Let now $x \in U$. Then $x \in C$ for some chain $C \subset U$ in X. If $y < x$ in U, then $y \in C$, ie. $C_x = U_x$. Indeed, y is contained in some chain D. If $D \subset C$, then $y \in C$. If $C \subset D$, then C

is a BS of D, and hence $y \in D_x \subset C$. So $U_x \subset C_x$. And it is clear that $C_x \subset U_x$. But then $u(U_x) = u(C_x) = x$, as needed.

So U is a chain, and also $U \cup \{u(U)\}$ is a chain, and consequently a subset of U (as U is the union of all chains). Hence $u(U) \in U$. But this means $u(U)$ is a maximal element of X. □

It is considerably simpler to deduce the Axiom of Choice from Zorn's Lemma. For the sake of completeness we include a proof:

Proposition. *Zorn's Lemma implies the Axiom of Choice.*[11]

Proof. Let X be a set of nonempty sets. Let $U = \bigcup_{A \in X} A$ be the union of all elements of X. We need to show that there is a choice function $f\colon X \to U$ such that $f(A) \in A$ for all $A \in X$.

A *partial choice function* (PCF) is a pair (Y, f), where $Y \subset X$ and $f\colon Y \to U$ such that $f(A) \in A$ for all $A \in Y$. Let P be the set of all partial choice functions. Note that P is nonempty: indeed, let $A \in X$ be arbitrary. Then A is not empty, so there is $a \in A$. We put $Y = \{A\}$ and $f\colon Y \to U$ as $f(A) = a$. Then (Y, f) is a PCF.

Moreover, P admits a partial order: we define $(Y, f) \leq (Z, g)$ if $Y \subset Z$ and $g \mid_Y = f$. It is easy to verify that this indeed is a partial order.

Let S be a totally ordered subset of P. We show that S has an upper bound as follows. If S is empty, take any PCF. Otherwise, let

$$Z = \bigcup_{(Y,f) \in S} Y$$

be the union of the domains of all PCFs in S. If $A \in Z$, and (Y, f) is any PCF in S such that $A \in Y$, then $f(A)$ is independent of (Y, f). Indeed, whenever $A \in Y \cap Y'$ where (Y, f), (Y', f') are PCFs in S, then $f(A) = f'(A')$. So we define $g\colon Z \to U$ as $g(A) = f(A)$ where (Y, f) is any PCF in S with $A \in Y$. By construction it is clear that (Z, g) is an upper bound for S.

By Zorn's Lemma, P contains a maximal element (Y, f), say. We need to show that $Y = X$. But that is clear: for if $A \in X$, and $A \notin Y$, then we could define $Z = Y \cup \{A\}$ and define $g(B) = f(B)$ for $B \in Y$ and $g(A) = a$ to obtain $(Z, g) > (Y, f)$ – contradicting the maximality of (Y, f). Thus, X has a choice function, namely, f. □

Remark. In the proof one step feels like cheating: the definition of the upper bound involved for each $A \in Z$, a "choice" of a PCF (Y, f) where $A \in Y$. Is this not an implicit application of the axiom of choice?

No, not really. What is (set theoretically) a function? A function $f\colon X \to Y$ is a subset $R \subset X \times Y$ such that for each $x \in X$ there is one and only one $(x, y) \in R$. (And then we write $f(x) = y$.) R is called the *graph* of f.

[11] Here is a subtle point. If you assume the Axiom of Choice to be true, then of course *any* statement implies it. However, what we are saying here is, if we were to be doing set theory, where we don't assume the Axiom of Choice to be true a priori, but rather assume Zorn's Lemma, then the Axiom of Choice is a consequence.

So we could have defined $g\colon Z \to U$ as follows: for any PCF (Y, f) in S, let $R_f \subset Y \times U \subset X \times U$ be its graph. Then we define $R = \bigcup_{(Y,f)\in S} R_f$, and what we saw above shows that R is the graph of a unique function $g\colon Z \to U$.

4. Linear transformations

In this chapter we will discuss the basic properties of linear transformations between vector spaces. The main result will be that every linear transformation is a matrix transformation in the suitable sense.

4.1. Basic definitions

In mathematics, sets are compared by studying the maps[1] between them. If our sets are vector spaces, say, then discussing maps between them makes the most sense if these maps actually preserve the structure. For more information on the terminology regarding functions please refer to the Appendix.

4.1 Definition. Let V, W be vector spaces (over the same field $\mathbb{F}$). A *linear transformation* or *linear map* or *homomorphism* (*of vector spaces*) from V to W is a map

$$T\colon V \to W$$

such that T is *linear*, that is,

a. For each $c \in \mathbb{F}$ and each $v \in V$ we have $T(cv) = cT(v)$.

b. For each $v, w \in V$ we have $T(v + w) = T(v) + T(w)$.

The set of all linear transformation between V and W will be denoted by $\mathrm{Hom}(V, W)$.

We often write Tv instead of $T(v)$ for the image of v. Also it is very common to write "Let T be a linear transformation $V \to W$ ..." instead of "Let $T\colon V \to W$ be a linear transformation ..." Also, often we write $T\colon V \to W, v \mapsto \ldots$ instead of $T\colon V \to W, T(v) = \ldots$" to specify the rule by which T is defined.

Before we discuss any properties let us first look at several examples.

4.2 Examples. a. There is hardly any mathematical structure without *trivial* example, and so is the case here: If V, W are vector spaces, then $\mathrm{Hom}(V, W)$ always contains the *zero transformation* that sends all elements in V to 0 in W. We will denote this particular transformation by 0. In particular, $\mathrm{Hom}(V, W)$ is never empty.

[1] For us the terms "map," "mapping," and "function" all mean the same thing, unless expressedly indicated otherwise.

b. If $V = W$, there is a distinguished linear transformation $\mathrm{id}_V : V \to V$ defined by $\mathrm{id}_V(v) = v$. This particular map is called the *identity transformation*.

c. Let A be any $m \times n$-matrix. Then the map

$$T_A : \mathbb{F}^n \to \mathbb{F}^m$$

defined by $T_A(X) = AX$ is a linear transformation. We say T_A is the *matrix transformation* associated to A.

Observations III and IV in Section 2.2 indeed guarantee that T_A is a linear transformation.

d. A variation of the previous example is the following: Let $a_1, a_2, \dots, a_n \in \mathbb{F}$ and let $V = \mathbb{F}^n$, and consider

$$f : V \to \mathbb{F}$$

defined by

$$f(x_1, x_2, \dots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n.$$

Then f is a linear transformation. We often also say that f is a *linear function* or *linear functional* on V. Unless the codomain is specified explicitly, usually a linear function on V is always understood to be a linear transformation $V \to \mathbb{F}$.

Notice that f can be realized as a matrix transformation: $f(x_1, x_2, \dots, x_n) = AX$ where $A = [\, a_1\, a_2\, \dots\, a_n \,]$ and and we identify $(x_1, x_2, \dots, x_n)$ with the column vector X with these entries.

4.1.1 Problem. Let $T : \mathbb{F} \to \mathbb{F}$ be a linear transformation. Show that there is $c \in \mathbb{F}$ such that for all x, $T(x) = cx$.

4.3 Example. Let V be a finite dimensional vector space of dimension n, let $\mathcal{B}$ be a basis. The map $T : \mathbb{F}^n \to V$ given by $T(X) = \mathcal{B}X$ that we discussed earlier is a linear transformation.

The following three examples use some calculus which you may or may not have seen yet.

4.4 Example. Let $I = [a, b] \subseteq \mathbb{R}$ be an interval. Let $V = \mathcal{C}^1(I)$ and let $W = \mathcal{C}^0(I)$. Then

$$D : V \to W$$

defined by

$$D(f) = f' = \frac{df}{dx}$$

is a linear transformation.

4.5 Example. Let $W = \mathcal{C}^0(I)$ and $V = \mathcal{C}^1(I)$. Define

$$\mathrm{Int} : W \to V$$

by

$$\mathrm{Int}(f)(x) = \int_a^x f(s)ds.$$

Then Int is a linear transformation (this uses the Fundamental Theorem of Calculus, in order to show that $\mathrm{Int}(f)$ is an element of $\mathcal{C}^1(I)$.).

4.6 Example. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be any function. We say f is *differentiable* at $x_0 \in \mathbb{R}^2$, if there is a linear transformation $T_f\colon \mathbb{R}^2 \to \mathbb{R}$ such that

$$f(x_0 + X) = f(x_0) + T_f(X) + o(X)$$

such that

$$\lim_{X\to 0} \frac{o(X)}{|X|} = 0.$$

Here $|X|$ is the length of X (usually defined as $\sqrt{x_1^2 + x_2^2}$), so that for all $\epsilon > 0$ there exists $\delta > 0$ such that whenever $|X| < \delta$ then $|\frac{o(X)}{|X|}| < \epsilon$. We will leave the details to future calculus classes).

4.7 Example. Returning to our previous example from Physics: in quantum mechanics, the *time development* of a state $|\psi\rangle \in \mathcal{H}$ is given by a linear transformation $H\colon \mathcal{H} \to \mathcal{H}$ (sometimes called the *Hamiltonian* operator of the system). This yields SCHRÖDINGER's equation which describes the change of state at time t:

$$\frac{d}{dt}|\psi(t)\rangle = -\frac{i}{\hbar}H|\psi(t)\rangle.$$

Sometimes a solution of this equation has the form

$$|\psi(t)\rangle = U(t)|\psi_0\rangle$$

where $|\psi_0\rangle$ is the original state of the system and $U(t)$ is a linear transformation (depending on the time t).

4.8 Example. Let us think of $\mathbb{R}^2$ as the Euclidean plane. A *Euclidean motion* is a distance preserving map $M\colon \mathbb{R}^2 \to \mathbb{R}^2$. If $M(0) = 0$, then M is a linear transformation. We won't prove this right now, but state for the record, that any rotation about the origin or reflection across a line through the origin is consequently a linear transformation.

Here we measure the *distance* of two points by $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ (which coincides with our intuitive notion coming from the Pythagorean Theorem), if $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

The same can be said for distance preserving maps from $\mathbb{R}^3 \to \mathbb{R}^3$.

4.9 Example. A very important example is the following: If $z \in \mathbb{C}$ is a complex number then $m_z\colon \mathbb{C} \to \mathbb{C}$ is a linear transformation, both, of the $\mathbb{C}$-vector space $\mathbb{C}$, but also of the $\mathbb{R}$-vector space $\mathbb{C}$.

Given two linear transformations $S, T\colon V \to W$ it is possible to form their *sum*: We define

$$S + T\colon V \to W$$

as $(S+T)(v) = S(v) + T(v)$ for $v \in V$. Similarly, for $T\colon V \to W$ and $c \in \mathbb{F}$ we define $cT\colon V \to W$ by $(cT)(v) = c(Tv)$ $(v \in V)$.

4.10 Theorem. *Let V, W be vector spaces over $\mathbb{F}$. Then for all $S, T \in \mathrm{Hom}(V, W)$ also*

$$S + T \in \mathrm{Hom}(V, W)$$

and for all $c \in \mathbb{F}$ and $T \in \mathrm{Hom}(V, W)$ also

$$cT \in \mathrm{Hom}(V, W).$$

With respect to these two operations, $\mathrm{Hom}(V, W)$ is a vector space over $\mathbb{F}$.

Proof. Let $S, T \in \mathrm{Hom}(V, W)$ and consider $v_1, v_2 \in V$. Then

$$(S+T)(v_1 + v_2) = S(v_1 + v_2) + T(v_1 + v_2)$$

which as S and T are linear is equal to

$$Sv_1 + Sv_2 + Tv_1 + Tv_2.$$

This now can be rewritten as

$$(Sv_1 + Tv_1) + (Sv_1 + Tv_2) = (S+T)v_1 + (S+T)v_2.$$

Now let $c \in \mathbb{F}$ and $v \in V$. Then

$$(S+T)(cv) = S(cv) + T(cv) = c(Sv) + c(Tv) = c(Sv + Tv) = c[(S+T)(v)].$$

The upshot is that $S + T \in \mathrm{Hom}(V, W)$. The case of cT is similar.

Finally, that $\mathrm{Hom}(V, W)$ is a vector space now follows almost verbatim as in the proof that $\mathcal{F}(X, \mathbb{F})$ is a vector space where X is a set (cf. Example 3.5). We leave the details to the reader, but observe that the additive identity in $\mathrm{Hom}(V, W)$ is the trivial homomorphism $0\colon V \to W$ that maps every $v \in W$ to 0; and for $T \in \mathrm{Hom}(V, W)$, we have $-T$ is the homomorphism that maps every $v \in V$ to $-(Tv) \in W$. □

4.11 Example. Let $A, B \in M_{m\times n}(\mathbb{F})$. Verify that $T_A + T_B = T_{A+B}$ and $cT_A = T_{cA}$ (in $\mathrm{Hom}(\mathbb{F}^n, \mathbb{F}^m)$).

Thus, the map $L\colon M_{m\times n}(\mathbb{F}) \to \mathrm{Hom}(\mathbb{F}^n, \mathbb{F}^m)$ defined by $L(A) = T_A$ is a linear transformation.

4.12 Definition. Let $S\colon U \to V$ and $T\colon V \to W$ be linear transformations (where U, V, W are vector spaces over $\mathbb{F}$). Then $TS\colon U \to W$ is defined as the composition $T \circ S$, that is

$$TS(u) = T(S(u)) \quad (u \in U).$$

4.1.2 Problem. Show that ST as defined above is a linear transformation.

The major properties of a linear transformation are summarized as follows:

4.13 Proposition. *Let $T\colon V \to W$ be a linear transformation. Then*

a. $T(0) = 0$ (*that is* $T(0_V) = 0_W$)*;*

b. $T(-v) = -T(v)$ *for all* $v \in V$.

c. *For any linear combination* $v = ca_1v_1 + c_2v_2 + \cdots + c_pv_p$ *we have*

$$T(v) = c_1T(v_1) + c_2T(v_2) + \cdots + c_pT(v_p).$$

This explains why one should think of linear transformation as *structure preserving* maps.

Proof.

a. $0 = 0 + 0$, so $T(0) = T(0 + 0) = T(0) + T(0)$. By the Cancelation Rule (Proposition 3.10 a.) we conclude that $T(0) = 0$.

b. By i), $T(0) = 0$. Hence $0 = T(0) = T(v + (-v)) = T(v) + T(-v)$. The uniqueness of the additive inverse implies that $T(-v) = -T(v)$.

   Notice we could have argued as well that $T(-v) = T((-1)v) = (-1)T(v) = -T(v)$.

c. This is an immediate induction on the number p of summands: if $p = 1$, then all it says is $T(c_1v_1) = c_1T(v_1)$ which is true because T is linear.

   So suppose for a given integer $p > 0$ the assertion of c. is true. Suppose $v = c_1v_1 + \cdots + c_{p+1}v_{p+1}$ is a linear combination of $p + 1$ elements of V. Then

$$T(v) = T((c_1v_1 + \cdots + c_pv_p) + c_{p+1}v_{p+1}) = T(c_1v_1 + \cdots + c_pv_p) + T(c_{p+1}v_{p+1})$$

   because T is linear. By the induction assumption we conclude that

$$T(v) = (c_1T(v_1) + \cdots + c_pT(v_p)) + c_{p+1}T(v_{p+1})$$

   which is the claim.

□

4.2. Kernel and image of a linear transformation

In the previous chapter we associated to a matrix A two subspaces, namely the column space and its null space. If course, we may as well think of $\mathrm{Col}(A)$ and $\mathcal{N}(A)$ as subspaces associated to T_A. In fact, any transformation has two subspaces associated to it.

4.14 Definition. Let $T\colon V \to W$ be a linear transformation. The *kernel* or *null space* of T, denoted $\mathcal{N}(T)$ is defined as

$$\mathcal{N}(T) = \{v \in V \mid T(v) = 0\}.$$

The *image* or *range* of T, denoted $\operatorname{im} T$ is defined as

$$\operatorname{im} T = \{w \in W \mid \text{there is some } v \in V \text{ such that } w = T(v)\}$$

4.15 Example.

a. The kernel of the identity transformation $\mathrm{id}_V\colon V \to V$ is the zero subspace: $\mathcal{N}(\mathrm{id}_V) = \{0\}$. The image of id_V is $\mathrm{im}(\mathrm{id}_V) = V$.

b. Let $T = 0 \in \mathrm{Hom}(V, W)$ be the zero transformation (ie. $T(v) = 0$ for all $v \in V$).

   Then $\mathrm{im}(T) = \{0\}$ and $\mathcal{N}(T) = V$.

4.16 Example.

a. The kernel of the matrix transformation T_A is precisely the null space $\mathcal{N}(A)$ of A. In the same spirit, the image of T_A is nothing but the column space $\mathrm{Col}(A)$.

b. Let $D\colon \mathcal{C}^2(\mathbb{R}) \to \mathcal{C}^0(\mathbb{R})$ be the differential operator

$$D(f) = f'' + mf$$

   where $m \in \mathbb{R}$. Then D is a linear transformation and consequently

$$\mathcal{N}(D) = \{f \in \mathcal{C}^2(\mathbb{R}) \mid D(f) = 0\}$$

   is the set of all functions that satisfy the differential equation $f'' + mf = 0$. In calculus we learn that if $m = 1$ then $\mathcal{N}(D) = \mathrm{Span}(\sin x, \cos x)$ is the two dimensional space spanned by the $\sin$ and $\cos$ functions.

   If on the other hand $m = 0$, then $\mathcal{N}(D) = \{f \in \mathcal{C}^2(\mathbb{R}) \mid \text{ there are } a, b \in \mathbb{R} \text{ such that } f(x) = ax + b\}$ is the space of polynomial functions of degree at most 1.

The previous example illustrates that the kernel of a linear transformation T is naturally a generalization of the solution set of a homogeneous system of linear equations. It also illustrates why we should not be surprised by the following proposition.

4.17 Proposition. *Let $T\colon V \to W$ be a linear transformation. Then $\mathcal{N}(T)$ is a subspace of V and $\operatorname{im} T$ is a subspace of W.*

Moreover, if $V = \mathrm{Span}(v_1, v_2, \ldots, v_p)$, then $\mathrm{im}(T) = \mathrm{Span}(T(v_1), T(v_2), \ldots, T(v_p))$.

Proof.

$\mathcal{N}(T)$ **is a subspace:** By Definition 3.11 we have to check three things. $\mathcal{N}(T)$ is nonempty, closed under addition, and is closed under scalar multiplication. By Proposition 4.13 c. the null space of T always contains 0 and is consequently nonempty. Next, let $v, w \in \mathcal{N}(T)$. Then $T(v) = T(w) = 0$. It follows that

$$T(v + w) = T(v) + T(w) = 0 + 0 = 0.$$

Thus, $v + w \in \mathcal{N}(T)$. Similarly, if $v \in \mathcal{N}(T)$ and $c \in \mathbb{F}$ then $T(cv) = cT(v) = c \cdot 0 = 0$ and so $cv \in \mathcal{N}(T)$.

$\mathrm{im}(T)$ **is a subspace:** Since V is nonempty, $\mathrm{im}(T)$ is nonempty (it contains at least $0 = T(0)$). Let $v, w \in \mathrm{im}(T)$. By definition this means that there are $v', w' \in V$ such that $T(v') = v$ and $T(w') = w$. Then

$$v + w = T(v') + T(w') = T(v' + w') \in \mathrm{im}(T).$$

Also if $v \in \mathrm{im}(T)$, $v = T(v')$, say, then $cv = cT(v') = T(cv') \in \mathrm{im}(T)$.

The image of a spanning set is a spanning set: Let $V = \mathrm{Span}(v_1, \ldots, v_p)$. It is clear that for $i = 1, 2, \ldots, p$, $T(v_i) \in \mathrm{im}(T)$ and hence by Lemma 3.17, $\mathrm{Span}(T(v_1), T(v_2), \ldots, T(v_2)) \subseteq \mathrm{im}(T)$.

For the reverse inclusion, let $w \in \mathrm{im}(T)$. Then $w = T(v)$ for some v. v, on the other hand, is equal to $v = c_1v_1 + \cdots + c_pv_p$ for some $c_1, c_2, \ldots, c_p \in \mathbb{F}$. Hence, by Proposition 4.13 c., we find that

$$w = T(v) = T(c_1v_1 + \cdots + c_pv_p) = c_1T(v_1) + \cdots + c_pT(v_p) \in \mathrm{Span}(T(v_1), \ldots, T(v_p)).$$

Consequently, $\mathrm{im}(T) \subseteq \mathrm{Span}(T(v_1), T(v_2), \ldots, T(v_p))$. Together, we have

$$\mathrm{im}(T) = \mathrm{Span}(T(v_1), T(v_2), \ldots, T(v_p)).$$

□

Injective maps Recall that a map $f\colon X \to Y$ between sets is *injective*, if $f(x) = f(y)$ only if $x = y$. In other words, $x \neq y$ implies $f(x) \neq f(y)$. For more details see the appendix.

The single most important relation between a linear transformation T and its kernel $\mathcal{N}(T)$ that to test whether T is injective, all we have to do is to compute $\mathcal{N}(T)$.

4.18 Proposition. *A linear transformation $T\colon V \to W$ is injective if and only if $\mathcal{N}(T) = \{0\}$.*

We often say the null space is *trivial* to indicate that it is equal to $\{0\}$.

Proof. The only if part is clear: If T is injective and $v \in \mathcal{N}(T)$ then $T(v) = 0 = T(0)$ implies that $v = 0$.

Conversely, suppose $\mathcal{N}(T) = \{0\}$. Let $v, w \in V$ be elements such that $T(v) = T(w)$. We have to show that $v = w$. Now

$$T(v) = T(w) \implies T(v) - T(w) = 0 \implies T(v + (-w)) = 0 \implies v - w \in \mathcal{N}(T).$$

Since the kernel only contains 0 this means $v - w = 0$, ie. $v = w$. □

Instead of verifying that the preimage of each and every vector in W contains at most one element, we only have to check this for the vector $0 \in W$.

Notice that in the case of $V = W$, we can compose linear transformations in $\operatorname{Hom}(V, V)$ and obtain again an element of $\operatorname{Hom}(V, V)$. We have seen this happening before: if $A, B \in M_n(\mathbb{F})$ are square matrices then also $AB \in M_n(\mathbb{F})$. Thus, on $\operatorname{Hom}(V, V)$ we do not only have an addition but also a multiplication of elements. As in the case of matrices it is not commutative. $\operatorname{Hom}(V, V)$, together with addition and composition of linear transformations, is what is called a *ring*. We will study this structure later more closely. For now, we introduce the same notion we did for matrices:

4.19 Definition. A linear transformation $T\colon V \to V$ is called *invertible* or *nonsingular* if there is a linear transformation S such that

$$TS = ST = \mathbf{1}.$$

S is called the *inverse* of T if it exists and is often denoted by T^{-1}.

Here $\mathbf{1}\colon V \to V$ is the identity transformation.

Note that since T is a map between two sets, T^{-1} already has a meaning (namely, as the inverse map for a bijective mapping), which might cause confusion. Luckily, this is not so:

4.20 Theorem. $T \in \operatorname{Hom}(V, V)$ *is invertible if and only if* T *is bijective, that is, if and only if* T *is injective and surjective.*

If T *is invertible, then* T^{-1} *is precisely the inverse mapping.*

Proof. It is well known that a map $f\colon X \to Y$ is bijective if and only if there is a map $g\colon Y \to X$ such that $f \circ g = \mathrm{id}_Y$ and $g \circ f = \mathrm{id}_X$ (equivalently, $f(g(y)) = y$ for all $y \in Y$ and $g(f(x)) = x$ for all $x \in X$).

Since $\mathbf{1} = \mathrm{id}_V$, and since the multiplication of linear transformations is the composition of maps, all we are saying here is that if T is bijective, then the inverse map T^{-1} (which exists as a map $V \to V$) is again a linear transformation. This is an easy exercise (and left as a homeowork problem). □

4.2.1 Problem. Finish the proof of this theorem: Show that if T is bijective, then the inverse map T^{-1} is again a linear transformation.

Warning. If V is an arbitrary vector space then it may happen that there are linear transformations $S, T \in \mathrm{Hom}(V, V)$ such tat $ST = \mathbf{1}$ but $TS \neq \mathbf{1}$.

As an example consider $V = \mathcal{P}(\mathbb{R})$ and $L_x \colon V \to V$ defined as $L_x(f) = xf$. Then L_x is an injective linear transformation: that L_x is linear is a simple exercise; that L_x is injective then follows from $\mathcal{N}(T) = \{f \mid xf(x) = 0 \forall x \in \mathbb{R}\} = \{0\}$.

For instance the constant function $1 \colon \mathbb{R} \to \mathbb{R}$ defined by $1(x) = 1$ is not in the image of L_x: all functions of the form $L_x(f)$ satisfy that $L_x(f)(0) = 0$.

This problem does not arise if V is *finite dimensional*.

4.2.2 Problem. Define $\mathrm{GL}(V) = \{T \in \mathrm{Hom}(V, V) \mid T \text{ is invertible}\}$. Show that $\mathrm{GL}(V)$ (together with the multiplication of linear transformations) is a group.

Among all linear transformations, the bijective ones stand out. Indeed, in terms of comparing two vector spaces V and W, they are as similar as possible as they can be, if there is a bijective linear transformation between them.

4.21 Definition. A linear transformation $T \colon V \to W$ is called an *isomorphism* if T is bijective.

We say V and W are *isomorphic* (to each other) if there exists an isomorphism $V \to W$.

So to test whether $T \colon V \to W$ is an isomorphism one has to check that $\mathcal{N}(T) = \{0\}$ and $\operatorname{im} T = W$.

4.2.3 Problem. Show that if $T \colon V \to W$ is an isomorphism, then the inverse map $T^{-1} \colon W \to V$ is again a linear transformation.

So the notion that V and W are isomorphic is really symmetric in V and W.

Isomorphic vector spaces cannot really be distinguished using their properties of vector spaces. Any law that holds in one holds in the other: we can think of an isomorphism between them as a way to translate formulas from one into the other.

The most surprising result in this area is the following:

4.22 Theorem. *Two finite dimensional vector spaces are isomorphic if and only if they have the same dimension.*

Proof. Let V and W be finite dimensional vector spaces.

If $T \colon V \to W$ is an isomorphism, it is easy to see (see it!) that T maps a basis of V to a basis of W. But this shows that $\dim V = \dim W$.

Conversely, if V and W have the same dimension, we can construct an isomorphism between them: pick a basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ of V and $\mathcal{C} = (w_1, w_2, \ldots, w_n)$ of W.

Define $T \colon V \to W$ by $T(x_1 v_1 + \cdots + x_n v_n) = x_1 w_1 + \cdots + x_n w_n$. This is a map that is clearly surjective (every linear combination of the elements of $\mathcal{C}$ is in the image). It is also a linear transformation (easy!) and injective because $\mathcal{N}(T) = \{0\}$. □

4.23 Corollary. *If $\dim V = n$ then V is isomorphic to $\mathbb{F}^n$.*

Proof. We could simply say that $\dim V = \dim \mathbb{F}^n = n$. However, here the isomorphism constructed in the proof of the previous theorem is explicit, once we choose a basis for V: If $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ is a basis then the map $\mathbb{F}^n \to V$ sending $X \to \mathcal{B}X$ is an isomorphism. (It is the unique map $\mathbb{F}^n \to V$ that sends e_i to v_i.) □

4.3. Linear transformations and bases

We have seen above that to each matrix $A \in M_{m\times n}(\mathbb{F})$ we obtain a matrix transformation $T_A \colon \mathbb{F}^n \to \mathbb{F}^m$. We will now show that indeed every linear transformation $T \colon \mathbb{F}^n \to \mathbb{F}^m$ is a matrix transformation. In fact we will show more: every linear transformation $T \colon V \to W$ (regardless of V, W) can be essentially reformulated as a matrix transformation which opens the door to explicit computations involving linear transformations.

Let V be a vector space with basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ and let W be a vector space with basis $\mathcal{C} = (w_1, w_2, \ldots, w_m)$. Given a linear transformation

$$T \colon V \to W$$

we can ask the following natural question: Let $v \in V$ be any vector. Can we compute $[T(v)]_{\mathcal{C}}$ if we know $[v]_{\mathcal{B}}$?

So let X be the coordinate vector of v with respect to $\mathcal{B}$; thus $X = [v]_{\mathcal{B}}$ (and $\mathcal{B}X = v$; see (3.18)). Then $v = x_1v_1 + \cdots + x_nv_n$ from which we deduce immediately that

$$T(v) = x_1T(v_1) + x_2T(v_2) + \cdots + x_nT(v_n) \tag{4.1}$$

according to Proposition 4.13. It follows that

$$[T(v)]_{\mathcal{C}} = x_1[T(v_1)]_{\mathcal{C}} + x_2[T(v_2)]_{\mathcal{C}} + \cdots + x_n[T(v_n)]_{\mathcal{C}} \tag{4.2}$$

according to Proposition 4.13 c. applied to the isomorphism $W \to \mathbb{F}^m$, that maps w to $[w]_{\mathcal{C}}$.

(4.2) could be restated as

$$[T(v)]_{\mathcal{C}} = AX$$

where

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [v]_{\mathcal{B}}$$

and

$$A = \big[\, [T(v_1)]_{\mathcal{C}}\; [T(v_2)]_{\mathcal{C}}\; \ldots\; [T(v_n)]_{\mathcal{C}} \,\big].$$

If we write Y for the coordinate vector of $T(v)$ with respect to $\mathcal{C}$ then $Y = AX$, where A is the matrix whose j-th column is the coordinate vector of the image of the j-th basis vector in $\mathcal{B}$.

$$[T(v)]_{\mathcal{C}} = A[v]_{\mathcal{B}}. \tag{4.3}$$

For this reason

> A is called the *matrix of T with respect to the bases $\mathcal{B}$ of V and $\mathcal{C}$ of W.*
>
> It is determined by either one of the three requirements that
>
> - The columns of A are the coordinate vectors of the images of the basis vectors of V with respect to $\mathcal{C}$.
> - If $v \in V$ has coordinate vector $X \in \mathbb{F}^n$, then $T(v) \in W$ has coordinate vector $Y = AX$ in $\mathbb{F}^m$.
> - A is the $m \times n$ matrix with entries a_{ij} defined by the requirement that
>
> $$T(v_j) = \sum_{i=1}^{m} a_{ij} w_i.$$
>
> We sometimes write $M_{\mathcal{B}}^{\mathcal{C}}(T)$ for the matrix of T.

A basis is a choice of coordinates on a vector space; this shows how we can compute the image $T(v)$ of a linear transformation using coordinates.

Let $A \in M_{m\times n}(\mathbb{F})$. Notice that if $\mathcal{B}$ and $\mathcal{C}$ are the standard bases of $\mathbb{F}^n$ and $\mathbb{F}^m$, respectively, then the matrix of T_A with respect to $\mathcal{B}$ and $\mathcal{C}$ is equal to A: indeed, for each $v \in \mathbb{F}^n$ it is immediate that $[v]_{\mathcal{B}} = v$ if $\mathcal{B}$ is the standard basis (and similar $[w]_{\mathcal{C}} = w$ for all $w \in \mathbb{F}^m$). Hence $[T_A(v)]_{\mathcal{C}} = [Av] = Av = A[v]_{\mathcal{B}}$. More importantly, the converse also holds: if $T\colon \mathbb{F}^n \to \mathbb{F}^m$ is any linear transformation, then $T = T_A$ where $A = M_{\mathcal{B}}^{\mathcal{C}}(T)$: indeed, $T(v) = [T(v)]_{\mathcal{C}} = A[v]_{\mathcal{B}} = Av$. (The reader is advised to carefully trace all statements in this paragraph and maybe even write them out as formulas involving the entries of v and $T(v)$, respectively.)

We conclude

> *Every* linear transformation $T\colon \mathbb{F}^n \to \mathbb{F}^m$ is a matrix transformation: $T = T_A$ where A is determined by the requirement that the ith column A_i of A is
>
> $$A_i = T(e_i).$$

Remark. Usually, if $T\colon V \to V$ is a linear transformation, we only choose one basis (for the domain and codomain). It makes usually (as always there may be exceptions) not a lot of sense to choose two different bases $\mathcal{B}, \mathcal{C}$ of V (so in a sense to choose different coordinates for v and $T(v)$).

So normally, whenever $V = W$, we also want that $\mathcal{B} = \mathcal{C}$: indeed, if $v, T(v)$ are both elements of the same vector space V, so we want to write them as a linear combination of the same basis.

Example. Let $V = \{f \in \mathcal{P}(\mathbb{R}) \mid \deg f \leq 5\}$. Let $D\colon V \to V$ be defined by $D(f) = f'$, ie.

$$D(a_0 + a_1 x + \cdots + a_5 x^5) = a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + 5a_5 x^4.$$

Let $\mathcal{B} = \mathcal{C} = (1, x, x^2, \ldots, x^5)$. Then

$$M^{\mathcal{B}}_{\mathcal{B}}(D) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Indeed, the image of the jth basis vector is $D(x^{j-1}) = (j-1)x^{j-2}$ (if $j > 1$ and 0 if $j = 1$. Thus, for $j = 2, 3, \ldots, 6$ we have $[D(x^{j-1})]_{\mathcal{B}} = (j-1)e_{j-2}$ and $[D(1)]_{\mathcal{B}} = 0$.

Example. Let $R\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the counter clockwise rotation by the angle α. Then R is a linear transformation, and with respect to the standard basis $\mathcal{E} = (e_1, e_2)$ we have

$$M^{\mathcal{E}}_{\mathcal{E}}(R) = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

4.24 Example. Let $c \in \mathbb{F}$ and let $T\colon V \to V$ be the linear transformation defined $T(v) = cv$. Let $\mathcal{B} = (v_1, v_2, \ldots, v_n)$. As per the remark above, we only choose one basis here.
Then

$$T(v_i) = cv_i$$

and hence

$$[T(v_i)]_{\mathcal{B}} = ce_i.$$

It follows that $M^{\mathcal{B}}_{\mathcal{B}}(T) = cI_n$. In particular, if $c = 0$ then $M^{\mathcal{B}}_{\mathcal{B}}(T) = 0$ is the $n \times n$ zero matrix and if $c = 1$, then $M^{\mathcal{B}}_{\mathcal{B}}(T)$ is equal to I_n.

Note that the matrix is independent of the basis $\mathcal{B}$ chosen. No matter what basis we choose, the matrix of this T will always be the same.

One can show (it is not that difficult), that the only linear transformation with this property are of the form $v \mapsto cv$.

4.25 Example. Let $T\colon V \to W$ be a linear transformation. Let $\mathcal{B}$ be a basis for V and $\mathcal{C}$ be a basis for W. Suppose $A = M^{\mathcal{C}}_{\mathcal{B}}(T)$.

a. The column space of A is in one-to-one correspondence with the image of T: indeed, if $Y \in \operatorname{Col}(A)$ then $Y = [w]_{\mathcal{C}}$ for a unique $w \in \operatorname{im} T$. Indeed, since $Y = AX$ for some $X \in \mathbb{F}^n$, $w = T(v)$ where $v = \mathcal{B}X$.

   Moreover, if $(Y_1, Y_2, \ldots, Y_p)$ is a basis for $\operatorname{Col}(A)$ then $(\mathcal{C}Y_1, \mathcal{C}Y_2, \ldots, \mathcal{C}Y_p)$ is a basis for $\operatorname{im} T$:

$$Y \in \operatorname{Col}(A) \iff \mathcal{C}Y \in \operatorname{im}(T).$$

b. The null space of A is in one to one correspondence to the kernel of T: indeed, if $AX = 0$ then X is the coordinate vector of $v = \mathcal{B}X \in \mathcal{N}(T)$. Similarly, if $(X_1, X_2, \ldots, X_q)$ is a basis of $\mathcal{N}(A)$, then $(\mathcal{B}X_1, \mathcal{B}X_2, \ldots, \mathcal{B}X_q)$ is a basis of $\mathcal{N}(T)$. This allows for explicit computations.

$$X \in \mathcal{N}(A) \iff \mathcal{B}X \in \mathcal{N}(T).$$

4.3.1 Problem. Let V, W be $\mathbb{F}$-vector spaces and let $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ be a basis for V and let $w_1, w_2, \ldots, w_n$ be arbitrary elements of W.

Show that there exists a unique linear transformation $T\colon V \to W$ such that $T(v_i) = w_i$ for $i = 1, 2, \ldots, n$.

(*Hint:* If $v \in V$ then $v = x_1v_1 + x_2v_2 + \cdots + x_nv_n$ for uniquely determined $x_i \in \mathbb{F}$. Define $T(v)$ as $T(v) = x_1w_1 + x_2w_2 + \cdots + x_nw_n$.)

4.26 Example. Let $V = \mathbb{R}^2$ and $v = (1, 1)$, $w = (-1, 1)$. By the previous problem, there exists a unique $T\colon V \to V$ such that $T(v) = w$ and $T(w) = v$: indeed, $\mathcal{B} = (v, w)$ is a basis for $\mathbb{R}^2$ (v, w are linearly independent) so apply Problem 4.3.1 to the $\mathcal{B}$, $W = V$, and $w_1 = w$ and $w_2 = v$.

Then

$$M^{\mathcal{B}}_{\mathcal{B}}(T) = \begin{bmatrix} & 1 \\ 1 & \end{bmatrix}$$

By the remarks above, T is a matrix transformation. That is, for each $X \in \mathbb{R}^2$, $T(X) = AX$. What is A? In other words, what is $M^{\mathcal{E}}_{\mathcal{E}}(T)$, where as usual $\mathcal{E} = (e_1, e_2)$ is the standard basis?

Applying the definition, A is given as

$$A = [\, T(e_1)\, T(e_2)\,].$$

Note that here there is no need to write $[T(e_i)]_{\mathcal{E}}$ because for any column vector $X \in \mathbb{R}^2$ we have $[X]_{\mathcal{E}} = X$.

Now

$$e_1 = \frac{1}{2}v - \frac{1}{2}w$$
$$e_2 = \frac{1}{2}v + \frac{1}{2}w$$

from which we conclude that

$$T(e_1) = \frac{1}{2}w - \frac{1}{2}v = -e_1$$
$$T(e_2) = \frac{1}{2}w + \frac{1}{2}v = e_2$$

We conclude that

$$A = \begin{bmatrix} -1 & \\ & 1 \end{bmatrix}$$

(which we also recognize as the matrix of the reflection in the x-axis, where we as usual identify $(x, y) \in \mathbb{R}^2$ with a point with these coordinates in the plane).

Remark. It is often crucial to be careful when picking a basis. Sometimes it is possible to simplify the matrix of a linear transformation T greatly that way. For instance, if $T\colon V \to V$ is a linear transformation, it may happen that when choosing $\mathcal{B}$ carefully that $M_{\mathcal{B}}^{\mathcal{B}}(T)$ is diagonal. Then it becomes rather easy to compute eg. the matrix of T^n (the $n$fold composition of T with itself).

One should think of the bases $\mathcal{B}$ and $\mathcal{C}$ as a mechanism of "translating" vectors in V and W respectively into corresponding vectors in $\mathbb{F}^n$ and $\mathbb{F}^m$. Any problem in V will be transferred into a problem in $\mathbb{F}^n$ where we hopefully can solve it. The following diagram hopefully clarifies this thought.

$$\begin{tikzcd}
v & V \arrow[r] \arrow[d, "T"] & \mathbb{F}^n \arrow[d, "T_A"] & X = [v]_{\mathcal{B}} \\
T(v) & W \arrow[r] & \mathbb{F}^m & {[T(v)]_{\mathcal{C}} = AX}
\end{tikzcd}$$

The main results in this area is the following:

4.27 Theorem. *Let V, W be vector spaces over $\mathbb{F}$ with bases $\mathcal{B}$ and $\mathcal{C}$ respectively, where V has dimension n and W has dimension m. Then the map*

$$M\colon \operatorname{Hom}(V, W) \to M_{m\times n}(\mathbb{F})$$

defined by $M(T) = M_{\mathcal{B}}^{\mathcal{C}}(T)$ is an isomorphism of vector spaces.

Proof. This is a homework problem; but compare Example 4.11 □

The ultimate reason why matrix multiplication is defined the way it is, is the following:

4.28 Theorem. *Let U, V, W be vector spaces with bases $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ respectively. Let $S\colon U \to V$ and $T\colon V \to W$ be linear transformations. Then the matrix of $TS\colon U \to W$ with respect to $\mathcal{B}_1$ and $\mathcal{B}_3$ is*

$$M_{\mathcal{B}_1}^{\mathcal{B}_3}(TS) = M_{\mathcal{B}_2}^{\mathcal{B}_3}(T) \cdot M_{\mathcal{B}_1}^{\mathcal{B}_2}(S)$$

where the right hand side means multiplication of matrices.

Proof. This is a straight forward computation. To avoid ridiculously convoluted notation let $C = M_{\mathcal{B}_1}^{\mathcal{B}_3}(TS)$, $B = M_{\mathcal{B}_2}^{\mathcal{B}_3}(T)$, and $A = M_{\mathcal{B}_1}^{\mathcal{B}_2}(S)$.

Let $\mathcal{B}_1 = (v_1, v_2, \ldots, v_n)$. Then the jth column of C is by definition

$$[TS(v_j)]_{\mathcal{B}_3} = [T(S(v_j))]_{\mathcal{B}_3} = B[S(v_j)]_{\mathcal{B}_2} = B(A[v_j]_{\mathcal{B}_1}).$$

Keeping in mind that $[v_j]_{\mathcal{B}_1} = e_j$ (the jth column of I_n) this means the jth column C_j of C is equal to

$$C_j = B(Ae_j) = BA_j$$

where A_j is the jth column of A. By the definition of matrix multiplication this means that

$$C = BA$$

as claimed. □

Corollary. *The matrix multiplication is associative.*

Proof. Let A, B, C be matrices of sizes $m \times n$, $n \times p$, $p \times q$, respectively.

Then $A(BC)$ and $(AB)C$ are both defined. Moreover, $A(BC)$ is the matrix of $T_{A(BC)}$ with respect to the standard bases of $\mathbb{F}^q$ and $\mathbb{F}^m$ respectively. Moreover, by the theorem, $T_A T_{BC}$ has matrix $A(BC)$ as well. Now T_{BC} has matrix BC as does $T_B T_C$. Hence $T_B T_C = T_{BC}$. It follows that $T_A T_{BC} = T_A(T_B T_C) = (T_A T_B)T_C = T_{AB}T_C$, which is immediate from evaluation: if $x \in \mathbb{F}^m$, then

$$(T_A(T_B T_C))(x) = T_A((T_B T_C)(x)) = T_A(T_B(T_C(x))) = (T_A T_B)(T_C(x)) = ((T_A T_B)T_C)(x).$$

□

5. The Determinant

We will now develop the theory of determinants: First and foremost, we want to develop numerical criteria for matrices to be invertible[1]. Also, determinants naturally appear two a priori unrelated contexts: they appear in substitution rules for integration in several variables; and they determine certain volumina.

The area function on the plane

In this subsection, let us identify $E = \mathbb{R}^2$ with the points of the Euclidean plane. Given two elements $v, w \in \mathbb{R}^2$, they determine a *parallelogram* in E: indeed, consider the figure with vertices $0, v, w, v + w$. Then the line through 0 and v is parallel to the line through w and $v + w$; similarly, the line through v and $v + w$ does not intersect the line through 0 and w (assuming v, w are linearly independent).

Let $A(v, w)$ denote the *area* of this parallelogram[2]. Instead of writing down a formula for $A(v, w)$ in terms of the entries of v, w, let us first discuss some of the main properties we would expect of such an area function of A: If $v = e_1$ and $w = e_2$, then the parallelogram is simply a square with sides of length 1; it seems natural to require

$$A(e_1, e_2) = 1. \tag{5.1}$$

Also, if we scale one of the sides of the parallelogram by a (positive) scalar λ, the area is multiplied by the same scalar:

$$A(\lambda v, w) = A(v, \lambda w) = \lambda A(v, w). \tag{5.2}$$

As you may recall from geometry, the area of a parallelogram only depends on its height (as measured as the distance between two parallel sides) and the length of a side. Thus, if we add any multiple of w to v, we still have the same area:

$$A(v + \lambda w, w) = A(v, w) \tag{5.3}$$

for all $\lambda \in \mathbb{R}$.

[1] This sounds tempting: instead of applying Gaussian elimination, you simply compute a number! However the computation of this number is so involved, that Gaussian elimination (done right, that is, only until we have an upper triangular matrix) is often faster.

[2] Strictly speaking, we would have to define what "area" means; we will avoid this, and simply pretend that we are experts in Euclidean geometry and well versed with notions of area and volume etc.

It turns out that these requirements completely determine A. In fact, it will follow that if v, w are linearly dependent, then $A(v, w) = 0$: indeed, if $v = \lambda w$, say, then $A(v, w) = \lambda A(v, v)$. But $A(v, v) = 0$ because $A(v, v) = A(v + v, v) = 2A(v, v)$.

One can show that there is one and only one way to define an area function that behaves like this, and the result is

$$A(v, w) = |ad - bc| \tag{*}$$

where for a real number x, $|x|$ denotes its absolute value and

$$v = \begin{bmatrix} a \\ b \end{bmatrix}, w = \begin{bmatrix} c \\ d \end{bmatrix}$$

Note that not once did we need an actual definition of the word "area" here. All we needed was an understanding of what our intuition (or Euclidean geometry) dictates that we should require of an area function.

That $A(v, w)$, defined as in (*), is actually an area function satisfying the above properties is easy to verify by straight forward computation. That it is the only one is not that hard. The reasoning follows similar arguments as we will employ in the next section for the determinant. We omit the details.

5.1. The definition and first properties of the determinant

5.1 Example. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

with *real* entries, say. Let us view the columns of A as points in the plane. Basic geometry tells us that if we want to write e_1 and e_2 (that is the unit vectors along the axis) in the form

$$x_1 \begin{bmatrix} a \\ c \end{bmatrix} + x_2 \begin{bmatrix} b \\ d \end{bmatrix}$$

then the two vectors $[\, a\, c\,]^T$ and $[\, b\, d\,]^T$ cannot be colinear (lying on one line). Two such vectors are colinear if and only if $ad = bc$ (cf. 2.1). Also, from a homework problem we recall that A is invertible if and only if $ad - bc \neq 0$; indeed, $AX = 0$ has only the solution $X = 0$ if and only if the two lines $ax + by = 0$ and $cx + dy = 0$ only intersect in $(0, 0)$.

For 2×2 matrices $\det A$ is defined as $ad - bc$ and called the *determinant* of the matrix A.

Note that this is about the only size of a matrix where computing a determinant is not a pain.

Also, if $\det A \neq 0$ then

$$A^{-1} = \frac{1}{\det A} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

The theoretical value of an expression like this is that for instance in the case of $\mathbb{F} = \mathbb{R}$, the entries of A^{-1} depend *continuously* (in fact differentiably) on the entries of A: they are rational functions in the entries of A. So if the entries of B are "close" to the entries of A, then also the entries of B^{-1} will be "close" to the entries of A^{-1}.

It is a fascinating thought to try and find a numerical criterion that determines whether an $n \times n$ matrix is invertible, and in fact to compute an inverse by a formula, rather than a long and tedious Gaussian elimination.

While we will now construct such a number for each matrix, and while we will find a formula for the inverse, it turns out that in practical terms, usually Gaussian elimination is still way faster to compute an inverse (or even determine whether a matrix is invertible). However, for theoretical purposes, the determinant and the resulting formulas are extremely important.

The concept itself of a determinant is not that hard. It is simply a function. However, it is not so easy to give a convincing reason how anyone could come up with this function in the first place. Here is an after the fact explanation on why the determinant does what it does:

Suppose we have a function $d\colon M_n(\mathbb{F}) \to \mathbb{F}$ such that $d(A) \neq 0$ if and only if A is invertible. What are properties of d we would naturally require?

Let $A_1, A_2, \ldots, A_{n-1}$ be $n-1$ row vectors in $M_{1\times n}(\mathbb{F})$. We could now ask for which row vectors X the $n \times n$ matrix

$$A_X := \left[\begin{array}{c} X \\ \hline A_1 \\ \hline A_2 \\ \hline \vdots \\ \hline A_{n-1} \end{array}\right] \tag{5.4}$$

invertible. Since we have the determinant, this is simply asking the question for which X is $f(X) := d(A_X)$ nonzero?

For instance, we would immediately deduce that $f(0_{M_{1\times n}(\mathbb{F})}) = 0$.

Also, if any two rows of A_X coincide then of course, A_X is not invertible and hence $f(X) = 0$. In particular, $f(A_1) = f(A_2) = \cdots = f(A_{n-1}) = 0$.

Also, if A_X and A_Y both are not invertible then so is A_{X+Y}: Indeed, we know that a matrix A is invertible if and only if its columns are linearly independent. We know that this happens if and only if its rows are linearly independent as well (because $\dim \mathrm{Row}(A) = \dim \mathrm{Col}(A)$). Now, if $A_1, A_2, \ldots, A_{n-1}$ are linearly dependent, then so are $X, A_1, A_2, \ldots, A_{n-1}$ for each X and hence $f = 0$ is the zero function (in particular, $f(X+Y) = 0$ for all X, Y). On the other hand, if $A_1, A_2, \ldots, A_{n-1}$ are linearly independent, then $f(X) = 0$ if and only if $X \in \mathrm{Span}(A_1, A_2, \ldots, A_{n-1})$. Thus, if $f(X) = 0$ and $f(Y) = 0$ then X, Y and hence also $X + Y \in \mathrm{Span}(A_1, A_2, \ldots, A_{n-1})$, so A_{X+Y} is not invertible which means $f(X+Y) = 0$. Thus, if $f(X) = f(Y) = 0$ then so is $f(X+Y)$.

Similarly, $A_{\lambda X}$ is not invertible, if A_X isn't so $f(\lambda X) = 0$ whenever $f(X) = 0$. And also $f(\lambda X) \neq 0$ whenever $f(X), \lambda \neq 0$.

All these properties of f would be naturally satisfied if $f(X+Y) = f(X) + f(Y)$ and if $f(\lambda X) = \lambda f(X)$.

If $d\colon M_n(\mathbb{F}) \to \mathbb{F}$ satisfies this property, we say d is *linear in the first row.*

Since nothing is special about the first row, it would be natural to require that d is linear in every row.

Similarly, the order of the rows of a matrix does not influence the fact whether it is invertible or not (since the reduced row echelon form does not depend on the order of the rows of A). Thus, if E is the elementary matrix of a row operation of Type I, then $d(EA) = 0$ if and only if $d(A) = 0$. It turns out that linearity in the rows and the fact that $d(A) = 0$ whenever to rows of A coincide already determine $d(EA)$: To see why this is true consider the following: For each X, Y, we have

$$\begin{aligned}
0 = d\left[\begin{array}{c} \vdots \\ \hline X+Y \\ \hline \vdots \\ \hline X+Y \\ \hline \vdots \end{array}\right] &= d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \vdots \\ \hline X+Y \\ \hline \vdots \end{array}\right] + d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \vdots \\ \hline X+Y \\ \hline \vdots \end{array}\right] \\
&= d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \vdots \\ \hline X \\ \hline \vdots \end{array}\right] + d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] + d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \vdots \\ \hline X \\ \hline \vdots \end{array}\right] + d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] \\
&= d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] + d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \vdots \\ \hline X \\ \hline \vdots \end{array}\right]
\end{aligned}$$

And so

$$d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] = -d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \vdots \\ \hline X \\ \hline \vdots \end{array}\right]$$

5.2 Theorem. *There is a unique function $d\colon M_n(\mathbb{F}) \to \mathbb{F}$ having the following three properties:*

a. *If A has two identical rows, then $d(A) = 0$.*

b. *d is linear in each row.*

c. $d(I) = 1$.

Before we prove this theorem let us first discuss what it actually tells us about d: in the following, we will assume that we are given a function $d\colon M_n(\mathbb{F}) \to \mathbb{F}$ that satisfies a. b. and c. of Theorem 5.2.

If A is diagonal, then we can immediately compute $d(A)$: Indeed, as d is linear in each row, if

$$A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix}$$

then

$$d(A) = a_1 d \begin{bmatrix} 1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix} = \cdots = a_1 a_2 \cdots a_n d(I) = a_1 \cdots a_n.$$

More generally, whenever E is a Type III elementary matrix then $d(EA) = cd(A)$ where c is the nonzero scalar used in the definition of E. We already discussed above that $d(EA) = -d(A)$ for all Type I elementary matrices. That leaves operations of Type II: suppose we want to add a times row i to row j in a square matrix, corresponding to the elementary matrix E. Then

$$d\left[\begin{array}{c} \vdots \\ \hline X + aY \\ \hline \\ \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] = d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \\ \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] + a \cdot d\left[\begin{array}{c} \vdots \\ \hline Y \\ \hline \\ \vdots \\ \hline Y \\ \hline \vdots \end{array}\right] = d\left[\begin{array}{c} \vdots \\ \hline X \\ \hline \\ \vdots \\ \hline Y \\ \hline \vdots \end{array}\right]$$

Note that here we used linearity in row i and the fact that d vanishes on matrices with identical rows.

5.3 Observation. *The function d transforms as follows with respect to row operations:*

a. *Type I operations change the sign*[3] *of d: $d(EA) = -d(A)$ if E is a Type I elementary matrix.*

[3] This is a very sloppy way of expressing this. If $\mathbb{F}$ is not the field of rational or real numbers, then there is usually no "sign" attached to any number. For instance, the complex numbers cannot be divided into "positive" and "negative" complex numbers such that this is compatible with the multiplicative structure (i.e. such that "positive times positive = positive, negative times negative = positive, and positive times negative = negative".

b. *Type II operations don't affect the value of d: $d(EA) = d(A)$ whenever E is a Type II elementary matrix.*

c. *A Type III operation affects d as follows: if E is the matrix corresponding to scaling a row by $c \neq 0$, then $d(EA) = cd(A)$.*

Notice that picking $A = I$ and using that $d(I) = 1$ these observations mean that $d(E) = -1$ if E is a Type I, $d(E) = 1$ if L is a Type II, and $d(E) = c$ if E is a Type III elementary matrix.

Remember: Keeping this in mind, Observation 5.3 can be restated as $d(EA) = d(E)d(A)$ whenever E is an elementary matrix.

What else does the theorem tell us? Let $A = [a_{ij}]$ be an $n \times n$ matrix and put

$$f_i = [\underbrace{0\,0\,\ldots\,0\,1}_{i}\,0\,\ldots\,0]$$

be the $1 \times n$ row vector with a 1 in column i and 0 everywhere else; that is, f_i is the ith row of I_n.

Note that the first row of A is then $a_{11}f_1 + a_{12}f_2 + \cdots + a_{1n}f_n$. Using the notation from (5.4) and applying linearity repeatedly we find that

$$\begin{aligned} d(A) = d(A_{a_{11}f_1}) + d(A_{a_{12}f_1}) + \cdots + d(A_{a_{1n}f_n}) \\ = a_{11}d(A_{f_1}) + a_{12}d(A_{f_2}) + \cdots + a_{1n}d(A_{f_n}). \end{aligned}$$

Now let us consider $B = A_{f_1}$ for the moment. Using linearity in the *second* row and denoting by B_X the matrix obtained from B by replacing the second row with a row vector X, we deduce, as above,

$$d(B) = a_{21}d(B_{f_1}) + a_{22}d(B_{f_2}) + \cdots + a_{2n}d(B_{f_n}).$$

But notice that $d(B_{f_1}) = 0$ because in the first and second row of B_{f_1} we have the same row vector, namely f_1. Now we can replace in each B_{f_j} the third row and continue until it is impossible to go on further. Doing the same to all A_{f_i} results in

$$d(A) = \sum c_P d(P) \tag{5.5}$$

where the sum ranges over certain matrices P satisfying the following properties:

a. Every entry of P is either 0 or 1.

b. Every row and column of P contains one and only one 1.

Matrices with this property are called *permutation matrices* (for reasons that will be clear in a minute).

Indeed, for $1 \leq k \leq n$ and $1 \leq i_1, i_2 \ldots, i_k \leq n$ let $A_{i_1,i_2,\ldots,i_k}$ be the matrix obtained from A by replacing the first k rows by $f_{i_1}, f_{i_2}, \ldots, f_{i_k}$.

Then, more formally, we get

$$
\begin{aligned}
(5.6)\quad d(A) = \sum_{i_1=1}^{n} a_{1i}d(A_{i_1}) &= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} a_{1i_1}a_{2i_2}d(A_{i_1,i_2}) \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_k=1}^{n} a_{1i_1}a_{2i_2}\cdots a_{ki_k}d(A_{i_1,i_2,\ldots,i_k}) \\
&= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_n=1}^{n} a_{1i_1}a_{2i_2}\cdots a_{ni_n}d(A_{i_1,i_2,\ldots,i_n}).
\end{aligned}
$$

Notice that indeed, if after we used linearity in each row of A, the matrices left are the ones with f_is for rows. Also, observe that if $i_k = i_\ell$ for some $k \neq \ell \leq n$ then

$$d(A_{i_1,i_2,\ldots,i_n}) = 0$$

because $A_{i_1,i_2,\ldots,i_n}$ will have two identical rows (namely rows k and ℓ will both be equal to f_{i_k}). Thus, we can omit all sequences $i_1, i_2, \ldots, i_n$ with duplicate entries.

If P is a matrix of the form $A_{i_1,i_2,\ldots,i_n}$ (where all the i_k are distinct), then no column contains more than one 1. On the other hand, each column must contain a 1 by construction. Hence the two properties; and in particular, P does not depend on A at all; it is completely determined by the sequence $i_1, i_2, \ldots, i_n$.

As a consequence, every row vector of the form f_i will appear as a row of P. So for P let us denote by σ_P the function $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ defined by $\sigma(i) = j$ if the f_i appears in row j of P. Notice that σ is a *permutation*, ie. a bijection of $\{1, 2, \ldots, n\}$ onto itself.

If $P = A_{i_1,i_2,\ldots,i_n}$ then $k = \sigma_P(i_k)$.

What is the coefficient c_P of P in (5.5)? If you carefully follow the construction, you will see that

$$c_P = a_{1i_1}a_{2i_2}\cdots a_{ni_n} = a_{\sigma_P(1)1}a_{\sigma_P(2)2}\cdots a_{\sigma_P(n)n}.$$

Moreover, *each* possible permutation matrix will appear exactly once.

How can we recover P given a permutation σ? We simply apply σ to the rows of the identity matrix: we move row i into row $\sigma(i)$. We say the P so constructed is the *permutation matrix associated to* σ and denote it by P_σ.

5.4 Definition. The set of all permutations of $\{1, 2, \ldots, n\}$ is denoted S_n and called *the symmetric group in n letters*. The set of all $n \times n$-permutation matrices is denoted by Π_n.

5.1.1 Problem. Show that if $\sigma, \mu \in S_n$ then also $\sigma \circ \mu \in S_n$. Show that for each $\sigma \in S_n$ there exists an inverse permutation σ^{-1} such that $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = \mathbf{1}$, where we denote the permutation that fixes each i by $\mathbf{1}$.

We also write $\sigma\mu$ instead of $\sigma \circ \mu$.

Warning: $\sigma\mu$ is the permutation obtained by *first* applying μ and *then* σ; not the other way around (which is also common in mathematical literature). As a reminder: the one closer to the argument is applied first: $\sigma\mu(i) = \sigma(\mu(i))$.

5.1.2 Problem. Show that $\Pi_n \subseteq \mathrm{GL}_n(\mathbb{F})$ is a subgroup. Also, check that $P_{\sigma\mu} = P_\sigma \cdot P_\mu$ for all $\sigma, \mu \in S_n$.

Can you prove that P_σ is orthogonal? That is, $P_\sigma^{-1} = P_{\sigma^{-1}} = P_\sigma^T$.

An elementary exercise shows that every permutation matrix may be obtained from I be a sequence of Type I operations. In fact the elementary Type I matrices are permutation matrices themselves. It follows that if P is a permutation matrix then $P = E_1E_2 \cdots E_p$ for p permutation matrices of Type I: σ_{E_i} will fix all but two integers and exchange the remaining two. Since we know that $d(E_i) = -1$ it follows that

$$d(P) = (-1)^p.$$

The following are all 2×2 permutation matrices:

$$\begin{bmatrix} 1 & \\ & 1 \end{bmatrix}, \begin{bmatrix} & 1 \\ 1 & \end{bmatrix} \tag{5.7}$$

For $n = 3$ there are already 6:

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}, \begin{bmatrix} 1 & & \\ & & 1 \\ & 1 & \end{bmatrix}, \begin{bmatrix} & 1 & \\ 1 & & \\ & & 1 \end{bmatrix}, \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix}, \begin{bmatrix} & & 1 \\ 1 & & \\ & 1 & \end{bmatrix}, \begin{bmatrix} & 1 & \\ & & 1 \\ 1 & & \end{bmatrix} \tag{5.8}$$

For the first, fifth, and sixth matrix p can be chosen to be equal 2; whereas for matrices $2, 3, 4$, $p = 1$ works.

In general there are $n!$ permutations and hence $n!$ permutation matrices in Π_n.

5.1.3 Problem. Use induction to show that

$$|\Pi_n| = |S_n| = n!.$$

It is tempting to use (5.5) or (5.6) as a definition for the determinant. And indeed one can do that. However, the problem here is that we would need an independent definition of $d(P)$. We have seen that if d exists then $d(P) = (-1)^p$ where p is the number of elementary Type I matrices that multiply together to give P. Unfortunately, this number p is not determined by P, so we would need to show that at least its parity[4] is determined by P. This can be done; and then one can *define* the determinant using what is known as Leibniz Formula:

$$d(A) = \sum_{\sigma \in S_n} d(P_\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n}. \tag{5.9}$$

[4] whether it is even or odd

In any event however, our discussion established that d is *unique* if it exists: indeed, if d, d' are two functions satisfying the three properties of Theorem 5.2, then

$$d(A) = \sum_{\sigma\in S_n} a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}d(P_\sigma) = \sum_{\sigma\in S_n} a_{\sigma(1)1}a_{\sigma(2)2}\cdots a_{\sigma(n)n}d'(P_\sigma) = d'(A)$$

because $d(P) = d'(P) = (-1)^p$ for any permutation matrix P.

Also for $n = 2$, this does give a well-defined determinant: here $d(P) = 1$ if $P = I$ and -1 otherwise. Hence

$$(5.10)\qquad d\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{\mathbf{1}(1)1}a_{\mathbf{1}(2)2} - a_{\sigma(1)1}a_{\sigma(2)2} = a_{11}a_{22} - a_{21}a_{12}$$

where $\sigma(1) = 2$ and $\sigma(2) = 1$. Of course, one still needs to verify that this so defined d actually does satisfy the properties mentioned in Theorem 5.2.

Also, for $n = 3$ we get the rather unpleasant formula

$$d(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}.$$

5.1.1. Proof of Theorem 5.2

Instead of the direct approach, that is, instead of defining the determinant by Leibniz Formula (5.9), we use a recursive definition. For $n = 1$, we define $d([\,a\,]) = a$.

Let $A \in M_n(\mathbb{F})$ be a a matrix (where $n \geq 2$). For $i = 1, 2, \ldots, n$ let A_i be the $n \times n$ matrix obtained from A by deleting the first column and i-th row. So if

$$A = \begin{bmatrix} \mathbf{a_{11}} & a_{12} & \dots & a_{1n} \\ \mathbf{a_{21}} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{a_{i1}} & \mathbf{a_{i2}} & \dots & \mathbf{a_{in}} \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{a_{n1}} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

Then

$$A_i = \begin{bmatrix} a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \dots & \vdots \\ a_{i-1,1} & a_{i-1,2} & \dots & a_{i-1,n} \\ a_{i+1,1} & a_{i+1,2} & \dots & a_{i+1,n} \\ \vdots & \vdots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

5.5 Definition. For $n = 1$ we define $d_1 \colon M_1(\mathbb{F}) \to \mathbb{F}$ by $d([a]) = a$.

Suppose $d_{n-1} \colon M_{n-1}(\mathbb{F}) \to \mathbb{F}$ has been defined. We define

$$d_n(A) = a_{11}d_{n-1}(A_1) - a_{21}d_{n-1}(A_2) + a_{31}d_{n-1}(A_3) - \cdots \pm a_{n1}d_{n-1}(A_n).$$

Thus

$$d_n(A) = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} d_{n-1}(A_i) \tag{5.11}$$

Example. For $n = 2$ this really amounts to the formula

$$d_2 \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad_1([d]) - cd_1([b]) = ad - cb = ad - bc.$$

For $n = 3$ one obtains the formula we found by Leibniz above. Concretely here eg.

$$d_3 \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} = 1 \cdot d_2 \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} - 2 \cdot d_2 \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix} + 1 \cdot d_2 \begin{bmatrix} 2 & 3 \\ 0 & 0 \end{bmatrix} = 0 - 2 \cdot (-1) + 0 = 2.$$

Observation I: By the Recursive Definition Theorem 1.15, there exists a unique collection of functions $d_n \colon M_n(\mathbb{F}) \to \mathbb{F}$ satisfying the Recursion 5.11 for all n.

Now we will prove the existence part of Theorem 5.2 by induction on n.

For the base case $n = 1$, it is clear that $d([1]) = 1$ and $d([ab]) = ab = ad[b]$ is linear, and hence d has all necessary properties.

Induction assumption: From now on we assume that an integer $n > 1$ is given and that it is known that the function d of Definition 5.5 satisfies Properties a. b. and c. of Theorem 5.2 for all $k \times k$-matrices where $k < n$.

In particular, d_k has all the properties exhibited so far: for instance it satisfies the observations in 5.3, as long as $k < n$.

Claim: If two rows in A coincide then $d(A) = 0$.

Proof. Suppose $i < j$ and rows i and j in A coincide. By induction, then, $d_{n-1}(A_k) = 0$ for all $k \neq i, j$ because A_k will also contain two identical rows. Thus

$$d_n(A) = \sum_{k=1}^{n} (-1)^{k+1} a_{k1} d_{n-1}(A_k) = (-1)^{i+1} a_{i1} d_{n-1}(A_i) + (-1)^{j+1} a_{j1} d_{n-1}(A_j)$$

How do A_i and A_j differ? Let $a_1, a_2, \ldots, a_n$ be the rows of A. Let a_i' be the rows of A with the first element removed. Then

$$A_i = \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_{i-1}' \\ a_{i+1}' \\ \vdots \\ a_j' \\ \vdots \\ a_n' \end{bmatrix} \qquad A_j = \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_i' \\ \vdots \\ a_{j-1}' \\ a_{j+1}' \\ \vdots \\ a_n' \end{bmatrix}$$

Since $a_i' = a_j'$, A_i is obtained from A_j by exchanging the rows $(i+1,i)$,$(i+1,i+2)$, …, $(j-2,j-1)$; each exchange changes d_{n-1} by a factor -1, and so $d_{n-1}(A_i) = (-1)^{j-i-1}d_{n-1}(A_j)$. This, together with the fact that $a_{i1} = a_{j1}$ shows that $d_n(A) = 0$. □

We can now also verify the second property:

Claim: d is linear in the rows.

Proof. Let A be an $n \times n$-matrix. We will now check that d is linear in row i. Let $a_1, a_2, \ldots, a_n$ be the rows of A as above. Suppose $a_i = X + Y$ for some $X, Y \in M_{1\times n}(\mathbb{F})$.

For simplicity let A_Z denote the matrix where a_i has been replaced by a row Z. Then

$$d_n(A) = d_n(A_X) + d_n(A_Y).$$

Indeed, we observe that $d_{n-1}(A_j) = d_{n-1}((A_X)_j) + d_{n-1}((A_Y)_j)$ by induction if $j \neq i$ and the summand corresponding to row i in $d_n(A)$ is

$$a_{i1}d_{n-1}(A_i) = (x_1 + y_1)d_{n-1}(A_i) = x_1 d_{n-1}((A_X)_i) + y_1 d_{n-1}((A_Y)_i).$$

$((A_X)_i = (A_Y)_i = A_i)$.

Similarly, if $a_i = \lambda X$ for some $\lambda \in \mathbb{F}$ and $X \in M_{1\times n}(\mathbb{F})$, then

$$d_n(A) = \lambda d_n(A)$$

because again $a_{i1} = \lambda x_1$ and $(A_X)_i = A_i$, so by induction for each summand of $d_n(A)$ corresponding to row $j \neq i$, $d_{n-1}(A) = \lambda d_{n-1}((A_X)_i)$. □

Example.

$$d\begin{bmatrix} 1 & 0 & 7 \\ 3+2 & 2+9 & 1+(-1) \\ 2 & 1 & 0 \end{bmatrix} = d\begin{bmatrix} 1 & 0 & 7 \\ 3 & 2 & 1 \\ 2 & 1 & 0 \end{bmatrix} + d\begin{bmatrix} 1 & 0 & 7 \\ 2 & 9 & -1 \\ 2 & 1 & 0 \end{bmatrix}$$

It remains to check that

Claim: $d_n(I_n) = 1$.

Proof. This follows immediately from the formula and induction:

$$d_n(I_n) = \sum_{i=1}^{n} (-1)^{i+1} \delta_{i1} d((I_n)_i).$$

Since δ_{i1} is nonzero only if $i = 1$, this is equal to $d(I_n) = d(I_{n-1}) = 1$. □

Summarizing, we have shown that there exists a function d_n that satisfies all three properties required in Theorem 5.2. Since the existence of a function d is now guaranteed also its uniqueness follows because we now can conclude that d satisfies the Leibniz Formula (5.9). □

5.1.2. Properties of the determinant

5.6 Definition. The unique function d that exists according to Theorem 5.2 is called the *determinant*, and we now write $\det A$ instead of $d(A)$.

Let us quickly recap what we already know

5.7 Facts. The determinant satisfies the following properties.

a. $\det(I_n) = 1$.

b. $\det(E) = -1$ if L is a Type I elementary matrix.

c. $\det(E) = 1$ if L is a Type II elementary matrix.

d. $\det(E) = c$ if L is a Type III elementary matrix corresponding to scaling with $c \in \mathbb{F}$.

e. $\det(EA) = d(E)d(A)$ whenever E is an elementary matrix.

f. $\det(A) = 0$ whenever two rows of A coincide.

5.1.4 Problem. Show that $\det(A) = 0$ whenever A contains a row of all zeros. Show that $\det(cA) = c^n \det A$.

We now turn to the main property of the determinant. The multiplicative rule.

5.8 Proposition. *Let A, B be two $n \times n$ matrices. Then*

$$\det(AB) = \det(A)\det(B).$$

Proof. **Step I:** A is invertible.

If A is invertible then we know that A is a product of elementary matrices, $A = E_1E_2\cdots E_p$, say. Because of Fact e.) above,

$$\begin{aligned}\det(AB) = \det(E_1)\det(E_2\cdots E_pB)\\ = \det(E_1)\det(E_2)\det(E_3\cdots E_pB) = \dots\\ = \det(E_1)\det(E_2)\cdots\det(E_p)\det(B)\end{aligned}$$

The same argument applied to A itself shows that $\det(A) = \det(E_1)\det(E_2)\cdots\det(E_p)$. Hence $\det(AB) = \det(A)\det(B)$.

Step II: A is not invertible.

If A is not invertible, then a row echelon form U of A must contain a row of all zeros. Now $A = E_1E_2\cdots E_pU$ where the E_i are elementary matrices. It follows that

$$\det(A) = \det(E_1)\det(E_2)\cdots\det(E_p)\det(U) = 0,$$

because U has a row of all zeros (see the problem above). Hence $\det(A)\det(B) = 0$. We have to show that $\det(AB) = 0$ as well.

Now UB also has a row of all zeros (why?). Thus, $\det(UB) = 0$. Moreover, $AB = (E_1E_2\cdots E_p)UB$ so

$$\det(AB) = \det(E_1)\det(E_2)\cdots\det(E_p)\det(UB) = 0.$$

□

5.9 Corollary (of proof)**.** *An $n \times n$ matrix is invertible if and only if $\det(A) \neq 0$.*

Proof. In the proof of the previous proposition we showed that if a row echelon form of A has a row of all zeros, then $\det(A) = 0$. But the former is equivalent to A being not invertible (by Theorem 2.20).

Conversely, if A is invertible then $AA^{-1} = I$ and so

$$1 = \det(I) = \det(AA^{-1}) = \det(A)\det(A^{-1}).$$

In particular, $\det(A) \neq 0$ and in fact $\det(A^{-1}) = \det(A)^{-1}$. □

Our definition of the determinant uses the first column. Nothing is special about the first column. For convenience, if A is an $n \times n$ matrix, let A_{ij} be the $(n-1)\times(n-1)$-matrix obtained by deleting the ith row and jth column of A. With this convention, we then find:

5.10 Proposition. *Let A be an $n \times n$-matrix. Let $1 \leq i \leq n$. Then*

$$\det(A) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det(A_{ij})$$

This formula is usually referred to as *expansion by the jth column.*

Proof. We could show that the function defined by the above formula also satisfies the properties of the determinant and hence must be equal to $\det A$. However, here we could also argue as follows:

Let E be the elementary matrix that corresponds to switching column j and 1. Thus, AE moves the column j into column 1. Then the (old) first column is now in column j. Now $\det(AE) = \det(A)\det(E) = -\det(A)$.

After $j-2$ column exchanges it is in the second column. We have a matrix A' with $\det A' = (-1)^{j-1}\det A$. Applying our usual formula for the determinant we obtain

$$\det A = (-1)^{j-1}\det A' = \sum_{i=1}^{n}(-1)^{i+j}\det A_{ij},$$

always noting that $A_{ij} = A'_i$. □

One of the most important properties of the determinant however is the following:

5.11 Proposition. $\det(A^T) = \det(A)$.

Proof. It is clear that A is invertible if and only if A^T is invertible. Hence $\det(A) = 0$ if and only if $\det(A^T) = 0$. We may therefore assume that A and A^T are invertible. If $A = E_1E_2\cdots E_p$ then

$$A^T = E_p^T E_{p-1}^T \cdots E_1^T.$$

A simple case by case study shows that for every elementary matrix $\det E = \det E^T$. Hence the claim. □

5.12 Corollary. *All properties of the determinant with respect to the rows are equally true with respect to the columns. In particular:*

a. $\det$ *is linear in the columns.*

b. *If two columns of A coincide then* $\det(A) = 0$.

c. *For each $i = 1, 2, \ldots, n$,* $\det(A) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det(A_{ij})$.

Proof. Any statement about rows and the determinant, when applied to $f(A) = \det(A^T)$ becomes a statement about columns and the determinant: Notice for instance that $f(A) = \det(A^T)$ is linear in the columns of A (which are the rows of A^T). But $f(A) = \det(A)$. □

We refer to c. as the *expansion of the determinant by the ith row.*

This can be used to some gain in situations where for example a column of a large matrix is only sparsely populated (e.g. has only one nonzero entry). Then the computation of the determinant is immediately reduced to computing one smaller determinant (as opposed to n).

We conclude by remarking that all statements in Observation 5.3 remain valid if we replace all row operations by column operations. In other words, switching two columns changes the determinant by a factor of -1. Adding a multiple of a column to another column doesn't affect the determinant at all, and finally, scaling a column by $c \in \mathbb{F}$ will multiply the determinant by the same scalar.

5.2. Excursion: More on the determinant

Leibniz formula can be rephrased as

$$\text{(LF)} \qquad \det(A) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}$$

where we define

$$\operatorname{sign}(\sigma) = \det(P_\sigma).$$

Notice that since we *proved* that the Leibniz Formula holds, this is not a circular thought. In fact we proved that the parity of the number of row swaps needed to obtain I_n from P_σ is independent of the actual sequence of row swaps used.

5.2.1 Problem. Show that

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sign}(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \ldots a_{n\sigma(n)}.$$

One may think that (LF) is a nice way of computing the determinant. This is true for 2×2 and 3×3 matrices. Beyond that, the formula becomes very large (it has $n!$ summands, a number that explodes: $5! = 125$, $100!$ is beyond the means of any computer.) The same is true for the recursive formula we used to define the determinant (which is just (LF) in disguise).

5.13 Remark. A more efficient way to compute the determinant is as follows: use row operations to reduce the matrix to an upper triangular matrix *without scaling any row* (this is always possible, why?). Then multiply the diagonal entries (and correct for any sign changes due to row swaps). This is a much faster algorithm.

5.2.2 Problem. Assuming an addition or multiplication on a computer takes time t. Estimate the time used to compute the determinant of an $n \times n$-matrix as a function of n, first using (LF), then using Gaussian elimination.

For theoretical purposes, however, (LF), and by extension the expansion by row or column, is the thing. It has the following remarkable consequence:

Recall the notation A_{ij} for the $(n-1) \times (n-1)$ matrix obtained by deleting row i and column j in A. Then

$$\sum_{k=1}^{n} (-1)^{k+j} a_{ik} \det(A_{jk}) = \begin{cases} \det(A) & i = j \\ 0 & i \neq j \end{cases}$$

To see why this is true observe that if $i = j$ then this is simply the expansion of the determinant by the ith row. If $i \neq j$, however, then this formula is the expansion of the determinant by the ith row of the matrix A' where A' is obtained from A by replacing the jth row with the ith row. But A' has two identical rows! Hence $\det(A') = 0$.

This motivates the following:

5.14 Definition. Let A be an $n \times n$ matrix. Its *classical adjoint* or simply *adjoint* or *adjungated*, denoted $\operatorname{Adj}(A)$ is defined as

$$\operatorname{Adj}(A) = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & \vdots & \dots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{bmatrix}$$

where

$$\alpha_{ij} = (-1)^{i+j} \det(A_{ji}).$$

(Note the swapped indices!)

As a consequence

5.15 Proposition. *For any $n \times n$ matrix A we have*

$$\operatorname{Adj}(A)A = A \operatorname{Adj}(A) = \det(A) I_n \tag{5.12}$$

Proof. The assertion is equivalent to saying that for each i, j,

$$\sum_{k=1}^{n} a_{ik}\alpha_{kj} = \delta_{ij} \det A = \sum_{k=1}^{n} \alpha_{ik} a_{kj}.$$

The first equality is simply the above formula. The second equality is the complete analogous reasoning using an expansion by the jth column of A. □

5.16 Corollary. *If A is invertible (i.e. if $\det(A) \neq 0$) then*

$$A^{-1} = \frac{1}{\det(A)} \operatorname{Adj}(A).$$

5.2.3 Problem. Let $M_n(\mathbb{Z}) \subset M_n(\mathbb{Q})$ denote the matrices with all integer entries. Show that if $A \in M_n(\mathbb{Z})$ is invertible then $A^{-1} \in M_n(\mathbb{Z})$ if and only if $\det(A) = \pm 1$.

5.17 Example. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

with $\det(A) = ad - bc \neq 0$. Then

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

To conclude this section we give one application.

5.18 Theorem (Cramer's Rule). *Let A be an invertible $n \times n$ matrix. For any $B \in \mathbb{F}^n$, the unique solution of $AX = B$ is equal to*

$$X = \frac{1}{\det(A)} \operatorname{Adj}(A)B.$$

Equivalently, The unique solution $X = [\, x_1 \, x_2 \, \dots \, x_n \,]$ is obtained by

$$x_i = \frac{\det(A_i(B))}{\det(A)}$$

where $A_i(B)$ is the $n \times n$ matrix obtained from A by replacing the ith column by B.

Proof. The first part follows immediately from the corollary. As for the second, notice that $\det(A_i(B)) = \sum_{j=1}^{n} (-1)^{i+j} b_j A_{ji}$, which is equal to

$$\sum_{j=1}^{n} b_j \alpha_{ij}$$

which is the i-th entry of $\operatorname{Adj}(A)B = \det(A)X$. □

Remark. Except for small n, Cramer's Rule is useless for actual computations. However, it is of tremendous importance for the following reason: it shows that the entries of A^{-1} are rational expressions in the entries of A. For instance this can be used to show that A^{-1} depends continuously on A.

5.3. Excursion: Determinants and eigenvalues

When analyzing a linear operator T on a vector space V, the notion of *eigenvectors* and *eigenvalues* are of paramount importance.

5.19 Definition. Let $T\colon V \to V$ be a linear operator. An *eigenvector* of T is a *nonzero* $v \in V$ such that $T(v) = \lambda v$ for some $\lambda \in \mathbb{F}$.

The scalar λ is called an *eigenvalue* of T, and then v is a *corresponding eigenvector*.

5.20 Example. Let

$$D = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 3 \end{bmatrix}$$

viewed as a real matrix. Let $T = T_D\colon \mathbb{R}^3 \to \mathbb{R}^3$. Then e_2 is an eigenvector for T with eigenvalue 2 since $T(e_2) = 2e_2$. Similarly e_1 is an eigenvector for eigenvalue 1 and e_3 is an eigenvector for eigenvalue 3.

Eigenvectors are useful, since in the example, the action of T on $\mathbb{R}^3$ is best understood componentwise: it scales vectors by 1, 2, 3 in the direction of e_1, e_2, e_3 respectively.

5.21 Definition. T is called *diagonalizable*, if there is a basis $\mathcal{B}$ of V such that $M_{\mathcal{B}}^{\mathcal{B}}(T)$ is a diagonal matrix.

Equivalently, T is diagonalizable if V has a basis consisting of eigenvectors for T.

5.22 Example. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ be given by the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

Then T has eigenvectors $v = e_1 + e_2$ (eigenvalue 3) and $w = e_1 - e_2$. (eigenvalue -1) . With respect to the basis $\mathcal{B} = (v, w)$, its matrix is

$$D = \begin{bmatrix} 3 & \\ & -1 \end{bmatrix}$$

When working with T, this suggests, we should be working with the basis $\mathcal{B}$ instead of the standard basis. Then T corresponds to multiplication with a diagonal matrix, which is considerably simpler that multiplying with the orginal A.

How do we find eigenvalues and eigenvectors? If $\dim V < \infty$, the crucial observation is that λ is an eigenvector if and only if $T - \lambda \operatorname{id}_V$ has a non-trivial kernel, that is, $T - \lambda \operatorname{id}_V$ is *not invertible*! So if A is the matrix of T with resptect to any basis $\mathcal{B}$, then λ is an eigenvalue if and only if

$$A - \lambda I$$

is not invertible. In other words, if and only if

$$\det(A - \lambda I) = 0.$$

So to find eigenvalues we need to solve $\det(A - \lambda I) = 0$ for λ. Equivalently we can solve $\det(\lambda I - A)$. In general, this is a polynomial equation of degree $\dim V$ (the size of the square matrix A).

5.23 Example. In Example 5.22 above $\det(\lambda I - A) = \lambda^2 - 2x - 3 = (\lambda + 1)(\lambda - 3)$.

Then the eigenvectors are found by solving $(A - \lambda I)X = 0$. Any nonzero solution are eigenvectors.

5.24 Example. Not all linear operators are diagonalizable.

a. The matrix (the operator given by the matrix)

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

viewed as a linear operator on $\mathbb{R}^2$ has no (real) eigenvalues: the equation to solve is $\lambda^2 + 1 = 0$.

This can be solved by passing to complex numbers!

b. The matrix

$$B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

has a single eigenvalue 0. It is a "double" root of the equation $\det(\lambda I - B) = \lambda^2 = 0$. The only eigenvectors are nonzero multiples of e_1, though, so there are not *enough* eigenvectors to form a basis. This cannot be solved by passing to the complex numbers.

5.25 Examples. Here are a few calculus inspired examples.

a. If we view differentiation as a linear operator $D\colon \mathcal{C}^\infty(I) \to \mathcal{C}^\infty(I)$, where $\mathcal{C}^\infty(I)$ is the space of infinitely many times differentiable functions on the interval I, many familiar functions can be recovered as eigenvectors of D or related operators: for instance

$$f(x) = e^k x$$

is an eigenvector of eigenvalue k: $D(f) = kf$. (Conversely, calculus shows that if $D(f) = kf$ then $f = ce^{kx}$ for some constant c.) $\cos x$ and $\sin x$ are eigenvectors of the operator D^2 of eigenvalue -1. If we write xD for the operator $f \mapsto xD(f)$, and if $I = (0, 1)$, say, then $\log x$ is an eigenvector for eigenvalue 0 of $D(xD)$:

$$D(xD)(\log x) = Dx\frac{1}{x} = D(1) = 0.$$

b. A simple example is also given by the *Euler operator* on $\mathcal{P}(\mathbb{R})$. It is defined as $E(f) = xf'$. Then the eigenvectors of E are of the form cx^i $(c \neq 0)$ and the eigenvalue of x^i is i.

5.26 Example. As mentioned before, in Quantum Mechanics, the state of a system is given by a vector $|\psi\rangle$ in a certain vector space $\mathcal{H}$.

Depending on the actual model for QM used, an *observable*, that is, an experimentally measurable quantity, corresponds to a linear operator A on $\mathcal{H}$. If a measurement is performed, *after* the measurement, the system is found in a state $|\psi_0\rangle$, which is an eigenvector for A for some eigenvalue λ, and λ is the actual value the measurement results in. (In the Kopenhagen model of QM, this is referred to as the *collapse of the wave function*: the system was in state $|\psi\rangle$ *before* the experiment, but is found with certain probabilities in each possible "eigenstate" (eigenvector) *after* the measurement.)

5.4. Excursion: More on permutations

We want to mention briefly some properties of S_n and Π_n that will be of recurring interest later on. Recall that a permutation matrix $P \in \Pi_n$ was defined as a matrix with exactly one 1 in each row and column and 0 everywhere else.

We observed that each P is obtained by a sequence of Type I operations from I_n and concluded for instance $\det(P) = (-1)^p$ where p is the number of row operations needed (this number is not uniquely determined).

How are σ and P_σ connected? Recall that σ_P was defined as $\sigma_P(i)$ is the position of e_i^T in P. Analyzing this we find that (using the matrix units e_{ij} from Section 2.2)

$$P_\sigma = \sum_{i=1}^{n} e_{\sigma(i)i}$$

Note that $e_{ij}e_k = e_i$ if $k = j$ and 0 otherwise. From this it immediately follows that

$$P_\sigma e_i = e_{\sigma(i)}. \tag{5.13}$$

If A is any $n \times n$ matrix (or any $n \times m$-matrix for that matter) then PA is obtained from A by applying σ to the rows of A: that is, row i is moved to row $\sigma(i)$.

Similarly, if A is a $m \times n$-matrix then AP is obtained from A by applying σ^{-1} to the columns of A: indeed, $AP = ((AP)^T)^T = (P^T A^T)^T$ and $P^T = P^{-1}$.

Recall that σ is a permutation, hence affords an inverse σ^{-1} defined by $\sigma(\sigma^{-1}(i)) = i$ and $\sigma^{-1}(\sigma(i)) = i$. (Notice that σ^{-1} is also the inverse of σ when S_n is viewed as a group.) While $Pe_i = e_{\sigma i}$ seems naturally enough, unfortunately we pay a price when computing

$$PX = \begin{bmatrix} x_{\sigma^{-1}(1)} \\ x_{\sigma^{-1}(2)} \\ \vdots \\ x_{\sigma^{-1}(n)} \end{bmatrix}.$$

The main observation of this section is the following:

$$P^{-1} = P^T$$

and

$$P_{\sigma^{-1}} = P_\sigma^{-1}$$

5.4.1 Problem. Prove the "main observation"

An immediate consequence is the fact that

5.27 Proposition. *Π_n is a subgroup of $\mathrm{GL}_n(\mathbb{F})$.*

Consider the bijection $\pi\colon S_n \to \Pi_n$ that sends a permutation σ to the associated permutation matrix P_σ (hence $\pi(\sigma) = P_\sigma$). We now have a group law on S_n and another one on Π_n. It turns out, the two are related: $\pi(\sigma\mu) = \pi(\sigma)\pi(\mu)$. Thus, it doesn't matter whether we compute the product in S_n first and then apply π, or if we first apply π and then multiply in Π_n. Maps between groups that have this property are called *group homomorphisms*. Since π is also bijective, there is no way to distinguish S_n and Π_n in group theoretic terms. We will make this somewhat vague notion more precise later on.

This is one of the reasons why we associated P_σ to σ rather than σ^{-1}. The map $\rho\colon S_n \to \Pi_n$ that sends σ to $P_{\sigma^{-1}}$ is *not* a homomorphism because $\rho(\sigma\mu) = \rho(\mu)\rho(\sigma) \neq \rho(\sigma)\rho(\mu)$ in general: this follows from the rule $(PQ)^{-1} = Q^{-1}P^{-1}$

Let $U_n \subseteq \mathrm{GL}_n(\mathbb{F})$ be the set of upper triangular matrices that have only 1 on the diagonal.

$$U_n = \left\{ \begin{bmatrix} 1 & * & * & * \\ & 1 & * & * \\ & & \ddots & * \\ & & & 1 \end{bmatrix} \right\}$$

5.4.2 Problem. Verify that U_n is a subgroup of $\mathrm{GL}_n(\mathbb{F})$.

Let us define an *upper triangular row operation* a row operation of Type II where we add a multiple of row i to to row j and $i < j$. Note that this is equivalent to saying that the corresponding elementary matrix is an element of U_n.

5.4.3 Problem. Show that U_n is *generated* by elementary matrices of Type II: show that every element in U is a product $L_1 L_2 \cdots L_p$ with L_i a Type II elementary matrix.
(*Hint:* Row reduce $A \in U_n$ to I by using only upper triangular row operations.)

If in our algorithm to find the reduced echelon form we omit any Type I and Type III row operationsl and only perform upper triangular row operations what do we end up with?

Let A be an $n \times n$ matrix. And assume A is invertible.

Since we cannot "touch" entries below any given entry we perform the following algorithm: In column 1 pick the lowest nonzero entry (such an entry exists because A is invertible, hence $\det A \neq 0$). Let's say the entry is in row i. With this entry erase all entries above, using only upper triangular row operations.

Now repeat the procedure in the second column as if row i was not there. That is, pick the lowest nonzero entry that is not in row i (such an entry exists, why?). Erase all entries above, *except* if applicable, the entry in row i. Never touch an entry in row i again... Let us call row i "fixed." Repeat this process with all columns. Because at each stage, $\det A \neq 0$, there must always be a nonzero entry in one of the rows that haven't been "fixed" for other wise we could obtain a zero column by applying some column operations of Type II.

The end result will be a matrix of the form

$$\begin{bmatrix} 0 & 0 & \bullet & * \\ \bullet & * & * & * \\ 0 & \bullet & * & * \\ 0 & 0 & 0 & \bullet \end{bmatrix}$$

Of course, this is only an example. The precise definition of the form will be as follows:

- Each row has a leading term.
- Each column contains the leading term of exactly one row.

Notice the positions of the leading terms are the positions of the 1s in a unique permutation matrix P_σ. So far we have passed from A to UA where U is some product of upper triangular elementary matrices in U_n. After applying P^{-1}, we may assume that the leading terms all are on the diagonal: $P^{-1}UA$ is upper triangular. Let D be the diagonal matrix with the same diagonal entries as $P^{-1}UA$. Then $D^{-1}P^{-1}UA$ has only 1s on the diagonal, that is $D^{-1}P^{-1}UA$ is an element of U_n (which itself is again a product of elementary matrices for upper triangular row operations. In fact what we have shown is the following

5.28 Theorem. *Every matrix A in $\mathrm{GL}_n(\mathbb{F})$ can be written as*

$$A = UPDV$$

where $U, V \in U_n$,$D \in D_n$, and $P \in \Pi_n$.

Proof. We just observed that there exist $U' \in U_n$, $D \in D_n$ and $P \in \Pi_n$ such that $V = D^{-1}P^{-1}U'A = V$ is an element of U_n. Hence $A = U'^{-1}PDV$ is as claimed, observing that $U = U'^{-1} \in U_n$. □

A slightly more careful analysis than what we did shows that P is uniquely determined (in fact, we did not have any choice when we picked the "fixed" row in each step, thus our algorithm suggests that P is unique). If we believe the uniqueness then we have arrived at what is known as *Bruhat decomposition* of GL_n :

$$\mathrm{GL}_n(\mathbb{F}) = \bigcup_{P \in \Pi_n} U_n D_n P U_n \tag{5.14}$$

where we formally write $U_nD_nPU_n$ for the set $\{UDPV \mid U, V \in U_n, D \in D_n\}$. The uniqueness of P forces that the sets $U_nD_nPU_n$ are *disjoint*. They are called the *Bruhat cells* and are of tremendous importance in studying $\mathrm{GL}_n(\mathbb{F})$ from a geometric viewpoint.

Part II.

MATH 227: Honours Linear Algebra II

6. Fields and polynomials

Before we can delve deeper into linear algebra, we have to omit the "linear" for a while and do just algebra. It is time to look a little more closely at the fields $\mathbb{F}$ we have available. In this chapter we will basically discuss two types of fields: the complex numbers and finite fields.

6.1. Equivalence relations

In Chapter 1 we constructed a finite field with 2 elements and we have seen a field with three elements in homework assignments. The purpose of this section is to introduce more finite fields. In fact we will show that for each prime p there is a field with p elements.

Along the way we will develop a very powerful method of constructing new algebraic objects out of old ones, a method that is used everywhere in modern mathematics (namely, quotients).

6.1.1. $\mathbb{F}_5$

Let us begin with an example: We want to construct a field $\mathbb{F}$ with 5 elements. We will use a completely naive approach here. Assume $\mathbb{F}$ exists. Then we know $\mathbb{F}$ has a 0 and 1 (cf. Chapter 1 and $0 \neq 1$. We have an addition as well, so there must be an element $1 + 1$ in $\mathbb{F}$. Not knowing anything else, let us call this elememnt 2, and we will assume that $2 \neq 0, 1$ (this is really an assumption at this point). Going on, we label the elements $1+1+1, 1+1+1+1$ by 3 and 4 respectively. But note that these elements are not integers, this is just a label. If we assume that the five elements so far created are all distinct, our field $\mathbb{F}$ does not contain any more elements. Thus, for the next guy, which would be $1+1+1+1+1$ we must make a choice: it must be equal to one of the five elements of $\mathbb{F}$. So which one should it be? It turns out we don't have a choice at all:

$$1+1+1+1+1 = 0.$$

Indeed, suppose $1+1+1+1+1 = 1 + \cdots + 1$ where on the right hand side we have 4 or fewer 1s. Then we can cancel all ones on the right to obtain $1 + \cdots + 1 = 0$ where now on the left we have *fewer* than 5 ones. This does not work: we assumed that $1, 2, 3, 4$ all are not equal to zero. So in fact we cannot have any positive number of 1 on the right hand side and hence "5" must equal 0 in $\mathbb{F}$.

If we now compute with our five elements $0, 1, 2, 3, 4$ *as if they were integers but with the additional "relation"* $5 = 0$, it turns out that we can write down complete addition and multiplication tables and the result will be a field: Let us pretend the addition and multiplication are similar to the integers with the additional rule that "$5 = 0$."

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\times$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Table 6.1.: The addition and multiplication tables for a field with five elements.

For example, $3 + 4 = 7$ (in the integers), but $7 = 2 + 5 = 2 + 0 = 2$, so "$7 = 2$" and we obtain $3 + 4 = 2$.

Similarly, $2 + 4 = 6 = 1$, $4 + 4 = 8 = 3$ and we get the following addition table:

Notice that the important property of these two tables is that in every row and column (with the exception of the multiplication with 0) *every* element appears exactly once as a result. Thus, for instance every nonzero element has a multiplicative inverse. We would now have to verify the associative and distributive laws (the commutative law is built into the tables).

The upshot in this subsection is that we obtain a field with 5 elements if we *"identify" integers that differ by a multiple of* 5. Thus, we treat eg. $-3, 2, 7$ all as equal, even though, as integers, they are certainly not.

Notice that we did make some choices in the constructions (for instance, we said that $1 + 1 \neq 0$). One can show however, that there is essentially no other field with five elements (up to relabeling of elements).

This was a very ad hoc construction, and it is not clear at this point, why this actually worked. We will now find a conceptual way of doing this.

6.1.2. Partitions and equivalence relations

The way to formalize our discussion is as follows: Imagine we are given a (nonempty) set X, eg. $X = \mathbb{Z}$. And for some reason, we would like to treat certain elements of X as equal.

A typical example might be $X = \mathbb{Z}$ and we want to think of two integers as equivalent if they have the same parity. So we partition the set of integers into two classes, the class of even and the class of odd integers. Similarly, we might want to treat congruent triangles as equal.

In other words, we have a (binary) *relation* $\sim$ among the elements of X, which is typically weaker than actual equality. That is, for any two $x, y \in X$, we have that

$$\text{either } x \sim y \text{ or } x \not\sim y$$

but never both, with the understanding that $x \sim y$ if and only if we would like to think of x and y as somehow equal. More formally, a binary relation on X is a subset $R \subseteq X \times X$ of the set of all ordered pairs[1] (x, y) with $x, y \in X$. Then

$$x \sim y :\iff (x, y) \in R.$$

(Note that the notation xRy instead of $x \sim y$ is also common.)

6.1 Examples.

a. $x \sim y$ if and only if $x = y$.
   Here $R \subseteq X \times X$ is the *diagonal*:
   $$R = \{(x, y) \in X \times X \mid x = y\} = \{(x, x) \mid x \in X\}.$$

b. $x \sim y$ for all $x, y \in X$.
   Here $R = X \times X$.

c. $x \sim y$ for *no* elements $x, y \in X$.
   Here of course $R = \emptyset$.

d. $X = \mathbb{R}$, $x \sim y$ if $x^2 = y^2$.
   Then $R = \{(x, y) \in \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \mid x^2 = y^2\}$.

e. Let $f \colon X \to Y$ be a function (where Y is some set). Put $x \sim y$ if $f(x) = f(y)$.
   Then $R = \{(x, y) \in X \times X \mid f(x) = f(y)\}$.

6.2 Definition. Let $(X, \sim)$ be a set X together with a binary relation $\sim$. The relation is called an *equivalence relation* if it has the following three properties:

a. it is *reflexive*: for each $x \in X$, $x \sim x$;

b. it is *symmetric*: for each $x, y \in X$, if $x \sim y$ then also $y \sim x$;

c. it is *transitive*: for each x, y.

If $x \sim y$ we sometimes say that x and y are *equivalent* (with respect to $\sim$).

6.1.1 Problem. Reformulate the conditions for an equivalence relation in terms of the set R.

6.3 Examples. The following are examples of equivalence relations.

a. $x \sim y$ if $x = y$.

b. $X = \mathbb{R}$, $x \sim y$ if $x^2 = y^2$.

[1] Recall that the "ordered" in "ordered pair" simply means that (x, y) is not the same as (y, x) (as long as $x \neq y$); that is, the order of the two elements matters.

c. $X = \mathbb{Z}$ and $a \sim b$ if both, a and b, are even or odd. (Convince yourself that this means $a \sim b$ iff $a - b$ is divisible by 2.)

d. $X = \mathbb{Z}$, $x \sim y$ if 5 divides $x - y$.

e. $X = \mathbb{R}$, and $x \sim y$ if $y = x^2$ is *not* an equivalence relation. It satisfies none of the properties.

f. If $X = \mathbb{F}_2$ and $x \sim t$ if $y = x^2$ *is* an equivalence relation as for all $x \in \mathbb{F}_2$ we have $x = x^2$.

In 6.1 the only relation that is not an equivalence relation is Example c. (if X is not empty). The relation is symmetric and transitive but not reflexive.

Given a set X with an equivalence relation, and given an element $x \in X$, it makes sense to define the *equivalence class* of x as the set

$$C_x = \{y \in X \mid y \sim x\}$$

of all elements in y that are equivalent to x with respect to the relation x. (One could think of this as grouping the elements of X together that share a certain (unspecified) property.)

Notice that since $\sim$ is symmetric we always have $y \in C_x$ if and only if $x \in C_y$. Also $x \in C_x$, because $\sim$ is reflexive.

In fact,

$$x \sim y \iff C_x = C_y.$$

Indeed, let $x \sim y$. Then transitivity implies that $C_x \subseteq C_y$; for if $z \in C_x$ then $z \sim x$, and so $x \sim y$ forces $z \sim y$ which means $z \in C_y$. Since $\sim$ is symmetric it follows that $C_y \subseteq C_x$ by the same argument (using that $y \sim x$). If, on the other hand, $C_x = C_y$, then clearly $x \in C_x = C_y$ implies that $x \sim y$.

A slightly more subtle property is the following:

> *Two equivalence classes are either equal or disjoint.*

Indeed, suppose $C_x \cap C_y \neq \emptyset$. Then there is $z \in C_x \cap C_y$ and so $z \sim x$ and $z \sim y$. But we just saw that this means $C_x = C_z = C_y$.

Thus, the subsets $C_x \subseteq X$ (for various x) have the following properties

a. They cover all of X: every $x \in X$ is contained in at least one such subset.

b. Two such subsets are either equal or disjoint.

Such families[2] of (nonempty) subsets with these properties are called *partitions* (of X). Note that a. and b. together imply that every $x \in X$ is contained in one and only one subset.

[2] A *family* of subsets is just a set of subsets. The word family is sometimes used to avoid the use of the word "set." A "set of subsets" might sound odd.

6.4 Theorem. *Let X be a nonempty set. The equivalence classes of an equivalence relation form a partition of X. Conversely, every partition arises in this way.*

Proof. We already proved the first statement of the the theorem in the remarks preceding it.

It remains to show that every partition is obtained from an equivalence relation. So let $\mathcal{P}$ be a family of subsets of X that form a partition. For $x \in X$ let $C_x \in \mathcal{P}$ be the unique element of the family that contains x. Then define $x \sim y$ if and only if $C_x = C_y$. It is easy to check that this is indeed an equivalence relation. Moreover, its equivalence classes are precisely the elements of $\mathcal{P}$. □

Remark. The elements of an equivalence class C_x are often called the *representatives* of C_x. Also, we often write $\overline{x}$ instead of C_x. Thus, x is a representative for $\overline{x}$. Depending on $\sim$ there may be many representatives for the same equivalence class.

Given a set X with equivalence relation $\sim$, we can form a new set $\overline{X}$ whose elements are precisely the equivalence classes. Thus, an element of $\overline{X}$ is a subset of X of the form C_x for some $x \in X$. We will now adopt the convention to write $\overline{x}$ instead of C_x. If $(X, \sim)$ is a set with equivalence relation, you will often find the notation $X/\sim$ for the set $\overline{X}$ of its equivalence classes. $\overline{X}$ is called the *quotient* of X by the relation $\sim$.

We have a natural mapping

$$\pi \colon X \to \overline{X}$$

defined by $\pi(x) = \overline{x}$. By definition the map is surjective. Then we have

$$x \sim y \iff \pi(x) = \pi(y).$$

Note that this establishes that *every* equivalence relation is of the form as in Example 6.1 d. (just put $Y = \overline{X}$ and $f = \pi$).

Sometimes it is helpful to think of the elements of $\overline{X}$ not as subsets of X but to work with them as if they were elements of X, but we now treat some of them as equal that may not be equal in X; or to think of the elements of $\overline{X}$ as parameterized by the elements of X.

Our example of $\mathbb{F}_5$ above fits in as follows: let $X = \mathbb{Z}$ and write $a \sim b$ if $a - b$ is divisible by 5. So we identify integers if they differ by multiples of 5. Notice that for each integer n we have that $n \sim x$ with $x \in \{0, 1, 2, 3, 4\}$. So $\overline{X} = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}$.

The integers $2, 7, 13, -3$ all are representatives of the same equivalence class (namely $\overline{2}$).

Rather than thinking of $\overline{i}$ as a subset of $\mathbb{Z}$ consisting of all integers of the form $i + 5k$, it is often more useful, to work with the elements of $\overline{X}$ as if they were integers, and thinking of them as equal if they differ by multiples of 5.

It is often useful to give a concrete description of the quotient $X/\sim$. The following examples illustrate this thought.

Examples. mbox

a. $X = \mathbb{R}^2$, and $x \sim y$, if x and y lie on a horizontal line (that is, $x - y \in \operatorname{Span}(e_1)$). Since every such line intersects the y-axis in a unique point, we find

$$X/\sim \leftrightarrow \left\{ \begin{bmatrix} 0 \\ t \end{bmatrix} \middle| t \in \mathbb{R} \right\} \leftrightarrow \mathbb{R}$$

So if we associate to each equivalence class this point of intersection (and identify the point with its only interesting entry t), we get an identification of $X/\sim$ with $\mathbb{R}$.

b. $X = \mathbb{R}$, and $a \sim b$ if $a - b \in \mathbb{Z}$. Here every $x \in \mathbb{R}$ is equivalent to a unique element of $[0, 1) = \{t \in \mathbb{R} \mid 0 \leq t < 1\}$. Taking into account that real numbers "close" to 1 are equivalent to real numbers "close" to 0, the most natural representation of the quotient is as $[0, 1]/\sim$ where $\sim$ on $[0, 1]$ is defined as $0 \sim 1$ and all other numbers are equivalent to only themselves.

This can be thought of as a line segment with the ends glued together, ie. a circle, usually denoted S^1. And indeed, the quotient map $\pi\colon X \to X/\sim$ can be realized as $\pi\colon X \to S^1 \subset \mathbb{R}^2$,

$$\pi(\alpha) = \begin{bmatrix} \cos(2\pi\alpha) \\ \sin(2\pi\alpha) \end{bmatrix}.$$

c. Modifying the previous example, and putting $X = \mathbb{R}^2$ and $(a, b) \sim (c, d)$ if $a-c, b-d \in \mathbb{Z}$, we get that every element of $\mathbb{R}^2$ is equivalent to a point in the unit square $\Gamma = \{(a, b) \mid 0 \leq a, b \leq 1\}$. Only elements on the boundary of the square can be equivalent, and it turns out, the quotient corresponds naturally to a square with opposing sides glued together, ie. a donought (or torus).

d. The integers can be constructed as quotients out of the set of pairs of natural numbers. Put $\mathbb{N} = \{1, 2, \dots\}$, and $X = \{(a, b) \mid a, b \in \mathbb{N}\}$. The fundamental idea here is that integers represent *differences* of natural numbers. That is, for (a, b) there is a corresponding integer representing $a - b$. We identify two pairs (a, b) and (c, d) if they have the same difference, i.e. we put $(a, b) \sim (c, d)$ if $a + d = b + c$. The quotient $X/\sim$ can then naturally be identified with the set of integers. (For example, the integer -1 is the equivalence class of the pair $(0, 1)$, or $(2, 3)$, $\dots$)

e. We are more familiar and naturally use the main idea of quotients (namely identifying objects that represent the same "property") with rational numbers: a rational number represents a *fraction* of two integers. We put

$$X = \{(a, b) \mid a, b \in \mathbb{Z}, b \neq\}$$

and define $(a, b) \sim (c, d)$ if $ad = bc$. Then (a, b) and (c, d) represent the same fractional relationship. The quotient $X/\sim$ may be naturally identified with the set $\mathbb{Q}$ of rational numbers. And we write a/b for the equivalence class of (a, b). Note we are very much used to the fact that $a/b = na/(nb)$ so the representation of rational numbers by means of pairs of integers is not unique, we *identify* symbols a/b and $2a/(2b)$, for instance.

f. Let $X = M_{m\times n}(\mathbb{F})$ and put $A \sim B$ if A is row-equivalent to B (that is, B is obtained from A by a sequence of elementary row operations).

   We know that every equivalence class contains a *unique* matrix in reduced row echelon form, so we can identify the quotient with the set of all $m \times n$ reduced row echelon forms.

   Alternatively, we observe that A and B are row-equivalent if and only if they have the same null-space (we introduced row operations precisely because they do not change the solutions of $AX = 0$; conversely it is not hard to see that for each possible subspace there is exactly one reduced row echelon form with that null space[3]). Thus, we may think of $X/\sim$ also as the set of all subspaces of $\mathbb{F}^n$.

Main Example

An important application is the following: Let A be an abelian group (cf. Definition 2.24). For simplicity we will write the group law additively (so the identity element will be denoted by 0).

For instance $A = (\mathbb{Z}, +)$, or $A = (V, +)$ where V is a vector space. These are the only two instances that will be important for us for quite a while.

Let $B \subseteq A$ be a subgroup. Recall, that this means $0 \in B$; $a, b \in B \implies a + b \in B$; and $a \in B$ implies $-a \in B$.

We can now introduce the following relation on A: we write

$$a \sim b :\iff b - a \in B.$$

We will now verify that this is an equivalence relation:

a. $a \sim a$: indeed, $a - a = 0 \in B$.

b. $a \sim b$ implies $b \sim a$: indeed, if $b - a \in B$ then also $-(b - a) = a - b \in B$.

c. $a \sim b$ and $b \sim c$ implies $a \sim c$: indeed, $b-a, c-b \in B$ means $(b-a)+(c-b) = c-a \in B$.

(Note how we used every single property of B being a subgroup.)

If $a \in A$, then it is common to denote its equivalence class as

$$a + B = \{a + b \mid b \in B\}.$$

A set of the form $a + B$ is also called a *coset*.

The quotient $\overline{A}$ by this equivalence relation is usually denoted by A/B ("A mod B" or "A modulo B").

[3] Indeed, the nullspace of a matrix determines the row space of the matrix (as the set of all row vectors that multiply to zero with the elements of the nullspace (why?)); matrices with the same row space are row equivalent (why?).

The crucial point is that A/B is again an abelian group, in a natural way.

Informally speaking, as with our example $\mathbb{F}_5$, what we can do is to compute in A/B as we would in A but with the additional relation that elements of B are treated as 0.

So given two elements x, y of A/B we define $x+y$ as follows: pick representatives of x and y, say, a and b, respectively. That is, pick $a, b \in A$ with $\overline{a} = x$ and $\overline{b} = y$. Then

$$x + y := \overline{a+b}$$

Using the notation for cosets introduced earlier, this means

$$(a+B) + (b+B) := (a+b) + B.$$

This is well defined[4]: indeed, if $a' \in x$ and $b' \in y$ are other representatives, then $a' = a + b_1$ and $b' = b + b_2$ with $b_1, b_2 \in B$ and hence $a' + b' = a + b + (b_1 + b_2)$ differs from $a + b$ by an element in B. Hence $(a' + b') + B = (a + b) + B$. So our definition is independent of the choice of representatives.

Now consider the map $\pi \colon A \to A/B$ that sends a to $\overline{a}$. It is immediate that $\pi(a+b) = \pi(a) + \pi(b)$. Indeed, this is precisely the definition. From this it follows easily that $+$ on A/B is commutative, associative, that $\pi(0)$ is an additive identity and that every element in A/B allows an inverse.

Indeed, for $\overline{a}, \overline{b}, \overline{c} \in A/B$ we have

$$\overline{a} + (\overline{b} + \overline{c}) = \overline{a} + \overline{b+c} = \overline{a + (b+c)} = \overline{(a+b)+c} = \overline{a+b} + \overline{c} = (\overline{a} + \overline{b}) + \overline{c}.$$

Similarly,

$$\overline{a} + \overline{0} = \overline{a+0} = \overline{a}.$$

And

$$\overline{a} + \overline{b} = \overline{a+b} = \overline{b+a} = \overline{b} + \overline{a}.$$

Also $\overline{-a}$ is clearly an additive inverse for $\overline{a}$ ($\overline{-a} + \overline{a} = \overline{-a+a} = \overline{0}$. Thus we have $-\overline{a} = \overline{-a}$.

In short, what we have proved is:

6.5 Proposition. *A/B together with the addition defined above, is an abelian group.*

A/B is called the *quotient group* of A by the subgroup B. In the following sections we will illustrate this concept at two examples.

Example. If we forget the multiplication, $(\mathbb{R}, +)$ is an abelian group and $\mathbb{Z} \subset \mathbb{R}$ is a subgroup. We have seen that we may identify $\mathbb{R}/\mathbb{Z}$ with the circle S^1, so this gives S^1 the structure of an abelian group. (Indeed, if we represent an element of S^1 by the angle it forms with the positive x-axis, then addition in S^1 just corresponds to adding angles.)

Similarly, $\mathbb{Z}^2 = \{(a,b) \in \mathbb{R}^2 \mid a, b \in \mathbb{Z}\}$ is a subgroup of $\mathbb{R}^2$ (where again, the group law is just addition). Since $\mathbb{R}^2/\mathbb{Z}^2$ is the torus, we get an addition on $\mathbb{R}^2/\mathbb{Z}^2$. This is just the componentwise addition in $S^1 \times S^1$, if we identify $\mathbb{R}^2/\mathbb{Z}^2$ with $S^1 \times S^1$.

These are first examples of what are called *Lie groups* (groups where the group law and inverse forming are "differentiable" in the appropriate sense).

[4] Well defined here means that the right hand side does not depend on which representatives a, b we chose.

6.2. Quotient spaces

Let V be an $\mathbb{F}$-vector space and let $W \subset V$ be a subspace. As mentioned, V and W are abelian groups and hence we obtain an abelian group V/W. Our goal is to turn this into a vector space.

All that is needed is to define a scalar multiplication. So let $c \in \mathbb{F}$ and let $v + W$ be a coset. We now define

$$c \cdot (v + W) := (cv) + W.$$

As above this does not depend on the choice of the representative v: indeed, if $v+W = v'+W$, then $v' = v + w$ for some $w \in W$ and hence $cv' = c(v+w) = cv + cw$. Since W is a subspace, $cw \in W$ and so

$$cv' + W = cv + W.$$

Notice that by construction we have $\pi \colon V \to V/W$, defined by $\pi(v) = v + W$, satisfies that $\pi(v_1 + v_2) = \pi(v_1) + \pi(v_2)$ and $\pi(cv) = c\pi(v)$ for all $v, v_1, v_2 \in V$ and $c \in \mathbb{F}$.

From this one may conclude two things: first, all the rules of a vector space are satisfies. Indeed, the scalar multiplication so introduced is associative and distributive, and $1(v + W) = v + W$ for all v. So V/W is actually a vector space! The space V/W is called the *quotient space*.

Second, because of this, π is a (surjective) linear transformation. Its kernel is equal to W. Recall that we proved that the null space of a linear transformation is always a subspace. This provides for a converse: every subspace is the kernel of a linear transformation.

Example. It is not easy to give a meaningful elementary example of a quotient space.

But here is one, important in Physics. We have observed before that quantum mechanics takes place in a vector space $\mathcal{H}$. As an example, we saw $\mathcal{H} = L^2(\mathbb{R}, \mathbb{C})$, the complex valued square integrable functions on the real line (which describes a one particle system in one dimension). It is convenient to think of $\mathcal{H}$ as a space of functions. However, for technical reasons, one wants to treat functions for which $\int_{\mathbb{R}} |f(x)|^2 dx = 0$ as equal to zero (the continuous ones certainly are). These functions form a subspace W, and $\mathcal{H} = L^2/W$ (in fact, usually L^2 is used to denote this quotient space).

Why are quotient spaces important? They provide for a universal treatment of the image of a linear transformation.

6.6 Theorem. *Let $T \colon V \to V'$ be a linear transformation. Let $W \subset V$ be a subspace. Suppose $W \subseteq \mathcal{N}(T)$. Then there is a unique linear transformation*

$$\overline{T} \colon V/W \to V'$$

such that

$$T = \overline{T} \circ \pi.$$

Proof. Let $\overline{v} \in V/W$. Then $\overline{v} = \pi(v)$, say. We need $\overline{T}(\overline{v}) = \overline{T}(\pi(v)) = T(v)$. So let us use this as a definition: $\overline{T}(\overline{v}) := T(v)$ where v is any representative of $\overline{v}$. Of course, we need to check that this makes sense: if $v' \in V$ is another representative of $\overline{v}$, then $T(v') = T(v)$ because $v' - v \in W \subseteq \mathcal{N}(T)$ which means that $0 = T(v - v') = T(v) - T(v')$. Thus, $\overline{T}$ as defined is a well defined map.

To see that $\overline{T}$ is a linear transformation is now immediate: for $\overline{v}, \overline{w} \in V/W$,

$$\overline{T}(\overline{v} + \overline{w}) = \overline{T}(\overline{v + w}) = T(v + w) = T(v) + T(w) = \overline{T}(\overline{v}) = \overline{T}(\overline{w})$$

Similarly, if $c \in \mathbb{F}$ and $\overline{v} \in V/W$,

$$\overline{T}(c\overline{v}) = \overline{T}(\overline{cv}) = T(cv) = cT(v) = c\overline{T}(\overline{v}).$$

□

One important consequence is the following:

6.7 Corollary. *(First Isomorphism Theorem) Let $T\colon V \to V'$ be a linear transformation. Then $\overline{T}\colon V/\mathcal{N}(T) \to \operatorname{im}(T)$ is an isomorphism.*

Proof. First note that $\overline{T}\colon V/\mathcal{N}(T) \to \operatorname{im}(T)$ is well defined by the theorem applied in case $W = \mathcal{N}(T)$: The theorem yields a map $\overline{T}\colon V/\mathcal{N}(T) \to V'$. Also, it is immediate from the construction of $\overline{T}$, that $\overline{T}$ and T have the same image (whenever $v' \in V'$ is of the form $T(v)$, then $v' = \overline{T}(\overline{v})$ and vice versa). Thus, $\overline{T}$ has image $\operatorname{im}(T)$ and consequently, by changing the codomain, we obtain a linear transformation $V/\mathcal{N}(T) \to \operatorname{im}(T)$, which we also denote by $\overline{T}$.

This $\overline{T}$ is clearly surjective. It remains to see that it is also injective. Recall that it suffices to verify that $\mathcal{N}(\overline{T}) = \{0\}$ (Proposition 4.18). So let $\overline{T}(\overline{v}) = 0$. Then if v is a representative for $\overline{v}$ we have $\overline{T}(\overline{v}) = T(v) = 0$ hence $v \in \mathcal{N}(T)$. But this means that $\overline{v} = 0$ in V/W and so $\mathcal{N}(\overline{T}) = \{0\}$. □

In many theoretical and practical considerations, one is naturally given a generating list for a vector space, but maybe a list with redundancies in the sense that it is not a basis. It is then often the case that one is also given a list of relations among the generators which completely describe the vector space. This is known as a description by *generators and relations*.

For instance we could have a vector space V generated by two elements v, w subject to the relations $v + w = 0$. This means

$$V \cong \tilde{V}/W$$

where $\tilde{V}$ is a vector space of dimension two with some basis $(\tilde{v}, \tilde{w})$ and W is the subspace generated by $\tilde{v} + \tilde{w}$. The cosets of $\tilde{v}$ and $\tilde{w}$ then correspond to v and w, respectively.

The First Isomorphism Theorem also provides the following interesting result:

> Let $T\colon V \to W$ and $S\colon V \to W'$ be two *surjective* linear transformations with $\mathcal{N}(T) = \mathcal{N}(S)$.
>
> Then $W \cong W'$.

Indeed, both W and W' are isomorphic to $V/\mathcal{N}(T)$, and the isomorphism is *canonical*, which is jargon for "natural," which in turn means something like that it is defined in a way independent of the particular case or particular choices.

Remark. The term "canonical" is not easily explained. There is no precise definition of its meaning. Rather it represents a judgement call or the taste of the person using it. Most of the time most mathematicians, however, will agree on its use.

For example, if V, W are two $\mathbb{F}$-vector spaces of the same dimension, then we have shown that they are isomorphic. But most people would not consider any such particular isomorphism as canonical (in general) as it requires the choice of two bases (one in V and one in W), and there usually is no "natural" choice of a basis of a vector space.

On the other hand, if $V = \mathbb{F}^n$ and $W = \mathcal{F}(I, \mathbb{F})$ where $I = \{1, 2, \ldots, n\}$. Then V and W are canonically isomorphic because this time, we do have sort of canonical bases: V has the standard basis $\mathcal{E}$ and W has the basis indexed by the elements of I (ie. $\delta_1, \delta_2, \ldots, \delta_n$ with $\delta_i(j) = \delta_{ij}$).

Example. Quotient spaces also arise naturally in the context of *dual spaces*.

For a vector space V let $V^* = \mathrm{Hom}(V, \mathbb{F})$ be the space of linear transformations from V to $\mathbb{F}$. V^* is called the dual space of V.

Now let $W \subseteq V$ be a subspace, and suppose $\dim V = n < \infty$. Note that there is no "canonical" way of embedding W^* into V^*. Rather than a map $W^* \to V^*$, we get a map $V^* \to W^*$ provided by *restriction*: for $\lambda \colon V \to \mathbb{F}$, consider $R(\lambda) \in W^*$ defined by $R(\lambda)(w) = \lambda(w)$, which makes sense because $W \subseteq V$. In other words, $R(\lambda) = \lambda|_W$ is the restriction of λ to W.

Verify that R is indeed a linear transformation $V^* \to W^*$. Moreover, R is surjective! Indeed, pick a basis for W and extend it to V. Any linear functional $W \to \mathbb{F}$ then can be extended to all of V by arbitrarily specifying values on the basis elements not in W.

What is the nullspace of R? $\mathcal{N}(R) = \{\lambda \in V^* \mid \lambda(w) = 0 \text{ for all } w \in W\}$.

The FIT now states that

$$W^* \cong V^*/\mathcal{N}(R)$$

and this identification is "natural." So the inclusion $W \subseteq V$ provides a description of W^* not as a subspace but as a quotient space of V^*.

On the other hand, suppose we want to describe the dual space of V/W. The FIT directly states that

$$(V/W)^* \cong \mathcal{N}(R).$$

Indeed, if $\overline{\lambda} \colon V/W \to \mathbb{F}$ is a function, then composing it with π we obtain a linear transformation $\overline{\lambda} \circ \pi \colon V \to \mathbb{F}$. Conversely, any element $\lambda \in \mathcal{N}(R)$ gives rise to an element $\overline{\lambda} \in (V/W)^*$ because $W \subseteq \mathcal{N}(\lambda)$.

Now $\lambda \mapsto \overline{\lambda}$ and $\overline{\lambda} \mapsto \overline{\lambda} \circ \pi$ are inverses of each other, hence isomorphisms.

Example. This is an advanced example (view it as an excursion). Let S be the 2-sphere, that is

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

On $\mathbb{R}^3$ we say a function $f\colon \mathbb{R}^3 \to \mathbb{R}$ is *polynomial* if it has the form

$$f(x,y,z) = \sum_{p,q,r} a_{pqr} x^p y^q z^r$$

(with only finitely many $a_{pqr} \neq 0$).

Let $A = \mathcal{P}(\mathbb{R}^3)$ be the set of all polynomial functions. Then A is an infinite dimensional vector space over $\mathbb{R}$.

We say a function $f\colon S \to \mathbb{R}$ is polynomial, if $f = h|_S$ for some $h \in \mathcal{P}(\mathbb{R}^3)$.

Let $\mathcal{P}(S)$ be the set of all polynomial functions. Again, $\mathcal{P}(S)$ is naturally an $\mathbb{R}$-vector space (by pointwise addition and scalar multiplication).

How to describe $\mathcal{P}(S)$? Notice that if we denote $f = x^2 + y^2 + z^2 - 1$ (as a function), then $f\,|_S = 0$. Also, if we have any polynomial function $h\colon S \to \mathbb{R}$, then how to describe h? Clearly, there is a formula $h = \sum a_{pqr} x^p y^q z^r$, but this is ambiguous because

$$\sum a_{pqr} x^p y^q z^r + f$$

defines the same function on S.

Thus, while the polynomial functions on $\mathbb{R}^3$ parameterize the polynomial functions on S, they do so in an ambiguous way.

Let $I(S) = \{h \in \mathcal{P}(\mathbb{R}^3) \mid h|_S = 0\}$. For instance $f \in I(S)$.

Consider the linear transformation $R\colon \mathcal{P}(\mathbb{R}^3) \to \mathcal{P}(S)$, defined by $R(h) = h|_S$. Then $I(S) = \mathcal{N}(R)$ and since by definition R is surjective, we get

$$\mathcal{P}(S) \cong \mathcal{P}(\mathbb{R}^3)/I(S).$$

$\mathcal{P}(S)$ together with addition and pointwise multiplication of functions, is what is called a *ring*. It turns out that the algebraic properties of $\mathcal{P}(S)$ are very closely connected to the geometric properties of the set S. The field of mathematics studying this connection is called *Algebraic Geometry*. Unfortunately, it would lead too far to delve any deeper into this subject.

6.3. Finite fields

At the moment the prime example of an equivalence relation is the following, which we will study now in greater detail:

6.8 Definition. Let $X = \mathbb{Z}$, let $n \in \mathbb{Z}$ be a *fixed* positive integer, and let $\sim$ be defined as $a \sim b$ if and only if $b - a$ is divisible by n. We usually write this as

$$a \equiv b \pmod{n}$$

("a congruent b mod n") and call the equivalence classes *congruence classes*.

The *set of equivalence classes* is usually denoted by $\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$ ("$\mathbb{Z}$ mod $n\mathbb{Z}$").

As an example, if $n = 2$, then there are precisely two congruence classes: C_0 and C_1 where C_0 is the set of even and C_1 is the set of odd integers.

Note that the notation $\mathbb{Z}/n\mathbb{Z}$ is consistent with our notation A/B if we write $n\mathbb{Z}$ for the subgroup of $(\mathbb{Z}, +)$ consisting of all integers divisible by n: $n\mathbb{Z} = \{na \mid n \text{ divides } a\}$. Verify that $n\mathbb{Z}$ is indeed a subgroup. Then it follows that $\mathbb{Z}/n\mathbb{Z}$ is again an abelian group.

As in the case of a vector space and a subspace, where we could also define a scalar multiplication, we can also define the multiplication of two congruence classes. The fundamental point here is that the congruence class of a product ab of integers only depends on the congruence classes of a and b, not the particular integers a, b.

To be precise, suppose $c \equiv a \pmod{n}$ and $d \equiv b \pmod{n}$. Then

$$ab \equiv cd \pmod{n} \tag{6.1}$$

Thus $\overline{ab} = \overline{cd}$. We can therefore define a multiplication of congruence classes as follows:

$$\overline{a} \cdot \overline{b} = \overline{ab}.$$

In words: given two congruence classes C, D, then in order to multiply them, pick $a \in C$ and $b \in D$ compute ab and take its congruence class. The important point is that this procedure is well-defined, i.e. independent of the choice of elements in C and D. It is important to be very comfortable with this construction because it will appear over and over again.

6.9 Example. Let $n = 8$. Then $\overline{2} \cdot \overline{4} = \overline{8}$. Now $\overline{2} = \overline{10}$ and $\overline{4} = \overline{-4}$. And indeed, $\overline{8} = \overline{10 \cdot (-4)} = \overline{-40}$.

Thus we have a (nonempty) set $\mathbb{Z}_n$ together with two operations and we may now proceed and check all our Axioms 1.1. It turns out that all with the possible exception of Axioms e. and f. are satisfied. For instance, that the operations are associative and commutative is an immediate consequence of their definition (and for the addition we have observed this when we proved that A/B is a group).

We show the associative law of the multiplication:

$$\overline{a} \cdot (\overline{b} \cdot \overline{c}) = \overline{a} \cdot \overline{bc} = \overline{a(bc)} = \overline{(ab)c} = \overline{ab} \cdot \overline{c} = (\overline{a} \cdot \overline{b}) \cdot \overline{c}$$

Similarly, we have identities for both, addition and multiplication:

$$\overline{0} + \overline{a} = \overline{0 + a} = \overline{a}$$

and

$$\overline{1} \cdot \overline{a} = \overline{1a} = \overline{a}.$$

We leave the remaining axioms to the reader. There is one and only one instance where Axiome. is violated: in case $n = 1$, we have that $\overline{0} = \overline{1}$, so the two identities are the same which is explicitly ruled out in the axioms. Thus, $\mathbb{Z}_1 = \{\overline{0}\}$ is not a field.

On the other hand, in many cases Axiom f. doesn't hold. For instance, if $n = 8$, then $\overline{24} = \overline{8} = \overline{0}$. But also $\overline{20} = \overline{0}$ and $\overline{4} \neq \overline{0}$ so Lemma 1.4 does not hold. And indeed, $\overline{2}$ does not posses an inverse.

Notice that in case $n = 2$, what we obtain is indeed a field, namely something that looks line $\mathbb{F}_2$; and if $n = 5$, what we obtain is precisely $\mathbb{F}_5$ as discussed at the beginning of this section.

Structures as $\mathbb{Z}_n$ that have two operations that are compatible are important and deserving of their own name.

6.3.1. Rings

6.10 Definition. A *ring* is a triple $(R, +, \cdot)$ where R is a nonempty set, and $+$ and $\cdot$ are *associative* binary operations on R, called the *addition* and *multiplication*, respectively, such that the following hold[5]:

a. $(R, +)$ is an abelian group (with identity element denoted by 0).

b. *Multiplicative identity.* There is an element, denoted 1, such that

$$1a = a1 = a$$

for all $a \in R$.

c. *Distributive Laws.* For $a, b, c \in R$

$$a(b + c) = ab + ac;$$

and

$$(a + b)c = ac + bc.$$

We usually simply write R instead of $(R, +, \cdot)$.

A ring R is *commutative* if the multiplication is commutative as well.

6.11 Remarks.

a. Note that we do not require the multiplication to be commutative.

b. In many textbooks what we defined above is called a *ring with identity* to indicate that it has a multiplicative identity. These books then do not require Axiom e. However, such rings are of no interest to us.

What we have shown so far about $\mathbb{Z}_n$ amounts to:

6.12 Theorem. *For each $n > 0$, $\mathbb{Z}_n$, with addition and multiplication as defined above, is a commutative ring with identity.*

Its additive identity is $\overline{0}$ and its multiplicative identity is $\overline{1}$.

[5] As usual we write ab instead of $a \cdot b$.

We will now always write 0 and 1 instead of $\overline{0}$ and $\overline{1}$ for the identities in $\mathbb{Z}_n$.

6.13 Examples. Besides $\mathbb{Z}_n$, there are a few other examples we came across:

a. Of course, $\mathbb{Z}$ itself is a (commutative) ring.

b. Any field is a (commutative) ring.

c. The set of $n \times n$ matrices $M_n(\mathbb{F})$ together with additin and multiplication of matrices as defined in Chapter 2 is a ring. If $n > 1$, it is *not* commutative.

d. If V is a vector space, then the set $\mathrm{Hom}(V, V)$ if linear transformations from V to V is a ring (with addition and composition of linear transformations).

e. If G is a finite group, and V an $\mathbb{F}$-vector space with a basis indexed by the elements of G (that is, for each $g \in G$, there is a unique element v_g in the basis), we can define a multiplication on V as follows: we put $v_g v_h = v_{gh}$, so that

$$\Big(\sum_{g\in G} c_g v_g\Big)\Big(\sum_{g\in G} d_g v_g\Big) = \sum_{g,h\in G} c_g d_h v_{gh}.$$

It is common, to write simply g instead of v_g. The resulting ring (with addition of V and this multiplication) is called the *group ring* of G. It is commutative if and only if G is abelian. It is important for understanding for instance the "representation theory" of G.

These examples should make it plausible, why such a structure deserves its own name, and why we study rings: if we can prove a theorem about arbitrary rings, it will hold in any one of these considerably different examples.

Remark. As in the case of fields the two identities in a ring are unique.

As are the solutions to equations of the form $a + x = b$. The proofs are the same as in the case of fields (and groups) and don't provide any new insights.

However, equations of the form $ax = b$ may have no or many solutions (even if $a \neq 0$). For instance in $\mathbb{Z}_8$, $\bar{2}x = 0$ has the two solutions $x = 0, \bar{4}$, whereas $\bar{3}x = \bar{5}$ has the unique solution $\bar{7}$ ($\bar{3}$ has a multiplicative inverse, namely $\bar{3}$, so $\bar{3}x = \bar{5}$ if and only if $\bar{3} \cdot \bar{3}x = x = \bar{3} \cdot \bar{5} = \bar{7}$.)

6.3.1 Problem. Let R be a ring. Show that if $0', 1'$ are additive and multiplicative identities then $0 = 0'$ and $1 = 1'$.

Also show that, given $a, b \in R$, $a + x = b$ has a unique solution for x.

What about the equation $ax = b$? Does it have a solution? Is it unique if it exists?

6.3.2 Problem. Is the set $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ together with addition and multiplication a ring? If so, prove it. If not, why not?

6.3.3 Problem. For which n is $M_n(\mathbb{F})$ commutative?

6.14 Proposition. *Let R be a ring. Let $a, b \in R$. Then the following rules hold:*

a. $0 \cdot a = a \cdot 0 = 0$.

b. $(-1) \cdot a = -a$. *More generally,* $(-b)a = -(ba) = b(-a)$.

c. $-(-a) = a$.

d. $(-a)(-b) = ab$.

e. $-(a+b) = -a + (-b)$.

Proof. The proof is identical to the corresponding facts in case R is a field. See Proposition 1.7 for details. □

6.3.2. $\mathbb{Z}_n$ when n is prime

We now turn to the question, when is $\mathbb{Z}_n$ a field? Recall that an integer p is called *prime* if $p \neq 0, \pm 1$ and if the only divisors[6] of p are $\pm 1, \pm p$.

Recall the following two fundamental facts:

Every integer other than $0, \pm 1$ is a product of primes: Even though, we all know this, a proof requires induction. We will show that any $n > 1$ is a product of primes. The case of negative integers then easily follows.

Now $n = 2$ is itself prime and hence a product of primes, covering the base case. Now suppose $n > 1$ is given, and for every integer a such that $1 < a < n$, a is a product of primes.

Then n is either prime itself, and we are done, or $n = ab$ with[7] $1 < a, b < n$. Then a, b both are products of primes by the induction assumption, and so is n. □

There are infinitely many primes: This is Euclid's famous theorem. Suppose there were only finitely many primes. Then there is also only a finite list of positive primes. Let $p_1, p_2, \ldots, p_k$ be a list of all positive primes. Then $N := p_1 p_2 \cdots p_k + 1$ is an integer. It is positive. However, it is not divisible by any prime: indeed, if a prime p divides N, then $|p|$ is positive and also a prime. Thus, $|p| = p_i$ for some i. Moreover, p_i divides N. But this clearly forces p_i to divide 1 which is absurd! Thus, N is not divisible by any prime, contradicting the previous paragraph. Thus, the list of primes must be infinite. □

If a, b are integers, not both zero, then the *greatest common divisor* of a and b, denoted $\gcd(a, b)$ (or simply but ambiguously by (a, b)) is defined as the greatest integer that divides[8] both a and b.

When studying integers (and modular properties), the following result is of fundamental importance.

[6] A *divisor* of an integer n is an integer m such that $n = qm$ for some $q \in \mathbb{Z}$.

[7] If n is not prime, then there is an integer other than $\pm 1, \pm n$ dividing n. So $n = ab$ with $b \neq \pm 1$ as well. Then also $n = |a||b|$ and it is clear that $1 < |a|, |b| < n$.

[8] An integer d *divides* an integer n, if d is a divisor of n.

6.15 Proposition. *Let a, b be integers.*

a. *There are integers u, v such that*

$$\gcd(a, b) = ua + vb.$$

b. *Suppose $a \neq 0$. Let d be an integer that divides both a and b, then d divides $\gcd(a, b)$.*

c. *Suppose $a \neq 0$ and $\gcd(a, b) = 1$. If a divides bc for some integer c, then a divides c.*

We defer the proof for a little while, since it is almost identical to the proof of a similar fact about polynomials. But note that b. and c. are easy consequences of a.

6.16 Corollary. *$\mathbb{Z}_n$ is a field if and only if n is prime.*

Proof. The only if part is clear (and we have discussed this above): suppose n is not prime. Then there are positive integers $a, b < n$ such that $n = ab$. It follows that in $\mathbb{Z}_n$ we have $\overline{ab} = \overline{a} \cdot \overline{b} = 0$. But $\overline{a}, \overline{b} \neq 0$ (because n does not divide a, b) which is impossible in a field since it violates the cancelation rule (Lemma 1.4).

More interesting is the converse: suppose n is prime. Let $\overline{a} \in \mathbb{Z}_n$ be nonzero. Then n does not divide any representative a of $\overline{a}$. In particular, $\gcd(a, n) = 1$ because n is prime. It follows that $1 = ua + vn$ for suitable u, v.

Modulo n this means $1 = \overline{ua + vn} = \overline{ua} + 0 = \overline{u} \cdot \overline{a}$ and $\overline{u}$ is a multiplicative inverse. □

6.3.4 Problem. Let $\mathbb{F}$ be a finite field.

For $n \in \mathbb{Z}$ and define $\overline{n} \in \mathbb{F}$ as $n \cdot 1$ where $n \cdot 1 = 1 + 1 + \cdots + 1$ (n summands) if $n > 0$, and $-((-n) \cdot 1)$ if $n < 0$, and 0 if $n = 0$.

Show that $E = \{\overline{n} \mid n \in \mathbb{Z}\}$ is a subfield of $\mathbb{F}$.

Show that E may be identified with a field of the form $\mathbb{Z}_p$ (p prime).

Show that $\mathbb{F}$ has p^k elements for some k.

6.4. Polynomials

In this section we want to introduce one more ring, namely the ring of polynomials. We all know what polynomial functions are. So why do we need a new definition? The problem are finite field. If $\mathbb{F}$ is a finite field, we can still define a polynomial function as a map $f \colon \mathbb{F} \to \mathbb{F}$ defined by a formula $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ (with $a_i \in \mathbb{F}$). However, there are vastly more formulas than there are functions!

For instance if $\mathbb{F} = \mathbb{F}_2$, there are precisely four distinct functions $\mathbb{F}_2 \to \mathbb{F}_2$. However, there are infinitely many formulas as above. Thus, there are many formulas, that define the same function. As an example, note that $f(x) = x^2$ and $f(x) = x$ both define the same function, and so $f(x) = x^2 - x$ is a complicated way of writing the constant zero function.

However, we will need an unambiguous way of talking about polynomial expressions, so we will need another definition of a polynomial.

How to approach this? Let us see what we want: we want that a polynomial is determined by its coefficients and conversely, that the coefficients unambiguously define a polynomial. Thus, let us simply define a polynomial that way.

6.17 Definition. Let $\mathbb{F}$ be a field. A *polynomial* with coefficients in $\mathbb{F}$ is a sequence[9] $f = (a_0, a_1, a_2, \dots)$ of elements $a_i \in R$ such that *only finitely many a_i are nonzero*, that is, there is $k > 0$ such that $a_i = 0$ for all $i > k$.

The elements a_i are called the *coefficients* of f.

It follows that two polynomials (with coefficients in $\mathbb{F}$) are equal if and only if the corresponding coefficients coincide.

We define an addition and multiplication of polynomials as follows:

Let $f = (a_0, a_1, a_2, \dots)$ and $g = (b_0, b_1, b_2, \dots)$ then

$$f + g = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \dots)$$

is the sequence whose ith coefficient is equal to $a_i + b_i$. More involved is the definition for the multiplication: we define $fg = (c_0, c_1, c_2, \dots)$ with

$$c_i = a_0 b_i + a_1 b_{i-1} + a_2 b_{i-2} + \cdots + a_i b_0 = \sum_{k=0}^{i} a_k b_{i-k}.$$

It is clear that both $f + g$ and fg are polynomials: the coefficients $a_i + b_i$ and c_i are zero for i sufficiently large.

6.18 Theorem. *The set of polynomials with coefficients in a field $\mathbb{F}$ form a commutative ring[10].*

Proof. We have seen before that $\mathcal{F}(\mathbb{N}_0, \mathbb{F})$ is a vector space. Since the addition of polynomials is actually the addition in this vector space, the axioms regarding the addition are clearly satisfied.

For instance the additive identity is the same as the additive identity of $\mathcal{F}(\mathbb{N}_0, \mathbb{F})$, namely the sequence $(0, 0, 0, \dots)$ of all zeros, which is a polynomial.

So let us focus on the multiplication: from the defining formula commutativity is clear since the multiplication in $\mathbb{F}$ is commutative. As for associativity: suppose f, g, h are three polynomials with coefficients a_i, b_i, c_i respectively.

Then the kth coefficient of $(fg)h$ is

$$\sum_{p+q=k} \Big(\sum_{r+s=p} a_r b_s \Big) c_q = \sum_{r+s+q=k} (a_r b_s) c_q.$$

[9] Formally, this is a function $\mathbb{N}_0 \to \mathbb{F}$.

[10] It is nowhere used that $\mathbb{F}$ is a field. The definition makes sense and the theorem holds, even in case $\mathbb{F}$ is just a commutative ring.

The kth coefficient of $f(gh)$, on the other hand, is

$$\sum_{p+q=k} a_p(\sum_{r+s=q} b_r c_s) = \sum_{p+r+s=k} a_p(b_r c_s)$$

which is the same (it does not matter how we label the indices), because of the associative law of the multiplication in $\mathbb{F}$.

The distributive law for the multiplication can also be checked by comparing coefficients: at position k, $f(g+h)$ has the coefficient

$$\sum_{p+q=k} a_p(b_q + c_q) = \sum_{p+q=k} a_p b_q + \sum_{p+q=k} a_p c_q$$

which is the coefficient of $fg + fh$ at the same position. □

Now consider the map from $\mathbb{F}$ into the set of polynomials that maps $a \in \mathbb{F}$ to the polynomial $a' = (a, 0, 0, \dots)$. It is clear that $(a+b)' = a' + b'$ and $(ab)' = a'b'$. Also, this map is clearly one to one. Its image, denoted $\mathbb{F}'$ for the moment, is a subfield of the ring of polynomials, which we identify[11] with $\mathbb{F}$. We call $\mathbb{F}'$ the set of constant polynomials and omit henceforth the supscript $'$.

Now let x be the polynomial $(0, 1, 0, 0, \dots)$. Then a straight forward computation shows that $x^2 = (0, 0, 1, 0, \dots)$ and more generally

$$x^i = (\underbrace{0, 0, \dots, 0, 1}_{i+1}, 0, \dots)$$

for $i > 0$.

Also, verify(!) that for each $a \in \mathbb{F}$ we have

$$ax^i = (a, 0, 0, \dots)(\underbrace{0, 0, \dots, 0, 1}_{i+1}, 0, \dots) = (\underbrace{0, 0, \dots, 0, a}_{i+1}, 0, \dots).$$

If $f = (a_0, a_1, \dots)$ is an arbitrary polynomial, we then find that

$$f = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k$$

where k is such that $a_i = 0$ for all $i > k$. If we require that $a_k \neq 0$ then f can be written uniquely in this form as long as $f \neq 0$.

From now on we denote the ring of polynomials with coefficients in $\mathbb{F}$ by $\mathbb{F}[x]$ (but note that it is irrelevant how we label x). $\mathbb{F}[x]$ is often also called the *polynomial ring in one variable over* $\mathbb{F}$. The term "variable" should be interpreted loosely, it has no specific meaning; it just reminds us that historically, polynomials were really thought of as functions.

What have we done now? We have now an object (namely, $\mathbb{F}[x]$) that behaves in exactly the same way as the polynomial functions on the real line do, but we avoid the difficulty sketched above when working with finite fields.

[11] Much in the same way as the field $\mathbb{R}$ may be identified with the set $\{(a, 0) \mid a \in \mathbb{R}\}$ of complex numbers.

Remark. Let $\mathcal{P}(\mathbb{F})$ denote the set of polynomial functions $\mathbb{F} \to \mathbb{F}$ (defined in the same way as for real numbers).

For $f \in \mathbb{F}[x]$, let $P_f : \mathbb{F} \to \mathbb{F}$ be defined as follows: if $f = a_0 + a_1 x + \cdots + a_n x^n$, then $P_f(a) = a_0 + a_1 a + a_2 a^2 + \cdots + a_n a^n$.

So for each $f \in \mathbb{F}[x]$ we can associate a polynomial function. One can show that the induced map $\mathbb{F}[x] \to \mathcal{P}(\mathbb{F})$ is surjective (in fact, that is immediate from the definition of $\mathcal{P}(\mathbb{F})$).

One can also show that this map is injective if and only if $\mathbb{F}$ is infinite.

It should be clear that computations in $\mathbb{F}[x]$ follow the familiar pattern of computations with polynomial functions.

Example. In $\mathbb{C}[x]$,

$$\begin{aligned}(x^3 + x + 1)(x^2 + 2x + i) &= x^5 + 2x^4 + ix^3 + x^3 + 2x^2 + ix + x^2 + 2x + i \\ &= x^5 + 2x^4 + (1+i)x^3 + 3x^2 + (2+i)x + i.\end{aligned}$$

6.19 Definition. Let $f = a_0 + a_1 x + a_2 x^2 + \cdots \in \mathbb{F}[x]$ be a nonzero polynomial.

Then the *degree* of f is $d \in \mathbb{N}_0$ where d is maximal such that $a_d \neq 0$. Thus $a_i = 0$ for all $i > d$ and $a_d \neq 0$.

We usually write $\deg f$ for d.

The zero polynomial $0 = 0 + 0x + 0x^2 + \ldots$ does not have a degree[12].

6.20 Proposition. *Let $f, g \in \mathbb{F}[x]$ be two nonzero polynomials. Then*

$$\begin{aligned}\deg(f+g) &\leq \max\{\deg f, \deg g\} \qquad \text{if } f+g \neq 0, \\ \deg(fg) &= \deg f + \deg g.\end{aligned}$$

Proof. The first statement is a straightforward exercise and left to the reader.

As for the second, let the coefficients of f be denoted by a_i and those of g by b_i. Then the kth coefficient c_k of fg is

$$\sum_{p+q=k} a_p b_q$$

Note that if $k > \deg f + \deg g$, in the sum we must have $p > \deg f$ or $q > \deg g$. Thus, $c_k = 0$. On the other hand, if $k = \deg f + \deg g$, then there is precisely one summand nonzero, namely $a_{\deg f} b_{\deg g}$ (which, as a product of nonzero field elements, is nonzero). In particular, $fg \neq 0$, and $\deg fg = \deg f + \deg g$. □

Remark. Convince yourself, that if $\deg f \neq \deg g$, then $\deg(f+g) = \max\{\deg f, \deg g\}$.

6.21 Corollary. *If in $\mathbb{F}[x]$ $fg = 0$, then $f = 0$ or $g = 0$. If $fg = fh$ and $f \neq 0$, then $g = h$.*

[12] It is also common to write formally $\deg 0 = -\infty$ with the understanding that this just means that $\deg 0 < \deg f$ for all $f \neq 0$.

Proof. We already proved the first part in the proof of the proposition.

So let $f, g, h \in \mathbb{F}[x]$ with $f \neq 0$ such that $fg = fh$. It then follows that $fg - fh = f(g-h) = 0$. Thus, $f = 0$ or $g - h = 0$. Since $f \neq 0$ this means $g - h = 0$, or $g = h$ as claimed. □

We all have painful memories of long division of polynomials. Here is a proper proof that it actually works.

6.22 Theorem (Division Algorithm)**.** *Let $f, g \in \mathbb{F}[x]$ with $g \neq 0$. There exist uniquely determined polynomials $Q, R \in \mathbb{F}[x]$ such that*

$$f = Qg + R$$

and such that $R = 0$ or $\deg R < \deg g$.

Proof. We first prove the existence of Q, R and worry about the uniqueness later.

This is a classical case for complete induction. First note that if $f = 0$, we can take $Q = R = 0$.

Let $f = a_0 + a_1x + \cdots + a_dx^d$ and suppose $g = b_0 + b_1x + \cdots + b_ex^e$ with $e = \deg g \geq 0$ and $b_e \neq 0$.

The base case is now $\deg f = 0$. If $\deg g > 0$ put $Q = 0$ and $R = f$ and we are done. Otherwise, $g = b_0$ and $f = a_0$ and $f = (b_0^{-1}a_0)b_0 + 0$ ie. $Q = b_0^{-1}a_0$ and $R = 0$.

Suppose $d > 0$ is given and suppose for polynomials f of degree strictly less than d, we can find Q and R (Induction Assumption).

Let f be a polynomial of degree d. If $e > d$, then $R = f$ and $Q = 0$ satisfies our requirements. Otherwise, consider

$$f_1 = f - a_e^{-1}b_dx^{d-e}g.$$

Then $f_1 = 0$ or $\deg f_1 < \deg f = d$. In either case there is Q_1, R_1 such that $f_1 = Q_1g + R_1$ and $R_1 = 0$ or $\deg R_1 < \deg g$.

Then

$$f = Q_1g + R_1 + a_e^{-1}b_dx^{d-e}g = (Q_1 + a_e^{-1}b_dx^{d-e})g + R_1.$$

Putting $R = R_1$ and $Q + Q_1 + a_e^{-1}b_dx^{d-e}$ then gives the result for polynomials of degree d.

Finally, suppose $f = Qg + R = Sg + T$ with R, T either zero or of degree strictly less than $\deg g$. Then

$$(Q - S)g = T - R.$$

Now if $T - R \neq 0$, then $\deg(T - R) < \deg g$ by Proposition 6.20. Then $(Q - S)g \neq 0$ and by the same proposition we get $\deg(T - R) = \deg(Q - S)g = \deg(Q - S) + \deg g \geq \deg g$. This is a contradiction and so $T = R$. But then also $Qg = Sg$ and we have $Q = S$ by Corollary 6.21. □

The set of polynomial functions $\mathcal{P}(\mathbb{R})$ on $\mathbb{R}$ consists of functions, ie. we can evaluate an element $f \in \mathcal{P}(\mathbb{R})$ at any given fixed $\xi \in \mathbb{R}$ to obtain a number $f(\xi) \in \mathbb{R}$. It turns out that we can do the same with polynomials.

6.23 Definition. Let $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \in \mathbb{F}[x]$ and $\xi \in \mathbb{F}$.

We define $f(\xi) \in \mathbb{F}$ as

$$f(\xi) = \sum_{i=0}^{n} a_i\xi^i$$

and call $f(\xi)$ the *value of f at ξ*.

Any $\xi \in \mathbb{F}$ such that $f(\xi) = 0$ will be called a *zero* or *root* of f.

6.4.1 Problem. Verify that for each $\xi \in \mathbb{F}$ and $f, g \in \mathbb{F}[x]$ we have

$$(f + g)(\xi) = f(\xi) + g(\xi)$$

and

$$(fg)(\xi) = f(\xi)g(\xi).$$

6.24 Proposition (Remainder Theorem)**.** *Let $f \in \mathbb{F}[x]$ and $\xi \in \mathbb{F}$. Then the remainder of the division of f by $g = x - \xi$ is equal to $f(\xi)$. In other words,*

$$f = (x - \xi)Q + f(\xi)$$

for some $Q \in \mathbb{F}[x]$.

In particular, $f = (x - \xi)Q$ for some Q if and only if $f(\xi) = 0$.

Proof. The second assertion is immediate from the first.

By the division algorithm, $f = (x - \xi)Q + R$ (applied to $g = x - \xi$).

Evaluating both sides at ξ and using Problem 6.4.1 we find that

$$f(\xi) = (\xi - \xi)Q(\xi) + R(\xi)$$

hence $f(\xi) = R(\xi)$. But also $R = 0$ or $\deg R < \deg g = 1$. In any event R is constant, ie. $R = R(\xi)$ and the claim follows. □

Let R be any commutative ring (think $R = \mathbb{Z}$ or $R = \mathbb{F}[x]$). For $a, b \in R$ we say a *divides* b, denoted $a \mid b$, if there is $c \in R$ such that $b = ac$. If that is the case, a is called a *divisor* or *factor* b and b is called a *multiple* of a. An element a of R is called a *unit*, if $a \mid 1$. (Thus, a is a unit if and only if it has a multiplicative inverse.) Notice that a unit divides any other element of R.

What we want to achieve here is a sound theory of factorization of polynomials. In other words, we want to be able to write a polynomial unambigously as a product of polynomials which themselves cannot be written as products of polynomials smaller degrees. As a guiding example you should keep in mind the ring of integers.

6.25 Definition. An element p of a ring R is called *prime* if p is not zero and p is not a unit and if the following holds:

> whenever $p \mid ab$ for some $a, b \in R$, then $p \mid a$ or $p \mid b$.

p is called *irreducible* if p is not zero nor a unit and if whenever $p = ab$ for some a, b then it follows that a or b is a unit.

Note that if $R = \mathbb{Z}$, the irreducible integers are exactly those that are commonly called prime integers. One has to work a little bit to show that prime integers are actually prime in the sense defined here as well.

6.26 Definition. Let $r_1, r_2, \ldots, r_n \in R$ be some elements, at least one of which is nonzero. A *greatest common divisor* (gcd) of $r_1, r_2, \ldots, r_n$ is an element $d \in R$ such that

$d \mid r_i$ for $i = 1, 2, \ldots, n$; and if $d' \mid r_i$ for all i then also $d' \mid d$.

Example. 3 is a greatest common divisor of 6 and 9, because it divies 6 and 9 and if n divides 6 and 9, then n divides also $9 - 6 = 3$.

We want to study these concepts now for polynomials. The first remark is that $f \in \mathbb{F}[x]$ is a unit if and only if $f \neq 0$ and f is "constant" that is $\deg f = 0$. (Can you prove this? Use Proposition 6.20.)

The following theorem says all there is to say about gcds of polynomials.

6.27 Theorem. *Let $f_1, f_2, \ldots, f_n \in \mathbb{F}[x]$ be not all zero.*

a. *A gcd for $f_1, f_2, \ldots, f_n$ exists.*

b. *Up to multiplication by a nonzero constant, this gcd d is unique and can be written as*

$$d = h_1 f_1 + h_2 f_2 + \cdots + h_n f_n$$

for suitable $h_i \in \mathbb{F}[x]$.

c. *d is indeed a greatest common divisor in the sense that among all common divsors, d has the largest degree.*

The proof is not hard but a little technical.

Proof. Let $I \subseteq \mathbb{F}[x]$ be the set of all polynomials of the form

$$\sum_i g_i f_i$$

(Compare this with the definition of Span in Chapter 3.)

I clearly contains $f_1, f_2, \ldots, f_n$. Also it has the following two additional properties:

a. If $p, q \in I$ then so is $p + q$.

b. If $q \in I$ and $p \in \mathbb{F}[x]$ then $pq \in I$.

(Compare this with the definition of a subspace.) Nonempty subsets of a ring with these property are called *ideals*.

Since not all f_i are zero, we have that I contains nonzero polynomials. By the Well Ordering Principle (for $\mathbb{N}_0$) it follows that there exists a nonzero polynomial $d \in I$ such that $\deg d \leq \deg p$ for all $p \in I$. Then

$$d = h_1 f_1 + \cdots + h_n f_n.$$

I claim that d is a gcd. Indeed, d is a common divisor: for let $f_i = Q_i d + R_i$, according to the division algorithm. Then $R_i = 0$ or $\deg R_i < \deg d$. Now $R_i = f_i - Q_i d \in I$ which forces $R_i = 0$ since d has minimum degree among all elements of I.

Now let d' be any common divisor of $f_1, f_2, \ldots, f_n$. Then clearly d' divides all elements of I: $f_i = g_i d'$ and so $d = \sum_i h_i f_i = (\sum_i h_i g_i) d'$ is divisible by d' which makes d a gcd.

Now let e be any gcd for $f_1, f_2, \ldots, f_n$. Then $e \mid d$ because e is a common divisor and d is a gcd. Similarly, $d \mid e$ by exchanging the roles of e and d. So we have $e = du$ and $d = ev$ for suitable u, v. It follows that $e = euv$ and hence $uv = 1$. This means u, v are units, ie. nonzero constants as claimed.

Now for the final assertion since every common divisor divides d, d clearly has maximum degree among all divisors. If e is any divisor of the same degree as d, then since e divides d, $d = eu$ it follows that $\deg u = 0$ and hence e is a gcd as well. □

Remark. Of course the theorem forces the question how we could compute the gcd, and how we could compute the h_i? Here is an outline in the case of two polynomials f, g.

> Convince yourself that the following holds: if $f = Qg + R$ according to the Division Algorithm then if $R \neq 0$ it holds that $\gcd(f, g) = \gcd(g, R)$.

This can be used to formulate what is known as the *Euclidean Algorithm*: The algorithm's input is a pair of nonzero polynomials a_0, b_0 with $\deg(a_0) \geq \deg(b_0)$. The output is $\gcd(a_0, b_0)$ in the form $ua_0 + vb_0$.

At each stage of the algorithm we have a_i, b_i with $\deg a_i \geq \deg b_i$. We then compute

$$a_i = Q_{i+1} b_i + R_{i+1}$$

If $R_{i+1} = 0$, this process stops. Otherwise we put $a_{i+1} = b_i$ and $b_{i+1} = R_{i+1}$. Then we have $\gcd(a_{i+1}, b_{i+1}) = \gcd(a_i, b_i) = \cdots = \gcd(a_0, b_0)$.

Note also that at each stage R_i is explicitly a linear combination (with polynomial coefficients) of a_0, b_0 (since if a_i, b_i are so is R_i and consequently also a_{i+1}, b_{i+1}).

Notice also that $\deg R_1 > \deg R_2 > \ldots$ which cannot go on indefinitely. Hence, finally, we must have a k such that $R_{k+1} = 0$. If that happens, b_k divides a_k and so $R_k = b_k = \gcd(a_k, b_k) = \gcd(a_0, b_0)$. Thus, the last nonzero remainder is the gcd[13].

6.28 Example.

[13] The very same procedure applies to computing the gcd of two integers: divide the larger by the smaller and repeat until the remainder is zero.

- Let $f = x^4 + 3x^3 + x^2$ and $g = 3x^2 + 4x + 1$ (both in $\mathbb{F}_7[x]$, where $\mathbb{F}_7 = \mathbb{Z}/7\mathbb{Z}$ is the field with 7 elements constructed earlier).
- $a_0 = f$ and $b_0 = g$.

$$\begin{aligned}
&x^4 + 3x^3 + x^2 && = (3x^2 + 4x + 1)(5x^2 + 6x) + x \\
&-(x^4 + 6x^3 + 5x^2) && \\
&=4x^3 + 3x^2 && \\
&-(4x^3 + 3x^2 + 6x) && \\
&=x &&
\end{aligned}$$

  So $a_1 = g$ and $b_1 = x$.
-

$$\begin{aligned}
&3x^2 + 4x + 1 = x(3x + 4) + 1 \\
&-3x^2 \\
&=4x + 1 \\
&-4x \\
&=1
\end{aligned}$$

  The remainder $R_2 = b_2 = 1$ which divides $a_2 = b_1 = x$ and so the

$$\gcd(f, g) = 1.$$

- Finally, observe that $b_1 = x = f - (5x^2 + 6x)g$,

$$\begin{aligned}
1 = R_2 = a_1 - (3x + 4)b_1 &= g - (3x + 4)(f - (5x^2 + 6x)g) \\
&= -(3x + 4)f + ((3x + 4)(5x^2 + 6x) + 1)g.
\end{aligned} \tag{6.2}$$

Now we have everything we need to make some very strong statements about polynomials. Here are some immediate consequences:

6.29 Lemma. *A polynomial $p \in \mathbb{F}[x]$ is prime if and only if it is irreducible.*

Proof. Indeed, suppose p is prime and $p = ab$ for some polynomials a, b. Since p divides p, it must divide ab and hence divides a or b as p is prime. We may assume that p divides a. Then $a = cp$ and so $p = cbp$. It follows that $cb = 1$ by the cancelation rule Corollary 6.21. Thus, b is a unit and so p is irreducible.

Let now p be irreducible and suppose p divides ab for some $a, b \in \mathbb{F}[x]$. Then $\gcd(a, p) = 1$ or $\gcd(a, p) = p$ (up to multiplication by a nonzero scalar).

In the second case we are done. In the first case, we may write $1 = ua + vp$ and conclude that $b = uab + vbp$, where p divides every summand on the right, and hence b. Hence p is prime. □

6.4.2 Problem. In every ring where we have a cancelation rule, a prime is always irreducible. However, the converse may not be true.

But show that also in $\mathbb{Z}$, every prime in the ring theoretic sense is a prime in the integer sense (ie. irreducible).

The main result of this section is the following:

6.30 Theorem (Unique Factorization Theorem)**.** *Let $f \in \mathbb{F}[x]$ be a nonzero polynomial. If f is not a unit* (*ie. not constant*) *then there are irreducible polynomials $p_1, p_2, \ldots, p_s$ such that*

$$f = p_1 p_2 \cdots p_s.$$

Furthermore, this factorization into irreducibles is unique in the sense that if

$$p_1 p_2 \cdots p_s = q_1 q_2 \cdots q_t$$

with q_i irreducible then $s = t$ and after a suitable renumbering, we have $q_i = c_i p_i$ for some nonzero scalar $c_i \in \mathbb{F}$.

Proof. The first part consists of showing that such a factorization exists. The proof of this is almost verbatim the same as the proof we have for integers on Page 144: We do induction on the degree.

The result is clear for polynomials of degree 1: they are all irreducible because if $f = pq$ then $\deg p + \deg q = 1$ and so p or q must have degree 0 and hence be a unit.

Now suppose the result has been shown for polynomials of degree strictly less than some given integer $n > 0$. Let f be a polynomial of degree n. Then f is either irreducible itself (and we are done), or $f = pq$ with p and q not units. Then $\deg p, \deg q > 0$ and so since $\deg p + \deg q = \deg f$ we find that $1 < \deg p, \deg q < n$. It follows that p and q both a products of irreducibles and hence so is f.

The uniqueness part is not much harder: we proceed by induction on the number s of factors. Suppose $s = 1$ and suppose $p_1 = q_1 q_2 \cdots q_t$. Then since p_1 is irreducible, and since the q_i are irreducible as well, this forces $t = 1$ and $q_1 = p_1$ as claimed: indeed, $p_1 = q_1(q_2 \cdots q_t)$ which means that q_1 or $(q_2 \cdots q_t)$ is a unit. But q_1 certainly isn't, and $q_2 \cdots q_t$, if present, has positive degree – a contradiction.

Now suppose $s > 0$ is given and for all products of at most $s - 1$ primes, the decomposition of the product is essentially unique (in the sense of the theorem).

Let

$$p_1 p_2 \cdots p_s = q_1 q_2 \cdots q_t.$$

Then $t \geq s$ because otherwise the induction assumption applies to t – forcing $s = t$, a contradiction.

Since p_s divides the left hand side, it must divide the right and so p_s divides $q_1 q_2 \cdots q_t$. Also, p_1 is irreducible and hence prime so it must divide one of the factors: $p_s \mid q_j$ for some j. After renumbering we may assume that $j = t$, and then, since q_t is irreducible, it follows that $q_t = c p_s$ for some nonzero c.

We thus have

$$p_1p_2\cdots p_s = q_1q_2\cdots q_{t-1}(cp_s)$$

and after canceling p_s on both sides we have

$$p_1p_2\cdots p_{s-1} = cq_1\cdots q_{t-1}.$$

The induction hypothesis now applies and we are done: $t-1=s-1$ and so $t=s$ after relabeling p_i is a multiple of q_i. □

6.4.3 Problem. Adapt this proof to show the Fundamental Theorem of Arithmetic: Every integer $n>1$ is a product of primes in an essentially unique way.

6.4.4 Problem. Show that a polynomial $f\in\mathbb{F}[x]$ with $\deg f=n>0$ has at most n distinct roots in $\mathbb{F}$.

6.4.5 Problem. Let $f,g\in\mathbb{F}[x]$ be nonzero such that $\gcd(f,g)$ is a unit in $\mathbb{F}[x]$. Show: Suppose $h\in\mathbb{F}[x]$ is divisible by both, f and g. Then h is divisible by fg.

Generalize this to arbitrarily many factors whose pairwise gcd is a unit.

6.5. More on complex numbers

Geometrically, if $z=a+bi$, then $\sqrt{a^2+b^2}$ is the length of the line segment from $(0,0)$ to $(a,b)=z$. This is the Pythagorean Theorem[14]

For this reason, we define the *absolute value* (also *modulus*) of a complex number to be

$$|z| = \sqrt{\mathrm{Re}(z)^2+\mathrm{Im}(z)^2} = \sqrt{z\overline{z}}.$$

The formula for the inverse of a nonzero complex number z now reads:

$$z^{-1} = \frac{\overline{z}}{|z|^2}$$

where we use the usual convention to write x/y instead of $y^{-1}x$. From the fact that $\overline{zw}=\overline{z}\cdot\overline{w}$ one readily deduces that

$$|zw| = |z||w|. \tag{6.3}$$

Indeed, since $\overline{zw}=\overline{z}\cdot\overline{w}$, we find that

$$\sqrt{(zw)(\overline{zw})} = \sqrt{(z\overline{z})(w\overline{w})} = |z|\cdot|w|$$

[14] Note however, that notions like "length," "angle," "area," or "line segment" don't have any meaning until we properly define them. One should view this simply as a heuristic *motivation* for the things to come. We will discuss this in greater detail later.

(recall that for two nonnegative real numbers a, b we always have $\sqrt{xy} = \sqrt{x}\sqrt{y}$; apply this with $x = z\overline{z}$ and $y = w\overline{w}$).

In particular, if z is a complex number with absolute value 1, then the absolute value of zw is the same as the absolute value of w.

To get some geometric intuition about complex numbers it is often helpful to visualize this as follows. But keep in mind that drawing pictures is no substitution for rigorous mathematics. In fact, these geometric pictures really are only helpers.

If $|w| = r$ and $|z| = 1$ this means that w and zw lie on a circle of radius r around the origin. In fact, one can show that if $|z| = 1$, there is a unique angle α such that for all complex numbers w, zw is obtained from w by a rotation of w by α about the origin. How do we find α? Well, $z = z \cdot 1$, so α is the angle that z encloses together with the x-axis (against the clock).

For example this implies that the multiplication with i corresponds to a rotation of 90 degrees[15]. Multiplying by $-i$ is rotation by 270 degrees.

If you know enough trigonometry you can deduce that for a complex number z of absolute value 1 there is a unique $\alpha \in [0, 2\pi)$ such that

$$z = \cos(\alpha) + \sin(\alpha)i$$

A proper proof of this fact would use a lot of calculus. Recall that the multiplication by z is a linear transformation of the $\mathbb{R}$-vector space $\mathbb{C}$ with matrix

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

with respect to the basis $(1, i)$ of $\mathbb{C}$. Since $a = \cos\alpha$, and $b = \sin\alpha$, this is indeed the matrix of a rotation (cf. Section 4.3).

Notice that since we have an absolute value for complex numbers, the usual notions of convergence of sequences makes sense in $\mathbb{C}$ as well: by definition, a sequence z_n of complex numbers converges, with limit $z_0 \in \mathbb{C}$, say, if the sequence

$$a_n := |z_n - z_0|$$

which is a sequence of nonnegative real numbers, converges to 0 in $\mathbb{R}$.

6.5.1 Problem. Show that a sequence z_n of complex numbers converges if and only if the two real sequences $a_n := \mathrm{Re}(z_n)$ and $b_n := \mathrm{Im}(z_n)$ both converge, and that if so, $\lim_{n\to\infty} z_n = a_0 + b_0 i$ where $a_0 = \lim_{n\to\infty} a_n$ and $b_0 = \lim_{n\to\infty} b_n$.

All related notions like convergence of series etc. now make sense. In particular, for every complex number z one might ask whether the series

$$\exp(z) := \sum_{n=0}^{\infty} \frac{1}{n!} z^n$$

[15] Rotation are always measured against the clock.

converges. The answer is yes. We usually write

$$e^z := \exp(z).$$

This can be used to define other exponentials: if $a \in \mathbb{C}$ is nonzero and $z \in \mathbb{C}$ one can define a^z as $\exp(z \ln a)$ where $\ln a$ is a complex number such that $e^{\ln a} = a$. However, $\ln a$ is *not uniquely determined*. One can show that *there exists no continuous logarithm on* $\mathbb{C}$ that is there is no continuous function $f\colon \mathbb{C} \setminus \{0\} \to \mathbb{C}$ such that $e^{f(z)} = z$. It would lead too far to investigate this further, however.

The usual rules for the exponential functions still apply:

$$e^{z+w} = e^z \cdot e^w.$$

Also, it is not hard to show that $\overline{e^z} = e^{\overline{z}}$, from which it easily follows that $\overline{e^{\alpha i}} = e^{-\alpha i}$ if $\alpha \in \mathbb{R}$. Also, $e^{-\alpha i} = (e^{\alpha i})^{-1}$ from which we deduce that

$$|e^{\alpha i}| = 1 \qquad \text{for all } \alpha \in \mathbb{R}.$$

It follows that $e^{\alpha i} = \cos(\beta) + \sin(\beta)i$ for some β. What is the relation between α and β? A beautiful old theorem (due to LEONHARD EULER) states that for the complex number z of absolute value 1 at angle β to the real axis, we have

$$z = e^{\beta i}$$

In other words, $\alpha = \beta + 2\pi n$ for some $n \in \mathbb{Z}$ and

$$\text{(Euler's Formula)} \qquad e^{\alpha i} = \cos(\alpha) + \sin(\alpha)i.$$

Note that it follows that $e^{\alpha i} = 1$ if and only if $\alpha \in 2\pi\mathbb{Z} := \{2\pi n \mid n \in \mathbb{Z}\}$.

6.31 Proposition. *Let $n > 0$ be an integer, and z a nonzero complex number. Then the equation $x^n = z$ has n distinct complex solutions.*

Proof. First suppose that $z = 1$. Then any solution of the equation is of absolute value 1: indeed, $1 = |z^n| = |z|^n$ and so $|z|$ itself is a real positive solution. The only such is 1 (since $\mathbb{R}$ is an ordered field).

Thus, every solution is of the form $e^{\alpha i}$. Obviously, the complex numbers

$$e^{\frac{2\pi k i}{n}} \qquad k = 0, 1, 2, \ldots, n-1$$

are n solutions of the equation. Also, they are distinct since if $e^{2\pi k i/n} = e^{2\pi \ell i/n}$, then $e^{2\pi(\ell-k)i/n} = 1$, which forces $(\ell - k)/n \in \mathbb{Z}$. But since $0 \leq k, \ell < n$, this means $\ell - k = 0$ (which uses the fact that if n divides a positive integer a, then $a > n$).

It is also clear that these are the only solutions: Any polynomial of degree n has at most n distinct roots (which follows from the prime factorization). Here one can also see this directly:

If x is any solution, then $x = e^{i\alpha}$ and so $e^{in\alpha} = 1$ which forces $n\alpha$ to be a multiple of 2π. We listed all such α in the interval $[0, 2\pi)$.

Finallt, if z is arbitrary, then $z = rz_0$ where $r = |z| > 0$ is real and z_0 has absolute value 1: $z_0 = z/|z|$. There exists a unique positive real number s with $s^n = r$ (Intermediate Value Theorem). Also $z_0 = e^{i\alpha}$, and so $x_0 = e^{i\alpha/n}$ is a solution of $x_0^n = z_0$. Together sx_0 is a solution of $x^n = z$, and if $x_1, x_2 \ldots, x_n$ are the distinct solutions of $x^n = 1$, we obtain the n distinct solutions

$$x_0x_1, x_0x_2, \ldots, x_0x_n$$

for $x^n = z$. There cannot be more solutions (again, prime factorization of $x^n - z$, or: if x is any solution, then x/x_0 is a solution of $x^n = 1$). □

We won't use this fact a lot but it should be mentioned in this context that

$$\cos\alpha = \operatorname{Re} e^{i\alpha} = \frac{1}{2}(e^{i\alpha} + e^{-i\alpha}) \sin\alpha \quad = \operatorname{Im} e^{i\alpha} = \frac{1}{2i}(e^{i\alpha} - e^{-i\alpha})$$

The usually painful trigonometric identities are now very elementary: For instance

$$\text{(6.4)} \quad \begin{aligned}\cos(\alpha+\beta) + \sin(\alpha+\beta)i &= e^{(\alpha+\beta)i} = e^{\alpha i}e^{\beta i} \\ &= (\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)) + (\cos(\alpha)\sin(\beta) + \sin(\alpha)\cos(\beta))i.\end{aligned}$$

Comparing real and imaginary parts on both sides then gives the usual identities for adding angles.

Similarly, De Moivre's Theorem is now now immediate: it states that $(\cos(\alpha) + \sin(\alpha)i)^n = \cos(n\alpha) + \sin(n\alpha)i$.

Why are we interested in solving the equation $x^2 + 1 = 0$? Taken by itself this is – while interesting – not exactly fascinating. The fascinating piece is the following remarkable fact: we can now solve *every* polynomial equation[16].

6.32 Theorem (Fundamental Theorem of Algebra; C.F. Gauss)**.** *Let $n \geq 1$ and $c_0, c_1, \ldots, c_n$ be complex numbers with $c_n \neq 0$. Then the equation*

$$c_nz^n + c_{n-1}z^{n-1} + \cdots + c_1z + c_0 = 0$$

has a solution $z \in \mathbb{C}$.

This is remarkable: even though we extended our number system ony slightly by making a single equation solvable, it turns out that we now can solve all equations in principle. Compare this with the transition from $\mathbb{Q}$ to $\mathbb{R}$: adding $\sqrt{2}$ to $\mathbb{Q}$ does not help to solve the equation $x^2 = 3$ for example.

The Fundamental Theorem of Algebra is the main reason why we are interested in the complex numbers. It is the natural place to do algebra in, even if we are mainly interested

[16] This is a little bit misleading. We cannot solve polynomial equations of degree five or higher by a formula; but solutions do exist.

in the real numbers. Unfortunately, there is no known proof of the Fundamental Theorem of Algebra that only uses algebra (in that sense the name is kind of a misnomer): all known proof require more or less calculus. The minimum requirements are a thorough understanding of limits and the intermediate value theorem. For this reason we will not discuss a proof here.

A final remark, which is important for theoretical purposes: if $\mathbb{F}$ is any field, one can always find a larger field $\overline{\mathbb{F}}$, called an *algebraic closure* of $\mathbb{F}$, such that $\mathbb{F} \subseteq \overline{\mathbb{F}}$ and such that in $\overline{\mathbb{F}}$ every polynomial equation with coefficients in $\overline{\mathbb{F}}$ has a solution, and finally no subfield of $\overline{\mathbb{F}}$ that contains $\mathbb{F}$ has this root property. Furthermore, one can show that such a closure is essentially unique. Thus, we can reformulate the Fundamental Theorem as $\overline{\mathbb{R}} = \mathbb{C}$. The general algebraic theory provides the inclusion $\mathbb{C} \subseteq \overline{\mathbb{R}}$. The crucial part, namely the reverse inclusion, can so far be proved only by transcendental means (i.e. using calculus).

6.5.1. Real and complex polynomials

As an application of the concepts above let us focus on polynomials with real and complex coefficients.

The fundamental Theorem of Algebra 6.32 has the immediate consequence:

6.33 Theorem. *The irreducible polynomials in $\mathbb{C}[x]$ are precisely the linear ones* (*the ones of degree* 1)*. Every polynomial f of degree $n > 0$ can be written uniquely in the form*

$$f = c(x - \xi_1)(x - \xi_2) \cdots (x - \xi_n)$$

where $\xi_1, \cdots, \xi_n$ are the roots of f with multiplicities and c is a nonzero constant.

Proof. It is clear that any polynomial of degree 1 is irreducible (this holds for polynomials over an arbitrary field; what would be the factors?).

Conversely, let f be irreducible. By the Fundamental Theorem of Algebra, f has a root ξ in $\mathbb{C}$. Thus, we can write

$$f = (x - \xi)Q$$

for some $Q \in \mathbb{C}[x]$. But as f is irreducible, and since $x - \xi$ is not a unit, this means Q is a unit, and hence constant.

The remaining assertion then follows from the Unique Factorization Theorem. □

As mentioned several times now, knowing that f has factors of degree 1 is not the same as actually finding these factors. A famous result due to Abel (and Ruffini) states that for polynomial of degrees 5 or higher there is no symbolic formula (involving only n-th roots and addition and multiplication, much like the quadratic formula), that produces a root. But of course there are numerical methods that produce roots to arbitrary exactness (and in fact even square roots of many rational numbers are irrational and hence only "computable" up to a certain degree of exactness).

6.5.2 Problem. Show that if $f \in \mathbb{R}[x]$ is a polynomial and if $z \in \mathbb{C}$ is a root of f, then so is $\overline{z}$.

Conclude that up to a constant nonzero factor the irreducible polynomials in $\mathbb{R}[x]$ are precisely of the form $x - a$ $(a \in \mathbb{R})$ and $x + bx + c$ with $b^2 - 4c < 0$.

6.34 Corollary. *Let $f \in \mathbb{R}[x]$ be a polynomial of odd degree. Then f has a root in $\mathbb{R}$.*

Proof. This follows from the problem: if $f = p_1 p_2 \cdots p_t$ is the factorization of f into irreducibles, then not all p_i can have degree 2 since $\deg f$ is odd. Thus, one of the p_i is linear. □

Remark. It is possible to deduce the fundamental theorem from this corollary, together with the statement that every positive real number has a square root. Also, this deduction is completely algebraic (using what is called Galois Theory). However, to show that square roots exist and to show that every polynomial of odd degree has a root, one needs calculus (in fact the Intermediate Value Theorem).

6.5.2. The field of rational functions.

In the homework assigments, we discovered a way to construct the rational numbers as equivalence classes of pairs of integers (a, b) with $b \neq 0$. The same approach works with polynomials over a field $\mathbb{F}$. To be precise consider the set F of all pairs (f, g) where $f, g \in \mathbb{F}[x]$ and $g \neq 0$. We write

$$(f, g) \sim (h, k) \quad \text{if } fk = gh.$$

It is not hard to see that $\sim$ is an equivalence relation and we write f/g for the equivalence class of (f, g). f/g is called the *fraction* of f and g. Recall from our earlier discussion of equivalence relations that $f/g = h/k$ if and only if $fk = gh$. We now define

$$\begin{aligned} f/g + h/k &= (fk + gh)/(gk) \quad \text{and} \\ (f/g) \cdot (h/k) &= (fh)/(gs). \end{aligned}$$

These definitions are independent of the actual chosen representatives for the fractions. If we denote the set of all fractions by $\mathbb{F}(x)$, then $\mathbb{F}(x)$ together with this addition and multiplication is easily seen to be a field: the verification of the ring axioms is straight forward, and $1_{\mathbb{F}(x)} = 1_{\mathbb{F}[x]}/1_{\mathbb{F}[x]}$ is a multiplicative identity which is not equal to $0 = 0/1$.

In fact, we may "embed" $\mathbb{F}[x]$ into $\mathbb{F}(x)$ by identifying $f \in \mathbb{F}[x]$ with the fraction $f/1$. Then it is a simple verification that $(f/1) + (g/1) = (f + g)/1$ and $(f/1)(g/1) = (fg)/1$ and moreover $\mathbb{F}[x] \to \mathbb{F}(x)$, mapping f to $f/1$ is injective.

Summarizing, it follows that $\mathbb{F}(x)$ is a field containing $\mathbb{F}[x]$: indeed, if $f/g \neq 0$ in $\mathbb{F}(x)$ then $f \neq 0$ and so $(f/g)^{-1} = g/f$.

Also note, under the identification $f \leftrightarrow f/1$ we have that $f/g = fg^{-1}$ in $\mathbb{F}(x)$.

$\mathbb{F}(x)$ is called the *field of fractions* of $\mathbb{F}[x]$, or also *quotient field*. $\mathbb{F}(x)$ is also often called (by abuse of language) the *field of rational functions in one variable*.

We leave the details here to the reader. The construction is completely analogous to the construction of rational numbers. For us the only important point is that $\mathbb{F}[x]$ is a subring of a field, which will be important in the next chapter: we then can talk about matrices with entries in $\mathbb{F}[x]$ as elements of $M_{m\times n}(\mathbb{F}(x))$. In case $\mathbb{F} = \mathbb{R}$, the field $\mathbb{R}(x)$ corresponds to the field of rational functions on $\mathbb{R}$ much in the same way that $\mathbb{R}[x]$ corresponds to the ring of polynomial functions on $\mathbb{R}$.

7. A single linear transformation

In the previous chapters we developed the general methods and concepts that we will now apply to study a single linear transformation of a vector space V. The high point will be the Jordan decomposition theorem and, as a consequence, the Jordan canonical form.

Convention. All our vector spaces in this chapter will always be finite dimensional unless explicitly stated otherwise.

7.1. Linear transformations and change of basis

Let V, W be finite dimensional $\mathbb{F}$-vector spaces with bases $\mathcal{B}$ and $\mathcal{C}$ respectively. In Section 4.3 we associated to each linear transformation $T\colon V \to W$ a matrix $A = M_{\mathcal{B}}^{\mathcal{C}}(T)$ such that

$$[T(v)]_{\mathcal{C}} = A[v]_{\mathcal{B}}.$$

Now we want to understand how the matrix A changes if we change the bases $\mathcal{B}$ and $\mathcal{C}$. For this it is very convenient to introduce some new notation. Recall, that if $\mathcal{B}$ has n elements, and if $X \in \mathbb{F}^n$, we write $\mathcal{B}X$ for the (unique) vector $v \in V$ for which $[v]_{\mathcal{B}} = X$. We generalize this now as follows: Let $L = (v_1, v_2, \dots, v_k)$ be a list of vectors in V and let $X \in \mathbb{F}^k$ with entries $x_1, x_2, \dots, x_k$. We write

$$LX := x_1 v_1 + x_2 v_2 + \cdots + x_n v_n.$$

Up to the order of $x_1 v_1$ (rather than $v_1 x_1$) this should remind you of the product of a row with a column vector, with L being the row, and X the column. But of course, L is a "row vector of vectors." If now A is any $k \times n$ matrix with entries in $\mathbb{F}$, we define

$$LA = (v_1, v_2, \dots, v_k)A := (w_1, w_2, \dots, w_n)$$

where $w_j = LA_j$ (A_i is the ith column of A). If $A = [a_{ij}]$, this means

$$w_j = \sum_{i=1}^{k} a_{ij} v_i$$

which again should remind you of the product of a row vector with a matrix (up to the order of the terms). The important straight forward fact here is that if B is any $n \times p$-matrix then

$$(LA)B = L(AB). \tag{7.1}$$

Change of basis matrix: Now suppose instead of the basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ of V, we choose a basis $\mathcal{B}' = (v_1', v_2', \ldots, v_n')$. With the above notation, there exists a *unique* $P \in M_n(\mathbb{F})$ that

$$\mathcal{B}' = \mathcal{B}P.$$

P is called the *change of basis matrix*, and it is determined by any one of the requirements:

- $\mathcal{B}' = \mathcal{B}P$.
- The jth column P_j of P is equal to

  $$P_j = [v_j']_{\mathcal{B}},$$

  the coordinate vector of the jth element of *the new basis with respect to the old basis*.
- $P = [p_{ij}]$ where the p_{ij} satisfy

  $$v_j' = \sum_{i=1}^{n} p_{ij} v_i.$$

How does a change of basis affect coordinate vectors? Let $v \in V$. Suppose $X' = [v]_{\mathcal{B}'}$ is the "new" coordinate vector. Note that X' is determined by $v = \mathcal{B}'X'$. Substituting $\mathcal{B}P$ and using (7.1), we conclude that

$$v = \mathcal{B}'X' = (\mathcal{B}P)X' = \mathcal{B}(PX').$$

Thus, $PX' = [v]_{\mathcal{B}}$ is the *old* coordinate vector of v. Let us denote it by X. Then P is also determined by the requirement

> The *old* coordinate vector X is equal to $X = PX'$ where X' is the *new* coordinate vector.

Unfortunately, if we know P (and hence $\mathcal{B}'$), we also need P^{-1} to "translate" from old coordinates into new coordinates: $X' = P^{-1}X$. P is indeed invertible: P^{-1} is the matrix of the change of basis from $\mathcal{B}'$ to $\mathcal{B}$: $\mathcal{B} = \mathcal{B}'P^{-1}$ (indeed, suppose $\mathcal{B} = \mathcal{B}'Q$, then $\mathcal{B} = \mathcal{B}'Q = \mathcal{B}PQ$ and this easily implies that $PQ = I$.)

Note that there is no standing convention, whether P or P^{-1} is called the change of basis matrix. In fact because the coordinates transform like $X' = P^{-1}X$, many people (including myself) use $Q = P^{-1}$ (and then "pay" for it by having to express the *old* basis in terms of the new: $\mathcal{B} = \mathcal{B}'Q$). It does not matter, as long as we are consistent. We adopt the convention of the textbook.

7.1 Example. Let $V = \{f \in \mathcal{P}(\mathbb{R}) \mid \deg f \leq n\}$ be the space of polynomial functions of degree at most n. We know that V is a vector space with basis $\mathcal{B}' = (1, x, \ldots, x^n)$ (Problem 3.3.4 and the example preceding it).

Let $\mathcal{B}$ be the basis constructed in said problem, namely given by $f_0, f_1, \ldots, f_n$ such that $f_i(j) = \delta_{ij}$. Thinking of $\mathcal{B}$ as the old, and $\mathcal{B}'$ as the new basis, we find that

$$\mathcal{B}' = \mathcal{B}P$$

with change of basis matrix P given by

$$P = [\,[1]_{\mathcal{B}}\,[x]_{\mathcal{B}} \ldots [x^n]_{\mathcal{B}}\,] = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 2^2 & \ldots & 2^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & n & n^2 & \ldots & n^n \end{bmatrix}$$

Unfortunately, the transformation of coordinate vectors is not as straight forward, since it requires the computation of P^{-1} (which means expressing the *old* basis in terms of the new one). Here this is not so bad, as it simply means expanding the formulas of the f_i and computing coefficients.

7.2 Example. Now let $V = \mathbb{F}^n$. Let $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ be a basis. Notice that since the v_i are column vectors we can form a matrix $[\mathcal{B}] := [\,v_1\,v_2\,\ldots\,v_n\,]$ with columns v_i. Recall that if $\mathcal{E}$ is the standard basis then $v_i = [v_i]_{\mathcal{E}}$ and hence $[\mathcal{B}]$ is the matrix of the change of basis from $\mathcal{E}$ to $\mathcal{B}$. Consequently, for each $v \in \mathbb{F}^n$, we have

$$[v]_{\mathcal{B}} = [\mathcal{B}]^{-1}v.$$

This shows that if we change the basis from $\mathcal{E}$ to $\mathcal{B}$, we have to compute $[\mathcal{B}]^{-1}$.

Now let $\mathcal{B}'$ be another basis of V. Then $\mathcal{B}' = \mathcal{B}P$. Note that this is equivalent to saying that

$$[\mathcal{B}'] = [\mathcal{B}]P$$

which now is an honest product of matrices. It now follows that

$$P = [\mathcal{B}]^{-1}[\mathcal{B}']$$

a formula that looks more interesting than it actually is.

So if eg. $V = \mathbb{R}^2$, and $v_1 = (1, 1)$, $v_2 = (-1, 1)$, and $\mathcal{B}' = (v_1, v_2)$, with $\mathcal{B} = \mathcal{E}$ the standard basis. Then

$$[\mathcal{B}'] = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

and the change of basis matrix for the change from $\mathcal{E}$ to $\mathcal{B}'$ is $P = [\mathcal{E}]^{-1}[\mathcal{B}'] = I_2^{-1}[\mathcal{B}'] = [\mathcal{B}']$. As for coordinates,

$$P^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

Then $[v]_{\mathcal{B}} = P^{-1}[v]_{\mathcal{E}} = P^{-1}v$.

To make matters worse, let us now consider arbitrary finite dimensional vector spaces V, W with bases $\mathcal{B}$ and $\mathcal{C}$ respectively and suppose $T\colon V \to W$ is a linear transformation with matrix $A = M_{\mathcal{B}}^{\mathcal{C}}(T)$ and switch bases from $\mathcal{B}$ to $\mathcal{B}' = \mathcal{B}P$ and $\mathcal{C}$ to $\mathcal{C}' = \mathcal{C}Q$. How does this affect the matrix? Let $A' = M_{\mathcal{B}'}^{\mathcal{C}'}(T)$. To simplify notation let us denote the coordinate vectors of v by X and X' (with respect to $\mathcal{B}$ and $\mathcal{B}'$, respectively), and the ones for $T(v)$ by Y and Y' (with respect to $\mathcal{C}$ and $\mathcal{C}'$ respectively).

Then recall that

$$Y = AX$$

and

$$Y' = A'X'.$$

Note also that $X = PX'$ and $Y = QY'$ from which we conclude that

$$Y' = Q^{-1}Y = Q^{-1}AX = (Q^{-1}AP)X'$$

and since this is true for all $X' \in \mathbb{F}^n$ and this determines A' it follows that

$$A' = Q^{-1}AP.$$

This has the following very powerful consequence for matrices:

7.3 Theorem. *Let $A \in M_{m\times n}(\mathbb{F})$ be a matrix of rank k. Then there is $Q \in \mathrm{GL}_m(\mathbb{F})$ and $P \in \mathrm{GL}_n(\mathbb{F})$ such that*

$$Q^{-1}AP = \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}$$

By the above the theorem is equivalent to the assertion of the following theorem, which one could view as a vector space version of the above.

7.4 Theorem. *Let $T\colon V \to W$ be a linear transformation between finite dimensional vector spaces. Then there exist a basis for V and a basis for W such that the corresponding matrix A of T has the form*

$$\begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix}$$

where $k = \dim \mathrm{im}(T)$.

Proof. This theorem was proved in a homework assignment. Let $(w_1, w_2, \dots, w_k)$ be a basis for $\mathrm{im}(T)$, and extend it to a basis $\mathcal{C} = (w_1, w_2, \dots, w_m)$ for all of W. Now for $i \leq k$, $w_i = T(v_i)$ for some $v_i \in V$. Let $v_{k+1}, v_{k+2}, \dots, v_n$ be basis for $\mathcal{N}(T)$. Then $(v_1, v_2, \dots, v_n)$ is a basis for V: it is linearly independent, Indeed, let $c_1v_1 + \cdots + c_nv_n = 0$. Then applying T, we get $c_1w_1 + c_2w_2 + \cdots + c_kw_k = 0$ and hence $c_1 = c_2 = \cdots = c_k = 0$. But then $c_{k+1}v_{k+1} + \cdots + c_nv_n = 0$ and so $c_{k+1} = c_{k+2} = \cdots = c_n = 0$ as well, since $(v_{k+1}, \dots, v_n)$ is a basis for $\mathcal{N}(T)$. If $v \in V$, then $T(v) = c_1w_1 + c_2w_2 + \cdots + c_kw_k$ by the definition of $(w_1, w_2, \dots, w_k)$, and so $v - c_1v_1 - c_2v_2 - \cdots - c_kv_k \in \mathcal{N}(T) = \mathrm{Span}(v_{k+1}, v_{k+2}, \dots, v_n)$. So $v \in \mathrm{Span}(v_1, v_2, \dots, v_n)$ and $\mathcal{B} = (v_1, v_2, \dots, v_n)$ is a basis. That the matrix $M_{\mathcal{B}}^{\mathcal{C}}(T)$ has the desired form follows directly from the way we constructed our bases. □

Applying this theorem to the matrix transformation $T_A \colon \mathbb{F}^n \to \mathbb{F}^m$ clearly shows Theorem 7.3.

Remark. These theorems essentially say that for the qualitative study of a linear transformation $T \colon V \to W$, the *rank* of T (or its matrix) encodes all information about T_A (that is, the dimension of its null-space, image), and up to a suitable choice of basis all linear transformations of the same rank look alike.

7.1.1 Problem. Prove the matrix version directly. (Think about row and column operations.)

7.2. Linear operators – eigenvalues and eigenvectors

Instead of discussing a general linear transformation $T \colon V \to W$ we now focus on the case where $V = W$: Given a finite-dimensional vector space V, a linear transformation $T \colon V \to V$ is called an *endomorphism* or *linear operator* (of V). We will denote by $\mathrm{End}(V)$ the set of all linear operators of V (which we also denoted by $\mathrm{Hom}(V, V)$ earlier). Recall that $\mathrm{End}(V)$ is a ring, a property that will be important soon.

Why is this situation different than what we discussed before? The main reason is the following: if we pick a basis $\mathcal{B}$ for V then we also have picked the basis $\mathcal{C}$ (of W): of course we don't want to pick two bases $\mathcal{B}$ and $\mathcal{C}$ of V and describe v with respect to $\mathcal{B}$ and $T(v)$ with respect to $\mathcal{C}$. If that was our intention then it is best not to think of the domain and codomain of T as equal but rather as some vector spaces of the same dimension.

Thus, if the domain and codomain of T are equal, we loose some freedom in terms of the change of basis: given a single basis $\mathcal{B} = (v_1, v_2, \dots, v_n)$ of V, the matrix of T with respect to $\mathcal{B}$ is defined as

$$A = [\,T(v_1)_{\mathcal{B}}\, T(v_2)_{\mathcal{B}}\, \dots\, T(v_n)_{\mathcal{B}}\,] \tag{7.2}$$

In other words, if X is the coordinate vector of v with respect to $\mathcal{B}$, then AX is the one of $T(v)$. If we change the basis from $\mathcal{B}$ to $\mathcal{B}' = \mathcal{B}P$, then A changes to

$$A' = P^{-1}AP \tag{7.3}$$

(since $\mathcal{B} = \mathcal{C}$ and $\mathcal{B}' = \mathcal{C}'$ in our usual formulas, also $P = Q$).

This motivates the following definition.

7.5 Definition. Let $A, B \in M_n(\mathbb{F})$ be two matrices. Then A and B are said to be *similar* (to each other) if there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that $B = P^{-1}AP$.

So A and B are similar, if they describe the same linear operator (on $\mathbb{F}^n$) with respect to different bases.

Remark. Notice that if $B = P^{-1}AP$ then $A = PBP^{-1}$ and so $A = Q^{-1}BQ$ for $Q = P^{-1}$. In that regard, similarity is a symmetric notion.

7.2.1 Problem. Show that similarity is an equivalence relation.

Similar matrices are often also called *conjugated* by the element P^{-1}.

The guiding question is now

> What is the simplest form a matrix A can be turned into by conjugation?

Since we force $\mathcal{B} = \mathcal{C}$ we lose the consequences of Theorem 7.3. A related question is: how can we understand a given linear operator T? Can we choose a basis such that its matrix becomes "simple" in a way?

7.6 Example. Consider the a matrix operator $T_D \colon \mathbb{R}^n \to \mathbb{R}^n$ where D is a diagonal matrix. Every possible question we might possibly be interested in regarding the operator T_D is immediately answerable.

A problem coming up frequently is the following: Assuming the vectors $X \in \mathbb{F}^n$ describe the possible states of some system. We know that the system is exposed to a dynamical process that changes the state X_m at time m to $X_{m+1} = DX_m$ at time $m+1$. So starting at a ground state X_0 we have $X_m = D^m X_0$.

What is the long term devopment of the state X_n? Does it converge to a *stable* state X_∞? Since D is diagonal this question can be readily answered:

$$X_m = \begin{bmatrix} d_1^m x_1 \\ d_2^m x_2 \\ \vdots \\ d_n^m x_n \end{bmatrix} \tag{7.4}$$

So eg. if $x_i \neq 0$ and $|d_i| > 1$ the system never converges.

Unfortunately, in most situations the matrix of such a dynamical process is not diagonal a priori. However, by finding a suitable basis, maybe it can be turned into a diagonal matrix: indeed, if $A = PDP^{-1}$ for a diagonal matrix A, then eg. $A^n = PD^nP^{-1}$, so up to a coordinate transformation the above analysis still goes through.

This motivates the following definition:

7.7 Definition. Let $A \in M_n(\mathbb{F})$. Then A is called *diagonalizable* if A is similar to a diagonal matrix (ie. if $P^{-1}AP$ is diagonal for some $P \in \mathrm{GL}_n(\mathbb{F})$).

Similarly, a linear operator $T \in \mathrm{End}(V)$ is called *diagonalizable* if with respect to some basis $\mathcal{B}$, its matrix $M_\mathcal{B}(T)$ is diagonal.

It will turn out that it is very useful to work with linear transformations rather than matrices since the linear transformation is always the same (no matter what basis we choose) whereas the matrix changes.

Many properties of matrices are in fact the same for similar matrices and then translate into properties of the respresented linear operator: some properties of a matrix A are actually properties of T (ie. independent of the choice of a basis). Here are two examples:

7.8 Definition. Let $T \in \mathrm{End}(V)$ be a linear operator with matrix A with respect to some basis $\mathcal{B}$. Then the *determinant* of T, denoted $\det T$, is defined as $\det T = \det A$.

Similarly, the *trace* of T is $\mathrm{trace}(T) = \mathrm{trace}(A)$.

Note that the change of basis formula (7.3) has the immediate consequence that this definition is independent of the particular choice of $\mathcal{B}$: If we picked another basis $\mathcal{B}'$ say, then the matrix of T would be $A' = P^{-1}AP$ where P is the change of basis matrix and hence

$$\det A' = \det(P^{-1}AP) = \det(P^{-1})\det(A)\det(P) = \det(P^{-1}P)\det(A) = \det(A).$$

Also, we have seen (in some homework assignment) that $\mathrm{trace}(A) = \mathrm{trace}(A')$.

As mentioned above, diagonal matrices are simple to handle. The existence of a basis $\mathcal{B}$ that the matrix of T becomes diagonal seems like a rather strong requirement (even though, over the complex numbers, *almost every* matrix is diagonalizable: the set of non-diagonalizable matrices is a set of measure zero). Assume that such a basis exists; then if v is a member of the basis, this means that $T(v) = \lambda v$ for some $\lambda \in \mathbb{F}$. Hence, if we want such a basis, we need plenty (n linearly independent ones) of vectors that satisfy such an equation. This motivates the following definition:

7.9 Definition. Let $T \in \mathrm{End}(V)$ be a linear operator. An *eigenvector* of T is a *nonzero* vector $v \in V$ such that $T(v) = \lambda v$ for some $\lambda \in \mathbb{F}$.

An *eigenvalue* (or *characteristic value* or *proper value*) of T is a scalar $\lambda \in \mathbb{F}$ such that $T(v) = \lambda v$ for at least one nonzero $v \in V$. Such a v is called an eigenvector *for* or *belonging to* λ.

Let $A \in M_n(\mathbb{F})$. An *eigenvalue* of A is an element $\lambda \in \mathbb{F}$ such that $AX = \lambda X$ has at least one nonzero solution $X \in \mathbb{F}^n$. Such an X is then called an *eigenvector* of A belonging to λ.

7.10 Remark. Note that eigenvectors are rarely unique. For instance if v is one, then so is cv for any $c \neq 0$ in $\mathbb{F}$. So unless $\mathbb{F} = \mathbb{F}_2$, this already gives several eigenvectors provided there is one. A more interesting non-uniqueness however is the fact that they may be linearly independent, too: Consider $\mathbf{1}$, the identity transformation of V. Then *every* nonzero vector $v \in V$ is an eigenvector for the eigenvalue 1. Note that 1 is the only eigenvector of $\mathbf{1}$.

A fancy way of saying that T is diagonalizable is therefore: T is diagonalizable if and only if there exists a basis consisting of eigenvectors.

In other words, an eigenvalue (resp. eigenvector) of A is nothing but an eigenvalue (resp. eigenvector) of T_A. Also, if $T \in \mathrm{End}(V)$ is any linear operator, and V has basis $\mathcal{B}$, then $v \in V$ is an eigenvector of T belonging to some eigenvalue λ if and only if $[v]_{\mathcal{B}}$ is an eigenvector of $M_{\mathcal{B}}(T)$ belonging to the same λ. In particular, the eigenvalues of T and its matrix with respect to any basis *are the same*.

7.11 Example. Let

$$R = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

be the matrix of a rotation on $\mathbb{R}^2$. Then unless R corresponds to rotations around 0 or π, R has no real eigenvalues and hence no real eigenvectors: indeed, no line through the origin is mapped by T_A to itself. So there cannot be any eigenvectors.

If we think of the same matrix R as an element of $M_2(\mathbb{C})$, then R has two eigenvalues, namely $e^{\alpha i}$ and $e^{-\alpha i}$, which again are distinct if α is not an integer multiple of π. It is easy to check that

$$\begin{bmatrix} 1 \\ -i \end{bmatrix}, \begin{bmatrix} 1 \\ i \end{bmatrix}$$

are linearly independent eigenvectors, so R is diagonalizable as a complex matrix.

What we learn from this is that the notion of eigenvectors and eigenvalues depends on the field: if we enlarge our number system, we may gain additional eigenvalues and eigenvectors.

Example. Consider the reflection S of $\mathbb{R}^2$ in a line L through the orgin. Then any vector $v \in L$ is fixed by S: $S(v) = v$, so any nnonzero vector $v \in L$ is an eigenvector of eigenvalue 1. On the other hand, if w is a nonzero element of the line through 0 which is orthogonal to L, then $S(w) = -w$ and so w is an eigenvector of eigenvalue -1. Such a reflection is therefore diagonalizable (as it has linearly independent eigenvectors, and hence a basis consisting of eigenvectors.

Given $T \in \mathrm{End}(V)$, with matrix A with respect to some basis $\mathcal{B}$, then $\lambda \in \mathbb{F}$ is an eigenvalue of A (and hence T) if and only if the matrix equation

$$(\lambda I - A)X = 0$$

has a nonzero solution (and any such solution will then be an eigenvector). We know, that such a solution exists if and only if $\lambda I - A$ is not an invertible matrix (cf. Theorem 2.20). By Corollary 5.9 this in turn is equivalent to saying that

$$\det(\lambda I - A) = 0. \tag{7.5}$$

Summarizing, we have shown:

7.12 Proposition. *Let $T \in \mathrm{End}(V)$ be a linear operator. Let $\lambda \in \mathbb{F}$. Then λ is an eigenvalue of T if and only if $\det(\lambda\mathbf{1} - T) = 0$.*

Proof. All we need to remark here is that if A is the matrix of T with respect to some basis $\mathcal{B}$, $\lambda I - A$ is the matrix of $\lambda\mathbf{1} - T$. □

Remark. One has to be a little careful here: as a function of λ, $f(\lambda) := \det(\lambda I - A)$ is clearly a polynomial function (indeed, the formula for the determinant is polynomial in the entries). It is tempting to think of f then as a polynomial, which is OK for large fields. However, for fields with n or less elements ($n = \dim V$), there is not a unique polynomial in $\mathbb{F}[x]$ of degree $\leq n$ with associated function f. We will remedy this situation shortly. We therefore rather want a unique polynomial associated to T than just a polynomial function.

Example. Let

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \in M_2(\mathbb{C})$$

which of course is the matrBecause of the close connection between matrices and linear operators we can adapt Definition 7.9 to matrices: ix of a rotation by 90 degrees (when viewed as a real matrix).

Now

$$A - \lambda I = \begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix}$$

and so $\det(A - \lambda I) = \lambda^2 + 1$ and we have the two eigenvalues $\lambda = i$ and $\mu = -i$.

To find the corresponding eigenvectors, we now have to solve the two linear systems

$$\begin{bmatrix} -i & -1 \\ 1 & -i \end{bmatrix} X = 0$$

and so

$$X = c \begin{bmatrix} i \\ 1 \end{bmatrix}$$

For $\mu = -i$, we need to solve

$$\begin{bmatrix} i & -1 \\ 1 & i \end{bmatrix} Y = 0.$$

and so

$$Y = c \begin{bmatrix} -i \\ 1 \end{bmatrix}$$

Any nonzero solution X will be an eigenvector for i and any nonzero Y will be one for $-i$.

Note also that if X, Y are two such vectors, then (X, Y) is a basis because they cannot be linearly dependent (why?), and T_A has matrix

$$\begin{bmatrix} i & \\ & -i \end{bmatrix}$$

with respect to this basis. Indeed, if $AX = iX$ and $AY = -iY$, $X, Y \neq 0$, then X cannot be a multiple of Y, since for any multiple of Y, cY, say we have $A(cY) = c(AY) = -ciY = (-i)(cY)$. So cY is never an eigenvector for i.

Example. If A is an $n \times n$-matrix that is upper triangular, then its top-left entry is an eigenvalue and e_1 is an eigenvector.

More generally, if the i-th column A_i of an $n \times n$ matrix A is of the form $A_i = \lambda e_i$ (e_i as usual being the ith column if I_n), then e_i is an eigenvector of A belonging to λ. Indeed, the ith column of a matrix is given by Ae_i, so $\lambda e_i = Ae_i$.

Finally, if A is upper triangular, then the entries of A on the diagonal are its eigenvalues: indeed, if $\lambda_1, \lambda_2, \dots, \lambda_n$ are the diagonal entries of A, then

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)$$

if

$$A = \begin{bmatrix} \lambda_1 & * & * & * \\ & \lambda_2 & * & * \\ & & \ddots & * \\ & & & \lambda_n \end{bmatrix}$$

We have seen in the example that (X, Y) were linearly independent. This is true in general:

7.13 Proposition. *Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be distinct eigenvalues of a linear operator $T \in \mathrm{End}(V)$ and let $v_1, v_2, \ldots, v_k$ be eigenvectors such that v_i belongs to λ_i. Then $(v_1, v_2, \ldots, v_k)$ is linearly independent.*

Proof. We proceed by induction on k. If $k = 1$, the result is obvious, since $v_1 \neq 0$ is linearly independent.

So suppose any list of k eigenvectors belonging to distinct eigenvalues is linearly independent. Suppose $v_1, v_2, \ldots, v_{k+1}$ is then a list of $k + 1$ eigenvectors for distinct eigenvalues. Let $c_1, c_2, \ldots, c_{k+1} \in \mathbb{F}$ such that $c_1v_1 + c_2v_2 + \cdots + c_{k+1}v_{k+1} = 0$.

If not all $c_i = 0$, there must be one, which is not zero (in fact *all* must be then nonzero by the induction hypothesis). After renumbering we can assume that this is c_{k+1}. Then

$$c_{k+1}v_{k+1} = -(c_1v_1 + \cdots + c_kv_k). \tag{7.6}$$

Applying T to both sides, we get

$$c_{k+1}\lambda_{k+1}v_{k+1} = -(c_1\lambda_1v_1 + \cdots + c_k\lambda_kv_k).$$

Now note that the vectors on the right hand side are linearly independent. By multiplying Equation (7.6) through with λ_{k+1} we have

$$c_{k+1}\lambda_{k+1}v_{k+1} = -(c_1\lambda_{k+1}v_1 + \cdots + c_k\lambda_{k+1}v_k).$$

By linear independence of the right hand side vectors we get for $i = 1, \ldots, k$, that

$$c_i\lambda_{k+1} = c_i\lambda_i$$

or

$$c_i(\lambda_{k+1} - \lambda_i) = 0$$

forcing $c_i = 0$ for $0 \leq i \leq k$ since $\lambda_i \neq \lambda_{k+1}$. This is impossible since at least one $c_i \neq 0$ (because $c_{k+1}v_{k+1} \neq 0$). □

7.14 Corollary. *Any $T \in \mathrm{End}(V)$ has at most $\dim V$ distinct eigenvalues.*

Moreover, if V has a basis consisting of eigenvectors, then any *eigenvalue λ of T has a basis element belonging to it.*

Proof. This is a good practice problem and left to the reader. □

7.15 Corollary. *If $T \in \operatorname{End}(V)$ has n distinct eigenvalues (where $\dim V = n$) then T is diagonalizable.*

Proof. Apply the proposition in case $k = n$ to conclude that any collection $(v_1, v_2, \ldots, v_n)$ of $n = \dim V$ eigenvectors belonging to these distinct eigenvalues is a basis. □

Note that the converse of this corollary is false, as the example of the identity matrix shows.

7.3. Invariant subspaces

From the previous section we gather that if $v \in V$ is an eigenvector for some linear operator T, then $\mathbb{F}v := \operatorname{Span}(v)$ is a subspace of V that is mapped to itself by T: $T(w) = \lambda w \in \operatorname{Span}(v)$ for all $w \in \operatorname{Span}(v)$.

7.16 Definition. Let $T\colon V \to V$ be a linear operator. A subspace $W \subseteq V$ is called T-*invariant* (or T-*stable*) or simply *invariant* (if T is understood), if

$$T(w) \in W \quad \text{for all } w \in W.$$

7.17 Examples.

a. For any T, the subspaces $\{0\}$ and V are invariant.

b. For any T, the subspaces $\mathcal{N}(T)$ and $\operatorname{im}(T)$ are invariant.

c. If $v \in V$ is an eigenvector, then $\operatorname{Span}(v)$ is invariant. Similarly, any subspace that is spanned by eigenvectors is invariant.

d. Generalizing the previous example, if λ is an eigenvalue of T, we put

$$E_\lambda := \mathcal{N}(T - \lambda \mathbf{1}).$$

This is a subspace consisting of all eigenvectors for λ plus the zero vector. It is invariant, since for every $v \in E_\lambda$, $T(v) = \lambda v \in E_\lambda$.

E_λ is called the *eigenspace* of T for λ.

e. As seen on an exam: If an $n \times n$ matrix A is upper triangular, then the subspaces

$$F_i = \operatorname{Span}(e_1, e_2, \ldots, e_i)$$

are T_A-stable.

7.18 Example. Let $T\colon V \to V$ be a linear operator and let $W \subseteq V$ be a subspace. Let W have basis $(w_1, w_2, \dots, w_k)$ and extend this to a basis $\mathcal{B} = (w_1, w_2, \dots, w_k, u_1, u_2, \dots, u_\ell)$ for V (which is possible by Proposition 3.46). Let M be the matrix of T with respect to this basis. Then we can write

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A, B, C, D are matrices of size $k \times k, k \times \ell, \ell \times k, \ell \times \ell$ respectively. To say that W is invariant is the very same as saying that $C = 0$:

Indeed, for any $X \in \mathbb{F}^{k+\ell}$ we can write

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

with $X_1 \in \mathbb{F}^k$ and $X_2 \in \mathbb{F}^\ell$. Then

$$MX = \begin{bmatrix} AX_1 + BX_2 \\ CX_1 + DX_2 \end{bmatrix}$$

If $w \in W$ then $X = [w]_{\mathcal{B}}$ has $X_2 = 0$. So

$$MX = \begin{bmatrix} AX_1 \\ CX_1 \end{bmatrix}$$

Of course, $T(w) \in W$ if and only if the X_2-part of $[T(w)]_{\mathcal{B}}$ is zero. Since w is arbitrary, this must hold for all $X_1 \in \mathbb{F}^k$ and this is true if only if $C = 0$.

Note that A in M only depends on the choice of the w_i whereas D only depends on the choice of the u_i.

This example is very helpful. We can say even more. Let us keep the notation of the example. Let $U = V/W$ be the quotient space. Notice that T induces a linear operator on U as follows: Consider $\pi\colon V \to U$ the linear transformation defined by $\pi(v) = \overline{v}$. Let $T' = \pi T$. Clearly $W \subseteq \mathcal{N}(T')$, so there is a unique linear transformation $\overline{T}\colon U \to U$ such that $\overline{T} \circ \pi = T'$ by Theorem 6.6. We say $\overline{T}$ is the *induced* operator on U. Check yourself that $\overline{T}$ is defined by

$$\overline{T}(v + W) = T(v) + W.$$

What can we say about $\overline{T}$? Recall that $\mathcal{C} = (\overline{u}_1, \overline{u}_2, \dots, \overline{u}_\ell)$ is a basis for U. With respect to this basis, the matrix of $\overline{T}$ is precisely D:

For this, note that if $v \in V$ has coordinate vector X split into X_1, X_2, as above, then $\pi(v)$ has coordinate vector X_2 with respect to $\mathcal{C}$: formally, using our list notation, we have

$$v = (u_1, u_2, \dots, u_\ell)X_2 + (w_1, w_2, \dots, w_k)X_1$$

and $(w_1, w_2, \dots, w_k)X_1 \in W$. Applying this to

$$\overline{T}(\overline{u}_i) = T'(u_i) = \pi(T(u_i)) = \mathcal{C}De_i,$$

because $T(u_i) = (u_1, u_2, \ldots, u_\ell)De_i + w$ with $w \in W$, and we find that $[\pi(T(u_i))]_{\mathcal{C}} = De_i$, which shows that the coordinate vector of $\overline{T}(\mathcal{C}X_2)$ is exactly DX_2. It is important that you become comfortable with this.

7.19 Example. Suppose $T = T_A \in \mathrm{End}(\mathbb{R}^2)$ where

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 3 \end{bmatrix}$$

Then clearly $W = \mathrm{Span}(e_1)$ is invariant. Let $U = \mathbb{R}^2/W$. Then

$$\overline{T}(X + W) = \overline{T}(x_2e_2 + W) = T(x_2e_2) + W = (3x_2e_2 + x_2e_1) + W = 3x_2e_2 + W$$

$\overline{T}(e_2 + W) = 3(e_2 + W)$ and $\overline{T}$ has "matrix" $[3]$ with respect to the basis $\overline{e}_2$

Sometimes we can achieve that also $B = 0$ in the matrix of T. To understand this situation, we make the following definition:

7.20 Definition. Let V be a vector space, and let $W_1, W_2, \ldots, W_k$ be subspaces. We define the *sum* of the W_i to be the subspace

$$W_1 + W_2 + \cdots + W_2 = \{w_1 + w_2 + \cdots + w_k \in V \mid w_i \in W_i \text{ for } i = 1, 2, \ldots, k\}.$$

(If we allowed infinite spanning sets we could say $W_1 + W_2 + \cdots + W_k = \mathrm{Span}(W_1 \cup W_2 \cup \cdots \cup W_k)$.)

We say V is the *direct sum* of the W_i, and write this as

$$V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$$

if

- V is equal to $W_1 + W_2 + \cdots + W_k$, and moreover,
- whenever $w_1 + w_2 + \cdots + w_k = 0$ with $w_i \in W_i$, then $w_1 = w_2 = \cdots = w_k$.

If the W_i satisfy this last condition, then we also say that the W_i are *independent*.

7.21 Examples.

a. Let $v_1, v_2, \ldots, v_k \in V$ and put $W_i = \mathrm{Span}(v_i)$. Then V is the direct sum of the W_i if and only if $V = \mathrm{Span}(v_1, v_2, \ldots, v_k)$ and the v_i are linearly independent.

b. Let $\mathbb{F} = \mathbb{R}$. Recall from some homework exercises that every matrix A in $M_n(\mathbb{R})$ can be written in a unique way as a sum $A = A_s + A_a$ where A_s is symmetric and A_a is skew-symmetric. So in particular, $0 = A_s + A_a$ means $A_s = A_a = 0$.

   If Sym_n is the space of all symmetric and Alt_n is the space of all alternating $n \times n$ matrices then

$$M_n(\mathbb{R}) = \mathrm{Sym}_n \oplus \mathrm{Alt}_n.$$

c. In Example 7.18, if we put $U = \mathrm{Span}(u_1, u_2, \dots, u_\ell)$, then

$$V = W \oplus U.$$

Recall the definition of a product space of two vector spaces. This can be easily extended to more than two spaces: if $V_1, V_2, \dots, V_k$ are vector spaces, define $V_1 \times V_2 \times \cdots \times V_k$ as the set of all $(v_1, v_2, \dots, v_k)$ with $v_i \in V_i$, together with componentwise addition and scalar multiplication.

$V_1 \times V_2 \times \cdots \times V_k$ is often called the *external direct sum* of the V_i, and some authors write $\oplus$ instead of $\times$ (and don't really distinguish between internal and external direct sums, see also the next proposition below).

The following proposition captures the most important properties of direct sums:

7.22 Proposition. *Let V be a vector space and let $V = W_1 + W_2 + \cdots + W_k$ be the sum of some subspaces. Then the following are equivalent:*

a. *V is the direct sum of these subspaces.*

b. *Every $v \in V$ can be written in a unique way as $v = w_1 + w_2 + \cdots + w_k$ with $w_i \in W$.*

c. *The map $S\colon W_1 \times W_2 \times \cdots \times W_k \to V$ defined by $S(w_1, w_2, \dots, w_k) = w_1 + w_2 + \cdots + w_k$ is an isomorphism.*

Proof. It should be clear that b. and c. are equivalent. The statement of b. is precisely saying that S is bijective, and the linearity of S is obvious.

Also, clearly b. implies a. because the statement of a. is precisely that 0 can be written in only one way as such a sum.

It remains to verify that a. implies b. But this is a standard trick: suppose

$$w_1 + w_2 + \cdots + w_k = w_1' + w_2' + \cdots + w_k'$$

with $w_i, w_i' \in W_i$ then subtracting the left from the right, and grouping by index, we find that

$$0 = (w_1' - w_1) + (w_2' - w_2) + \cdots + (w_k' - w_k).$$

Since $w_i' - w_i \in W_i$, it follows that $w_i' - w_i = 0$ if a. holds. But this means $w_i' = w_i$ for all i, what we had to show. □

See also Problems 3.3.9 and 3.3.11.

7.3.1 Problem. Let $V = V_1 \times V_2 \times \cdots \times V_p$. For each i identify a subspace V_i' of V, isomorphic to V_i, such that $V = V_1' \oplus V_2' \oplus \cdots \oplus V_p'$.

If V is the direct sum of the subspaces $W_1, W_2, \ldots, W_k$ and if $v = w_1 + w_2 + \cdots + w_k$ (with $w_i \in W_i$) we sometimes say that w_i is the *component* of v along W_i. Notice that if $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$ then the *projection* $P_i\colon V \to W_i$ which associates to each v its component w_i along W_i is a surjective linear transformation. Also, in a decomposition $V = U \oplus W$ we often say that U is the *complement* of W.

Why are we interested in direct sums? Our strategy for understanding a linear operator T will be to "decompose" V into a direct sum $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ of invariant subspaces, and moreover, do this in a way such that we perfectly understand what T does to V_i: if V_i is invariant, then the restriction of T_i to V_i determines an element of $\mathrm{Hom}(V_i)$. This opens the door to two approaches: induction on the dimension of V, because hopefully V_i is not all of V and second, to choosing the V_i in a clever way.

Remark. If T is diagonalizable, and if $(v_1, v_2, \ldots, v_n)$ is a basis consisting of eigenvectors and $V_i = \mathrm{Span}(v_i)$, then $V = V_1 \oplus V_2 \oplus \cdots \oplus V_n$ is a decomposition into invariant subspaces.

We conclude this section by observing that if we can write $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ and if V_i has basis $\mathcal{B}_i$, then V has basis $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_k$. Now this is a decomposition into invariant subspaces if and only if T has matrix of the form

$$A = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{bmatrix}$$

where A_i is an $n_i \times n_i$ matrix and $n_i = \dim V_i$, and all other entries are equal to zero. Moreover, if this is the case then A_i is the matrix of $T|_{V_i}$ with respect to $\mathcal{B}_i$. Indeed, this is a generalization of Example 7.18: in the above notation, the complement $U = \mathrm{Span}(u_1, u_2, \ldots, u_\ell)$ is invariant if and only if $B = 0$.

7.3.2 Problem. Compare this to the notion of diagonalizable.

7.4. The minimal and characteristic polynomials

Now we will employ our knowledge of polynomials. Recall that if $\dim V = n$ and $\mathcal{B}$ is a basis for V, then the map $\mathrm{End}(V) \to M_n(\mathbb{F})$ associating to T its matrix $M_{\mathcal{B}}(T)$ is an isomorphism (cf. Theorem 4.27). In particular, both spaces have the same dimension. (In fact, this map is what is called a ring isomorphism, since it is bijective and also takes products to products and $\mathbf{1}$ to I.)

Notice that our matrix units e_{ij} are the images of the linear transformations T_{ij} defined by $T_{ij}(v_k) = \delta_{kj} v_i$ (where $\mathcal{B} = (v_1, v_2, \ldots, v_n)$).

Given $T \in \mathrm{End}(V)$, with matrix $A \in M_n(\mathbb{F})$, consider the powers $\mathbf{1}, T, T^2, \cdots \in \mathrm{End}(V)$ (they have matrices $I, A, A^2, \ldots$ respectively). Since $\dim \mathrm{End}(V) = n^2$, these powers cannot all be linearly independent. In particular, any n^2+1 of them must be linearly dependent, and so there must be $c_0, c_1, \dots, c_{n^2} \in \mathbb{F}$ *not all zero* such that

$$c_0\mathbf{1} + c_1T + c_2T^2 + \cdots + c_{n^2}T^{n^2} = 0$$

and equivalently,

$$c_0I + c_1A + c_2A^2 + \cdots + c_{c^2}A^{n^2} = 0.$$

If we define $f \in \mathbb{F}[x]$ as $f = c_0 + c_1x + \cdots + c_{n^2}x^{n^2}$, we can formally write

$$f(T) = 0$$

and also

$$f(A) = 0.$$

To make this a precise notion, let $f = \sum_i a_ix^i \in \mathbb{F}[x]$ be an arbitrary polynomial and let $T \in \mathrm{End}(V)$ be any operator and let $A \in M_n(\mathbb{F})$ be any matrix.

We then put

$$f(A) = \sum_i a_iA^i = a_0I + a_1A + a_2A^2 + \cdots + a_kA^k$$

and

$$f(T) = \sum_i a_iT^i = a_0\mathbf{1} + a_1T + a_2T^2 + \cdots + a_kT^k$$

where $k \geq \deg f$.

Compare this definition with the definition of $f(\xi)$ for $\xi \in \mathbb{F}$ in the previous chapter.

7.4.1 Problem. Check the following statements, where X is either a linear transformation in $\mathrm{End}(V)$ or a matrix in $M_n(\mathbb{F})$:

a. If $f, g \in \mathbb{F}[x]$ then $(fg)(X) = f(X)g(X)$ and $(f+g)(X) = f(X) + g(X)$.

b. If $f \in \mathbb{F}[x]$ then $f(X)$ and X *commute* that is $Xf(X) = f(X)X$.

7.23 Definition. Let $T \in \mathrm{End}(V)$. A nonzero polynomial $m \in \mathbb{F}[x]$ is called the *minimal polynomial* of T, if

a. The leading coefficient of m is 1: $m = x^r + a_{r-1}x^{r-1} + \cdots + a_1x + a_0$.

b. $m(T) = 0$.

c. m has minimum degree among all polynomials satisfting b.

7.24 Theorem. *Let $T \in \mathrm{End}(V)$ be a linear operator on a finite dimensional vector space V.*

a. *The minimal polynomial of T exists and is unique.*

b. *If $f \in \mathbb{F}[x]$ is any polynomial such that $f(T) = 0$, then $m \mid f$.*

c. *If* $m = x^r + a_{r-1}x^{r-1} + \cdots + a_1 x + a_0$*, then* $\mathbf{1}, T, T^2, \ldots, T^{r-1}$ *are linearly independent in* $\mathrm{End}(V)$.

Proof. It is clear that m exists: we have shown above that there are nonzero polynomials f such that $f(T) = 0$. In fact we have shown that there is one of degree at most n^2. Thus, let f be such a polynomial of minimum degree, r, say. Of course, if $f(T) = 0$ then also $(cf)(T) = 0$ for all $c \in \mathbb{F}$, so by scaling we may achieve that the coefficient of x^r in f is 1 and then define $m = f$.

Let now g be any polynomial such that $g(T) = 0$. By the Division Algorithm we may write

$$g = Qm + R$$

with $R = 0$ or $\deg R < \deg m$. We find

$$g(T) = Q(T)m(Y) + R(T)$$

Since $m(T) = g(T) = 0$, we get

$$R(T) = g(T) - Q(T)m(Y) = 0$$

as well. Since $\deg m$ is minimal among all such polynomials, this forces $R = 0$ proving b.

To finish a. let M be another minimal polynomial. Then $M(T) = 0$ and so $m \mid M$. Since both M and m have the same degree this means that $M = cm$ for some (nonzero) $c \in \mathbb{F}$. Since also the highest coefficients in both M and m are 1, this means $c = 1$ and $m = M$.

To show c. we just observe that if $\mathbf{1}, T, \ldots, T^{r-1}$ were linearly dependent, then there would be $c_0, c_1, \ldots, c_{r-1}$ not all zero such that $c_0\mathbf{1} + c_1 T + \cdots + c_{r-1}T^{r-1} = 0$ contradicting that $\deg m = r$. □

Note that the theorem holds for an arbitrary $n \times n$ matrix A if we exchange the word linear transformation by "matrix" everywhere. In fact, if T has matrix A with respect to some basis, then the minimal polynomial of T is also the minimal polynomial of A (in the sense that it satisfies the three properties of Definition 7.23. We usually write m_T, respectively, m_A for the minimal polynomials of T and A.

7.25 Examples.

a. The minimal polynomial of 0 is simply x. The minimal polynomial of $\mathbf{1}$ is $x - 1$. More generally the minimal polynomial of $\lambda\mathbf{1}$ is $x - \lambda$.

b. Let T be a linear transformation with matrix

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$$

Note that if $\lambda = \mu$, then $m_A = (x - \lambda)$ since then $A = \lambda I$.

Otherwise, m_A cannot be linear, since $A + cI$ will always have at least one nonzero diagonal entry. So $\deg m_A \geq 2$. However, A clearly satisfies

$$(A - \lambda I)(A - \mu I) = \begin{bmatrix} 0 & \\ 0 & \mu - \lambda \end{bmatrix} \begin{bmatrix} \lambda - \mu & \\ & 0 \end{bmatrix} = 0$$

So we have a polynomial of degree 2 and it follows that $m_A = (x - \lambda)(x - \mu)$.

Let $T \in \operatorname{End}(V)$ be a *nilpotent* operator, that is, there exists $k > 0$ such that $T^k = 0$.

Then the minimal polynomial of T is x^k where k is minimal such that $T^k = 0$. Indeed, if $m = x^k$, then $m(T) = 0$. Also, the minimal polynomial must be a divisor of x^k. But these are all of the form x^ℓ and hence $m_T = m$ since $T^\ell \neq 0$ for all $0 < \ell < k$.

There is another polynomial which is of utmost importance: Recall that an eigenvalue λ of a matrix A satisfies the equation $\det(\lambda I - A) = 0$.

Recall, that $\mathbb{F}[x] \subseteq \mathbb{F}(x)$ and $\mathbb{F}(x)$ is a field, so all our linear algebra applies to matrices with entries in $\mathbb{F}(x)$. In particular, if we have a matrix in $M_n(\mathbb{F}(x))$ we can compute its determinant (and again obtain an element of $\mathbb{F}(x)$).

Moreover note that, if $A \in M_n(\mathbb{F}(x))$ has all entries in $\mathbb{F}[x]$, then $\det A \in \mathbb{F}[x]$: this is clear from the Leibniz Formula 5.9 in Section 5.1 (keep in mind that $\det P_\sigma = \pm 1$).

7.26 Definition. Let $A \in M_n(\mathbb{F})$ be a square matrix. The *characteristic polynomial* of A, denoted p_A, is defined as $p_A = \det(xI - A)$ where $xI - A$ is viewed as an element of $M_n(\mathbb{F}(x))$.

Similarly, if $T \in \operatorname{End}(V)$, then its characteristic polynomial p_T is defined as $p_T = p_A$ where A is the matrix of T with respect to any basis.

Remark. Note that $p_A \in \mathbb{F}[x]$ because $xI - A$ has polynomial entries.

Also, p_T is well defined since if A, A' represent T with respect to bases $\mathcal{B}$ and $\mathcal{B}' = \mathcal{B}P$ then note that P is also invertible when viewed as an element of $M_n(\mathbb{F}(x))$, and so

$$xI - A' = xI - P^{-1}AP = P^{-1}(xI - A)P$$

and so

$$\det(xI - A') = \det(P^{-1}(xI - A)P) = \det(xI - A).$$

7.27 Proposition. *The roots of p_T in $\mathbb{F}$ are precisely the eigenvalues of T in $\mathbb{F}$.*

Proof. We have shown that λ is an eigenvalue if and only if $\det(A - \lambda I) = 0$. Of course this is equivalent to saying that $\det(\lambda I - A) = 0$. Thus, the statement follows if it is true that $\det(\lambda I - A) = p_A(\lambda)$. This is indeed true (but *not* trivial): it does not matter whether we first substitute λ for x and then compute the determinant, or if we first compute the determinant and then substitute λ. □

7.28 Example.

$$R = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

$p_R = x^2 - 2\cos(\alpha)x + 1$.

Note that for any 2×2 matrix A we have

$$p_A = x^2 - \operatorname{trace}(A)x + \det(A).$$

The main result of this section is

7.29 Theorem (Cayley-Hamilton Theorem)**.** *Let $T \in \operatorname{End}(V)$ be a linear operator. Then*

$$p_T(T) = 0.$$

In particular, $m_T \mid p_T$ and so $\deg m_T \leq \deg p_T = \dim V$.

Proof. We will first prove the following statement:

Claim: Let $A \in M_n(\mathbb{F})$ be a matrix such that p_A decomposes into linear factors. That is,

$$p_A = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the (not necessarily distinct) eigenvalues of A.

Then $p_A(A) = 0$.

We will prove the claim by induction on n. For $n = 1$ the assertion is clear: $A = [\lambda_1]$ and $p_A = x - \lambda_1$ so $p_A(A) = A - \lambda_1 I = 0$.

Now let $n > 0$ be given and assume that the claim holds for all $n \times n$ matrices whose characteristic polynomial decomposes as stated.

Let $A \in M_{n+1}(\mathbb{F})$ and suppose $p_A = (x-\lambda_1)(x-\lambda_2)\cdots(x-\lambda_{n+1})$. Let v be an eigenvector belonging to λ_{n+1}, and choose any basis for $\mathbb{F}^n$ containing v as first vector. If P is the matrix of the change of basis, then

$$A' = P^{-1}AP = \begin{bmatrix} \lambda_{n+1} & * \\ 0 & B \end{bmatrix}$$

where $B \in M_n(\mathbb{F})$ has characteristic polynomial $p_B = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n)$. (This follows from the fact that $\det(xI - A) = (x - \lambda_{n+1}) \det(xI_n - B) = (x - \lambda_{n+1})p_B$.)

Also note that as $p_{A'} = p_A$ and since $p_A(A) = p_{A'}(PA'P^{-1}) = Pp_{A'}(A')P^{-1}$ it suffices to show that $p_{A'}(A') = 0$.

Now check that

$$p_{A'}(A') = (A' - \lambda_{n+1}I)p_B(A') = \begin{bmatrix} 0 & * \\ 0 & B - \lambda_{n+1}I_n \end{bmatrix} \begin{bmatrix} * & * \\ 0 & p_B(B) \end{bmatrix}$$

Here $*$ stands for some unspecified entries. By induction, we know that $p_B(B) = 0$. But then also

$$\begin{bmatrix} 0 & * \\ 0 & B - \lambda_{n+1}I_n \end{bmatrix} \begin{bmatrix} * & * \\ 0 & 0_{n,n} \end{bmatrix} = 0$$

The claim is now proved.

If p_A does not decompose into linear factors (ie. if p_A has irreducible factors without roots in $\mathbb{F}$), then we may always extend $\mathbb{F}$ and find a field L such that $\mathbb{F} \subseteq L$ and such that p_A has n roots in L (with multiplicities). Then $A \in M_n(L)$, so A has also a characteristic polynomial, q_A, say, as an element of $M_n(L)$. But $q_A = \det(xI - A)$ is the same, whether we think of A as an element of $M_n(\mathbb{F})$ or as an element of $M_n(L)$ because $xI - A$ has entries in $\mathbb{F}[x] \subseteq L[x]$. Thus, $q_A = p_A$. The claim applies to q_A: $q_A(A) = 0$ in $M_n(L)$, but so $p_A(A) = 0$.

Finally, the assertion for T now follows; if $\mathcal{B}$ is any basis for V, then $p_A(A) = 0$ where A is the marix of T with respect to $\mathcal{B}$. This of course implies that $p_A(T) = 0$. But sine $p_A = p_T$, the theorem is proven. □

Remark. In the proof we mentioned that we can always enlarge $\mathbb{F}$ to decompose a polynomial. To see how this goes suppose $f \in \mathbb{F}[x]$ does not decompose into linear factors. Let p be an irreducible factor of f without root in $\mathbb{F}$. Let $I = \{hp \mid h \in \mathbb{F}[x]\}$, which is an ideal in $\mathbb{F}[x]$. From homework problems we know that $L := \mathbb{F}[x]/I$ is a field containing (a copy of) $\mathbb{F}$ ($c \in \mathbb{F}$ may be identified with $\overline{c} \in L$) and that $p \in \mathbb{F}[x] \subseteq L[x]$ has root $\overline{x}$ in L.

Now we decompose f in $L[x]$ then f has at least one root more than before. This way we continue until we arrive at a field L' containing $\mathbb{F}$ where f has n roots ($n = \deg f$), where we count roots with multiplicities.

Example. If A is an 2×2 matrix then

$$A^2 - \mathrm{trace}(A) \cdot A + \det(A) \cdot I = 0.$$

7.5. Irreducible and cyclic subspaces

Let us fix V and a linear operator T on V. As mentioned before, we want to arrive at a well understood decomposition of V into T-invariant subspaces. A good starting point is the case when V has *no* invariant subspaces except of course V and $\{0\}$.

7.30 Definition. $V \neq \{0\}$ is called *irreducible* if the only invariant subspaces of V are V and $\{0\}$.

V is called *indecomposable* if V cannot be written as $V = V_1 \oplus V_2$ with both, $V_1, V_2 \subsetneq V$ invariant. Of course, a not indecomposable space is called *decomposable*.

Finally, V is called *cyclic* if there is a vector $v \in V$ such that v is spanned by $v, Tv, T^2v, \ldots$

It is easy to see that irreducible implies indecomposable. Unfortunately, the converse is not true, creating a lot of headache (this has wide ramifications if one studies more than just one linear operator simultaneously).

It is also not hard to see that irreducible implies cyclic, but again, the converse fails.

7.5.1 Problem. Let $v \in V$. Show that $W = \mathrm{Span}(v, Tv, T^2v, \ldots, T^{n^2}v)$ is an invariant cyclic subspace and conclude that $V = W$ if V is irreducible and $v \neq 0$.

7.31 Examples.

a. Let $T\colon V \to V$ be a diagonalizable linear transformation. Then V is irreducible if and only if $\dim V = 1$.

   V is decomposable if $\dim V > 1$.

   V is cyclic if and only if all eigenvalues are distinct.

b. Let $R\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a rotation by α, where $\alpha \notin \pi\mathbb{Z}$ (is not an integer multiple of π).

   Then $V = \mathbb{R}^2$ is irreducible.

   It is also cyclic: let $v \neq 0$, then $v, R(v)$ are linearly independent and hence form a basis.

c. Let $D\colon V \to V$ be defined as $D(f) = f'$ where $V = \{f \in \mathcal{P}(\mathbb{R}) \mid \deg f \leq n\}$.

   Then V is indecomposable but V is not irreducible.

   V is cyclic: indeed, $V = \operatorname{Span}(x^n, D(x^n), \ldots, D^n(x^n))$.

7.5.2 Problem. Let
$$U = \begin{bmatrix} 1 & 1 \\ & 1 \end{bmatrix}$$
and consider $T = T_U$ on $\mathbb{C}^2$. Show that $\mathbb{C}^2$ is cyclic and indecomposable but not irreducible.

How can we construct T-invariant subspaces of V? Here is an elementary (but rather powerful) implicit one: given any polynomial $f \in \mathbb{F}[x]$, define
$$V^f = \{v \in V \mid f(T)v = 0\} = \mathcal{N}(f(T)).$$
Since $Tf(T) = f(T)T$ in $\operatorname{End}(V)$, it is immediate that V^f is invariant: if $v \in V^f$ then $Tv \in V^f$. Indeed, $f(T)Tv = Tf(T)v = T(0) = 0$.

Another (explicit) way is alluded to in the definition of cyclic. If $v \in V$ we define $\langle v \rangle$ as
$$\langle v \rangle = \operatorname{Span}(v, Tv, T^2v, \ldots, T^{n^2-1}v)$$
(Note that since $\deg m_T \leq n^2$, it is enough to consider powers up to T^{n^2-1}. All higher powers are in the span of the lower ones.) $\langle v \rangle$ is a T-invariant subspace and called the *invariant subspace generated by* v. How to describe $\langle v \rangle$ and how "high" do we really need to go in terms of powers of T?

Of $v \in V$, let $I_v \subseteq \mathbb{F}[x]$ be defined as
$$I_v = \{f \in \mathbb{F}[x] \mid f(T)v = 0\}.$$
I_v is often called the *annihilator* of v. It is an ideal of $\mathbb{F}[x]$ (cf. the proof of Theorem 6.27). (We could show that every ideal of $\mathbb{F}[x]$ is of the form $f\mathbb{F}[x] = \{fg \mid g \in \mathbb{F}[x]\}$; but we will proceed directly.)

It is clear that $f, g \in I_v$ implies that $f + g \in I_v$. Also $f \in I_v$ and $g \in \mathbb{F}[x]$ means $gf \in I_v$. Also not that $m_T \in I_v$ so I_v contains nonzero elements. Let $f \in I_v$ be a polynomial of minimum degree in I_v. We also assume that f has highest coefficient equal to 1. If $h \in I_v$ is any other polynomial then

$$h = Qf + R$$

and since $h(T)v = Q(T)f(T)v + R(T)v$ it follows that also $R(T)v = 0$ and so $R \in I_v$. Since R cannot have degree less than $\deg f$, it follows that $R = 0$ and so

$$I_v = \{gf \mid g \in \mathbb{F}[x]\}.$$

(We just proved $\subseteq$, and $\supseteq$ is clear.)

It also follows that f is unique: any other element of I_v of the same degree is of the form cf with $c \neq 0$ in $\mathbb{F}$. But then the "highest coefficient equal 1" condition implies that there is only one such. f is called the *order*[1] of the element $v \in V$.

7.32 Example. Let U be as in Problem 7.5.2, and consider $T_U\colon \mathbb{R}^2 \to \mathbb{R}^2$ Note that e_1 is an eigenvector belonging to 1. So the order of e_1 is $x - 1$: indeed, $(U - I)v = 0$.

What about e_2? Note that $\mathbb{R}^2$ is cyclic and $\langle e_2 \rangle = \mathbb{R}^2$. $Ue_2 = e_1 + e_2$. We know that $p_T(U)e_2 = 0$, $p_T = (x-1)(x-1)$. Since $p_T(U) = 0$ it follows that the order divides $P_T = (x-1)^2$. It is not $x - 1$ since e_2 is not an eigenvector: $(U - I)e_2 = e_1$, from which we conclude that $(U - I)^2 e_2 = 0$.

The order therefore is equal to $(x-1)^2$.

7.33 Proposition. *Let $v \in V$ with order f. Then the following holds:*

a. $f \mid m_T$.

b. $\deg f = \dim\langle v \rangle$.

c. *Let $T_{\langle v \rangle}$ denote the restriction of T to $\langle v \rangle$. Then $f = m_{T_{\langle v \rangle}}$.*

d. *If $v \neq 0$, $\langle v \rangle$ is irreducible if and only if f is irreducible.*

Proof. Part a. is clear because $m_T \in I_v$ as observed above.

As for the dimension assertion: a down to earth explanation is the following: let $f = a_0 + a_1 x + \cdots + a_{r-1}x^r + x^r$. Then $v, Tv, \ldots, T^{r-1}v$ must be linearly independent; otherwise there was a nonzero polynomial g of degree $r - 1$ or less that $g(T)v = 0$. Thus, the dimension of $\langle v \rangle$ is at least r. However, if $w \in \langle v \rangle$, then there is a polynomial $g \in \mathbb{F}[x]$ such that $g(T)v = w$. This follows from the definition of $\langle v \rangle$: $w = b_0 v + b_1 Tv + \cdots + b_k T^k v$ for suitable b_i which means that $g = b_0 + b_1 x + \cdots + b_k x^k$ will suffice. Write $g = Qf + R$ as usual and

[1] The use of the term "order" could be motivated as follows: as a consequence of the First Isomorphism Theorem applied to the linear transformation $E\colon \mathbb{F}[x] \to \langle v \rangle$, $E(g) = g(T)v$, one obtains an isomorphism $\overline{E}\colon F[x]/I_v \to \langle v \rangle$. Now f plays the same role for $\mathbb{F}[x]/I_v$ as does n in $\mathbb{Z}/n\mathbb{Z}$; and n is the order (ie. number of elements) of the abelian group $\mathbb{Z}/n\mathbb{Z}$, and the order of the element $1 \in \mathbb{Z}/n\mathbb{Z}$ as defined in a homework problem: $1 + 1 + \cdots + 1 = 0$ (n summands). Note that we write $\mathbb{Z}/n\mathbb{Z}$ additively and hence one would formally write na for $a + a + \cdots + a$ (n summands). In that sense $n \cdot 1 = 0$ as is $\overline{f} \cdot 1 = 0$ in $\mathbb{F}[x]/I_v$ (which is also a ring))

observe that $R = 0$ or $\deg R \le r$ and we get $w = R(T)v$. Hence $w \in \operatorname{Span}(v, Tv, \dots, T^{r-1}v)$ and the dimension of $\langle v\rangle$ is at most r. This shows b.

To prove c. first observe that $f(T)$ "kills" everything in $\langle v\rangle$. Indeed, we just showed that every element w of $\langle v\rangle$ is of the form $g(T)v$ for some $g \in \mathbb{F}[x]$ and hence $f(T)w = f(T)(g(T)v) = (f(T)g(T))v = g(T)(f(T)v) = 0$. So $m_{T_{\langle v\rangle}}$ divides f (because of course $f(T)w = f(T_{\langle v\rangle})w$). On the other hand $m_{T_{\langle v\rangle}} \in I_v$ since $v \in \langle v\rangle$ so $f \mid m_{T_{\langle v\rangle}}$; since both have highest term 1, they are equal.

The most interesting assertion is d. Suppose f is irreducible and let $W \subseteq \langle v\rangle$ be a nonzero invariant subspace. We have to show that $W = \langle v\rangle$. Pick any nonzero $w \in W$ and consider $\langle w\rangle$. Clearly w has also an order, g, say. Since $\dim W \le \dim\langle v\rangle = r$ it follows by b. applied to $\langle w\rangle$ that $\deg g = \dim W \le r$. Also note that $f \in I_w$ because $f(T)u = 0$ for all $u \in \langle v\rangle$ (can you prove this?). Thus, $g \mid f$. But f is irreducible and g is not a unit so $\deg g = \deg f$ and $r = \dim W$. Hence $W = \langle v\rangle$.

For the converse suppose $\langle v\rangle$ is irreducible and let $p \mid f$, $f = pq$, say. Consider $W = \langle v\rangle^p$, the nullspace of $p(T_{\langle v\rangle})$. Notice that W is T-stable, so there are two cases: $W = \{0\}$, or $W = \langle v\rangle$. If $W = \{0\}$, $p(T_{\langle v\rangle})$ is an invertible linear operator, and so since $f(T_{\langle v\rangle}) = 0$, we must have that $q(T_{\langle v\rangle}) = 0$. For degree reasons it then follows that $\deg q = \deg f$ (f is the minimal polynomial), which in turn means p is constant and hence a unit as needed. If on the other hand $W = \langle v\rangle$, then $p(T_{\langle v\rangle}) = 0$ and it follows that $\deg p = \deg f$ because f is the minimal polynomial of $T_{\langle v\rangle}$. □

7.34 Corollary.

a. *V is cyclic if and only if* $\deg m_T = \dim V$.

b. *V is irreducible if and only if m_T is irreducible and* $\deg m_T = \dim V$.

Proof. This is a homework problem. But as a hint: If V is irreducible it is cyclic. □

Note because of Proposition 7.33, we understand $T_{\langle v\rangle}$ very well:

Remark. Let $V = \langle v\rangle$ be cyclic where v has order $f = b_0 + b_1x + \cdots + b_{d-1}x^{d-1} + x^d$. We have seen in the proof of Proposition 7.33 b. that $\mathcal{B} = (v, Tv, \dots, T^{d-1}v)$ is a basis for V. Because $f(T)v = 0$, we have

$$T^d v = -b_0v - b_1Tv - \cdots - b_{d-1}T^dv.$$

Thus,

$$M_{\mathcal{B}}(T) = \begin{bmatrix} 0 & & & & -b_0 \\ 1 & 0 & & & \vdots \\ & 1 & \ddots & & \vdots \\ & & \ddots & 0 & -b_{d-2} \\ & & & 1 & -b_{d-1} \end{bmatrix} \tag{7.7}$$

This matrix is called the *companion matrix* of the polynomial f.

7.5.3 Problem. Suppose V is cyclic with minimal polynomial $f = p^e$ where p is irreducible and $e > 0$.

Show that if $w \in V$ then the order of w is p^g with $g \leq e$.

Show that if the order of w is p^g with $g < e$ then there is $u \in V$ such that $w = p(T)^{e-g}u$.

(*Hint:* For the second claim: Suppose $V = \langle v \rangle$ (ie. the order of v is p^e). Then $w = h(T)v$ for some unique $h \in \mathbb{F}[x]$ with $\deg h < \deg p^e$ (division by remainder).

Now $0 = p^{e-g}(T)w = p^{e-g}(T)h(T)v$ so it follows that p^e divides $p^{e-g}h$, which means that p^g divides h. In fact, it follows that $h = p^{e-g}q$ for some $q \in \mathbb{F}[x]$. The claim then follows with $u = q(T)v$.

Fill in the details.)

7.6. The Jordan canonical form

Let now $m_T = p_1^{e_1}p_2^{e_2}\cdots p_r^{e_r}$ be the decomposition of the minimal polynomial into irreducible factors: that is, p_i is irreducible, and $\gcd(p_i, p_j) = 1$ if $i \neq j$. We also assume that $e_i > 0$ for all i. The main goal is to connect this factorization of m_T to properties of T.

7.35 Definition. Let $f \in \mathbb{F}[x]$ be a polynomial. We define

$$V(f) := \{v \in V \mid f(T)^\ell v = 0 \text{ for some } \ell > 0\}.$$

Note that ℓ in the definition may depend on v.

7.6.1 Problem. Show that if f is irreducible, then $V(f)$ is the set of all $v \in V$ whose order is a power of f.

7.36 Examples.

a. T is nilpotent if and only if $V(x) = V$.

b. If $m_T = p_1^{e_1}p_2^{e_2}\cdots p_r^{e_r}$ with $e_i > 0$, then $V(p_1p_2\cdots p_r) = V$.

c. If $V = \mathbb{F}^2$, and
$$A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$$
with $T = T_A$, then $V(x-a) = E_a$ is the eigenspace of A for a if $a \neq c$: The nullspace of $A - aI$ is E_a by definition, and is spanned by e_1. Note that $(A - aI)^k \neq 0$ for all k, so $V(x-a) \neq V$.

If $a = c$, on the other hand, then $V(x-a) = V$, as then $A - aI$ is nilpotent.

d. If λ is an eigenvalue of T, the space $V(x-\lambda)$ is called the *generalized eigenspace* for λ. Note that always $E_\lambda \subset V(x-\lambda)$ but the two may differ.

We will need the following observations below:

7.37 Facts.

a. If $\gcd(f, m_T) = 1$, then $f(T)$ is invertible.

b. Let $V = V_1 \oplus V_2 \oplus \cdots \oplus V_r$ is a direct sum decomposition where all V_i are invariant with minimal polynomial of $T|_{V_i} = m_i$, say. If the m_i are coprime, then
$$m_T = m_1 m_2 \cdots m_r.$$

Proof. We may find $p, q \in \mathbb{F}[x]$ such that $1 = pf + qm_T$. Then $\mathbf{1} = p(T)f(T) + q(T)m_T(T) = p(T)f(T)$, so $f(T)$ is invertible with inverse $p(T)$. This shows a..

To see that b) holds, it is clear that m_T divides $m_1 m_2 \cdots m_r$, because $m_1(T)m_2(T) \cdots m_r(T) = 0$. Also since $m(T)$ is zero on V_i, $m_i \mid m$. Since they are corpime, it follows that $m_1 m_2 \cdots m_r \mid m$ as well (e.g. from the uniqueness of prime factorization). □

7.38 Proposition. *Let $T\colon V \to V$ be a linear operator, and $f \in \mathbb{F}[x]$.*

a. *$V(f)$ is invariant.*

b. *If g is coprime to f, then $V(f) \cap V(g) = \{0\}$.*

c. *If $m_T = p_1^{e_1} \cdots p_r^{e_r}$, then*
$$V = V(p_1) \oplus V(p_2) \oplus \cdots \oplus V(p_r)$$
and $V(p_i) \neq \{0\}$. Moreover, the minimal polynomial of T restricted to $V(p_i)$ is $p_i^{e_i}$.

Proof.

a. Let $v \in V(f)$. Then $f(T)^\ell v = 0$ for some ℓ. Then also $Tf(T)^\ell v = 0$. But $Tf(T)^\ell = f(T)^\ell T$, so $f(T)^\ell T(v) = 0$ which means $T(v) \in V(f)$.

b. Let $v \in V(f) \cap V(g)$. Then $f(T)^\ell v = g(T)^m v = 0$ for some ℓ, m. Note that also f^ℓ, g^m are coprime. Hence there are $a, b \in \mathbb{F}[x]$ such that
$$1 = af^\ell + bg^m.$$
Then $v = \mathbf{1}v = (a(T)f(T)^\ell + b(T)g(T)^m)v = a(T)f(T)^\ell(v) + b(T)g(T)^m(v) = 0.$

c. This is similar to the previous statement.

Let $W = \operatorname{Im}(p_1(T)^{e_1}) = p_1(T)^{e_1}(V)$. Then W is an invariant subspace. Let $f = p_2^{e_1} p_3^{e_2} \cdots p_r^{e_r}$. Then $f(T)|_W = 0$, so $m_{T|_W}$ divides f. In particular, $p_1(T)^{e_1}$ is invertible on W, since $\gcd(f, p_1^{e_1}) = 1$. On the other hand, $V(p_1) = \mathcal{N}(p_1(T)^{e_1})$ because the order of any element in $V(p_1)$ is a power of p_1 and divides m_T. It follows that $V(p_1) \cap W = \{0\}$. Also note that $\operatorname{Im}(f(T)) \subset V(p_1)$, because $p_1(T)^{e_1}(f(T)v) = m_T(T)v = 0$ for all v. If we write $\mathbf{1} = a(T)p_1(T)^{e_1} + b(T)f(T)$, then
$$v = \mathbf{1}v = a(T)p_1(T)^{e_1}(v) + b(T)f(T)(v) \in W + V(p_1).$$

We have shown that $V = V(p_1) \oplus W$. Note that $m_{T|_W} \mid f$ and so is not equal to m_T, so $W \neq V$, and $V(p_1) \neq \{0\}$. Since W is invariant, and $\deg m_{T|_W} < \deg m_T$, we may proceed by induction on $\deg m_T$ to conlude that W is the direct sum of some $V(p_i)$. By part b. of the above Fact, m_T is the product of all the minimal polynomials of T on the occurring $V(p_i)$. Hence all p_i must occur, and the proof is finished, and moreover $p_i^{e_i}$ is the minimal polynomial of T on $V(p_i)$.

□

7.39 Example. Suppose m_T decomposes into linear factors. Then V is the direct sum of the generalized eigenspaces. In aprticular, if the matrix of T with respect to some basis is upper triangular, this applies.

We have seen that every eigenvalue is a root of m_T. But this also shows that every root of m_T is an eigenvalue (which we also know by Caley-Hamilton): indeed, if λ is a root of m_T, then $V(x-\lambda) \neq \{0\}$. But on $V(x-\lambda)$, $T - \lambda\mathbf{1}$ is not invertible and hence has a nonzero nullspace.

We will now study what we can say about the $V(p_i)$ individually.

7.40 Theorem. *Let $T \colon V \to V$ be a linear operator with $m_T = p^e$ where p is irreducible. Then V is the direct sum of cyclic subspaces. That is,*

$$V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_r \rangle$$

where the order of v_i is p^{e_i}, and $(e_1+e_2+\cdots+e_r)\deg p = \dim V$, and moreover $e = \max_i\{e_i\}$.

Proof. The order of any element of V is p^f for some f, so all that is left to show is that V allows such a decomposition.

We proceed by induction on $n = \dim V$. If $n = 1$, then $V = \operatorname{Span}(v)$ and $N = 0$ as claimed, proving the base case.

Now suppose for all spaces of dimension $< \dim V$ the theorem has been proven.

There is an element $v_1 \in V$ of order p^e. (Indeed, $p(T)^{e-1}(v) \neq 0$ for some v and then the order, which divides p^e must be equal to p^e.) Let $W = \langle v_1 \rangle$. This is a subspace of dimension $\deg p^e = e \deg p$. If $V = W$ we are done. Otherwise, T gives rise to a linear operator $\overline{T}$ on V/W (by defining $\overline{T}\overline{w} = \overline{T(w)}$). Note that $m_{\overline{T}} = p^f$ for some $f \leq e$:

$$p(\overline{T})^e(\overline{v}) = 0$$

for all $\overline{v} \in V/W$. Since $\dim V/W < \dim V$, the induction hypthesis applies and there are $v_2, v_3, \ldots, v_r$ in V such that

$$V/W = \langle \overline{v_1} \rangle \oplus \langle \overline{v_2} \rangle \oplus \cdots \oplus \langle \overline{v_r} \rangle.$$

We will now change the v_i such that $\langle v_i \rangle \cap W = \{0\}$. To achieve this let p^{e_i} be the order of $\overline{v_i}$ $(i \geq 2)$. Then the order of v_i is p^{f_i} with $f_i \geq e_i$: indeed, $p^f(T)(v) \neq 0$ for $f < e_i$ because it is not zero mod W. If $f_i > e_i$, then $p(T)^{e_i}(v_i) = w_i \in W$ is not zero. Note also that $p(T)^{e-e_i}(w_i) = 0$.

Claim: There is $u_i \in W$ such that $w_i = p(T)^{e_i} u_i$. To see why this is true, recall that every element w of $W = \langle v_1 \rangle$ is of the form $w = f(T)v_1$ for some $f \in \mathbb{F}[x]$ (it is a linear combination of elements of the form $T^i(v_1)$). So $w_i = f(T)v_1$ for some f. And $p^{e-e_1}(T)(w_i) = 0$. This means $p^{e-e_1}(T)f(T)v_1 = 0$. Thus, the order of v_1, which is p^e, divides $p^{e-e_1}f$. As p is irreducible this means p^{e_1} divides f: $f = p^{e_1}g$ and we put $u_i := g(T)v_1$. Then

$$w_i = f(T)v_1 = p^{e_1}(T)g(T)v_1 = p^{e_1}(T)u_i.$$

This proves the claim. □

Then $w_i = p(T)^{e_i} u_i$ for some $u_i \in W$ (this is true also if $w_i = 0$, ie. $f_i = e_i$; in this case pick $u_i = 0$). We now define

$$v_i' = v_i - u_i.$$

Then $\overline{v_i} = \overline{v_i'}$. Also, the order of v_i' is exactly p^{e_i}: $p^{e_i}(T)(v_i') = w_i - p(T)^{e_i}u_i = 0$. We now replace the v_i by the v_i'. Thus we are in the following situation: we have $v_2, v_3, \ldots, v_r \in V$ such that

- $V/W = \langle \overline{v_1} \rangle \oplus \langle \overline{v_2} \rangle \oplus \cdots \oplus \langle \overline{v_r} \rangle$.
- the order of v_i is p^{e_i}.

It now follows that $W \cap \langle v_i \rangle = \{0\}$. Indeed, let $w \in \langle v_i \rangle \cap W$. Then $\overline{w} = 0$. But $\langle v_i \rangle$ and $\langle \overline{v_i} \rangle$ have the same dimension, and $x \mapsto \overline{x}$ is a surjective linear transformation between them, so it is an isomorphism.

We are done: suppose $w_1 + w_2 + \cdots + w_r = 0$ with $w_i \in \langle v_i \rangle$. Then $\overline{w_1 + w_2 + \cdots + w_r} = 0$, and so

$$\overline{w_1} + \overline{w_2} + \cdots + \overline{w_r} = 0.$$

Now the spaces $\langle \overline{v_i} \rangle$ are independent in V/W, so $\overline{w_i} = 0$. Thus, $w_i \in W$. But this means $w_i \in W \cap \langle v_i \rangle$, and so $w_i = 0$ for $i = 2, 3, \ldots, r$. Hence also $w_1 = 0$, and we are done. □

Example. Let $V = \mathbb{F}^7$ and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix}$$

The proof of Theorem 7.40 suggests the following approach to find the v_i.

- The minimal polynomial of A is $m_A = x^3$ (A is nilpotent and 3 is clearly the smallest k such that $A^k = 0$). Start with the standard basis $(e_1, e_2, \ldots, e_7)$ to find an element of order x^3: The orders of $e_1, e_2, \ldots, e_7$ are

$$x, x^2, x^3, x, x^2, x, x^2$$

so $v := e_3$ has order x^3.

- Next, $W = \langle v \rangle = \mathrm{Span}(e_1, e_2, e_3)$. And $V/W \cong \mathbb{F}^4$ with basis $(\overline{e}_4, \overline{e}_5, \overline{e}_6, \overline{e}_7)$. Note that here (this depends on A of course!) the matrix of $\overline{T}_A \colon \mathbb{F}^4 \to \mathbb{F}^4$ is the south-east 4×4 block of A, given by

$$\begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}$$

Its minimal polynomial is x^2, and we find we can take $\overline{v}_2 = \overline{e}_4$, $\overline{v}_3 = \overline{e}_5$, and $\overline{v}_4 = \overline{e}_7$, then

$$V/W = \langle \overline{v}_2 \rangle \oplus \langle \overline{v}_3 \rangle \oplus \langle \overline{v}_4 \rangle = \mathrm{Span}(\overline{e}_4) \oplus \mathrm{Span}(\overline{e}_5) \oplus \mathrm{Span}(\overline{e}_6, \overline{e}_7).$$

- Next, we choose $v_2 = e_4$, $v_3 = e_5$, $v_4 = e_7$ (this is a choice, every element of $e_4 + W$ would do for v_1).

Then the orders are x, x^2, x^2, and note that the order of v_3 is x^2, whereas the order of $\overline{v}_3$ is x.

And indeed, $Av_3 = e_1 \neq 0 \in W$.

- For each of the v_i for which the order of v_i is not equal to the order of $\overline{v}_i$ we need to compute some u_i and replace v_i by $v_i - u_i$.

Here this is only v_3. $Av_3 = e_1 \in W$, and note that $A^2(Av_3) = A^3 v_3 = 0$ so Av_3 is in $\mathcal{N}(A^2) \cap W$. There is $w \in W$ such that $Av_3 = Aw$ and indeed, $w = e_2$ will work: $Ae_2 = e_1 = Av_3$. So we replace $v_3 = e_5$ by $e_5 - e_2$.

Now it follows that $v_1 = v, v_2, v_3, v_4$ have orders x^3, x, x, x^2, and

$$V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \langle v_3 \rangle \oplus \langle v_4 \rangle.$$

Note that we can now choose a basis of V consisting of the $T^j v_i$ (where j is less than the power of the order of v_i): With respect to the basis

$$(A^2 e_3, Ae_3, e_3, e_4, e_5, Ae_7 = e_6 + e_1, e_7)$$

T_A has matrix

$$\begin{bmatrix} 0 & 1 & & & & & \\ & 0 & 1 & & & & \\ & & 0 & & & & \\ & & & 0 & & & \\ & & & & 0 & & \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{bmatrix}$$

This is of the form

$$\begin{bmatrix} C_{x^3}^T & & & \\ & C_x^T & & \\ & & C_x^T & \\ & & & C_{x^2}^T \end{bmatrix}$$

Where C_f denotes the companion matrix of $f \in \mathbb{F}[x]$ (cf. (7.7) on Page 184).

- Finally observe that the number of companion matrices is precisely the dimension of the nullspace of A (ie. $\dim V^x$ in our notation above; x plays the role of p in the lemma and has degree 1).

  Next, the number of companion matrices for powers at least 2 is equal 2, and indeed, $\dim \mathcal{N}(A^2) = \dim \mathcal{N}(A) + 2 = 6$.

  Finally, the number of companion matrices for powers at least 3 of x is 1 and indeed, $\dim \mathcal{N}(A^3) = 7 = 6 + 1$.

Also observe that to find the v_i we could have proceeded as follows: identify the basis elements with order x^3 (here there is just one, v say). Compute a basis for $\langle v \rangle$, extend Av to a basis for $\mathcal{N}(A^2)$. Identify the basis elements of order x^2 (which would be Av and e_7).

Now compute $\langle v \rangle \oplus \langle e_7 \rangle$ (which is a direct sum (why?)). Extend A^2v, Ae_7 to a basis for $\mathcal{N}(A)$, and we are done.

7.41 Corollary. *Let T and V be as in Theorem 7.40. Then $p_T = p^{e_1+e_2+\cdots+e_r}$.*

Proof. If we choose a basis compatible with the decomposition of $V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_r \rangle$, then the matrix of T has the form

$$\begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_r \end{bmatrix}$$

where A_i is the matrix of $T_{\langle v_i \rangle}$. Then $p_T = p_{A_1} p_{A_2} \cdots p_{A_r}$ and it suffices to show that $p_{A_i} = p^{e_i}$. By Cayley-Hamilton applied to $\langle v_i \rangle$, we know that $p^{e_i} = m_{T_{\langle v_i \rangle}} = p_{T_{\langle v_i \rangle}} = p_{A_i}$ (cf. Corollary 7.34). □

7.42 Corollary (Elementary Divisor Theorem)**.** *Let $T \in \operatorname{End}(V)$ have minimal polynomial $m_T = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$ for coprime p_i and $e_i > 0$. Then there are $v_1, v_2, \ldots, v_t \in V$ such that*

$$V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_t \rangle$$

and the order of v_i is p_j^f for some j and $f \leq e_j$, and for each j there is at least one v_i with order $p_j^{e_j}$. Moreover p_T is the product of the orders of all the v_i. In particular p_T and m_T have the same irreducible divisors.

Proof. We know that $V = V(p_1) \oplus V(p_2) \oplus \cdots \oplus V(p_s)$. This is just an application of Theorem 7.40 and Corollary 7.41 to $V(p_i)$, and the observation that p_T is the product of the characteristic polynomials of all the characteristic polynomials of its restrictions to the various $V(p_i)$, analogous to Corollary 7.41. □

The occurring orders of the form p^e in the Elementary Divisor Theorem are called the *elementary divisors*.

Remark. This corollary shows that m_T divides p_T: indeed, m_T is the product of all the $p_i^{e_i}$ where e_i is maximal (one for each i), whereas p_T is the product of all orders. This does imply the Cayley-Hamilton Theorem. Of course we used it in the proof of Corollary 7.41.

However, we could deduce the Cayley-Hamilton theorem from the two corollaries by first showing that for a companion matrix A of f, $p_A = m_A = f$.

7.43 Fact. Let $V = \langle v_1\rangle \oplus \langle v_2\rangle \oplus \cdots \oplus \langle v_r\rangle$ according the Elementary Divisor Theorem. That is, the order of v_i is p^e for some irreducible polynomial p and $e > 0$. Then the occurring orders are uniquely determined (that is, in any other decompositon of $V = \langle w_1\rangle \oplus \cdots \oplus \langle w_s\rangle$ of this form, the number of w_i with a given order is the same as the number of v_i with a given order.

Proof. T determines the numbers $\dim \mathcal{N}(p_i(T)^e)$ where $1 \leq e \leq e_i$. These numbers in turn determine the number of v_js of order $p_j^{f_j}$ with $f_j \geq e$:

Indeed, the null space of $p_j^e(T)$, by definition, is a subspace of $V(p_j)$, so let us assume that $V = V(p_j)$. If we write $V(p_j) = \langle v_1\rangle \oplus \langle v_r\rangle \oplus \cdots \oplus \langle v_r\rangle$, then because this is a direct sum of invariant subspaces, if $v = w_1 + w_2 + \cdots + w_r$ with $w_i \in \langle v_i\rangle$, $p_j(T)^{f_j}(v) = 0$ if and only if $p_j(T)^e(w_i) = 0$. We hence see that

$$\mathcal{N}(p_j(T)^e) = (\mathcal{N}(p_j(T)^e) \cap \langle v_1\rangle) \oplus (\mathcal{N}(p_j(T)^e) \cap \langle v_2\rangle) \oplus \cdots \oplus (\mathcal{N}(p_j(T)^e) \cap \langle v_r\rangle).$$

Now, if the order of v_i is $p_j^{e_i}$, then

$$\dim \mathcal{N}(p_j(T)^e) \cap \langle v_i\rangle = \begin{cases} \dim\langle v_i\rangle = e_i \deg p_j & \text{if } e \geq e_i \\ e \deg p_j & \text{otherwise} \end{cases}$$

Indeed, we have seen that $\mathcal{N}(p_j(T)^e) = \langle p_j(T)^{e_i - e} v\rangle$ (cf. the claim in the proof of Theorem 7.40), and this is a vector with order p_j^e.

If we know all the numbers $\dim \mathcal{N}(p_j(T)^e)$ where $e = 1, 2, \ldots, \dim V$, then r (the number of v_i in $V(p_j)$) is determined by $r \deg p_j = \dim \mathcal{N}(p_j(T))$. Next, the number r_2 of v_j with order at least p_j^e, is then determined by $r_2 \deg p_j + \dim \mathcal{N}(p_j(T)) = \dim \mathcal{N}(p_j(T)^2)$, and so on.

It follows that

$$\dim \mathcal{N}(p(T)^{e+1}) = \dim \mathcal{N}(p(T)^e) + n(e) \deg p$$

where $n(e) = |\{i \mid e_i > e\}$. Now from the dimensions, we can determine the numbers $n(e)$, and knowing $n(e)$ for all $e > 0$ is the same as knowing all e_i (check this). □

We now focus on the most important special case: Let $T \in \mathrm{End}(V)$ and suppose p_T decomposes into linear factors, that is

$$p_T = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n)$$

where $n = \dim V$ and the λ_i are the eigenvalues (not necessarily distinct). For example, this applies if $\mathbb{F} = \mathbb{C}$, or if T is nilpotent.

The elementary divisor theorem gives rise to a very nice decomposition of V. Let $\mu_1, \mu_2, \dots, \mu_k$ be the distinct eigenvalues of T (ie. the μ_i are simply the distinct λ_i).

As mentioned before, we then have

$$V = V(x - \mu_1) \oplus \cdots \oplus V(x - \mu_k)$$

Let $\mu = \mu_i$, then $V(x - \mu) = \mathcal{N}(A - \mu I)^e$ for some e. In fact, we can always take $e = \dim V$, because restricted to $V(x-\mu)$, $N := T - \mu\mathbf{1}$ is a nilpotent linear transformation: restricted to $V(x-\mu)$, the minimal polynomial of T is $(x-\mu)^e$ for some e (and $e \le \dim V(x-\mu) \le \dim V$), and hence $N^e = 0$ on $V(x-\mu)$. Also note that any element $v \in \mathcal{N}(N^e)$ must be an element of $V(x-\mu)$ by definition, and so it follows

$$V(x - \mu) = \mathcal{N}(N^e).$$

Note that $\dim V(x-\mu)$ is exactly the multiplicity[2] m of μ as a root of p_A (exercise!). Thus, we can always put $e = m$.

$$V(x - \mu) = \mathcal{N}(T - \mu\mathbf{1})^m.$$

Because of this, $V(x-\mu)$ is often called the *generalized eigenspace* for μ. It generalizes the concept of the *eigenspace* for μ, which is defined as $\mathcal{N}(T - \mu\mathbf{1})$, the set of all eigenvectors belonging to μ together with the zero vector.

Recall that $V(x-\mu) = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_r \rangle$ for some v_i. Also recall that if v_i has order $(x-\mu)^{e_i}$ then

$$(v_i, Tv_i, \dots, T^{e_i - 1} v_i)$$

is a basis for $\langle v_i \rangle$.

7.44 Fact. $v_i, Nv_i, \dots, N^{e_i-1} v_i$ form a basis for $\langle v_i \rangle$ as well.

(Indeed, $\langle v_i \rangle$ is N-stable. And the order of v_i with respect to N is simply x^{e_i}. Thus, $(v_i, Nv_i, \dots, N^{e_i-1} v_i)$ is a basis for $\langle v_i \rangle$ (as it is linearly independent has the right number of elements). Alternatively, let $\langle v_i \rangle_N$ be the cyclic subspace with respect to N. Then $\langle v_i \rangle_N \subset \langle v_i \rangle$ because $\langle v_i \rangle$ is N-stable. On the other hand, since $T = N + \mu_i \mathbf{1}$, $\langle v_i \rangle_N$ is T-stable, forcing $\langle v_i \rangle \subset \langle v_i \rangle_N$, so the two spaces are the same.)

But N is *nilpotent* on $\langle v_i \rangle$ and its matrix with respect to that basis is then simply

$$\begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

[2] If f is a polynomial in $\mathbb{F}[x]$, the *multiplicity* of a root $\alpha \in \mathbb{F}$ of f is m, where m is the largest integer such that $(x-\alpha)^m$ divides f.

Reordering the basis in the opposite order, and adding back $\mu\mathbf{1}$ we find that the matrix of T restricted to $\langle v_i\rangle$ is the $e_i \times e_i$ matrix

$$J(\mu) = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix}$$

Such a matrix is called a *Jordan block*. Summarizing, there is a basis for V such that the matrix of T has the form

$$\begin{bmatrix} J_1(\nu_1) & & & \\ & J_2(\nu_2) & & \\ & & \ddots & \\ & & & J_\ell(\nu_\ell) \end{bmatrix}$$

where $J_i(\nu_i)$ is a Jordan block corresponding to the eigenvalue ν_i. Note that each eigenvalue appears *at least* once, but may be more often. This matrix is said to be in *Jordan canonical form*.

Applying this to a matrix $A \in M_n(\mathbb{F})$ whose minimal (or characteristic) polynomial decomposes, we find that there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that $A' = P^{-1}AP$ is in Jordan canonical form. A' is then called the Jordan canonical form of A.

7.6.2 Problem. Let A be a complex $n \times n$ matrix with distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ $(k \leq n)$.

Show that $m_A = (x - \lambda_1)^{n_1} \cdots (x - \lambda_k)^{n_k}$ where n_i is the size of the *largest* Jordan block for λ_i.

7.6.3 Problem. Show that the Jordan canonical form of a matrix A is uniquely determined by the various numbers $\dim \mathcal{N}(A - \lambda_i I)^e$ (for $e \geq 1$) (uniquely up to reordering of the Jordan blocks).

Thus to find the Jordan canonical form, we only need to know these numbers.

7.6.1. The case A nilpotent

As an application of the elementary divisor theorem we will now study what we can say about a nilpotent matrix $A \in M_n(\mathbb{F})$: $A^k = 0$ for some k. If we choose k minimal, then $m_A = x^k$. Also note that $A^n = 0$ and as a consequence of the elementary divisor theorem, $p_A = x^n$ (if V is the direct sum of the $\langle v_i\rangle$ of order x^{e_i}, then $p_A = x^{\sum_i e_i}$ and $\dim V = n = \sum e_i$).

In this case, our task is simple: $V = \mathbb{F}^n$, and all elementary divisors are of the form x^{e_i} (such that $\sum e_i = n$). But how to find them? We look at our algorithm: Note that $V(x) = V$. Because x has degree 1, the algorithm becomes slightly more pleasant (see also Figure 7.1):

- For $i = 1, 2, \ldots, k$ compute a basis $\mathcal{B}_i$ for $\mathcal{N}(T^i)$ such that $\mathcal{B}_1 \subseteq \mathcal{B}_2 \subseteq \cdots \subseteq \mathcal{B}_k$.

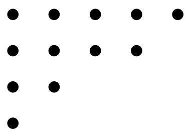


Figure 7.1.: The first row symbolizes $\mathcal{B}_1$, the next $\mathcal{B}_2 - \mathcal{B}_1$, and so on. If $v \in \mathcal{B}_{i+1} - \mathcal{B}_i$ is symbolized by a bullet, then Tv is symbolized the by the bullet above it and is an element of $\mathcal{B}_i$.

- For each $i = k-1, k-2, \ldots, 2$, let $v_1, v_2, \ldots, v_\ell$ be elements of $\mathcal{B}_{i+1} - \mathcal{B}_i$ that have not been replaced in a previous step. For $j = 1, 2, \ldots, \ell$ do the following:

  > Replace in $\mathcal{B}_1$ (using the Exchange Lemma 3.38) one element by $A^{i-1}v_j$. Then replace the same element in each of the $\mathcal{B}_s$ $s = 1, 2, \ldots, k$. Then, using the exchange lemma, replace the same element in each $\mathcal{B}_2, \mathcal{B}_3, \ldots, \mathcal{B}_k$ by $A^{i-2}v_j$, and continue, until each $\mathcal{B}_s$ contains $A^{i-t}v_j$ for $i > t \geq s$. (Each such replacement involves first expressing $A^{i-s}v_j$ by the elements of $\mathcal{B}_s$.) Never replace an element that has been replaced at a previous step (ie. is of the form $A^i v_j$).

- Finally, rearrange the basis such that the vectors of the form $A^i v_j$ for $i = 1, 2, \ldots$ appear consecutively in reverse order.

7.45 Lemma. *Let $A \in M_n(\mathbb{F})$ be nilpotent. Then there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that*

$$P^{-1}AP = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{bmatrix}$$

where each J_k has the form

$$J(0) = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}$$

Proof. Apply the Elementary Divisor Theorem. □

7.46 Example. Let

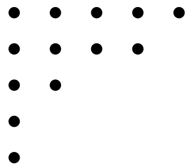
$$A = \begin{bmatrix} 1 & 2 & 1 \\ -1 & -2 & -1 \\ 1 & 2 & 1 \end{bmatrix} \in M_3(\mathbb{R}).$$

Then A is nilpotent: $A^2 = 0$: indeed, obviously $\operatorname{rank} A = 1$, and $\operatorname{Col}(A) \subseteq \mathcal{N}(A)$. Since $A^2 = 0$, it follows that the size of the largest block of the form $J(0)$ as above is at most 2. On the other hand, since $A \neq 0$, there must be at least one of size 2 and we get that

$$A \sim A' = \begin{bmatrix} 0 & 1 & 0 \\ & 0 & 0 \\ & & 0 \end{bmatrix}$$

Here we could put $v_1 = e_1$, and $v_2 = e_1 - e_3$. Then if $P = [\mathcal{B}]$ is the change of basis matrix for the change to the new basis $\mathcal{B} = (Av_1, v_1, v_2)$, we have $P^{-1}AP = A'$. (Note, $Av_2 = 0$, and $A^2 v_1 = 0$.)

Example. Here is an example how the algorithm to find the v_i would work in practice. Suppose the diagram according to Fig. 7.1 is



This means, $V = \mathbb{F}^{13}$ and $A^5 = 0$ such that

$$\dim \mathcal{N}(A) = 5, \dim \mathcal{N}(A^2) = 9, \dim \mathcal{N}(A^3) = 11, \dim \mathcal{N}(A^4) = 12, \dim \mathcal{N}(A^5) = 13.$$

Let

$$\begin{aligned} \mathcal{B}_1 &= (u_1, u_2, \ldots, u_5), \\ \mathcal{B}_2 &= \mathcal{B}_1 \cup (v_1, v_2, v_3, v_4), \\ \mathcal{B}_3 &= \mathcal{B}_2 \cup (w_1, w_2), \\ \mathcal{B}_4 &= \mathcal{B}_3 \cup (x), \text{ and} \\ \mathcal{B}_5 &= \mathcal{B}_4 \cup (y) \end{aligned}$$

be bases of $\mathcal{N}(A), \mathcal{N}(A^2), \mathcal{N}(A^3), \mathcal{N}(A^4), \mathcal{N}(A^5) = V$, respectively.

Now consider, Ay, A^2y, A^3y, A^4y and note that only $A^5y = 0$. It follows that $Ay \in \mathcal{N}(A^4)$, but $Ay \notin \mathcal{N}(A^3)$. Thus, $Ay \in \operatorname{Span}(\mathcal{B}_4)$ but $Ay \notin \operatorname{Span}(\mathcal{B}_5)$. Thus, if expressed as a linear combination of $\mathcal{B}_4$, Ay must have a coefficient in x and hence by the exchange lemma we may replace x by Ay in $\mathcal{B}_4$; we can do the same in $\mathcal{B}_5$. Similarly, A^2y is in $\operatorname{Span}(\mathcal{B}_3)$ and must have a nonzero coefficient in at least one of w_1, w_2. So we may replace one of them (say, w_1) by A^2y and still have a basis, and do the same in $\mathcal{B}_4, \mathcal{B}_5$. Repeating, we replace in all $\mathcal{B}_i$ one element among the u_i, by A^4y, in $\mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5$ one element among the v_i by A^3v, in $\mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5$ we replace w_1 by A^2y, and finally in $\mathcal{B}_4, \mathcal{B}_5$ we replace x by Ay.

After these subsitutions, in $\mathcal{B}_4 - \mathcal{B}_3$, all elements are of the form Av for some $v \in \mathcal{B}_5$, so we do nothing.

$\mathcal{B}_3 - \mathcal{B}_2 = (A^2y, w_2)$. So we replace in $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5$ one element among the remaining u_i (and *not* the element A^4y) by A^2w_2. This is possible, because A^2w_2 cannot be an element of $\langle y \rangle$ so cannot be a multiple of A^4y (if it were, $w_2 - cA^2y \in \mathcal{N}(A^2)$ for some c, and so w_2 and A^2y are not linearly independent mod $\mathcal{N}(A^2)$, but they are by construction).

We then replace one element among the v_i by Aw_2 in $\mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5$, always using the exchange lemma. So a possible outcome after these steps would be

$$\begin{aligned}
\mathcal{B}_1 &= (A^4y, A^2w_2, u_3, u_4, u_5),\\
\mathcal{B}_2 &= \mathcal{B}_1 \cup (A^3y, Aw_2, v_1, v_4),\\
\mathcal{B}_3 &= \mathcal{B}_2 \cup (A^2y, w_2),\\
\mathcal{B}_4 &= \mathcal{B}_3 \cup (Ay), \text{ and}\\
\mathcal{B}_5 &= \mathcal{B}_4 \cup (y)
\end{aligned}$$

The final step would now involve to replace first one element among u_2, u_3, u_5 by Av_1 in all bases; and then one of the remaining two by Av_4.

At the very end, we must make sure that in our bases the A^iv_j appear consecutively in reverse order (descending powers of A).

The upshot of all this is, that it boils down to solving a lot of equations of the form $[\mathcal{B}_i]X = Y$ (which is exactly what we need to do in order to check with respect to which element of $\mathcal{B}_i$, Y has a nonzero coefficient.

7.47 Example. Let

$$A = \begin{bmatrix} 2 & 2 & 0 & 1 \\ 0 & 7 & 2 & 3 \\ 0 & 0 & 2 & 0 \\ 0 & -10 & -4 & -4 \end{bmatrix} \in M_3(\mathbb{C}).$$

Let us find its Jordan canonical form. First, its characteristic polynomial is $(x-2)^3(x-1)$.

Note that in this case $V(x-1) = V^{x-1}$ is exactly the eigenspace, and spanned by

$$v_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -2 \end{bmatrix}$$

Now $V(x-2)$ must be 3 dimensional.

$$A' := A - 2I = \begin{bmatrix} 0 & 2 & 0 & 1 \\ 0 & 5 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & -10 & -4 & -6 \end{bmatrix} \to \begin{bmatrix} 0 & 0 & 4 & 1 & \\ 0 & 2 & 0 & 1 & \\ & & & & 0 \\ & & & & 0 \end{bmatrix}$$

Note that $\mathcal{N}(A')$ is spanned by

$$e_1 \quad \text{and} \quad v_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \\ -4 \end{bmatrix}$$

Now we know that there are two Jordan blocks for this eigenvalue. The diagram according to Fig. 7.1 (as applied to A') is then

$$\begin{matrix} \bullet & \bullet \\ \bullet & \end{matrix}$$

And indeed,

$$A'^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -5 & -2 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & 10 & 4 & 6 \end{bmatrix} \; A'^3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -5 & -2 & -3 \\ 0 & 0 & 0 & 0 \\ 0 & 10 & 4 & 6 \end{bmatrix} = A'^2$$

So A'^2 and A'^3 have the same null space, spanned by

$$e_1, u = \begin{bmatrix} 0 \\ 3 \\ 0 \\ -5 \end{bmatrix} \text{ and } w = \begin{bmatrix} 0 \\ 2 \\ -5 \\ 0 \end{bmatrix}$$

Note that $v_2 = 1/5(4u - w)$. By the exchange lemma, $\mathcal{N}(A'^2)$ has basis

$$(e_1, v_2, w)$$

(or (e_1, v_2, u)), which contains our basis for $\mathcal{N}(A')$. The orders (with respect to A') of v_2, e_1, w are, respectively, x, x, x^2. Our algorithm now says we should replace v_1 or e_1 by $A'w$. Note that

$$A'w = \begin{bmatrix} 4 \\ 0 \\ 0 \\ 0 \end{bmatrix} = 4e_1.$$

So we must replace e_1 by $4e_1$. So the basis now is $(4e_1 = A'u, v_2, u)$. Finally, we need to rearrange so that the basis is $(v_2, 4e_1, u)$.

With respect to the basis $(v_1, v_2, 4e_1, u)$, T_A has matrix

$$\begin{bmatrix} 1 & & & \\ & 2 & & \\ & & 2 & 1 \\ & & & 2 \end{bmatrix}$$

which is in Jordan canonical form.

7.6.4 Problem. Suppose a 5×5 matrix A has characteristic polynomial $(x-2)^2(x-7)^3$. What are the possibilities for its Jordan canonical form?

7.6.5 Problem. Let A be a matrix whose characteristic polynomial decomposes into linear factors. Let $(x-\lambda)^m$ divide p_A (and m is maximal with that property). So m is the multiplicity of the root λ of p_A.

Show that
$$\dim V(x-\lambda) = m.$$

7.6.6 Problem. Show that two complex matrices are similar iff their Jordan canonical forms coincide.

7.6.7 Problem. Show that for $A \in M_n(\mathbb{C})$, A and A^T have the same Jordan canonical form. Conclude that A and A^T are similar.

Here is a list of facts, connecting the minimal polynomial, the characteristic polynomial, and the Jordan canonical form over the complex numbers. (The same holds for all algebraically closed fields.)

7.48 Facts. Let $A \in M_n(\mathbb{C})$ with minimal polynomial m_A and characteristic polynomial p_A. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of A.

For each i let $m_i = \dim V(x-\lambda_i)$ and n_i be the largest size of a Jordan block for λ_i.

a. $p_A = (x-\lambda_1)^{m_1}(x-\lambda_2)^{m_2}\cdots(x-\lambda_1)^{m_k}$.

b. $m_A = (x-\lambda_1)^{n_1}(x-\lambda_2)^{n_1}\cdots(x-\lambda_k)^{n_k}$.

c. $m_A \mid p_A$ and hence $p_A(A) = 0$ (alternate proof of Cayley-Hamilton).

d. A is diagonalizable if and only if all roots of m_A are distinct, ie. iff all $n_i = 1$.

e. $\mathbb{C}^n$ is cyclic for T_A iff $m_A = p_A$ (ie. if $n_i = m_i$ for all i).

f. The maximal number of linearly independent eigenvectors belonging to λ_i is the number of Jordan blocks for that same eigenvalue.

g. The dimension of the null space of A is the number of Jordan blocks for the eigenvalue $\lambda = 0$ and of course equal to 0 if $\lambda = 0$ is not an eigenvalue.

h. A is nilpotent iff its only eigenvalue is 0.

i. The trace of A is $\sum_i m_i\lambda_i$ and the determinant of A is $\lambda^{m_1}\cdots\lambda_k^{m_k}$.

Proof. Exercise. Here is a hint: all these facts only depend on the Jordan canonical form of A. So you may assume that A is in Jordan canonical form. □

Remark. In the algorithm for finding the JCF, we sometimes have to extend a basis of a subspace to a basis of some other subspace.

So suppose $\mathcal{B} = (w_1, w_2, \ldots, w_r)$ is a basis for a subspace $W \subset \mathbb{F}^n$, and $W \subset U$, where $U = \operatorname{Span}(u_1, u_2, \ldots, u_s) \subset \mathbb{F}^n$ is also a subspace. How to find a basis $\mathcal{C}$ for W that contains $\mathcal{B}$?

Here is one way: Let $A = \begin{bmatrix} w_1 & w_2 & \ldots & w_r \end{bmatrix}$, and $B = \begin{bmatrix} u_1 & u_2 & \ldots & u_s \end{bmatrix}$. Then let R be the reduced row echelon form of

$$\begin{bmatrix} A & | & B \end{bmatrix}.$$

Since the RREF has a left-ward bias, R will be of the form

$$\begin{bmatrix} e_1 & e_2 & \ldots & e_r & B' \end{bmatrix}$$

where B' is some matrix. This is true because R will have the RREF of A on the left, which is given by the e_i. Thus, the columns of B which correspond to pivot columns in R, give the additional elements (among the u_i) needed to extend $\mathcal{B}$.

So for instance to find a basis for all of $\mathbb{F}^n$, that contains $\mathcal{B}$, apply this to $\begin{bmatrix} A & | & I_n \end{bmatrix}$.

Example. Find a basis for $\mathbb{R}^4$ that contains

$$v = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ and } w = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

To do this, compute

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

From which we see that columns $1, 2, 3$, and 5 are pivot colums. Thus the basis is (v, w, e_1, e_3). And indeed, since $e_2 \in \operatorname{Span}(v, w)$, we cannot take e_2.

7.6.2. *Excursion: the invariant factor theorem

The main result of this section is to give the prime factorization of the characteristic polynomial p_T in $\mathbb{F}[x]$ a meaning in terms of decomposing V into invariant subspaces. The upshot will be:

7.49 Theorem (Invariant Factor Theorem)**.** *Let $T \in \operatorname{End}(V)$. Then there exist $v_1, v_2, \ldots, v_p$ of orders $d_1, d_2, \ldots, d_p$, respectively, such that*

$$V = \langle v_1 \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_p \rangle$$

and $d_1 \mid d_2 \mid d_3 \mid \cdots \mid d_p$.

Moreover, the d_i are uniquely determined.

It is not hard to see that this Theorem is logically equivalent to the so called *Elementary Divisor Theorem* stated and proved below.

Note that in general the d_i are not irreducible. If we translate this into matrix language, (by choosing a basis of V) we get

Example. Let us translate what the elementary divisor theorem has to say about matrices: If $A \in M_n(\mathbb{F})$ the theorem applied to the matrix transformation T_A says the following:

Matrix version: Let $A \in M_n(\mathbb{F})$, then there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that

$$P^{-1}AP = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_p \end{bmatrix} \tag{7.8}$$

where A_i is the *companion matrix* of a polynomial $d_i \in \mathbb{F}[x]$ and $d_1 \mid d_2 \mid \cdots \mid d_p$:

$$A_i = \begin{bmatrix} 0 & \dots & 0 & -a_0 \\ 1 & \ddots & \vdots & \vdots \\ & \ddots & 0 & -a_{n-2} \\ & & 1 & -a_{n-1} \end{bmatrix}$$

where $d_i = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$.

7.6.8 Problem. Show that if A is the companion matrix of $f = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$, $p_A = f$.

Method 1: Direct approach by induction:

$$\det(xI - A) = \det \begin{bmatrix} x & \dots & 0 & a_1 \\ -1 & \ddots & \vdots & \vdots \\ & \ddots & x & a_{n-2} \\ & & -1 & x + a_{n-1} \end{bmatrix} - (-1) \cdot \det \begin{bmatrix} 0 & \dots & 0 & +a_0 \\ 1 & x & \vdots & \vdots \\ & \ddots & x & a_{n-2} \\ & & 1 & x + a_{n-1} \end{bmatrix}$$

Note that the first determinant is the characteristic polynomial of the companion matrix of $a_1 + a_2x + \cdots + a_{n-1}x^{n-2} + x^{n-1}$. Thus, the result follows if the second determinant is a_0.

For this, observe that it is the determinant of an $(n-1) \times (n-1)$-matrix and hence

$$\det \begin{bmatrix} 0 & \dots & 0 & +a_0 \\ 1 & x & \vdots & \vdots \\ & \ddots & x & a_{n-2} \\ & & 1 & x + a_{n-1} \end{bmatrix} = (-1)^{n-2}a_0 \det \begin{bmatrix} -1 & x & & \\ & -1 & x & \\ & & \vdots & x \\ & & & -1 \end{bmatrix} = (-1)^{n-2}a_0(-1)^{n-2} = a_0$$

as needed.

Of course, to do a proper induction, we must verify the base case $n = 1$ (which is easy). In our proof we also rely on a computation of a determinant of an $(n-2) \times (n-2)$-determinant, which makes sense only if $n > 2$. Thus, we should also check the case $n = 2$ directly. But that is easy.

Method 2: Here is a conceptual method using Caley-Hamilton: Note that if $T = T_A \colon \mathbb{F}^n \to \mathbb{F}^n$, then $\mathbb{F}^n$ is cyclic. Indeed, $e_1, e_2 = Te_1, \ldots, e_n = Te_{n-1} = T^{n-1}e_1$ form a basis.

Notice also that $T^n e_n = -a_0 e_1 - a_1 e_2 - \cdots - a_{n-1}e_n$, so that

$$\sum_{i=0} a_i T^i e_1 = 0$$

which means that the order of e_1 divides f. But $\langle e_1 \rangle = \mathbb{F}^n$ so the order of e_1 is the minimal polynomial of T and has degree n. It follows that $m_T = f$. But also $p_T = f$ because $m_T \mid p_T$ and p_T has degree n.

Before we look at some examples let us now prove the Invariant Factor Theorem.

Proof of Theorem 7.49. We know that $V = V(p_1) \oplus V(p_2) \oplus \cdots \oplus V(p_t)$. Also, we can further decompose each of the $V(p_i)$ according to Theorem 7.40. So

$$V(p_i) = \langle v_{i1} \rangle \oplus \langle v_{i2} \rangle \oplus \cdots \oplus \langle v_{ir_i} \rangle$$

with $v_1, v_2, \ldots, v_{ir_i}$ elements of orders $p_i^{e_{i1}}, p_i^{e_{i2}}, \ldots, p_i^{e_{ir_i}}$, respectively. Also we may assume that $e_{i1} \geq e_{i2} \leq \cdots \geq e_{ir_i}$. The trick consists of grouping various of the v_{ij} appropriately. In fact, there is one and only one way to do this: Let us define $e_{ij} = 0$ whenever $j > r_i$ (just to give $p_i^{e_{ij}}$ some a meaning for all j). Let $D_1 = p_1^{e_{11}} p_2^{e_{21}} \cdots p_t^{e_{t1}}$ be the product of the "maximal" orders in each $V(p_i)$. Put $D_2 = p_1^{e_{12}} p_2^{e_{22}} \cdots p_t^{e_{t2}}$, and continue

$$D_j := p_1^{e_{1j}} p_2^{e_{2j}} \cdots p_t^{e_{tj}}$$

for $j = 1, 2, \ldots, \max_i r_i =: p$. Then clearly $D_p \mid D_{p-1} \mid \cdots \mid D_1$, so we define $d_j := D_{p+1-j}$. What vector should we associate to d_j? Again, define $v_{ij} = 0$ if $j > r_i$ and for $j = 1, 2, \ldots, p$ put

$$w_j = v_{1j} + v_{2j} + \cdots + v_{tj}.$$

Claim: The order of w_j is D_j.

Suppose $f(T)w_j = 0$; thus $f(T)w_j = \sum_i f(T)v_{ij}$. Also, $f(T)v_{ij} \in \langle v_{ij} \rangle \subseteq V(p_i)$ because $V(p_i)$ is T- and hence $f(T)$-invariant.

Necessarily this means that $f(T)v_{ij} = 0$. In particular, this means that the order of v_{ij} divides f. Hence $p_i^{e_{ij}}$ divides f (which is true also if $v_{ij} = 0$ for then its order is simply 1).

Taken together this means D_j divides f (since the p_i are coprime; cf. Problem 6.4.5). On the other hand, $D_j(T)w_j = 0$, because $p_i^{e_{ij}}(T)$ kills all v_{ij}. This shows the claim. □

Now all that remains to show is that

$$V = \langle w_1 \rangle \oplus \langle w_2 \rangle \oplus \cdots \oplus \langle w_p \rangle.$$

But note that the right hand side is really a direct sum because, V is the direct sum of all the $\langle v_{ij} \rangle$. It thus suffices to show that its dimension is equal to $\dim V$. But this follows easily from the fact that $\sum_{i=1}^{p} \deg D_i = \deg p_T = \dim V$. □

Note as a consequence we have that $v_{ij} \in \langle w_j \rangle$.

7.6.3. *Excursion: Finding the v_i

This leaves us with the question of how we can find the v_i of Corollary 7.42 or Theorem 7.40. It turns out that with the tools available to us, carefully analyzing the proof of Lemma 7.40 is the thing to do:

- Suppose we are given a linear transformation $T\colon V \to V$. By choosing a basis, we immediately translate into a problem of a matrix transformation $T_A\colon \mathbb{F}^n \to \mathbb{F}^n$ so we may assume that $V = \mathbb{F}^n$ and $T = T_A$.

- Next, we compute p_A and its prime decomposition $p_A = p_1^{b_1} \cdots p_t^{b_t}$. Note that we could now compute m_A, but it turns out this is unnecessary.

- Now, for each i, compute $V(p_i) = \mathcal{N}(p_i(A)^k)$ for some $k \geq 0$. In fact, since $p_i^{b_i}$ is the characteristic polynomial of $T|_{V(p_i)}$, it follows

  $$V(p_i) = \mathcal{N}(p_i(A)^{b_i}),$$

  a space of dimension $\deg p_i^{b_i} = b_i \deg p_i$.

- We may now assume that A is in block form with blocks A_i along the diagonal, the matrix of T on $V(p_i)$ with respect to some basis of our choice (computed in the previous step). So for the remainder we assume that $V = V(p_i)$ and $A = A_i$.

- Given a basis of V, identify an element v of maximal order (equal to the minimal polynomial of A on V) $p_i^{a_i}$. Let $d = \deg p_i^{a_i}$.

  Compute $v, Av, A^2v, \ldots, A^{d-1}v$ and extend this to a basis of V. This also gives us a basis of V/W where $W = \mathrm{Span}(v, Av, \ldots, A^{d-1}v)$.

- Apply the algorithm to V/W, resulting in $\overline{v}_2, \overline{v}_3, \ldots, \overline{v}_r$. Choose $v_i \in V$ such that $v_i + W = \overline{v}_i$.

  Note that the orders of v_i is of the form $p_i^{e_i}$ for some e_i but may be bigger than the order $p_i^{\overline{e}_i}$ of $\overline{v}_i$.

Observe: The following set is a basis of V:

$$(v,\dots,A^{d-1}v,v_2,Av_2,\dots,A^{d_2-1}v_2,\dots,v_r,\dots,A^{d_r-1}v-r)$$

where $d_i=\deg(p_i^{\overline{e}_i})-1$ (why?).

The matrix of T_A with respect to this basis has the form

$$\begin{bmatrix} C_1 & * & * & * & * \\ & C_2 & & & \\ & & & \ddots & \\ & & & & C_r \end{bmatrix}$$

where C_j is the companion matrix of p_i^d (for $j=1$) and of $p_i^{\overline{e}_i}$ (for $i>1$).

The remaining task is now to clear the positions with a $*$.

- So suppose $p_i(T)^{\overline{e}_i}v_i\neq 0$. According to our proof we need to find $w_i\in\langle v_1\rangle$ such that $p(T)^{\overline{e}_i}w_i=p(T)^{\overline{e}_i}v_i$. This corresponds to solving a system of linear equations.

  Replace v_i by v_i-w_i and we are done.

This is not a really efficient method. There are better algorithms which use more ring theory.

7.7. *Excursion: Applications of the elementary divisor theorem

In this section we investigate various applications of the EDT both within and without mathematics.

7.7.1. *Excursion:The rational canonical form.

Suppose $A\in M_n(\mathbb{F})$. Applying the invariant factor theorem to A we get the following:

We can write $V=\mathbb{F}^n=\langle v_1\rangle\oplus\cdots\oplus\langle v_p\rangle$ with the d_i order of v_i and $d_1\mid d_2\mid\cdots\mid d_p$ as in (7.8). Then A is similar to a matrix

$$A'=\begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_p \end{bmatrix}$$

with A_i the companion matrix for d_i. The matrix A' is called the *rational canonical form* of A. It is uniquely determined by A. What can we use this for? Here are two applications:

Theorem (Cayley-Hamilton Theorem). $p_A=d_1d_2\cdots d_p$. *In particular* $m_A=d_p$ *divides* p_A *and* $p_A(A)=0$.

Proof. We know that $p_{A_i} = d_i$. Also, $p_{A'} = p_{A_1} p_{A_2} \cdots p_{A_p}$. Since $p_A = p_{A'}$ we are done. (It is clear that $m_A = d_p$ since $d_p(A)w_i = 0$ for all i.) □

Here is an even more striking application:

Theorem. *Let $\mathbb{F} \subseteq L$ be an inclusion of fields. Suppose $A, B \in M_n(\mathbb{F}) \subseteq M_n(L)$. Suppose $A \sim B$ over L. Then $A \sim B$ over $\mathbb{F}$.*

In other words, if there is $P \in \mathrm{GL}_n(L)$ such that $A = PBP^{-1}$, then there is also $Q \in \mathrm{GL}_n(\mathbb{F})$ such that $A = QBQ^{-1}$.

Proof. The rational canonical forms of A and B do not depend on the field.

Indeed, Let A' be the rational canonical form of A over $\mathbb{F}$. Then $A \sim A'$ in $M_n(\mathbb{F})$. Since A' is in rational canonical form, and since $A' \in M_n(L)$, A' is also the rational canonical form of A as an element of $M_n(L)$. If B' is the rational canonical form of B (over $\mathbb{F}$), the same reasoning shows that B' is the rational canonical form of B in $M_n(L)$ and since $A \sim B$ as elements of $M_n(L)$ it follows that $A' = B'$. But this means $A \sim B$ in $M_n(\mathbb{F})$ since $\sim$ is a transitive relation (it is an equivalence relation). □

The main advantage of the invariant factor theorem over the elementary divisor theorem is that the d_i do not depend on the field $\mathbb{F}$ (ie. if we enlarge $\mathbb{F}$, the d_i remain the same). But the elementary divisors change since irreducible polynomials may no longer be irreducible over a large field.

*Excursion: The Jordan decomposition of a linear transformation

As a consequence of the Jordan canonical form, we obtain the following decomposition for linear transformations of complex vector spaces:

Theorem (Jordan decomposition theorem)**.** *Let $T \in \mathrm{End}(V)$ where V is a complex vector space of dimension n. There exist $D, N \in \mathrm{End}(V)$ with D diagonalizable and N nilpotent such that $DN = ND$ and*

$$T = D + N.$$

Proof. Suppose A is the matrix of T with respect to some basis. After a suitable change of basis we may assume that A is in Jordan canonical form. Let D be the linear transformation whose matrix is given by the diagonal part of A and let N be the transformation whose matrix is $A - D$. Then $N^n = 0$ and also $DN = ND$ because the matrices commute. D is diagonalizable since its matrix is diagonal. □

7.7.1 Problem. Show that the transformations D, N are uniquely determined.

7.7.2. *Excursion: Systems of homogeneous linear differential equations

In many applications a system of differential equations of the following form appears:

$$\begin{aligned} x_1' &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ x_2' &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ &\vdots \\ x_n' &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{aligned} \tag{7.9}$$

where the $a_{ij} \in \mathbb{C}$ and the x_i are (unknown) functions $I \to \mathbb{C}$ where I is some interval in $\mathbb{R}$ (or all of $\mathbb{R}$). Note that we could restrict ourselves also to real valued functions (and the a_{ij} to real numbers).

Such a system is called an *homogeneous system of linear differential equations with constant coefficients* ("homogeneous" because on the right hand side, there are no nonzero constants.) To solve such a system means to specify n functions x_i such that the equations (7.9) hold (for each $t \in I$). Note that a function $f \colon I \to \mathbb{C}$ is *differentiable* if $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ both are differentiable functions on I. If this is the case f' is defined as $f'(t) = \mathrm{Re}(f)'(t) + \mathrm{Im}(f)'(t) \cdot i$.

(Note this is consistent: f is differentiable if and only if for all $t_0 \in I$,

$$\lim_{t \to t_0} \frac{f(t) - f(t_0)}{t - t_0}$$

exists (in $\mathbb{C}$).)

Note that we can write

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

for the associated function $I \to \mathbb{C}^n$ with value $[x_1(t)\, x_2(t)\, \ldots\, x_n(t)]^T$ at $t \in I$. If we define differentiation of such a function componentwise then to solve (7.9) means to find X such that

$$X' = AX.$$

where A is the $n \times n$ matrix with coefficients a_{ij}.

Generalizing a little bit further suppose $A(t)$ is a matrix depending on $t \in I$ (that is its entries are functions $I \to \mathbb{C}$). Its *derivative* $A'(t)$ is defined by componentwise differentiation.

7.50 Lemma (Leibniz rule). *Let $A, B \colon I \to M_n(\mathbb{C})$ be two differentiable matrix valued functions. Then*

$$(AB)'(t) = A'(t)B(t) + A(t)B'(t).$$

In particular, if $A'(t)$ and $A(t)$ commute for all $t \in I$, $(A^k)'(t) = kA^{k-1}(t)A'(t)$.

Proof. Calculus. (Differentiate each entry of AB.) The final assertion follows by induction. □

Let us now be "naive:" Having the Leibniz rule in calculus gave us the exponential function: solving $f'(t) = af(t)$ gives rise to $f(t) = e^{at}$, with e^x defined as $e^x = \sum_{n\geq 0} 1/n!x^n$. Let's see what happens we we define

$$e^A := \exp(A) := \sum_{k\geq 0} \frac{1}{k!} A^k.$$

First of all what does this mean? Of course it should mean that

$$e^A = \lim_{K\to\infty} \sum_{k=0}^{K} \frac{1}{k!} A^k.$$

To define this, we define a sequence in $M_n(\mathbb{C})$ as *convergent*, if it converges componentwise. From this, it follows immediately that *every Cauchy sequence converges*. Note that if a is an upper bound for the absolute value of an entry of A then the entries of A^2 are bounded by na^2, and inductively, those of A^k are bounded by $n^{k-1}a^k$ which is certainly bounded by $(na)^k$. Thus to show that the above is a Cauchy sequence observe that the maximum entry of

$$\sum_{k=K_1}^{K_2} \frac{1}{k!} A^k$$

is bounded by

$$\sum_{k=K_1}^{K_2} \frac{1}{k!} (na)^k$$

which is arbitrarily small if K_1 is large enough.

Slightly more technical, but not really hard, is to show that

$$F(t) := \exp(tA) = \sum_{k\geq 0} \frac{1}{k!} t^k A^k$$

is differentiable on I with derivative

$$F'(t) = \sum_{k\geq 0} \frac{1}{k!} k t^{k-1} A^k = A \exp(At).$$

Also, it is not difficult to verify that if $AB = BA$ then

$$e^{A+B} = e^A e^B.$$

(Essentially the same proof that shows this in the case of real or complex numbers, also works here.)

From this it follows that e^A is invertible and

$$(e^A)^{-1} = e^{-A}.$$

7.51 Theorem. *Let I contain 0. The columns of e^{At} form a basis of the solution space of* (7.9).

In particular, for each $X_0 \in \mathbb{C}^n$ there exists a unique solution X such that $X(0) = X_0$.

Proof. Let X be any solution. We will first show that then $X = e^{tA}X(0)$. So put $Y(t) = e^{-tA}X$. Then $Y'(t) = -Ae^{-tA}X(t) + e^{-tA}AX(t) = 0$ because e^{-tA} and A commute. But this means Y is constant; $Y(0) = X(0)$; and $X(t) = e^{tA}Y$. This shows that the columns of e^{tA} form a generating set of the solution space.

Also, suppose $X_0 \in \mathbb{C}^n$ is such that$e^{tA}X_0 = 0$. Then since e^{tA} is invertible (for each t), it follows that $X_0 = 0$ so e^{tA} has linearly independent columns (even pointwise!). □

Note if we replace A by a similar matrix $B = P^{-1}AP$ then the solution X behave like a coordinate vector, that is, the solutions of $Y' = BY$ are of the form $P^{-1}X$ where X solves $X' = AX$.

In particular, if A is diagonalizable (we say the system is *decoupled*), solving it is easy: if A is similar to D then the solutions of $Y' = YD$ are given by the columns of

$$e^{tD} = \begin{bmatrix} e^{\lambda_1 t} & & & \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ & & & e^{\lambda_n t} \end{bmatrix}$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvectors of A. It then follows that the the solutions X are of the form PY.

What if A is not diagonalizable? In this case, we may use the Jordan canonical form. Suppose B is in Jordan canonical form. It suffcices to treat each Jordan block individually. Now check for yourself that

$$(tJ)^k = \begin{bmatrix} \lambda^k t^k & k\lambda^{k-1}t^k & \binom{k}{2}\lambda^{k-2}t^k & \dots & t^k & 0 & \dots & 0 \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ & & & \ddots & \ddots & \ddots & \ddots & t^k \\ & & & & \ddots & \ddots & \ddots & \vdots \\ & & & & & \ddots & \ddots & \binom{k}{2}\lambda^{k-2}t^k \\ & & & & & & \ddots & k\lambda^{k-1}t^k \\ & & & & & & & \lambda^k t^k \end{bmatrix}$$

from which one easily deduces that for a $n \times n$ Jordan block J,

$$e^{tJ} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & \frac{1}{2!}t^2e^{\lambda t} & \dots & \frac{1}{(n-1)!}t^{n-1}e^{\lambda t} \\ & \ddots & \ddots & \ddots & \vdots \\ & & e^{\lambda t} & te^{\lambda t} & \frac{1}{2!}t^2e^{\lambda t} \\ & & & e^{\lambda t} & te^{\lambda t} \\ & & & & e^{\lambda t} \end{bmatrix}$$

Note that this can also be obtained as follows: Write $J = D + N$ with $D = \lambda I$ diagonal. Then since $DN = ND$ we have

$$e^{tJ} = e^{tD}e^{tN}$$

but N is nilpotent, so $N^n = 0$ (n being the size of the J) and hence

$$e^{tN} = I + tN + \frac{t^2}{2}N^2 + \dots + \frac{t^{n-1}}{(n-1)!}N^{n-1}.$$

$$e^{tJ} = \begin{bmatrix} e^{\lambda t} & & & \\ & e^{\lambda t} & & \\ & & \ddots & \\ & & & e^{\lambda t} \end{bmatrix} \begin{bmatrix} 1 & t & \dots & \frac{t^{n-1}}{(n-1)!} \\ & \ddots & \ddots & \vdots \\ & & 1 & t \\ & & & 1 \end{bmatrix}$$

Finally, note that for $A \in M_n(\mathbb{C})$ and $P \in \mathrm{GL}_n(\mathbb{C})$ we have

$$e^{P^{-1}AP} = \sum_{k=0}^{\infty} \frac{1}{k!}(P^{-1}AP)^k = \sum_{k=0}^{\infty} \frac{1}{k!}P^{-1}(A^k)P = P^{-1}(\sum_{k=0}^{\infty} \frac{1}{k!}A^k)P.$$

(Note that "formally" this is clear, but one needs to show that if S_K is a convergent sequence of matrices with limit S_∞, say, then so is $P^{-1}S_KP$ and the limit is $P^{-1}S_\infty P$. But this is easy.

Thus, to compute e^{tA} for any matrix A, we compute the Jordan canonical form $A' = P^{-1}AP$ of A (and P as well) and then have

$$e^{tA} = Pe^{tA'}P^{-1}.$$

The matrix function $e^{tA'}$ we can compute by observing that if

$$A' = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_r \end{bmatrix}$$

with Jordan blocks J_i, then

$$e^{tA'} = \begin{bmatrix} e^{tJ_1} & & & \\ & e^{tJ_2} & & \\ & & \ddots & \\ & & & e^{tJ_r} \end{bmatrix}$$

Remark. Note that if v is an eigenvector for $A \in M_n(\mathbb{C})$, then $e^A v = (\sum_{k\geq 0} \frac{1}{k!} A^k)v = \sum_{k\geq 0} \frac{1}{k!}(A^k v)$, which is (that needs a proof similar to the the proof of $e^{P^{-1}AP} = P^{-1}e^A P$). But this is

$$\sum_{k\geq 0} \frac{1}{k!} \lambda^k v = e^\lambda v.$$

So the numbers e^λ are eigenvalues of e^A. In fact, if A is in Jordan canonical form we know that e^A is upper triangular, and then $\det e^A = e^{\lambda_1} e^{\lambda_2} \cdots e^{\lambda_n}$ where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the (not necessarily distinct) eigenvalues of A, and the e^{λ_i} are precisely the eigenvalues of e^A. In particular,

$$\det e^A = e^{\operatorname{trace}(A)}.$$

7.52 Example. Before we study a concrete example let us see how this helps us to solve a differential equation of higher order.

For this, suppose we want to solve

$$y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_1 y^{(1)} + a_0 y = 0 \tag{7.10}$$

where we define

$$y^{(i)} = \frac{d^i y}{dt^i}$$

for the ith derivative of $y = y(t)$.

An equation as in (7.10) is called a *linear differential equation of order n with constant coefficients*.

We can turn this into a system of equations of order 1 by solving the following system (where we put $x_1 = y$):

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= x_3 \\ &\vdots \\ x_{n-1}' &= x_n \\ x_{n-1}' &= -a_{n-1}x_{n-1} - a_{n-2}x_{n-2} - \cdots - a_1 x_2 - a_0 x_1 \end{aligned} \tag{7.11}$$

If $X(t)$ is any solution of this system, then $y = x_1$ will be a solution of the original differential equation of order n.

But note the matrix A of the (7.11) is

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & -a_1 & \ldots & -a_{n-1} \end{bmatrix}$$

which we recognize as the transpose of the companion matrix of $f = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. In particular, $p_A = p_{A^T} = f$. Thus, the columns of e^{tA} will form a basis of the solution space (as a subspace of the space of $\mathbb{C}^n$-valued functions on $\mathbb{R}$). Let

$$e^{tA} = [X_1 \, X_2 \, \ldots \, X_n]$$

where $X_i = X_i(t)$ is a function with values in $\mathbb{C}^n$.

Then the solution space for the solutions y of the original problem has a basis given by the first entries of all the X_i. And these entries will be linearly independent (since they determine all the remaining entries of the X_i), so if x_{ij} is the j-th entry of X_i, and if $\sum_i c_i x_{i1}(t) = 0$ for all t, then

$$0 = \sum_i c_i x'_{i1}(t) = \sum_i c_i x_{i2}(t)$$

and so on so in the end we get

$$\sum_i c_i X_i(t) = 0$$

for all t, but the X_i are linearly independent and hence the c_i are all zero.

How do these entries look like? Let A' be the Jordan canonical form of A. Now note that since A and A^T have the same eigenvalues, and moreover the dimensions $\dim \mathcal{N}(A - \lambda I)^e$ and $\dim \mathcal{N}(A^T - \lambda I)^e$ all coincide since $(A^T - \lambda I)^e = ((A - \lambda I)^e)^T$, it follows that A and A^T have the same Jordan canonical form.

Recall that if $B \in M_n(\mathbb{C})$ is the companion matrix of a polynomial, then $\mathbb{C}^n$ is cyclic for T_B (cf. the remarks following the Invariant Divisor Theorem 7.49; see also 7.48). Thus, the Jordan canonical form will contain a single block for each eigenvalue.

Let J be the Jordan canonical form of A. Then $A = PJP^{-1}$ and so

$$e^{tA} = Pe^{tJ}P^{-1}.$$

We are only interested in the first row of e^{tA} (as this gives us a basis for the solution space of our original problem). To shorten a lengthy discussion observe that the entries if e^{tJ} are of the form $\frac{t^k}{k}e^{\lambda t}$ where k is at most the multiplicity of $\lambda - 1$. Thus, the first row of e^{tA} will be given by linear combinations of the following n functions

$$e^{\lambda_1 t}, te^{\lambda_1 t}, \ldots, t^{m_1 - 1}e^{\lambda_1 t}, \ldots, e^{\lambda_k t}, \ldots, t^{m_k - 1}e^{\lambda_k t}$$

Since our solution space is n-dimensional, as we determined, it must be actually equal to the span of these functions and indeed, every function here is a solution of the original problem.

7.53 Example. Let us consider the harmonic oscillator. That is, a system that obeys an equation

$$y^{(2)} + ky = 0.$$

(For instance $y(t)$ could be the elongation of a spring, then $y^{(2)}$ is the acceleration at time t, and hence proportional to the force, which in turn is proportional to the elongation.)

Here

$$A = \begin{bmatrix} 0 & 1 \\ -k & 0 \end{bmatrix}$$

The characteristic polynomial is $x^2 + k$. In most practical applications, $k > 0$ (the force goes against elongation; energy is preserved!). So the eigenvalues are $\pm i\sqrt{k}$ which are complex. Let us call $\omega = \sqrt{k}$ the *frequency*. By the above, we obtain the two solutions

$$e^{\pm i\omega t}.$$

But our original problem was real! And indeed, if the matrix A has real entries, then there is a basis for the solution space in the real valued functions: simply pick real and imaginary parts (which will also be solutions).

So recall if $a + bi \in \mathbb{C}$, then

$$e^{a+bi} = e^a e^{bi} = e^a(\cos b + i \sin b).$$

Here we are left with

$$\cos(\omega t) \quad \text{and} \quad \sin(\omega t).$$

7.54 Example. As a final example let us consider a harmonic oscillator that is also dampened. That is, the acceleration of the particle is decreased by a factor proportional to the speed of the particle. In other words, we have

$$y^{(2)} + dy^{(1)} + ky = 0.$$

Usually here $d, k > 0$ (higher speed (ie. a higher value of $y^{(1)}(t)$) should mean lower not faster acceleration). This is for instance the model of a mass point attached to a spring with constant k, where the point is moving through a viscous fluid.

The characteristic polynomial is

$$x^2 + dx + k$$

which has zeros $\lambda_{1/2} = -d/2 \pm \sqrt{d^2/4 - k}$. Now let us consider several cases: if $k > d^2/4$, then let us put $\omega = \sqrt{k - d^2/4}$, then the solutions are of the form $x(t) = ae^{-dt/2}e^{i\omega t} + be^{-dt/2}e^{-i\omega t}$. If only real solutions are interesting then this means

$$x(t) = e^{-dt/2}(a \cos \omega t + b \sin \omega t).$$

So the solutions are oscillating with exponentially decreasing amplitude (provided the damping is positive).

If $k < d^2/4$ then both eigenvalues are real, and the solution is of the form

$$x(t) = ae^{\lambda_1 t} + be^{\lambda_2 t}$$

so here nothing oscillates and the motion quickly dies (note $\lambda_i < 0$ unless $k = 0$).

Finally, if $k = d^2/4$ we only have a single eigenvalue, namely $\lambda = -d/2$. Again, in many practical applications, $d > 0$. Then a typical solution has the form

$$x(t) = e^{-dt/2}(a + bt).$$

So here nothing "oscillates," really.

7.55 Example. Let us finish this discussion with an example of a first order coupled system. Let $X(t)$ denote some population at time t,

$$X(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} \in \mathbb{R}^2$$

where the population is partitioned into two classes which interact. For instance, $x(t)$ could be the number of adults and $y(t)$ the number of offspring at time t. Or, $x(t)$ could be the number of healthy people and $y(t)$ the number of people suffering from some illness. In line with this example let us assume that per time, $py(t)$ infected persons die and that $hy(t)$ persons become healthy again. Finally suppose healthy people become ill at a rate $vx(t)$ (which here is independent of the number of people already having the illness). This is of course an extremely simplified model.

Thus we have

$$y'(t) = vx(t) - (p + h)y(t).$$

And then necessarily,

$$x'(t) = -vx(t) + hy(t).$$

Thus,

$$X'(t) = AX(t)$$

where

$$A = \begin{bmatrix} -v & h \\ v & -p - h \end{bmatrix}$$

Note that here $p, v, h \geq 0$. Let us assume that the virus is benign in the sense that $p = 0$. Then $p_A = x^2 + (v + h)x$. So that the eigenvalues are

$$0 \text{ and } -(v + h).$$

Note that an eigenvector for $-(v+h)$ is given by $v_1 = e_1 - e_2$. Assuming $v, h > 0$, $v_2 = he_1 + ve_2$ is an eigenvector for 0.

Thus, if we put

$$P = \begin{bmatrix} 1 & h \\ -1 & v \end{bmatrix}$$

then

$$e^{tA} = P \begin{bmatrix} e^{-(v+h)t} & \\ & 1 \end{bmatrix} P^{-1}$$

Note that the span of the columns of e^{tA} is the same as the span of the columns of $e^{tA}P$ since P is invertible and right multiplication corresponds to column operations.

Thus the columns of $e^{tA}P$ also form a basis of the solution space and these columns are given by

$$e^{-(h+v)t}\begin{bmatrix}1\\-1\end{bmatrix} \text{ and } \begin{bmatrix}h\\v\end{bmatrix}$$

Thus, if we start with a healthy population say, (so $X(0)=e_1$) then, as $e_1 = \frac{v}{v+h}v_1 + \frac{1}{v+h}v_2$.

$$X(t) = e^{tA}e_1 = \frac{ve^{-(v+h)t}}{v+h}v_1 + \frac{1}{v+h}v_2.$$

Thus, for large t, we will approximate the state

$$\frac{1}{v+h}v_2 = \begin{bmatrix}\frac{h}{h+v}\\\frac{v}{v+h}\end{bmatrix}.$$

What happens if $p>0$? Then $p_A = x^2 + (v+p+h)x + vp$ and we get eigenvalues

$$\lambda_{1/2} = \frac{-(v+p+h) \pm \sqrt{v^2+p^2+h^2+2hp+2vh-2vp}}{2}$$

Note that both eigenvalues are negative and that this clearly shows that humanity is doomed (as now $X(t)\to 0$ for large t) if there exists a virus abiding by this model with $p>0$.

Of course this shows the limits of the model (for instance, t is a continuous parameter, so we get infinitely many values for $X(t)$ but of course $X(t)$ counts something, so the entries should be integers, but there are also more serious concerns: people are born, h, v, p all may be time dependent etc.).

8. Bilinear forms

In this chapter, we will study the notion of a *bilinear form*, mainly on vector spaces over $\mathbb{R}$ and $\mathbb{C}$.

8.1. Definition

The prototype of any bilinear form is the *dot-product* on $\mathbb{R}^n$: For two column vectors, $X, Y \in \mathbb{R}^n$, one defines

$$(X \cdot Y) = X^T Y$$

where the right hand side is a 1×1 matrix and hence may be identified with a real number.

A straight forward computation shows that the dot product has the following properties: It is *bilinear*: for all $X, Y, Z \in \mathbb{R}^n$ and all $c \in \mathbb{R}$ we have

$$((X + Y) \cdot Z = (X \cdot Z) + (Y \cdot Z)$$

$$(X \cdot (Y + Z)) = (X \cdot Y) + (X \cdot Z)$$

and

$$((cX) \cdot Y) = (X \cdot (cY)) = c(X \cdot Y).$$

In addition, the dot product has two more properties: it is *symmetric*, ie.

$$(X \cdot Y) = (Y \cdot X)$$

and *positive definite*: $(X \cdot X) \geq 0$ for all X and $(X \cdot X) = 0$ if and only if $X = 0$.

Because of this latter property, we can define the *length* of a vector $X \in \mathbb{R}^n$ as

$$|X| := \sqrt{(X \cdot X)}.$$

As outlined in several homework assignments, this allows us to do geometry in $\mathbb{R}^n$: We can define the *angle* between two nonzero vectors simply as a real number θ such that

$$(X \cdot Y) = |X||Y| \cos \theta.$$

Such a θ exists and is defined up to adding integer multiples of 2π since for all X, Y we have[1]

$$|(X \cdot Y)| \leq |X||Y|.$$

[1] As for a proof, note that it suffices to check that $(X \cdot Y)^2 \leq (X \cdot X)(Y \cdot Y)$. The left hand side is equal to $(\sum_i x_i y_i)^2 = \sum_{i,j} x_i x_j y_i y_j$ whereas the right hand side is $(\sum_i x_i^2)(\sum_i y_i^2) = \sum_{i,j} x_i^2 y_j^2$. For each pair

Also, the *distance* between X and Y is then defined as

$$d(X,Y) = |X - Y|$$

which satisfies the *triangle inequality*: $d(X,Y) \leq d(X,Z) + d(Z,Y)$ for all $X, Y, Z \in \mathbb{R}^n$.

With these definitions in place, one can show that $\mathbb{R}^2$, together with this concept of angle and length satisfies all axioms of plane Euclidean geometry. Every statement that can be proven in Euclidean geometry, holds for points, lines, circles etc. in $\mathbb{R}^2$ (and mutatis mutandis for $\mathbb{R}^3$ as well).

We will now focus on various aspects in the definition of the dot product and generalize it to arbitrary vector spaces.

8.1 Definition. Let V be a (complex or real) vector space. A *bilinear form* is a function

$$\langle\ ,\ \rangle \colon V \times V \to \mathbb{F} \quad (= \mathbb{R} \text{ or } \mathbb{C})$$

associating to a pair (v, w) an element $\langle v, w\rangle \in \mathbb{F}$ such that

a. $\langle u + v, w\rangle = \langle u, w\rangle + \langle v, w\rangle$ for all $u, v, w \in V$;

b. $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle$ for all $u, v, w \in V$; and

c. $\langle cv, w\rangle = \langle v, w\rangle = \langle v, cw\rangle$ for all $c \in \mathbb{F}$, $v, w \in V$.

If in addition

$$\langle v, w\rangle = \langle w, v\rangle$$

for all v, w we call the bilinear form *symmetric*.

8.2 Examples.

a. Let $B \in M_n(\mathbb{R})$ be arbitrary. Define

$$\langle X, Y\rangle := X^T BY.$$

(Note that the dot product is just the case $B = I$.)

b. $V = M_n(\mathbb{R})$ and

$$\langle A, B\rangle = \operatorname{trace}(AB).$$

This is a symmetric bilinear form!

(k, ℓ) with $1 \leq k < \ell \leq n$, on the two sides we will have two summands corresponding to $(i, j) = (k, \ell)$ and $(i, j) = (\ell, k)$. On the left, this means $x_k x_\ell y_k y_\ell + x_\ell x_k y_\ell y_k = 2x_k x_\ell y_k y_\ell$ and on the right this means $x_k^2 y_\ell^2 + x_\ell^2 y_\ell^2$. The remaining summands all correspond to cases where $i = j$ (and they are the same on the left and on the right namely $x_i^2 y_i^2$). Subtracting the left from the right we find that the result is a sum of summands of the form

$$x_k^2 y_\ell^2 + x_\ell^2 y_k^2 - 2x_k x_\ell y_k y_\ell = (x_k y_\ell - x_\ell y_k)^2 \geq 0.$$

□

c. $V = \mathcal{C}(I)$.

$$\langle f, g\rangle = \frac{1}{|I|}\int_I f(x)g(x)dx = \frac{1}{b-a}\int_a^b f(x)g(x)dx.$$

where $I = [a, b]$ and $|I| = b - a$.

d. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function. Recall we said f is *differentiable* at $p \in \mathbb{R}^n$, if there is a linear transformation $f'(p)\colon \mathbb{R}^n \to \mathbb{R}$, such that $f(p+X) = f(p) + f'(p)(X) + o(X)$ with $\lim_{X\to 0}\frac{o(X)}{|X|} = 0$.

Now what about the *second derivative*? Consider the case that there is a bilinear form H_p such that

$$f(p+X) = f(p) + f'(p)(X) + \frac{1}{2}H_p(X,X) + o(X)$$

with $\lim_{X\to 0}\frac{o(X)}{|X|^2} = 0$.

For those of you knowing about partial derivatives, assuming they exist (to a high enough order), H_p does exist and the matrix of H_p with respect to the standard basis is given by $H = [h_{ij}]$ with

$$h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(p).$$

H is called the *Hessian* of f (at p). It is a symmetric $n \times n$ matrix. If $n = 1$, then $H_p(X,X) = f''(p)X^2$.

Note that for linear transformation it has been very useful to choose a basis and work with matrices. The same works here.

8.3 Definition. Let V be a vector space with basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ and let $\langle\ ,\ \rangle$ be a bilinear form on V. The matrix of $\langle\quad\rangle$ with respect to $\mathcal{B}$ is defined as the $n \times n$ matrix B whose entry b_{ij} at position (i, j) is defined by

$$b_{ij} = \langle v_i, v_j\rangle.$$

8.4 Lemma. *The matrix B of $\langle\ ,\ \rangle$ has the following property: let $v, w \in V$ be arbitrary, then*

$$\langle v, w\rangle = [v]_{\mathcal{B}}^T B[w]_{\mathcal{B}}. \tag{8.1}$$

Proof. The condition holds clearly $v, w \in \mathcal{B}$: indeed, $e_i^T B e_j = b_{ij}$ for any matrix. Since $b_{ij} = \langle v_i, v_j\rangle$, the assertion is true for $v, w \in \mathcal{B}$.

It now follows for arbitrary v, w by computing $v = \sum_i c_i v_i$ and $w = \sum_i d_i v_i$, we have

$$\langle v, w\rangle = \sum_i\sum_j c_i d_j \langle v_i, v_j\rangle = \sum_{i,j} c_i b_{ij} d_j = [v]_{\mathcal{B}}^T B[w]_{\mathcal{B}}.$$

□

Note that since for any matrix $A = [a_{ij}]$ we have $a_{ij} = e_i^T A e_j$, the matrix B of a bilinear form is uniquely determined. Conversely, given a basis $\mathcal{B}$, any $B \in M_n(\mathbb{F})$ determines a bilinear form $\langle\ ,\ \rangle$ on V by means of (8.1).

Example. The matrix of the dot product is I_n.

8.1.1 Problem. Show that $\langle , \rangle$ is symmetric if and only if its matrix with respect to any basis is.

As in the case of a linear transformation, let us now check how the matrix B changes if we change the basis $\mathcal{B}$. To simplify notation let v, w be vectors with coordinate vectors X, Y with respect to $\mathcal{B}$. Let $\mathcal{B}' = \mathcal{B}P$ be another basis such that v has coordinate vector $X' = P^{-1}X$ and w has vector $Y' = P^{-1}Y$. Let B' be the matrix of $\langle\ ,\ \rangle$ with respect to B'. Then we have

$$X^T BY = \langle v, w\rangle = X'^T B'Y'$$

substituting $X = PX'$ and $Y = PY'$ we get

$$X'P^T BPY' = X'^T B'Y'.$$

Since X, Y and hence X', Y' range over all elements of $\mathbb{F}^n$ it follow2s that

$$B' = P^T BP.$$

We will now discuss in various cases, what the simplest form of B is that can be achieved by a clever choice of basis.

8.2. Inner product spaces

In this section V is always a finite dimensional vector space over $\mathbb{R}$.

8.5 Definition. An *inner product space* is a vector space V together with a *symmetric* bilinear form $\langle\ ,\ \rangle$ such that

$$\langle v, v\rangle \geq 0$$

for all $v \in V$ and $\langle v, v\rangle = 0$ if and only if $v = 0$. Such a form is called *positive definite*.

A symmetric matrix $B \in M_n(\mathbb{R})$ is *positive definite* if the bilinear form $X^T BY$ is positive definite.

If V is an inner product space, its positive definite symmetric bilinear form is usually referred to as its *inner product* or also *scalar product* (not to be confused with the scalar multiplication!).

8.6 Theorem. *Let $B \in M_n(\mathbb{R})$. The following are equivalent.*

a. *B is symmetric and positive definite.*

b. *B is the matrix of the dot product with respect to some basis of $\mathbb{R}^n$.*

c. *$B = P^T P$ for some invertible matrix $P \in \mathrm{GL}_n(\mathbb{R})$.*

Note that we already established the equivalence of b. and c. because $P^T P = P^T I P$, so the matrices of this form exactly are those occurring as matrices of the dot product with respect to the different bases of $\mathbb{R}^n$.

Also, if B is the matrix of the dot product with respect to some basis then B is symmetric and positive definite since the dot product is. It thus remains to show a. implies b.. The trick for this is to prove the following nice theorem:

8.7 Theorem (Gram-Schmidt algorithm)**.** *Let V be an inner product space. Then there is a basis $\mathcal{B} = (v_1, v_2, \dots, v_n)$ such that the bilinear form of V has matrix I. In other words, $\mathcal{B}$ satisfies*

$$\langle v_i, v_j \rangle = \delta_{ij}.$$

A basis as in the theorem is called an *orthonormal basis* and the vectors v_i are called orthonormal.

Proof. Let $\mathcal{B}_1 = (v_1, v_2, \dots, v_n)$ be an arbitrary basis of V. We will describe an algorithm to find an orthonormal basis. For this we will successively replace elements of $\mathcal{B}_1$ by elements $w_1, w_2, \dots$

To begin with, we replace v_1 by $w_1 = cv_1$ where

$$c = \frac{1}{\sqrt{\langle v_1, v_1 \rangle}}.$$

Note that c is defined since $\langle v_1, v_1 \rangle > 0$. Then we replace v_2 by $w_2' = v_2 - \langle w_1, v_2 \rangle w_1$. Note that we now have $\langle w_1, w_2' \rangle = 0$. Finally, by scaling w_2' (with $1/\sqrt{\langle w_2', w_2' \rangle}$) we arrive at w_2 such that $\langle w_2, w_2 \rangle = 1$. Note that $(w_1, w_2, \dots)$ is still a basis. We keep going:

Suppose we have an intermediate basis $(w_1, w_2, \dots, w_k, v_{k+1}, \dots, v_n)$ where $\langle w_i, w_j \rangle = \delta_{ij}$ as long as $i, j \leq k$. Then compute

$$w = v_{k+1} - \langle w_1, v_{k+1} \rangle w_1 - \langle w_2, v_{k+1} \rangle w_2 - \cdots - \langle w_k, v_{k+1} \rangle w_k.$$

Note that

$$\langle w_i, w \rangle = \langle w_i, v_{k+1} \rangle - \langle w_i, v_{k+1} \rangle \langle w_i, w_i \rangle = 0.$$

We replace v_{k+1} by

$$w_{k+1} := \frac{1}{\sqrt{\langle w, w \rangle}} w.$$

Note that by the exchange lemma 3.38, $(w_1, w_2, \dots, w_{k+1}, v_{k+2}, \dots, v_n)$ is still a basis. This process of course stops once $k = n$ and we are done. □

8.8 Example.
$$B = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \in M_2(\mathbb{R})$$
is positive definite iff $a > 0$ and $ac - b^2 > 0$.

Let $\langle\ ,\ \rangle$ be the bilinear form defined by B. Suppose B is positive definite. Then we can apply Gram-Schmidt to the standard basis $\mathcal{E} = (e_1, e_2)$. In the first step, we need the square root of $a = \langle e_1, e_2 \rangle$ which therefore must be positive.

In the second step, we need the square root of $\langle e_2 - \frac{b}{a}e_1, e_2 - \frac{b}{a}e_1 \rangle$ which means the square root of $c - b^2/a$. Thus $c - b^2/a$ is positive which forces $ac - b^2 > 0$ as well (since $a > 0$).

Conversely, if these two numbers are positive (and hence Gram-Schmidt is possible) we end up with an orthonormal basis. But then the form is positive definite and hence B is.

This provides the desired implication from a. to b. in Theorem 8.6.

Proof of Theorem 8.6. Let B be any positive definite symmetric matrix. Then B is the matrix of the dot product with respect to some basis of $\mathbb{R}^n$.

Indeed, define $\langle X, Y \rangle = X^T BY$. Then this is a positive definite symmetric bilinear form turning $\mathbb{R}^n$ into an inner product space. Thus, there is a basis $\mathcal{B}$ of $\mathbb{R}^n$ such that the matrix of this form becomes I.

Equivalently, if $P = [\mathcal{B}]$, then $I = P^T BP$. Then the columns of P^{-1} form a basis with respect to which the dot product has matrix B. □

Note that as an important consequence of the Gram-Schmidt process, whatever is true for the usual $\mathbb{R}^n$ with the dot product, also is true in a general vector space V with an inner product. In that sense, there is essentially (up to a choice of a suitable basis) only one inner product space of a given dimension.

A (finite dimensional) inner product space is also called an *Euclidean space* because we can do Euclidean geometry in it. As in the case of $\mathbb{R}^n$ (in fact it follows from that case) we have the following important *Cauchy-Schwarz inequality*:
$$|\langle v, w \rangle| \leq |v||w|$$
were we define $|v|$ as
$$|v| := \sqrt{\langle v, v \rangle}.$$
This then may be used to define the *angle* between v and w as the real number θ such that
$$\langle v, w \rangle = |v||w| \cos\theta.$$
Also, if we put $E = \mathrm{Span}(v, w)$, then E (together with the inner product) behaves exactly as the Euclidean plane. We can now talk about the distance between two vectors in an arbitrary inner product space.

> **Every property of the usual dot-product holds true for any symmetric positive definite bilinear form**: we simply choose an orthonormal basis and then our arbitrary bilinear form becomes the usual dot-product of coordinate vectors.

Indeed, a rotation on a plane E in an inner product space V is a linear transformation $R\colon E \to E$ such that, with respect to some orthonormal basis (v_1, v_2) for E, R has matrix of the form

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

which we recognize as the matrix of a rotation around the origin in $\mathbb{R}^2$.

For instance this allows us to define rotations in $\mathbb{R}^3$: A *rotation with pole* $v \in \mathbb{R}^3$ is a linear transformation $T\colon \mathbb{R}^3 \to \mathbb{R}^3$ such that v is an eigenvector of T for eigenvalue 1 and T maps the plane[2] $E = \{w \in \mathbb{R}^3 \mid (w \cdot v) = 0\}$ to itself, and finally, $T|_E$ is a rotation on that plane.

Let V be an inner product space. We say a linear transformation $T \in \operatorname{Hom}(V)$ is *orthogonal* (with respect to the chosen inner product) if $\langle T(v), T(w)\rangle = \langle v, w\rangle$ for all $v, w \in V$.

8.2.1 Problem. Show that T is orthogonal iff $M_{\mathcal{B}}(T)$ is orthogonal whenever $\mathcal{B}$ is an orthonormal basis of V.

8.9 Example. Let A be an invertible $n \times n$ matrix and let $S = A^T A$. Then S gives rise to an inner product. Let $g \in \mathrm{GL}_n(V)$ is orthogonal with respect to this form if and only if

$$g^T S g = S.$$

Indeed, for $X, Y \in \mathbb{R}^n$, we have

$$\langle gX, gY\rangle = (gX)^T S (gY) = X^T (g^T S g) Y$$

which equals $X^T S Y$ (for *all* X, Y) only if $S = g^T S g$. Since $S = A^T A$ this means that $(g^T A^T)(Ag) = A^T A$, or,

$$(AgA^{-1})^{-1} = (AgA^{-1})^T$$

ie. AgA^{-1} is orthogonal.

The set of all orthogonal transformations of V form a *subgroup* of $\mathrm{GL}(V)$, the so called *orthogonal group* $\mathrm{O}(V)$ of V.

If $V = \mathbb{R}^n$ with the dot-product, then by the above, $\mathrm{O}(V)$ may be identified with

$$\mathrm{O}_n(\mathbb{R}) = \{A \in \mathrm{GL}_n(\mathbb{R}) \mid A^T A = I\}.$$

This group is the correct object to discuss (and define) symmetries of a figure in $\mathbb{R}^2$, or $\mathbb{R}^3$. One can show (it is not that hard) that *any* distance preserving function $f\colon \mathbb{R}^n \to \mathbb{R}^n$, that is

$$|f(v) - f(w)| = |v - w|$$

[2] That this is a plane is proved in Proposition 8.2 below. But it should also be clear since E is the solution space of a single nonzero linear equation.

for all $v, w \in \mathbb{R}^n$, is of the form

$$f(v) = Av + b$$

where $A \in \mathrm{O}_n(\mathbb{R})$ and $b \in \mathbb{R}^n$ is some vector. Functions of this type are called *Euclidean motions*.

If $\Gamma \subseteq \mathbb{R}^2$ is any geometric figure (eg. a triangle, circle, square, rectangle, regular polygon, $\ldots$) then its *symmetry group* $\mathrm{Sym}(\Gamma)$ is defined as the set of all Euclidean motions that map Γ to itself, that is the set of all f such that $f(\Gamma) = \Gamma$. Note that if Γ is bounded, no translation will satisfy this condition. If we then put the centre of mass of Γ to the origin of $\mathbb{R}^2$, the symmetry group is a subgroup of the orthogonal group.

For instance for the circle S centered at 0, $\mathrm{Sym}(S) = \mathrm{O}(V)$, and for a regular n-gon Γ, centered at 0, $D_n := \mathrm{Sym}(\Gamma)$ is called the *dihedral group* with $2n$ elements. It consists of n rotations (by $2\pi/n$) and n reflections.

If Γ is unbounded, there may be translations in its symmetry group (think wallpaper). One can show that there are precisely 17 different classes of symmetries of wallpapers.

Remarks.

a. Let $\mathcal{B}$, $\mathcal{B}' = \mathcal{B}P$ both be bases of an inner product space V where $\mathcal{B}$ is orthonormal. Then $\mathcal{B}'$ is also orthonormal iff P is an orthogonal matrix.

   Indeed, the matrix of the bilinear form with respect to $\mathcal{B}$ is I, so $\mathcal{B}'$ is orthonormal iff $P^T I P = I$, ie. P is orthogonal.

b. Let $A \in M_n(\mathbb{R})$ be a matrix. Then A is orthogonal if $A^T A = I$. If this is the case then A is invertible and $A^{-1} = A^T$ and it follows that also $AA^T = I$.

c. Again, let $A \in M_n(\mathbb{R})$. Let us write $A = [A_1\, A_2\, \ldots\, A_n]$ with columns $A_i \in \mathbb{R}^n$.

   Notice that the formula for the matrix multiplication (row times column) and the definition of A^T imply that the entry of $A^T A$ at position i, j is given by $(A_i \cdot A_j)$ and hence A is orthogonal iff

$$(A_i \cdot A_j) = \delta_{ij}.$$

   Another way of saying this is that A is orthogonal if and only if the columns of A form an orthonormal basis of $\mathbb{R}^n$. (Indeed, if A is orthogonal, A is invertible and so the columns form a basis, orthonormal by the above; the converse is also clear.)

d. Suppose $A \in \mathrm{O}_2(\mathbb{R})$. Then there are $a, b \in \mathbb{R}$ such that $a^2 + b^2 = 1$ and

$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \text{ or } A = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

   In the first case ($\det A = 1$), A is a rotation, in the second case ($\det A = -1$), A is a reflection.

For later reference if $E \subseteq V$ is a subspace, let us define

$$E^\perp := \{v \in V \mid \langle v, w\rangle = 0 \text{ for all } w \in E\}.$$

$E^\perp$ is called the *orthogonal complement* of E and is a subspace of V.

Note that

$$\dim E^\perp = \dim V - \dim E.$$

To see this choose an orthonormal basis $\mathcal{B}_1$ for E and extend it to a basis $\mathcal{B}$ for all of V. After applying Gram-Schmidt to $\mathcal{B}$, we may assume that $\mathcal{B}$ is othonormal, and still $\mathcal{B}_1 \subseteq \mathcal{B}$ because we did not touch $\mathcal{B}_1$ during Gram-Schmidt (as it was orthonormal in the first place).

Let E' be the span of all vectors in $\mathcal{B}$ that are not in $\mathcal{B}_1$. Then clearly $V = E \oplus E'$ and $\dim E' = \dim V - \dim E$ (as $\mathcal{B} - \mathcal{B}_1$ is a basis for E').

Note that $E' = E^\perp$: indeed, $E' \subseteq E^\perp$ is clear because the basis of E' is contained in $E^\perp$. As for $E^\perp \subseteq E'$ let $v \in E^\perp$. We may write v (uniquely) as

$$v = w + w'$$

with $w \in E$ and $w' \in E'$. Then $0 = \langle w, v\rangle = \langle w, w\rangle + \langle w', w\rangle = \langle w, w\rangle$ since $w' \in E' \subseteq E^\perp$. It follows that $\langle w, w\rangle = 0$ which means $w = 0$. Thus $v \in E'$ and we are done.

Finally, let us now discuss the Elementary Divisor Theorem as it pertains to orthogonal transformations.

8.10 Theorem. *Let $T \in \mathrm{O}(V)$ be an orthogonal linear transformation.*

Then

$$V = W_1 \oplus W_2 \oplus \cdots \oplus W_r$$

where each W_i is irreducible *and $\dim W_i \leq 2$, and $W_i \perp W_j$ if $i \neq j$, ie. $W_i \subseteq W_j^\perp$.*

Moreover, if $\dim W_i = 1$, then it corresponds to an eigenvector of eigenvalue 1 or -1. If $\dim W_i = 2$, then T is a rotation *in the plane W_i, ie. has matrix*

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}$$

with respect to any orthonormal basis of W_i and the characteristic polynomial of $T|_{W_i}$ is $(x - e^{i\alpha})(x - e^{-i\alpha})$.

Proof. Let $W \subseteq V$ be an irreducible subspace. Indeed, if V itself is irreducible, $W = V$, otherwise, there is a smaller invariant nonzero subspace W_1. If this is irreducible, fine, otherwise, there is $W_2 \subsetneq W_1$ and so on. This process must stop since $\dim V$ is finite. Thus, there must be an irreducible subspace. (Which also follows from the elementary divisor theorem, by the way).

By the remarks preceding the theorem we may write $V = W \oplus W^\perp$. Since T is orthogonal and W is invariant, also $W^\perp$ is invariant, and T restricted to $W^\perp$ is obviously still orthogonal (the inner product is the restriction of the inner product of V). Since $W \neq \{0\}$, $\dim W^\perp < \dim V$, so by induction on $\dim V$, $W^\perp$ is the direct sum of pairwise orthogonal irreducible subspaces. And clearly W is orthogonal to each of them. Thus, we have

$$V = W_1 \oplus W_2 \oplus \cdots \oplus W_r$$

with each W_i irreducible and $W_i \perp W_j$.

To finish, it thus suffices to analyze a single irreducible subspace $W = W_i$. Since it is irreducible, $W = \langle v \rangle$ is cyclic. Also the order f of v is an irreducible polynomial (cf. Proposition 7.33) and $\dim W = \deg f$. But $\mathbb{F} = \mathbb{R}$, so f has degree 1 or 2 (see Problem 6.5.2). In the first case, v is an eigenvector. In the second case $f = (x - \lambda)(x - \overline{\lambda})$ where λ is a complex number. We will show below that $|\lambda| = 1$, so that $\lambda = e^{i\alpha}$ for some α.

If W has dimension 2, then T is still orthogonal when restricted to W. Thus, with respect to an orthonormal basis of W_i, T has an orthogonal matrix. Such a matrix has determinant ± 1 $(1 = \det(A^T A) = \det(A)^2)$. So for the two eigenvalues $\lambda, \overline{\lambda}$ we have

$$\lambda\overline{\lambda} = |\lambda|^2 = \pm 1.$$

But this means $|\lambda| = 1$, $\lambda = e^{i\alpha}$ and our transformation has in effect determinant 1 on W. Pick an orthonormal basis $\mathcal{B} = (v_1, v_2)$ for W. Let $A = [A_1\, A_2]$ be the matrix of $T|_W$ with respect to $\mathcal{B}$. Then the restriction of $\langle\ ,\ \rangle$ to W_i becomes the dot-product of coordinate vectors, so the matrix A is an orthogonal 2×2 matrix of determinant 1. Let $A_1 = [a\, b]^T$. Then $a^2 + b^2 = 1$, so there is $\beta \in \mathbb{R}$ such that $a = \cos\beta$ and $b = \sin\beta$. Notice that it follows that $A_2 = [-b\, a]^T$ or $[b\ -a]^T$: any of these two spans the line orthogonal to A_1; and these two are the only elements of length 1. Now the determinant condition asserts that $A_2 = [-b\, a]^T$ as claimed. Finally, since the eigenvalues of A are $e^{\pm i\beta}$, it follows that $\alpha = \pm\beta$ and we are done in this case.

If $\dim W = 1$ it corresponds to an eigenvector v, say. Then

$$\langle v, v \rangle = \langle Tv, Tv \rangle = \langle \lambda v, \lambda v \rangle = \lambda^2 \langle v, v \rangle.$$

Since $\langle v, v \rangle \neq 0$ this means that $\lambda^2 = 1$ and hence $\lambda = \pm 1$ as claimed. □

A beautiful consequence is the following theorem.

8.11 Theorem. *If $n = 2$ or 3, then the linear transformations corresponding to rotations are precisely the given by those $A \in M_n(\mathbb{R})$ that are orthogonal and have determinant 1.*

Proof. If $A \in \mathrm{O}_n(\mathbb{R})$ has determinant 1, then by the previous theorem, if $n = 2$, $\mathbb{R}^2$ is either irreducible (and T_A is a rotation), or $\mathbb{R}^2$ is the direct sum of two lines spanned by orthogonal eigenvectors for eigenvalues ± 1. By the fact that $\det A = 1$, the only choices are that both eigenvalues must be equal hence $A = I$ or $A = -I$, both of which correspond to rotations (by 0 or 180 degrees, respectively). If $n = 3$, then by the previous theorem,

$$\mathbb{R}^3 = E \oplus L$$

with $\dim E = 2$ and $\dim L = 1$, or

$$\mathbb{R}^3 = L_1 \oplus L_2 \oplus L_3$$

and all these spaces are perpendicular to each other. In the first case we must have that L is spanned by an eigenvector for eigenvalue 1, and T_A is a rotation. In the second case, the number of eigenvalues -1 must be even, so either $A = I$, or, after reordering if necessary, $L_1 \oplus L_2$ spans a plane on which T_A is multiplication by -1, hence it is a rotation with pole (contained in) L_3 and angle 180 degrees.

That any rotation has a matrix which is orthogonal and of determinant one, is a straightforward consequence of the definition. □

*Excursion: Symmetric bilinear forms

Let V be a vector space of dimension n, and let $\langle\ ,\ \rangle$ be a symmetric bilinear form. What we will discuss in this section will hold for arbitrary fields $\mathbb{F}$ unless mentioned otherwise.

We say $\langle\ ,\ \rangle$ is *nondegenerate* if $\langle v, w\rangle = 0$ for all w in V implies that $v = 0$.

Problem. For $v \in V$, denote by $\lambda_v \in V^* = \mathrm{Hom}(V, \mathbb{F})$ the map $\lambda_v(w) = \langle v, w\rangle$.

Show that $v \mapsto \lambda_v$ is a linear transformation.

Show that $\langle\ ,\ \rangle$ is nondegenerate iff this transformation is an isomorphism.

Note that even if the form is nondegenerate it need not be positive definite. For instance, take the form given by $-I_n$.

Classifying symmetric bilinear forms is an important topic of modern abstract algebra, in particular over fields other than $\mathbb{R}$ or $\mathbb{C}$ (where everything is understood). By "classifying" we mean the following question: If we fix $V = \mathbb{F}^n$, we say two symmetric bilinear forms given by the matrices A, B are *equivalent*, if there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that $A = P^T BP$. In other words, by a change of basis we can transform one form into the other.

Thus, "classifying" refers to determining how many equivalence classes there are. Over the complex numbers for instance, one can show that there are exactly $n + 1$ equivalence classes of symmetric bilinear forms on $\mathbb{C}^n$: two forms are equivalent iff their matrices have the same rank. Over the real numbers the situation is already more complex (no pun intended), and over many fields the situation is not completely understood.

First, let us discuss, when a form is nondegenerate. As in the case of an inner product, the notion of an *orthogonal complement* is useful albeit not as useful in general.

Definition. Let $\langle\ ,\ \rangle$ be a symmetric bilinear form on V.

For any subspace $W \subseteq V$ let

$$W^\perp = \{v \in V \mid \langle w, v\rangle = 0 \text{ for all } w \in W\}.$$

Clearly $W^\perp$ is a subspace of V.

The *radical* of $\langle\ ,\ \rangle$ is defined as $V^\perp$.

Remark. Note that $\langle\ ,\ \rangle$ is nondegenerate iff $V^\perp = \{0\}$.

One needs to be careful not to be carried away by one's intuition coming from inner products. In general $W^\perp \cap W \neq \{0\}$ and also $\dim W^\perp \neq \dim V - \dim W$ in general.

Example. On $\mathbb{R}^n$, let $\langle\ ,\ \rangle$ be defined as

$$\langle v, w \rangle = v^T A w$$

for some symmetric $A \in M_n(\mathbb{R})$.

Then A is nondegenerate if and only if A is invertible. Indeed, $(\mathbb{R}^n)^\perp = \mathcal{N}(A)$.

(The same holds if $\mathbb{F}$ is arbitrary.)

The following are some useful facts about the complement.

Proposition. *Let V be a vector space with a given symmetric bilinear form. Let W be a subspace.*

a. $\dim W^\perp \geq \dim V - \dim W$.

b. $W \subseteq W^{\perp\perp}$.

c. *If the form restricted to W is nondegenerate, then $V = W \oplus W^\perp$.*

Finally, in a. *and* b. *we have equality if the form is nondegenerate.*

Proof.

a. Let $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ be a basis of V such that the first k vectors form a basis for W. Then $W^\perp$ is defined by the k linear equations $\langle v_i, x \rangle = 0$. The rank nullity theorem then asserts that $\dim W^\perp \geq n - k$ (any matrix with k rows has at most k pivots).

   If the form is nondegenerate, then the k linear functionals $\lambda_{v_i} = \langle v_i, \cdot \rangle$ are linearly independent. In other words, the map

$$T\colon V \to \mathbb{F}^k$$

   defined by

$$T(v) = \begin{bmatrix} \langle v_1, v \rangle \\ \langle v_2, v \rangle \\ \vdots \\ \langle v_k, v \rangle \end{bmatrix}$$

   is surjective (indeed, if not, there is a subspace $H \subseteq \mathbb{F}^k$ with $\dim H = k-1$ containing the image of T; such a subspace is the solution space of a single linear equation[3] $\sum_i c_i x_i = 0$. But this then means

$$\sum_{i=1}^{k} c_i \langle v_i, v \rangle = 0$$

[3] For instance, extend some basis $(v_1, v_2, \ldots, v_{k-1})$ of H to a basis $(v_1, v_2, \ldots, v_k)$ of $V = \mathbb{F}^k$. Then if $v \in V$, $v = \sum_i d_i v_i$, so $v \in H$ iff $d_k = 0$. Note that $v \mapsto d_k$ is a linear transformation $\mathbb{F}^k \to \mathbb{F}$. Its matrix with respect to the standard basis is $[c_1\, c_2\, \ldots\, c_k]$ for some $c_i \in \mathbb{F}$.

for all v and hence the λ_{v_i} are linearly dependent – a contradiction). The rank nullity theorem now says that $\dim \mathcal{N}(T) = n - k = \dim V - \dim W$ as claimed.

b. This is obvious from the definition. If the form is nondegenerate then $\dim W^\perp = \dim V - \dim W$ and hence $\dim (W^\perp)^\perp = \dim W$ so we must have $W = (W^\perp)^\perp$.

c. Clearly $W \cap W^\perp = \{0\}$ since $\langle w, w' \rangle = 0$ for all $w \in W$ implies that $w' = 0$ (if $w' \in W$) since the form is nondegenerate on W. Thus V contains $W \oplus W^\perp$ as a subspace. But then by a. we must have $\dim V \leq \dim W + \dim W^\perp$ and hence equality[4]. Thus $W \oplus W^\perp = V$.

□

Note that if the form is nondegenerate it need not remain so when restricted to a subspace. In fact this can happen already when the subspace is one-dimensional. Let $v \in V$. We say v is *isotropic* if $v \neq 0$ and if $\langle v, v \rangle = 0$. Otherwise (but v still nonzero), v is called *anisotropic*.

Proposition. *Suppose $\mathbb{F}$ contains $\mathbb{Q}$. Suppose the form on V is not identical zero. Then there exists an anisotropic vector.*

Proof. Suppose all vectors in V are isotropic. We have to show that the form is identically zero. Now let $v, w \in V$ be arbitrary. Then

$$0 = \langle v + w, v + w \rangle = \langle v, v \rangle + 2\langle v, w \rangle + \langle w, w \rangle = 2\langle v, w \rangle.$$

Now since $\mathbb{F}$ contains $\mathbb{Q}$, $2 \neq 0$ and so $\langle v, w \rangle = 0$. □

(It follows from the proof that the proposition remains valid as long as $2 \neq 0$ in $\mathbb{F}$. That is, $\mathbb{F}_2 \not\subseteq \mathbb{F}$.)

As a consequence we immediately deduce:

Corollary. *Let $v \in V$ be anisotropic and $L = \operatorname{Span}(v)$. Then*

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

An *orthogonal basis* for V is a basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ such that for all $i \neq j$ we have $\langle v_i, v_j \rangle = 0$. Note that we make no assumption about $\langle v_i, v_i \rangle$ which is allowed to be zero for instance.

Proposition. *If $\mathbb{Q} \subseteq \mathbb{F}$, V admits an orthogonal basis.*

[4] Here is an argument not using dimenson: Let $v \in V$ be arbitrary. Then $\ell_v(w) := \langle v, w \rangle$ is a linear function on W and hence en element of W^*. Thus, there is a unique $w' \in W$ such that $\ell_v(w) = \langle w', w \rangle$ for all $w \in W$ and so $w'' = v - w' \in W^\perp$. It follows that $v = w' + w''$.

Proof. This is essentially Gram-Schmidt without normalizing. However, let us give a slightly different argument.

If the form is identically zero, nothing is to do. Otherwise, let v be an anisotropic vector. Then we have seen that $V = W \oplus W^{\perp}$ where $W = \mathrm{Span}(v)$. Now $\dim W < \dim V$ so by induction, $W^{\perp}$ admits an orthogonal basis. Adding v to this basis we obtain an orthogonal basis for W. □

What does this say about symmetric matrices?

Corollary. *Let $S \in M_n(\mathbb{F})$, where $\mathbb{F} \supseteq \mathbb{Q}$, be a symmetric matrix. Then there exist $P \in \mathrm{GL}_n(\mathbb{F})$ such that*

$$P^T S P$$

is diagonal.

Proof. If $\mathcal{B}$ is an orthogonal basis for the form given by S (with respect to the standard basis) then the matrix of the form (ie. $P^T S P$ where $P = [\mathcal{B}]$) is diagonal. □

We will find a see below that if $\mathbb{F} = \mathbb{R}$, P may be chosen to be orthogonal. If $\mathbb{F} = \mathbb{R}$, we can say a little more: Suppose $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ is an orthogonal basis. What about the values $\langle v_i, v_i \langle$ (ie. the diagonal entries of the matrix of our form)?

Let $c = \langle v_i, v_i \rangle$. If $c > 0$ we may replace v_i by $\frac{1}{\sqrt{c}} v_i$. If $c < 0$ we may replace v_i by $\frac{1}{\sqrt{-c}} v_i$, and if $c = 0$ we do nothing. Thus, after reordering we can achieve that the matrix of the form has the shape

$$\begin{bmatrix} I_p & & \\ & -I_q & \\ & & 0_r \end{bmatrix}$$

with $n = p+q+r$. The triple (p, q, r) is called the *signature* of the bilinear form. The following theorem asserts that this makes sense (ie. is independent of the choice of orthogonal basis).

Theorem (Sylvester's Inertia Theorem)**.** *The numbers p, q, r are uniquely determined by the bilinear form.*

Proof. Let $A \in M_n(\mathbb{R})$. Then there is P such that $P^T A P$ has the diagonal form with signature p, q, r. What can we say about p, q, r. Let $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ be the basis given by the columns of P.

First $r = \dim V^{\perp}$. Indeed, $V^{\perp}$ is given by the null space of A. Its dimension is of course independent of the basis for V that we choose. And its dimension is always the dimension of the nullspace of the matrix representing the form with respect to that basis. Thus, r is the nullity of A and hence depends only on A.

Next, I claim, p is the maximum dimension d of a subspace $W \subseteq \mathbb{R}^n$ on which the form is positive definite. Note that if we put $W = \mathrm{Span}(v_1, v_2, \ldots, v_p)$ then the form is positive definite on W. Thus $p \leq d$. Next, suppose W' is a subspace so that the form is positive definite on W'. Let us write $\mathbb{R}^n = W \oplus W^{\perp}$ with $W^{\perp} = \mathrm{Span}(v_{p+1}, \ldots, v_n)$.

Consider the projection $T\colon W' \to W$ (which maps $w+w^\perp$ to w). If $\dim W > p$, by the rank nullity theorem, $\mathcal{N}(T) \neq \{0\}$ and thus $W' \cap W^\perp$ cannot be zero. Thus, there is $w \in W' \cap W^\perp$ nonzero. But then $w^T A w \leq 0$ – a contradiction since $w^T A w > 0$ as the form is positive definite on W'.

It follows that also p (and hence also q) are uniquely determined by the matrix A. □

Example. An important symmetric bilinear form is given by the matrix

$$M = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & -1 \end{bmatrix}$$

which is the bilinear form used in physics in connection with space time (physicists replace -1 by $-c^2$ where c is the speed of light). Its signature is $(3,1,0)$.

Think of first three coordinates x_1, x_2, x_3 on $\mathbb{R}^4$ as space variables, and the last coordinate as time t. Special relativity asserts that the farthest distance you can travel in a time t is t (where, mathematicians that we are, we assume $c = 1$), and hence that if you travel from (x_1, x_2, x_3, t_1) to (y_1, y_2, y_3, t_2) in space time (which means that the time you need is $t_2 - t_1$), then

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 - c^2(t_2 - t_1)^2 \leq 0.$$

Note that this is exactly $(P - Q)^T M (P - Q)$ where

$$P = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ t_1 \end{bmatrix} \text{ and } Q = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ t_2 \end{bmatrix}$$

Thus, if we are at $(0,0,0,0)$ "here and now," then all our future will take place with the cone given by $q(P) \leq 0$ where $q(P) = P^T M P$. Also note that light will always move along the "spheres" given by $q(P) = 0$.

Similarly all our history (and the history of everything that is at the same point in space time) happened at points where $q(P) \geq 0$.

8.3. Hermitian forms and the spectral theorem

The concept of a symmetric bilinear form for complex vector spaces is unfortunately more involved: consider the bilinear form given by the identity matrix: it turns out that for $n > 1$ there are plenty of nonzero $X \in \mathbb{C}^n$ such that

$$X^T X = 0.$$

For instance consider

$$X = e_1 + ie_2.$$

The reason here is that $\sum x_i^2 = 0$ of course has nonzero complex solutions for $n > 1$ (this is another consequence of the Fundamental Theorem of Algebra: Pick $x_1, x_2, \ldots, x_{n-1}$ arbitrary (but not all zero). Then $f(t) := t^2 + (x_1^2 + \cdots + x_{n-1}^2)$ is a polynomial function which must have a zero $x_n \in \mathbb{C}$.)

However, there is a remedy to this situation. For $\mathbb{C}^n$, the following definition provides for a solution. For $A \in M_n(\mathbb{C})$ define $A^* = (\overline{A})^T$ where $\overline{A}$ is obtained from A by replacing its entries by their complex conjugates. We call A^* also the *adjoint* of A (not to be confused with the adjungated matrix in Chapter 5). Note that for instance

$$X^*X = \overline{x}_1 x_1 + \cdots + \overline{x}_n x_n = \sum_i |x_i|^2 \geq 0$$

is a nonnegative real number. It is zero if and only if all $|x_i| = 0$, ie. iff $X = 0$. If we define $\langle X, Y\rangle := X^*Y$, then this is no longer a bilinear form but it is close enough: the only difference is that $\langle cX, Y\rangle = \overline{c}\langle X, Y\rangle$ (and this happens only with respect to the first argument, not the second). We call this the *standard hermitian form* on $\mathbb{C}^n$.

Examples.

a. Note that if $A \in M_n(\mathbb{C})$, then A^* is the *unique* matrix B such that $\langle v, Aw\rangle = \langle Bv, w\rangle$ for all $v, w \in \mathbb{C}^n$.

b. The adjoint behaves very similar to the transpose: $(A + B)^* = A^* + B^*$ and $(AB)^* = B^*A^*$ for all A, B such that $A + B$ and/or AB are defined. But note that $(cA)^* = \overline{c}A$.

c. A matrix is called *hermitian* or *self-adjoint* if $A^* = A$.

8.12 Definition. A *hermitian form* on a complex vector space V is a function

$$\langle\ ,\ \rangle \colon V \to \mathbb{C}$$

such that

a. it is linear with respect to the second argument: $\langle v, w_1 + w_2\rangle = \langle v, w_1\rangle + \langle v, w_2\rangle$ and $\langle v, cw\rangle = c\langle v, w\rangle$ for all $v, w, w_1, w_2 \in V$ and $c \in \mathbb{C}$.

b. it is conjugate linear with respect to the first argument: $\langle v_1 + v_2, w\rangle = \langle v_1, w\rangle + \langle v_2, w\rangle$ and $\langle cv, w\rangle = \overline{c}\langle v, w\rangle$ for all $v_1, v_2, w \in V$ and $c \in \mathbb{C}$.

c. $\langle v, w\rangle = \overline{\langle w, v\rangle}$

If in addition, $\langle v, v\rangle \geq 0$ and $\langle v, v\rangle = 0$ only if $v = 0$, then we say $\langle\ ,\ \rangle$ is a *hermitian inner product* or *hermitian scalar product*.

Remark. Note that some authors prefer that a hermitian form is conjugate linear with respect to the *second* argument rather than the first.

8.13 Examples.

a. Let $V = \mathbb{C}^n$ and $A \in M_n(\mathbb{R})$ be any symmetric matrix. We define

$$\langle X, Y \rangle := X^* AY.$$

Note that this is a hermitian inner product if and only if A is positive definite (why?).

b. Let

$$V = \mathcal{C}([0, 2\pi], \mathbb{C}) = \{f \colon [0, 2\pi] \to \mathbb{C} \mid f \text{ is continuous}\}.$$

Define

$$\langle f, g \rangle := \frac{1}{2\pi} \int_0^{2\pi} \overline{f(x)} g(x) dx.$$

c. In Quantum Mechanics, the state space $\mathcal{H}$ is equipped with a hermitian scalar product such that for two states ψ_1, ψ_2 the probability for a system in state ψ_2 to be measured to have state ψ_1 is equal to $|\langle \psi_2, \psi_1 \rangle|^2$.

d. Related to the previous problem, in Analysis, *Hilbert spaces* play an important role: a Hilbert space is a complex vector space H together with a hermitian inner product, such that H is *complete* in the sense that every Cauchy sequence[5] $v_n \in H$ converges.

8.3.1 Problem. Show that a real symmetric matrix A is a positive definite hermitian matrix iff it is a positive definite as a real symmetric matrix.

(That is, show that $X * AX > 0$ if $X \neq 0$ in $\mathbb{C}^n$ provided you know that $X^T AX > 0$ whenever $X \in \mathbb{R}^n$ is not zero.)

We now mimic our treatment of symmetric bilinear forms. If $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ is a basis of V, define the *matrix of* $\langle\ ,\ \rangle$ to be $H = [h_{ij}] \in M_n(\mathbb{C})$ defined by

$$h_{ij} = \langle v_i, v_j \rangle.$$

Note that the definition of a hermitian form asserts that $h_{ji} = \overline{h_{ij}}$, or, in other words,

$$H^T = \overline{H}$$

which means that

$$H^* = H,$$

ie. H is hermitian.

As before, if $v, w \in V$ then

$$\langle v, w \rangle = [v]^*_{\mathcal{B}} H [w]_{\mathcal{B}}.$$

[5] A sequence $v_n \in H$ is a *Cauchy sequence* if for every $\varepsilon > 0$ there exists an integer $N > 0$ such that for all $n, m > N$, $|v_n - v_m| < \varepsilon$. Here $|v| = \sqrt{\langle v, v \rangle}$.

8.14 Lemma. *Let $\mathcal{B}' = \mathcal{B}P$ be another basis with change of basis matrix $P \in \mathrm{GL}_n(\mathbb{C})$. Let H be the matrix of $\langle \ , \ \rangle$ with respect to $\mathcal{B}$ and H' be the matrix of $\langle \ , \ \rangle$ with respect to $\mathcal{B}'$. Then*

$$H' = P^*HP.$$

Proof. This is straight forward and very similar to the proof of the change of basis formula for bilinear forms. We omit the details. □

8.15 Example. The diagonal entries of a hermitian matrix are *real*. Thus,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is hermitian iff $a, d \in \mathbb{R}$ and $c = \overline{b}$. Thus the 2×2 hermitian matrices are precisely the matrices of the form

$$\begin{bmatrix} r & z \\ \overline{z} & r \end{bmatrix} \quad (r \in \mathbb{R}, z \in \mathbb{C}).$$

Observe that the set of all hermitian 2×2 matrice is **not** a subspace of $M_2(\mathbb{C})$. (It is however, a subspace if we view $M_2(\mathbb{C})$ as a real vector space by allowing scalar multiplication with real numbers only.)

8.3.1. Hermitian inner products

As in the case of positive definite symmetric forms on real vector spaces, hermitian inner products admit *orthonormal bases*: Let V be a complex vector space with hermitian form $\langle \ , \ \rangle$. Then a basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ is called orthonormal if $\langle v_i, v_j \rangle = \delta_{ij}$. In other words, $\mathcal{B}$ is orthonormal if and only if the matrix of the form is equal to I. Obviously, if an orthonormal basis exists it follows from the fact that the standard hermitian form is an inner product that $\langle \ , \ \rangle$ is an inner product as well. The converse is also true.

8.16 Theorem. *Let V be a vector space with hermitian inner product. Then V admits an orthonormal basis.*

Proof. Just apply the Gram-Schmidt process to any basis. □

Also, as in the case of an inner product of real vector spaces, every subspace $W \subseteq V$ admits an *orthogonal complement* defined by

$$W^{\perp} = \{w \in V \mid \langle v, w \rangle = 0 \text{ for all} v \in V\}.$$

One proves easily that $W \cap W^{\perp} = \{0\}$ and that $W \oplus W^{\perp} = V$.

Note that if on V a positive definite form is given, then we can define the *length* of $v \in V$ as $|v| = \sqrt{\langle v, v \rangle}$ as in the case of an inner product space. Similarly, we can define the *angle* between v, w as a real number θ such that

$$\langle v, w \rangle = |v||w| \cos \theta.$$

That this makes sense is an important consequence of the Cauchy-Schwarz inequality which also holds: $|\langle v, w\rangle| \leq |v||w|$. (To prove it, pick an orthonormal basis.)

Remark. If $V = \mathbb{C}$ is one-dimensional, the standard inner product is $\langle z, w\rangle = \overline{z} \cdot w$. Note that if $z = a + bi$ and $w = c + di$ then $\langle z, w\rangle = (ac + bd) + (ad - bc)i$. If $z = w$ this becomes $a^2 + b^2$ and hence the length of z is just the absolute value of z as a complex number.

Note, similar to the orthogonal matrices, the matrices of a change of basis between orthonormal bases play a special role: since the form has matrix I with respect to both bases, the change of basis matrix P satisfies

$$P^* I P = I$$

which means

$$P^* P = I$$

Such a matrix is called *unitary*. It is elementary to check that

$$\mathrm{U}_n(\mathbb{C}) = \{P \in M_n(\mathbb{C}) \mid P^* P = I\}$$

form a subgroup of $\mathrm{GL}_n(\mathbb{C})$, the so called *unitary group*.

Example. A matrix is unitary iff its columns form an orthonormal basis for $\mathbb{C}^n$. For instance the matrices

$$\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \text{ and } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

are unitary.

8.3.2 Problem. Show that a diagonal matrix D is unitary if and only if its diagonal entries all have absolute value 1.

8.3.2. Spectral theorem for hermitian matrices

Let V be a complex vector space with a fixed hermitian inner product. A linear transformation $T \colon V \to V$ is called *hermitian* iff $\langle v, Tw\rangle = \langle Tv, w\rangle$ for all $v, w \in V$.

8.3.3 Problem. Let $\mathcal{B}$ be an orthonormal basis and let T be a linear operator on V. Show that T is hermitian if and only if $M_{\mathcal{B}}(T)$ is.

8.17 Lemma. *The eigenvalues of a hermitian operator are real.*

Proof. Let λ be an eigenvector and let v be an eigenvector. Then $\langle v, v\rangle \neq 0$ is a positive real number. Moreover, it follows that

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

Thus, $\lambda = \overline{\lambda}$. □

8.18 Theorem (Spectral theorem)**.** *Let T be a hermitian operator on V. Then there is an orthonormal basis for V with respect to which the matrix of T is diagonal.*

Proof. T has an eigenvector, v, say belonging to some eigenvalue λ. Let $W = v^{\perp} = \{w \in V \mid \langle v, w\rangle = 0\}$. Then T maps W to itself (ie. W is invariant).

Moreover, $\langle\ ,\ \rangle$, restricted to W defines an hermitian inner product on a complex vector space of strictly smaller dimension. It is immediate that $T|_W$ is hermitian. By induction, there is an orthonormal basis $\mathcal{B}'$ of W such that $T|_W$ has a diagonal matrix. Add v to the basis and the theorem is proved. □

Example. In Quantum Mechanics, physical observables correspond to hermitian operators on a Hilbert space. If a system is in state ψ, then after measurement of an observable with operator H, say, it will be found in the state ψ_λ (which is an eigenvector for the *real* eigenvalue λ of H) with probability $|\langle\psi_\lambda, \psi\rangle|^2$; the actual quantity that is measured (eg. the energy of the system) is the eigenvalue λ.

8.3.3. Spectral theorem for symmetric matrices

Let V be a vector space with a nondegenerate symmetric bilinear form. An operator T on V is called *symmetric* if for all $v, w \in V$ we have

$$\langle Tv, w\rangle = \langle v, Tw\rangle.$$

8.3.4 Problem. Show that T is symmetric if and only if with respect to a (hence any) orthogonal basis $\mathcal{B}$ of V, $M_{\mathcal{B}}(T)$ is symmetric.

The aim of this section is to prove:

8.19 Theorem (Spectral theorem for symmetric transformations)**.** *Let T be a symmetric linear operator on V. Then T is diagonalizable with orthogonal eigenvectors.*

Proof. Let $\mathcal{B}$ be any orthonormal basis for V. Then the matrix $A = M_{\mathcal{B}}(T)$ of T is symmetric. Observe that a *real* symmetric matrix is hermitian! By Lemma 8.17 this implies that all its eigenvalues are real. Thus, any symmetric operator on an inner product space has real eigenvalues (and hence eigenvectors).

We can now repeat the proof of the spectral theorem for hermitian matrices verbatim for symmetric operators on an inner product space (always replacing "hermitian" by "symmetric" and "hermitian scalar product" by "inner product." □

Note that this shows that T is symmetric iff there is an orthonormal basis $\mathcal{B} = (v_1, v_2, \ldots, v_n)$ such that for all $v \in V$,

$$T(v) = \lambda_1\langle v_1, v\rangle v_1 + \lambda_2\langle v_2, v\rangle v_2 + \cdots + \lambda_n\langle v_n, v\rangle v_n.$$

(Here $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of T.)

Note also that applying the theorem to T_A where A is a real symmetric matrix this implies that there is an orthogonal matrix P such that $P^{-1}AP$ is diagonal. Note that since P is orthogonal, then also P^TAP is diagonal, so this gives another proof that the symmetric bilinear form on $\mathbb{R}^n$ defined by A (as X^TAY) admits a basis such that the matrix of the form is diagonal (we called such a basis an *orthogonal basis* in the excursion on symmetric bilinear forms).

Also, as long as we stick to orthonormal bases (for the dot-product), so that all change of basis matrices are orthogonal, the matrices for T_A and for the symmetric form (given by A for the standard basis) always coincide. This is by no means true for arbitrary bases!

8.3.4. *Excursion: Spectral theorem for normal transformations

A complex matrix A is called *normal* if A and A^* commute. For instance hermitian ($A = A^*$) and unitary matrices ($A^* = A^{-1}$) are normal. But there are also matrices that are neither (for instance $A = \lambda I$ where λ is not real and not of absolute value 1).

If P is unitary and A is normal, then also P^*AP is normal: indeed, $(P^*AP)^* = P^*A^*P$ and so

$$(P^*A^*P)(P^*AP) = P^*A^*AP = (P^*AP)(P^*AP).$$

Since the unitary matrices are precisely the change of basis matrices between orthonormal bases, the following definition makes sense:

8.20 Definition. Let V be a complex vector space with hermitian inner product. A linear operator $T\colon V \to V$ is called *normal*, if the matrix of T with respect to a (and hence any) orthonormal basis is normal.

The most general form of the spectral theorem is:

8.21 Theorem. *A normal linear transformation T is diagonalizable: There is an orthonormal basis of V consisting of eigenvectors of T.*

Proof. Let $\mathcal{B}$ be any orthonormal basis of V, so that the inner product becomes the standard inner product and the matrix A of T is normal.

Let X be an eigenvector of A belonging to λ, say. We may normalize X so that $X^*X = 1$. Let $W = X^\perp = \{Y \in \mathbb{C}^n \mid Y^*X = 0\}$. For any $Y \in W$, we have $X^*(A^*Y) = (AX)^*Y = \lambda X^*Y = 0$ so A^* maps W to itself. On the other hand,

$$(A^*X)^*(A^*X) = X^*AA^*X = X^*A^*AX = \lambda\overline{\lambda}X^*X = \lambda\overline{\lambda} \geq 0 \in \mathbb{R}.$$

Now $X^*(A^*X) = (AX)^*X = \overline{\lambda}$. We can write $A^*X = cX + Y$ with $Y \in W$ and so $c = \overline{\lambda}$. It follows that

$$(A^*X)^*(A^*X) = \overline{\lambda}\lambda + Y^*Y$$

and hence $Y^*Y = 0$ ie. $Y = 0$. This implies $A^*X = \overline{\lambda}X$ and so $AY \in W$ for all $Y \in W$: $X^*(AY) = (A^*X)^*Y = \overline{\lambda}X^*Y = 0$. Now W corresponds to an invariant subspace W' (the space of all $\mathcal{B}Y$ with $Y \in W$) of V and $T_|W'$ is also normal. By induction on the dimension, we can assume that W' has an orthonormal basis of eigenvectors, and we are done. □

8.22 Corollary. *A matrix $A \in M_n(\mathbb{C})$ is normal if and only if there is a unitary matrix P such that P^*AP is diagonal.*

8.23 Corollary. *If A itself is unitary, then there is a unitary P such that P^*AP is unitary and diagonal.*

8.3.5. *Excursion: The Hopf-Fibration

The Hopf-Fibration is a beautiful geometric phenomenon: Let us define the n-sphere S^n as

$$S^n = \{(x_1, x_2, \ldots, x_{n+1}) \mid \sum_i x_i^2 = 1\} \subset \mathbb{R}^{n+1}.$$

So S^1 is the standard unit circle and S^2 is the actual standard unit sphere. Now the Hopf-Fibration is a surjective function $H\colon S^3 \to S^2$, such that for each $t \in S^2$, its preimage "looks" like a circle S^1.

One way to realize the Hopf-Fibration is as follows: Note that S^3 can be realized inside $\mathbb{C}^2 = \mathbb{R}^4$ as the set

$$S^3 = \{x \in \mathbb{C}^2 \mid |x| = 1\}.$$

Let $\mathrm{SU}_2(\mathbb{C}) = \{g \in \mathrm{U}_2(\mathbb{C}) \mid \det g = 1\}$, the *special unitary group*.

8.24 Lemma. *Let $v \in S^3 \subset \mathbb{C}^2$. Then there is a unique vector $w \in \mathbb{C}^2$ such that $\begin{bmatrix} v & w \end{bmatrix} \in \mathrm{SU}_2(\mathbb{C})$. Conversely, every element of $\mathrm{SU}_2(\mathbb{C})$ is of this form.*

Proof. Let

$$g = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

be in $\mathrm{SU}_2(\mathbb{C})$. We then gave the following conditions: the columns of g form an orthonormal basis and the determinant is 1:

$$\begin{aligned} \bar{a}c + \bar{b}d &= 0 \\ \bar{a}a + \bar{b}b &= 1 \\ \bar{c}c + \bar{d}d &= 1 \\ ad - bc &= 1 \end{aligned}$$

The second and third equation means that the two columns v, w of g are elements of S^3. Let now $v \in S^3$ be arbiyrary. Let w be a vector of length 1 such that $\langle v, w \rangle = 0$. As this is a non-trivial linear equation, the set of all vectors of length 1 with this property is of the form $\{\lambda w \mid |\lambda| = 1\}$. For which values of c is $\begin{bmatrix} v & cw \end{bmatrix}$ in $\mathrm{SU}_2(\mathbb{C})$? If $v = [\,a\,b\,]^T$ and $w = [\,c\,d\,]^T$, then

$$\lambda(ad - bc) = 1,$$

so $\lambda = (ad - bc)^{-1}$ is uniquely determined. (Note, $ad - bc \neq 0$ because v, w are linearly independent.) □

It follows that $\mathrm{SU}_2(\mathbb{C})$ is in a natural one-to-one correspondence to S^3.

Let $h \in \mathrm{SU}_2(\mathbb{C})$ be an element with eigenvalues $i, -i$. By the spectral theorem there is a unitary matrix g such that

$$ghg^{-1} = \begin{bmatrix} i & \\ & -i \end{bmatrix}$$

Note that the columns of $g^{-1} = g^*$ are eigenvectors for h. Multiplying one of them by a nonzero complex number (of absolute value 1) if necessary, we can achieve that $\det g = 1$, so that $g \in \mathrm{SU}_2(\mathbb{C})$. Define

$$f\colon\ \mathrm{SU}_2(\mathbb{C}) \to \mathrm{SU}_2(\mathbb{C})$$

as $f(g) = gxg^{-1}$. With the notation above,

$$f(g) = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} i & \\ & -i \end{bmatrix} \begin{bmatrix} \overline{a} & \overline{b} \\ \overline{c} & \overline{d} \end{bmatrix} = \begin{bmatrix} ai & -ci \\ bi & -di \end{bmatrix} = \begin{bmatrix} (|a|^2 - |c|^2)i & (a\overline{b} - c\overline{d})i \\ (b\overline{a} - d\overline{c})i & (|b|^2 - |d|^2)i \end{bmatrix}$$

Now observe that $|a|^2 - |c|^2 = -(|b|^2 - |d|^2)$ since

$$|a|^2 + |b|^2 = |c|^2 + |d|^2 = 1.$$

So if

$$f(g) = \begin{bmatrix} \alpha i & \beta i \\ \overline{\beta} i & -\alpha i \end{bmatrix}$$

Then $\alpha \in \mathbb{R}$. Consider $(\alpha, \mathrm{Re}(\beta), \mathrm{Im}(\beta))$. Then this is an element of S^2, because

$$1 = \det f(g) = \alpha^2 + \overline{\beta}\beta = \alpha^2 + \mathrm{Re}(\beta)^2 + \mathrm{Im}(\beta)^2.$$

So let $S = \{h \in \mathrm{SU}_2(\mathbb{C}) \mid h \text{ has eigenvalues } \pm i\}$. Then S is precisely the set of elements of the form $f(g)$ for some g. Moreover, for $(a, b, c) \in S^2$, the matrix

$$\begin{bmatrix} ai & b + ci \\ b - ci & -ai \end{bmatrix}$$

is in S, and S and S^2 are in bijection to each other. If, by these correspondences, we identify S^3 with SU_3 and S^2 with S, we obtain a surjective map $H\colon S^3 \to S$. This is a realization of the Hopf fibration. Indeed, let $s \in S$. Then

$$H^{-1}(s) = \{g \in \mathrm{SU}_3(\mathbb{C}) \mid f(g) = s\} = g_0 H^{-1}(x) = \{g_0 h \mid H(h) = x\},$$

where $g_0 \in H^{-1}$ is an arbitrary fixed element. The fibers $H^{-1}(s)$ hence all "look" the same, they are of the form $g_0 H^{-1}(x)$, which is in bjection to $H^{-1}(x)$. What is $H^{-1}(x)$?

$$H^{-1}(x) = \{g \in \mathrm{SU}_3(\mathbb{C}) \mid gxg^{-1} = x\} = \left\{ \begin{bmatrix} \lambda & \\ & \overline{\lambda} \end{bmatrix} \middle|\, |\lambda| = 1 \right\}.$$

Now the set $\{\lambda \in \mathbb{C} \mid |\lambda| = 1\}$ is the unit circle S^1, if we identify $\mathbb{C}$ with $\mathbb{R}^2$ in the usual fashion. Thus, $H^{-1}(x)$ is a copy of S^1 in $\mathrm{SU}_3(\mathbb{C})$. In fact, it is a subgroup D, (essentially a copy of $\mathrm{SU}_1(\mathbb{C}) = \{z \in \mathbb{C} \mid \overline{z} = z^{-1}\} = S^1$. So the fibers are translated copies of D.

Recall from an assignment, that if $H \subset G$ are subgroups, we do have a quotient set G/H (whose elements are precisely the set gH). Here, this shows that any $H^{-1}(gxg^{-1}) = gD$, and we obtain a one-to-one correspondence $\mathrm{SU}_2(\mathbb{C})/D \to S = S^2$.

By the above, $\mathrm{SU}_2(\mathbb{C})$ can be though of as a group structure on S^3, whereas $\mathrm{SU}_1(\mathbb{C})$ (or D) is a group structure on S^1. A deep theorem in analysis states, that there is no such group structure on S^2.

8.4. *Excursion: Quadratic Forms

8.25 Definition. Let $\mathbb{F}$ be either $\mathbb{Q}, \mathbb{R}, \mathbb{C}$. A *quadratic form* on $\mathbb{F}^n$ is a function $q\colon \mathbb{F}^n \to \mathbb{F}$ such that

$$q(x_1, x_2, \dots, x_n) = \sum_{i \le j} c_{ij} x_i x_j$$

for some $c_{ij} \in \mathbb{F}$.

8.26 Example. On $\mathbb{R}^2$, a quadratic form is a function $ax^2 + bxy + cy^2$.

The *matrix* of a quadratic form q is defined as the $n \times n$-matrix $A = [a_{ij}]$ where

$$a_{ij} = \begin{cases} c_{ii} & \text{if } i = j \\ \frac{1}{2} c_{ij} & \text{if } i < j \\ \frac{1}{2} c_{ji} & \text{if } i > j. \end{cases}$$

It is the unique symmetric $n \times n$ matrix A such that for all $x \in \mathbb{F}^n$,

$$q(x) = x^T A x.$$

Conversely, if A is a symmetric $n \times n$-matrix then $q(x) = x^T A x$ defines a quadratic form on $\mathbb{F}^n$.

Let $\mathcal{B} = \mathcal{E}P$ be any basis for $\mathbb{F}^n$. Then for $y = P^{-1}x$ (the coordinate vector for x with respect to $\mathcal{B}$), we have

$$q(x) = x^T A x = (Py)^T A (Py) = y^T (P^T A P) y =: q'(y).$$

So if we treat the entries of y as the new variables, we can "simplify" q.

8.27 Theorem. *Let q be a quadratic form on $\mathbb{R}^n$. Then there is an orthonormal basis for $\mathbb{R}^n$ such that*

$$q'(y) = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$

So with respect to a suitable coordinate system, a quadratic form has no *cross terms*.

Proof. Let P be an orthogonal matrix such that P^TAP is diagonal where A is the matrix of q. This exists by the spectral theorem. Theh q' has the desired form. □

After reordering the basis, we may assume that $\lambda_1, \lambda_2, \dots, \lambda_p > 0$, and $\lambda_{p+1}, \lambda_{p+2}, \dots, \lambda_{p+r} < 0$. Note if we allow coordinates that are non orthonormal, we can simplify further: if $\lambda_i > 0$, we may "absorb" $\sqrt{\lambda_i}$ in y_i, and if $\lambda_i < 0$, we may absorb $\sqrt{-\lambda_i}$ into y_i. More precisely, replace P by PD, where

$$D = \begin{bmatrix} \sqrt{\lambda_1} \\ & \sqrt{\lambda_2} \\ & & \ddots \\ & & & \sqrt{\lambda_p} \\ & & & & \sqrt{-\lambda_{p+1}} \\ & & & & & \ddots \\ & & & & & & \sqrt{-\lambda_{p+q}} \\ & & & & & & & 1 \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}$$

Then

$$(PD)^TA(PD) = \begin{bmatrix} I_p \\ & -I_q \\ & & 0_{n-(p+q)} \end{bmatrix}$$

Note by Sylvester's Inertia Theorem, p, q are uniquely determined. So

$$q' = y_1^2 + \dots + y_p^2 - y_{p+1}^2 - \dots - y_{p+q}^2.$$

A. Notations and conventions

This chapter is intended at introducing some of the background material needed throughout the class. We discuss two topics, sets and maps.

A.1. Sets

We will throughout adopt the "naive" point of view on set theory: a set is a collection of objects. For any (mathematical) object x and a set S there are two and only two possibilities: $x \in S$ (x is an element or member of S) or $x \notin S$ (x is not an element of S). If we are careful, then we can avoid the problems that led to the development of axiomatic set theory[1].

We will use "set", "collection", "class", and "family" as interchangeable notions. A (finite) "list" or sequence will usually mean an ordered set: to be precise, by a list $x_1, x_2, \ldots, x_k$ of k elements in a set X we mean the *function* $f\colon \{1, 2, \ldots, k\} \to X$ with the property $f(i) = x_i$ (for more on functions see below). Similarly, an infinite sequence $x_1, x_2, \ldots$ in X is simply a map $f\colon \mathbb{N} \to X$ where $f(i) = x_i$.

A set can be specified by listing all of its elements. For instance $S = \{a, b\}$ is the set containing the elements a and b. Note, that it may well be that $a = b$. To define a set we often use the following notation:

$$\{x \mid \text{ statements about } x\}$$

is by definition the set of all objects x for which the statements on the right hand side are true. For instance,

$$S = \{x \mid x \in \mathbb{R}; x^2 = 1\}$$

[1] Problems in set theory arise if one tries to claim the existence of too large sets: The famous "Russell's paradox" is the following. Let S be the set of all sets that do not contain itself at as an element: $T \in S$ if and only if T is a set that does not contain itself as an element. That's right, a priori nothing bars a set from being an element of itself. Of course this sounds a little weird. Anyway, the problem is as follows: Suppose S exists. Then we may ask the question: Is S an element of S? If yes, then, as $S \in S$ means "S does not contain itself as an element" we arrive at a contradiction. We smartly conclude that therefore, we must have that S is not a member of S. But this translates into "S is an element of S." – what now? The answer is that S cannot be a set; by extension, there is no set of all sets. To avoid such problems, we usually stick to very concrete sets and shun phrases like "set of all objects" or things like that; we will always restrict the admissible objects to being members of a (sometimes not explicitly specified) given set. There is a colloquial rephrasing of Russell's paradox: In a faraway village the barber proudly claims he shaves every villager who does not shave himself. Does the barber shave himself?

is the set of all real numbers for which $x^2 = 1$. Thus, $S = \{-1, 1\} \subseteq \mathbb{R}$ is a *subset* of $\mathbb{R}$. That is, every element of S is also an element of $\mathbb{R}$. Often, the same set S is also denoted by

$$S = \{x \in \mathbb{R} \mid x^2 = 1\}.$$

Examples of sets are manifold: for instance, in two-dimensional Euclidean geometry, lines are sets of points; in fact they are subsets of the plane which itself is a set of points. In three dimensions, planes in space are sets of points. The intersection of two lines in the plane is usually a set consisting of a single point (with two exceptions: if the lines are parallel or equal, the intersection is either empty or equal to the original line).

A.1.1. Quantifiers

Often in mathematics, quantifiers are used to shorten statements. In fact, every mathematical statements usually can be rephrased using only symbols and quantifiers.

The "for all" quantifier: It is very common to abreviate the statement "for all X" by "$\forall X$." Strictly speaking, it is usually used in the following form: if we want to say, that "for all x that satisfy a condition A, the statement B is true" we usually write $\forall x$ such that $A : B$. For instance, "for all integers x, x is a real number" we would write

$$\forall x \in \mathbb{Z} : x \in \mathbb{R}.$$

Similarly, to say a set S is a subset of a set T is the same as asserting that

$$\forall x \in S : x \in T.$$

The for all quantifier may be iterated: so "$\forall X \forall Y : \dots$" means "$\dots$" holds for each object X and each object Y.

Often we are sloppy and we often write the forall-quantifier at the end: For instance we might want to say that "$v+w \in W$ for all $v, w \in W$." You will find that also as $v+w \in W \forall v, w \in W$. More formally one would write

$$\forall v, w \in W : v + w \in W$$

or even

$$\forall v \in W \forall w \in W : v + w \in W.$$

The "exists" quantifier: The statement "there exists X such that A is true" is often abreviated as $\exists X : A$. So for instance, the statement "there exists a real number" (ie. "the set of real numbers is not empty") might be written as

$$\exists x : x \in \mathbb{R}.$$

Similarly, "there exists a real number x such that $x^2 = 1$" could be written as

$$\exists x : x \in \mathbb{R}, x^2 = 1.$$

It is more common to write

$$\exists x \in \mathbb{R} : x^2 = 1.$$

to express the same thing.

Sometimes we want to say more than just that there exists an object but that there exists a *unique* object: this is abreviated by $\exists!$: "there exists a unique positive real number whose square is equal to 2" becomes

$$\exists! x : x \in \mathbb{R}, x > 0, x^2 = 2$$

or

$$\exists! x \in \mathbb{R} : x > 0, x^2 = 2 \tag{A.1}$$

The "does not exist" quantifier is very similar, except that it expresses the exact opposite: To say that a set S is equal to the empty set is the same as asserting:

$$\nexists x : x \in S.$$

Often the two quantifiers are mixed. For instance, the statement (A.1) could be rephrased as

$$\exists x \in \mathbb{R} : x > 0, x^2 = 2, \forall y \in \{y \in \mathbb{R} \mid y > 0, y^2 = 2\} : y = x$$

which is barely readable. As a final remark, the order of quantifiers is very important. For instance the statement

$$\forall v \in V : \exists w \in V : v + w = 0$$

means that for each v there exists some element w such that $v + w = 0$. However,

$$\exists w \in V : \forall v \in V : v + w = 0$$

means: there exists some $w \in V$ such that $v + w = 0$ for all $v \in V$. Where is the difference? In the first case, w *depends on* v whereas in the second instance, w is independent of v.

A.2. Functions

Recall that a *map* or *mapping* or *function*[2] f from a set (nonempty) X to a (nonempty) set Y is a *rule* that assigns to each $x \in X$ one and only one $y \in Y$. We write $f\colon X \to Y$ to indicate that f assigns to an element of X an element of Y. X is called the *domain* of f and Y is called the *target* or *codomain*.

[2] For us a *function* is not necessarily a map whose codomain is a set of numbers; it can be any set.

We usually write $f(x)$ to denote the element of Y that x is being "mapped to" by f.

The target of a map f is not to be confused with its *image* or *range* which is usually denoted by $f(X)$. It is the set

$$f(X) = \{y \in Y \mid \text{ there is some } x \in X \text{ such that } y = f(x)\}.$$

As an example consider $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Then the domain and target of f is the set $\mathbb{R}$ of real numbers. Its range however is the set $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \mid x \geq 0\}$ of nonnegative real numbers, which of course is a proper[3] subset of $\mathbb{R}$.

If $f\colon X \to Y$ is any mapping, then f *induces* a map $\overline{f}\colon X \to Z$ whenever $Z \subseteq Y$ is a subset that contains $f(X)$: whenever $f(X) \subseteq Z$. It is defined by the same "rule," that is for $x \in X$, $\overline{f}(x) = f(x)$. However, since its codomain is Z and not Y, it is a *different function from f* unless $Z = Y$. This is a subtle point but we have to raise it at least once. Often, however, we will not distinguish between f and $\overline{f}$. So for instance the map $g\colon \mathbb{Q} \to \mathbb{Q}_{\geq 0}$ given by $g(x) = x^2$ is a different function from the function f above.

If $Z \subset X$ is a subset, then a function $f\colon X \to Y$ defines a function $f\mid_Z\colon Z \to Y$ called the *restriction of f to Z*. Again, for $z \in Z$,

$$f \mid_Z (z) = f(z).$$

Notice that this map differs from f in two ways: not only has it a different domain, it may even have a different range. For instance if $f\colon \mathbb{Q} \to \mathbb{R}$ is given by $f(x) = x^2$, and if $Z = \{1\}$, then $f\mid_Z\colon Z \to \mathbb{R}$ differs substantially from the original f.

Here are two natural constructions of new sets, given a function $f\colon X \to Y$. If $S \subseteq X$ is any subset, define its image as

$$f(S) = \{y \in Y \mid \text{ there is some } x \in S \text{ such that } f(x) = y\}.$$

The astute reader will recognize this as the range of $f\mid_S$.

If $T \subseteq Y$ is any subset, we define its *preimage* or *inverse image*

$$f^{-1}(T) = \{x \in X \mid f(x) \in T\}.$$

If T is a one point set, that is, if $T = \{t\}$ for a unique $t \in Y$, one often writes $f^{-1}(t)$ instead of $f^{-1}(\{t\})$.

Notice that even though we think of a function f as a "rule" this rule may be far from explicit. If $f\colon X \to Y$ is a map between sets of numbers, there may not be an explicit formula to compute $f(x)$. For instance consider the following function $f\colon \mathbb{R} \to \mathbb{Q}$:

$$f(x) = \begin{cases} 0 & x \in \mathbb{Q} \\ 1 & x \notin \mathbb{Q} \end{cases}$$

[3] A subset $Z \subseteq Y$ is *proper* if it is not equal to Y.

The *graph* of a map $f\colon X \to Y$ is the set

$$\Gamma_f = \{(x, y) \mid x \in X, y = f(x)\} \subseteq X \times Y.$$

Notice that it has the following properties:

a. For each $x \in X$, there is $y \in Y$ such that $(x, y) \in \Gamma_f$.

b. If (x, y) and (x, y') are elements of Γ_f then $y = y'$.

Conversely, suppose a subset $\Gamma \subseteq X \times Y$ satisfies a. and b. Then there is a unique function $f\colon X \to Y$ such that $\Gamma = \Gamma_f$. Why?

For each x we have to define $f(x) \in Y$. Now, by a. for each x there is a $y \in Y$ such that $(x, y) \in \Gamma$. Moreover, by b. this y is uniquely determined. We therefore put $f(x) = y$. This defines a map $f\colon X \to Y$. And $\Gamma_f = \Gamma$.

For this reason, one could *define* a function using its graph. We could identify a function $f\colon X \to Y$ with the triple (X, Y, Γ_f) (which is often done in set theory).

A.2.1. Injective, surjective, and bijective functions

In the following let X, Y be arbitrary sets.

A map $f\colon X \to Y$ is called *injective* or *one-to-one* if for all $x, y \in X$, $f(x) = f(y)$ only if $x = y$. In other words, if $f(x) = f(y)$ then necessarily $x = y$, and $f(x) \neq f(y)$ if $x \neq y$.

f is called *surjective* or *onto* if for each $y \in Y$ there is $x \in X$ such that $f(x) = y$.

We can rephrase this as follows:

- f is injective, if for each $y \in Y$, the equation $f(x) = y$ has *at most* one solution for $x \in X$.
- f is surjective, if for each $y \in Y$, the equation $f(x) = y$ has *at least* one solution for $x \in X$.

A function $f\colon X \to Y$ is *bijective* if f is both, injective and surjective.

Thus, f is bijective, if and only if for each $y \in Y$, there is exactly one $x \in X$ such that $f(x) = y$.

A.1 Lemma. *Let $f\colon X \to Y$ be a function. Then f is bijective if and only if there exists $g\colon Y \to X$ such that $g \circ f = \mathrm{id}_X$ and $f \circ g = \mathrm{id}_Y$.*

Here for any set S, $\mathrm{id}_S\colon S \to S$ is the function defined by $\mathrm{id}_S(s) = s$.

Proof. Let first f be bijective. Define $g\colon Y \to X$ as follows. For each $y \in Y$, there is a unique $x \in X$ such that $f(x) = y$. Let $g(y)$ be that x. Then by definition $f(g(y)) = y$ for all $y \in Y$ and hence $f \circ g = \mathrm{id}_Y$. Also, if $x \in X$, and if $y = f(x)$, then $g(f(x)) = x$: indeed, let $y = f(x)$, then $g(y)$ is the unique element of X such that $f(g(y)) = y$. But x is such an element, so $g(y) = x$. It follows $g \circ f = \mathrm{id}_X$.

Now suppose g exists. We have to show that f is injective and surjective. Let $x, x' \in X$ such that $f(x) = f(x')$. Then, applying g to both sides, $g(f(x)) = g(f(x'))$. The left hand side is equal to x whereas the right hand side is equal to x', so $x = x'$. Consequently, f is injective. Let $y \in Y$ be arbitrary. Then $f(g(y)) = y$ so $x = g(y) \in X$ is an element such that $f(x) = y$. Hence f is surjective. □

A.2 Remark. If f is bijective, the map g that exists by the previous Lemma is usually denoted by f^{-1} and is called the *inverse* function[4] of f.

A.3 Example.

a. Let $f\colon \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$. Then f is neither injective nor surjective: indeed, $f(1) = f(-1) = 1$, so f is not injective. And there is no $x \in \mathbb{R}$ such that $f(x) = -1$, so f is not surjective.

b. Let $f\colon \mathbb{R}_{>0} \to \mathbb{R}$ be defined by $f(x) = x^2$, where $\mathbb{R}_{>0}. = \{x \in \mathbb{R} \mid x > 0\}$. Now f is injective: if $f(x) = f(x')$ then $x^2 = x'^2$ and hence $(x - x')(x + x') = 0$ which means that $x = x'$ or $x = -x'$. However, $x = -x'$ is impossible, if both x and x' are positive, so $x = x'$.

   However, f is not surjective, because $f(x) = -1$ has no solution.

c. Let $f\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ be defined by $f(x) = x^2$. Now f is injective and surjective. Indeed, the same argument as in b. shows that f is injective. f is surjective, because for every positive real number y there is a positive real number x such that $x^2 = y$ (this is an application of the Intermediate Value Theorem in calculus). The function f^{-1} is commonly written as $f^{-1}\colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$, $f^{-1}(y) = \sqrt{y}$.

This example also shows that the domain and codomain of a function are important parts of its definition. It is not reasonable to treat the functions in a., b., c. as equal, even though they are defined using the same rule; they have very different properties.

[4] This leads to an unfortunate break down of notation: if $y \in Y$, $f^{-1}(y)$ denotes both, the inverse image of y, which is, in the notation of the previous lemma the set $\{g(y)\}$, *and* the value of f^{-1} at y, which is $g(y)$.

$X := Y$	X is defined as being equal to Y. ":=" is sometimes used to emphasize the fact that this is defines X.
$a \in S$	a is an element of the set S.
$a \notin S$	a is not an element of the set S.
$S = T$	The sets S and T are equal, i.e. they have the same elements: $a \in S$ if and only if $a \in T$.
$S \subset T$ or $S \subseteq T$	The set S is a subset of the set T: $a \in S$ implies that $a \in T$.
$S \supset T$ or $S \supseteq T$	Same as $T \subset S$.
$S \subsetneq T$:	The set S is a subset of the set T, but S is not equal to T.
$S \cup T$	The union of the sets S and T: $a \in S \cup T$ if and only if $a \in S$ *or* $a \in T$.
$S \cap T$	The intersection of the sets S and T: $a \in S \cap T$ if and only if $a \in S$ *and* $a \in T$.
$S - T$	The *difference* of the sets S and T, also denoted the *complement of* T *in* S: $a \in S - T$ if and only if $a \in S$ but $a \notin T$. This is also written $S \setminus T$ in the literature.
$\{a\}$	A set containing a single element, a.
$\{a_1, a_2, \dots\}$	A set containing the (finite or infinite) list of elements $a_1, a_2, \dots$. The a_i need not be distinct: the sets $\{a, a\}$ and $\{a\}$ are identical.
$\{a \mid$ statements about $a\}$	The set of all "objects" a for which the "statements" are true. Often the a are assumed to be elements of some (huge, unmentioned) set.
$\{a \in S \mid$ statements about $a\}$	The set of all a that are elements of S and for which the "statements" are true. Equal to $\{a \mid a \in S;$ statements about $a\}$.
$P(S)$	The *power set* of S: $P(S)$ is the set of subsets of S. Formally, $P(S) = \{T \mid T \subset S\}$.
$\emptyset$	The empty set. It has no elements.
$S \times T$	The cartesian product $\{(a, b) \mid a \in S, b \in T\}$ of two sets S and T.
S^n	The cartesion product of n copies of S: $S^n = \{(a_1, a_2, \dots, a_n) \mid a_i \in S\}$.

Symbols and notations commonly used in these notes.

$\lvert S\rvert$	If S is a set, then $\lvert S\rvert$ is the *cardinality* of S, that is, the number of elements in S.
$\#S$	$\lvert S\rvert$ if S is a set.
$\mathcal{F}(S,T)$	The set of all maps from S to T: $\mathcal{F}(S,T) = \{f \mid f\colon S \to T\}$.
$f\colon S \to T$	A map between two sets S and T: S is the domain of f and T the codomain; for each $a \in S$, we have a uniquely determined $f(a)$ in T.
id_S	The identity mapping on a set S: $\mathrm{id}_S\colon S \to S$ is defined by $\mathrm{id}_S(s) = s$. Often we omit the subscript S, if the context is clear.
$\mathrm{id}\colon S \to S$	Same as $\mathrm{id}_S\colon S \to S$.
$f \circ g$	The composition of two mappings: if $g\colon S \to T$, $f\colon T \to U$ are maps, then $f \circ g\colon S \to U$ is the map sending an $s \in S$ to $f \circ g(s) = f(g(s))$.
$f^{-1}(T)$	Given $f\colon S \to T$, $f^{-1}(T) = \{s \in S \mid f(s) \in T\}$.
$f^{-1}(t)$	Given $f\colon S \to T$, $f^{-1}(t) = \{s \in S \mid f(s) = t\}$. *Also:* if f is *bijective*, then f^{-1} is a function $T \to S$ and $f^{-1}(t)$ is then the unique element s_0 of S for which $f(s_0) = t$. In this case, $f^{-1}(t)$ may mean both s_0 and $\{s_0\}$.
$f^{-1}\colon T \to S$	Inverse of function $f\colon S \to T$; only exists if f is bijective. $f^{-1}(f(s)) = s$ and $f(f^{-1}(t)) = t$ for all $s \in S$ and $t \in T$. Equivalently: $f \circ f^{-1} = \mathrm{id}_T$ and $f^{-1} \circ f = \mathrm{id}_S$.
$\mathbb{N}$	The set of natural numbers $\{1, 2, \dots\}$.
$\mathbb{N}_0$	The set of natural numbers including 0: $\{0, 1, 2, \dots\}$.
$\mathbb{Z}$	The set of integers $\{0, \pm 1, \pm 2, \dots\}$.
$\mathbb{Q}$	The set of rational numbers $\mathbb{Q} = \{\frac{a}{b} \mid a, b \in \mathbb{Z}; b \neq 0\}$.
$\mathbb{R}$	The set of real numbers.
$\mathbb{C}$	The set of complex numbers.
$\mathbb{F}$	An arbitrary field.
$\mathbb{F}_2$	The field with two elements.
$[a_{ij}]$	Matrix whose entries are denoted by a_{ij} (the dimensions of the matrix is implicit).
$\sum_{i=p}^{q} a_i$	Sum of the a_i where i ranges over all integers from p to q. It is equal to $\underbrace{a_p + a_{p+1} + \cdots + a_q}_{n \text{ times}}$. The a_i are all elements of a set with an associative operation denoted by $+$ (e.g. they may be elements of a field, they may be matrices of the same size over the same field, etc.). If $q < p$ the sum is *empty* and by convention equal to 0 (where 0 is the identity element of $+$).
$\lvert x\rvert$	absolute value of a real or complex number.
$\lVert v\rVert$	the length of a vector; also called its *norm*.
$M_{m\times n}(\mathbb{F})$	The set of all $m \times n$ matrices with entries in the field $\mathbb{F}$.
$M_n(\mathbb{F})$	Same as $M_{n\times n}(\mathbb{F})$.
$\mathrm{GL}_n(\mathbb{F})$	The set of invertible $n \times n$ matrices with entries in $\mathbb{F}$.

Symbols and notations commonly used in these notes, cont'd.

$\forall X$	For all objects X (usually the objects are restricted to being elements in a set).
$\forall X$ conditions on X : statements about X	For every object X that satisfies the ”conditions on X” the ”statements about X” are true.
$\exists X$: statements about X	There exists an object X such that the ”statements about X” are true.
$\nexists X$: statements about X	There exists no object X such that the ”statements about X” are true.

Symbols and notations commonly used in these notes, cont'd.

Index