UNIVERSITY OF VIENNA
VO LINEAR ALGEBRA
2023W–2024S
LECTURE NOTES

VLADIMIR KAZEEV

2024-06-24 14:37

Preface

These lecture notes were in large part developed for the course MATH 104 “Applied Matrix Theory” (what a title!) as I taught it in the winter quarter of 2019 at the Department of Mathematics of Stanford University. I would like to thank those students and colleagues who have helped in improving the notes by providing valuable, specific remarks. In this regard, I express my particular gratitude to Alexander Dunlap, Adam Gurary, Joy Hsu, Hongtao Sun, Katerina Velcheva (all from Stanford) and to Mark Strempel, Clemens Schütz, Mavilo Bozkurt, Maksim Bjelić, David Kan, Julian Pigall, Elias Fuchs, Tim Paulis, Sara Dedic, Janis Asprion, Melanie Arlt, Thomas Aumaier, Anna Eberl, Anonymous, Moritz Perl, Marit Raubuch, Jan Baldreich, Kevin Untersmeier, Alexander Tarasov, Sophie Schuhböck, Adrian Weishörndl, Michele Sereni, Martin Schmidt, Sebastian Schmutzhard-Höfler, Helena Auböck, Stefan Renoldner, Jakob Fischer, Gabriel Pflügl, Jochen Nemetschek, Ivonne Gattringer, Tristan Berthold, Yauheni Pakhomau, Denis Konov, Lena Stadlmair, Peter Fritz, Borsa Kopper, Amelie Hochgerner and Moritz Gallauner (all from UniVie).

Vladimir Kazeev
24th June 2024
Vienna

Contents

Chapter I. Introduction

§ I.1. Some mathematical language and notations

This section serves to set up certain mathematical notations, notions and associated properties that are standard for the whole of mathematics and are not specific to the course. To some extent, they should be familiar to the enrolled students from high school or from the STEOP course; in any case, this section serves to introduce the reader to the notations used throughout the course and also to provide material for reference in the following chapters.

§ I.1.1. Sets

A *set* is a well-defined collection of distinct objects. In the expression *well defined*, *well* means *correctly* (not just *clearly*), so saying that a collection is *well defined* means that any object one can imagine either belongs to the collection or not. Actually, this definition of a *well-defined collection* is itself not well defined (look up paradoxes of naive set theory), but we shall pretend that it is so — by working only with sets for which the above characterization is well defined. The relationships "x belongs to X" and "x does not belong to X", where X is a set and x is something that can belong or not belong to the set, are denoted by "$x \in X$" and "$x \notin X$".

For example, consider the set of students enrolled in `Linear Algebra 1` at this moment according to `u:space`. Generously idealizing that outdated system, we may assume that the instructor can, at any moment, open the list and verify for any object (such as "apple", "Jane Doe" or "Max Muster") whether it belongs to the set or not.

In mathematical notations, sets are often defined by listing their elements in curly brackets; for example, we may write $X = \{0, 1, ☺\}$ to define X as the set consisting of the objects denoted by symbols "0", "1" and "☺". What these symbols actually mean and whether the objects they refer to are distinct is the *context* and not part of the definition of X. It may happen that some or all of the symbols refer to the same object; in any case, they all should be well defined in some way. If $☺ = 0 \neq 1$, then X is a *two-element set* since it contains exactly two distinct elements.

Sets *per se* are not ordered: a set is nothing else than the information about what belongs to it and what does not, regardless of the order in which the elements are listed in any particular definition. For example, the expressions $\{0, 1, ☺\}$, $\{1, ☺, 0\}$ and $\{0, 1, ☺, 1, 1, ☺, 0\}$ represent the same set. When in that case the symbols 0, 1 and 2 denote the integers they usually denote and $☺ = 2$, the elements of the same set, $X = \{0, 1, 2\}$, can be ordered as integers; however, strictly speaking, this order is not an intrinsic characteristic of X as a set but a manifestation of the additional structure introduced in X together with arithmetic operations.

The *empty set*, denoted by $\varnothing$, is the unique set that does not contain anything: $\varnothing = \{\,\}$.

If X and Y are sets such that $x \in Y$ holds for every $x \in X$, then X is called a *nonstrict subset* (often just *subset*) of Y and Y is called a *nonstrict superset* (often just *superset*) of X. This is denoted by writing "$X \subseteq Y$" or "$Y \supseteq X$". If additionally $X \neq Y$, then X is called a *strict subset* (or *proper subset*) of X and Y is called a *strict superset* (or *proper superset*) of X. These equivalent expressions of the same relation between X and Y are denoted by writing "$X \subset Y$" or "$Y \supset X$".

For the purpose of defining a set, a very practical alternative to listing all of its elements (which is often inconvenient or even impossible) is specifying the set in terms of a larger set, collecting all elements of the larger set satisfying a certain logical condition (called *predicate*). For example, the set $Y = \{1, 2, 3\}$ can be defined using the superset $X = \{0, 1, 2, 3, 4, 5, 6, 7\}$ as follows:

$$Y = \{x \in X \text{ such that } |x - 2| \leq 1\}\,.$$

For notational convenience, the expressions “such that”, “for which” and “with the property that” are often denoted by a colon. Using this convention, the above definition can be rewritten as follows:

$$Y = \{x \in X \colon |x-2| \leq 1\} .$$

§ I.1.2. Set-theoretic operations

The *union* of two sets X and Y, denoted by $X \cup Y$, is the set that consists of all elements of X and of all elements of Y:

$$X \cup Y = \{z \colon z \in X \text{ or } z \in Y\} .$$

The *intersection* of two sets X and Y, denoted by $X \cap Y$, is the set that consists of all elements of X that are also elements of Y:

$$X \cap Y = \{z \colon z \in X \text{ and } z \in Y\} .$$

Any two sets X and Y such that $X \cap Y = \varnothing$ are called *disjoint.*

The *difference* of two sets X and Y, denoted by $X \setminus Y$, is the set that consists of all elements of X that are not elements of Y:

$$X \setminus Y = \{z \colon z \in X \text{ and } z \notin Y\} .$$

§ I.1.3. Natural numbers. Ellipsis in set definitions

The set of natural numbers is denoted by $\mathbb{N}$:

$$\mathbb{N} = \{1, 2, 3, \ldots\} ,$$

where the ellipsis means that we consider the infinite (welcome ∞) iterative construction of natural numbers each step of which consists in adding 1 to the largest listed one and including the result in the list.

§ I.1.4. Nonnegative integers

The set of nonnegative integers is denoted by $\mathbb{N}_0$:

$$\mathbb{N}_0 = \{0\} \cup \mathbb{N} = \{0, 1, 2, \ldots\} .$$

§ I.1.5. Integers. Colon in set definitions

The set of integers is denoted by $\mathbb{Z}$:

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\} = \{0\} \cup \mathbb{N} \cup \{-n \colon n \in \mathbb{N}\} ,$$

where we see equivalent ways of writing the same set. On the right-hand side, in the definition of the latter component (“negative natural numbers”), the colon stands for “such that” and means that we take *all* objects of the form $-n$ with $n \in \mathbb{N}$. In any such a construction, the definition of $-n$ for every $n \in \mathbb{N}$ is external to the construction and is supposed to be available.

§ I.1.6. Rational numbers

Any two integers p and q are called *co-prime* if they have no common divisor larger than one. The set of (real) rational numbers is denoted by $\mathbb{Q}$:

$$\mathbb{Q} = \big\{(p, q) \colon p \in \mathbb{Z},\ q \in \mathbb{N},\ p \text{ and } q \text{ are co-prime}\big\} .$$

An element $(p, q) \in \mathbb{Q}$ is often written as a fraction:

$$\frac{p}{q} \quad \text{or, equivalently,} \quad p/q .$$

These two notations are supposed to be initially understood as alternative graphical expressions for (p, q). Note that the above definition explicitly requires that p and q are co-prime. This detail reflects that, for any co-prime $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, all pairs (pr, qr) with $r \in \mathbb{N}$, i.e., all fractions

$$\frac{pr}{qr} \quad \text{with} \quad r \in \mathbb{N}\,,$$

are to be identified with (considered to be equal to) the one corresponding to $r = 1$. To ensure that the equality relation between two elements of $\mathbb{Q}$ means what it should mean for fractions, it is standard to either (i) explicitly exclude "duplicate" fractions from the very beginning, as we do by requiring co-primality, or (ii) identify all distinct pairs representing "equal" fractions at a later stage.

With operations of addition and multiplication suitably defined on pairs of elements of $\mathbb{Q}$, the fraction notation acquires the meaning of division corresponding to the multiplication operation. The suitability of addition and multiplication can be understood in terms of our usual experience with fractions but can be formalized, as we do in definition II.1.1.1 below.

§ I.1.7. Function, functional, mapping, map. Arguments and values. Domain and co-domain. Image

A *function*, *functional*, *mapping* or *map* is a rule that matches *every* element of one set (*argument*) with *exactly one* element of another set (*value*). The sets are called the *domain* and *co-domain* of the function, respectively. Just as sets, functions are denoted by various symbols, which, in mathematics, are often chosen among Latin or Greek letters, lowercase or uppercase. The domain and co-domain are intrinsic characteristics of the function (rule); without these two sets, the function (rule) does not make sense.

In mathematical notations, one often writes $f\colon X \to Y$ to say that f is a function with domain X and co-domain Y. To *evaluate* a function, we add a parenthesis with an argument to its mathematical symbol (name). Such an expression is to be understood as the value of the function at the specified argument and is *not to be confused* with the function (rule) itself.

The set of all possible values of f,

$$f(X) = \{f(x)\colon x \in X\} \subseteq Y\,,$$

is called the *image of X under f*.

A function can be defined in many ways: by the explicit indication of its value for every possible argument, by reduction to known functions ("using a formula") or implicitly. Consider the following examples.

(a) The function $f\colon X \to Y$ with $X = \{0, 1, \text{☺}\}$ and $Y = \{\text{😎}, \text{🌳}, \text{☕}\}$ (whatever the pictographs mean, 0 and 1 being the integer zero and one) given by

$$f(0) = \text{😎}, \quad f(1) = \text{🌳} \quad \text{and} \quad f(\text{☺}) = \text{☕}.$$

Note that, in order for this definition to be *correct* and for the function f to be thereby *well defined*, it is required that (i) $\text{☕} = \text{😎}$ whenever $\text{☺} = 0$ and (ii) $\text{☕} = \text{🌳}$ whenever $\text{☺} = 1$. Otherwise, several *distinct* elements of Y are defined as the values of f at *identical* elements of X, which renders the definition *incorrect* and the function, *ill-defined*.

(b) The function $f\colon \mathbb{N} \to \mathbb{N}_0$ given by

$$g(n) = n - 1 \quad \text{for all} \quad n \in \mathbb{N}\,.$$

We consider it is important, for clarity and correctness, to write "for all $n \in \mathbb{N}$" in definitions as right above. First, n is a "local" variable: it has no meaning before it appears out of the blue on that line, and we should specify — in some way or another — what n means. Second, defining $g\colon \mathbb{N} \to \mathbb{N}_0$ means defining $g(n)$ as an element of $\mathbb{N}_0$ *for all* $n \in \mathbb{N}$ (not for *some* n from *some unspecified set*), so this is precisely what we have to do to give a correct definition.

Note that $g(m) = m - 1$ and $g(n) = n - 1$ with $m, n \in \mathbb{N}$ are *numbers*, distinct or identical depending on whether $m = n$. These numbers are the *values* of g at m and n respectively, and not the function itself.

(c) Consider the function $h\colon \mathbb{N} \to \mathbb{N}$ given by

$$\bigl(h(n) - n\bigr)\bigl(h(n) + 1\bigr) = 0 \quad \text{for all} \quad n \in \mathbb{N}\,.$$

Under the assumption that $n \in \mathbb{N}$ and $h(n) \in \mathbb{N}$, the above equation is equivalent to $h(n) = n$. So we could define the same function $h\colon \mathbb{N} \to \mathbb{N}$ by setting

$$h(n) = n \quad \text{for all} \quad n \in \mathbb{N}\,.$$

Replacing $\mathbb{N}$ with $\mathbb{Z}$ (including the domain and co-domain) renders the first (implicit) definition incorrect (there are two candidates in $\mathbb{Z}$ for $h(n)$ with $n \in \mathbb{Z}$: -1 and n), while the second (explicit) remains correct.

§ I.1.8. Function equality

Let X, Y be sets. We say that two functions $f, g : X \to Y$ are equal (and write $f = g$) meaning that

$$f(x) = g(x) \quad \text{for all} \quad x \in X\,.$$

§ I.1.9. Tuple

A *tuple* with $n \in \mathbb{N}$ components (elements), also referred to as an n-tuple, is an ordered list of n objects (i.e., an n-term sequence). A tuple consisting of objects $x_1, \ldots, x_n$, whatever they are, is denoted by $(x_1, \ldots, x_n)$ or, for brevity, by $(x_i)_{i=1}^n$.

The difference from a set is that elements enter into a tuple together with their positions, so that identical elements at different positions are not identified. Two n-tuples $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are considered equal if and only if $x_i = y_i$ for every $i \in \{1, \ldots, n\}$. For example, $(0, 1, \text{☺}) \neq (1, \text{☺}, 0)$ at least because $0 \neq 1$.

§ I.1.10. Tuples as functions

An n-tuple $(x_1, \ldots, x_n)$ with $n \in \mathbb{N}$ elements $x_1, \ldots, x_n$ from a set X can be naturally identified with the function $\phi\colon \{1, \ldots, n\} \to X$ given by

$$\phi(k) = x_k \quad \text{for each} \quad k \in \{1, \ldots, n\}\,.$$

This is precisely what we mean by saying that *tuples are functions*.

§ I.1.11. Function composition

Consider sets $X, Y, \tilde{Y}, Z$ and functions $\phi\colon X \to Y$ and $\psi\colon \tilde{Y} \to Z$. Assume additionally that $Y \subseteq \tilde{Y}$. Then the equality

$$(\psi \circ \phi)(x) = \psi\bigl(\phi(x)\bigr) \quad \text{for each} \quad x \in X$$

defines a function $\psi \circ \phi\colon X \to Z$, referred to as the *composition* (or *superposition*) of ϕ and ψ. The order matters: it may easily occur that the composition of the same two function in reverse order is not defined. That $\psi \circ \phi$ denotes applying ϕ first and ψ, second (and not the other way round) is a convention, known as the prefix notation.

§ I.1.12. Associativity of function composition

Compositions of functions are functions and can therefore be composed with other functions. For any functions ϕ, ψ, χ such that the compositions $\psi \circ \phi$ and $\chi \circ \psi$ are well defined (which means two inclusions for domains and co-domains), one notes that the compositions $\chi \circ (\psi \circ \phi)$ and $(\chi \circ \psi) \circ \phi$ are both also well defined (i.e., that the corresponding inclusions for domains and co-domains hold). Furthermore, using the pointwise definition of composition, one finds that

$$\big(\chi \circ (\psi \circ \phi)\big)(x) = \chi\big(\psi\big(\phi(x)\big)\big) = \big((\chi \circ \psi) \circ \phi\big)(x) \quad \text{for all} \quad x \in X\,,$$

so that $\chi \circ (\psi \circ \phi) = (\chi \circ \psi) \circ \phi$. This means that composition is an *associative operation*; thanks to this fact, we do not need to put parentheses to indicate the order of composition and may write just $\chi \circ \psi \circ \phi$.

§ I.1.13. Transformation

For any set X, any function $f \colon X \to X$ (i.e., with identical sets used as domain and co-domain) is called a *transformation* of X.

§ I.1.14. Identity transformation

The *identity transformation* of any set is the function that maps every its element into itself. Depending on the context, we will denote the identity transformation of a set X by id, id_X or $\mathsf{id}_{X\to X}$. Using the latter notation, we can formally define the identity transformation of X by

$$\mathsf{id}_{X\to X}(x) = x \quad \text{for all} \quad x \in X\,.$$

§ I.1.15. Inverse function

Consider a function $\phi \colon X \to Y$. Assume that $\psi \colon Y \to X$ is such a function that

$$\psi \circ \phi = \mathsf{id}_{X\to X} \quad \text{and} \quad \phi \circ \psi = \mathsf{id}_{Y\to Y} \tag{I.1.15.1}$$

and that $\chi \colon Y \to X$ is another function satisfying the same conditions. Then, using the associativity of composition, we obtain

$$\chi = \chi \circ \mathsf{id}_{Y\to Y} = \chi \circ (\phi \circ \psi) = \chi \circ \phi \circ \psi = (\chi \circ \phi) \circ \psi = \mathsf{id}_{X\to X} \circ \psi = \psi\,,$$

i.e. $\chi = \psi$, which means that there is at most one function ψ satisfying (I.1.15.1).

When a function ψ satisfying (I.1.15.1) exists, ϕ is called *invertible* and ψ is called the *inverse* of ϕ and is denoted by ϕ^{-1}. Note that the English language often forces us to explicitly indicate, by the choice of the article, whether we are certain of uniqueness or not. Our solution above is to show uniqueness *before* introducing the notion of inverse function. A frequently practiced alternative consists in defining *an* inverse of ϕ without stipulating the uniqueness of such a function by the use of the definite article and then showing the uniqueness thereof and immediately switching to the definite article, talking thenceforth about *the* inverse of ϕ.

Clearly, the relation between ϕ and ϕ^{-1} is symmetric: by the same definition, if ϕ is invertible, then ϕ^{-1} is also so and ϕ is the inverse of ϕ^{-1}.

The latter statement admits the following curious interpretation. In the same context as above, with two sets X and Y, it is natural to consider the sets of all invertible functions from X to Y and of all invertible functions from Y to X:

$$\mathcal{X} = \big\{\phi \colon X \to Y \text{ such that } \phi \text{ is invertible}\big\}$$

and

$$\mathcal{Y} = \big\{\psi \colon Y \to X \text{ such that } \psi \text{ is invertible}\big\}\,.$$

For the two sets X and Y, the operation of function inversion induces the functions $\mathsf{inv}_{\mathcal{X}} \colon \mathcal{X} \to \mathcal{Y}$ and $\mathsf{inv}_{\mathcal{Y}} \colon \mathcal{Y} \to \mathcal{X}$ given by

$$\mathsf{inv}_{\mathcal{X}}(\phi) = \phi^{-1} \quad \text{for all} \quad \phi \in \mathcal{X} \qquad \text{and} \qquad \mathsf{inv}_{\mathcal{Y}}(\psi) = \psi^{-1} \quad \text{for all} \quad \psi \in \mathcal{Y}\,.$$

For these functions, the statement made in the previous paragraph means that the function $\mathsf{inv}_{\mathcal{X}}$ is invertible and the function $\mathsf{inv}_{\mathcal{Y}}$ is its inverse. Obviously, one can use $\mathcal{X}$ and $\mathcal{Y}$ in place of X and Y and consider all invertible functions between $\mathcal{X}$ and $\mathcal{Y}$, and the entire construction can be iterated as many times as needed or even infinitely.

§ I.1.16. Arrows for implications. Necessary and sufficient conditions. Criteria

The symbols "$\Leftarrow$", "$\Rightarrow$" and "$\Leftrightarrow$" are reserved for *implications*, which indicate relations between logical statements and therefore constitute logical statements themselves. These symbols usually can be replaced with the following respective wordings.

(i) "(A) $\Leftarrow$ (B)" reads (A) "is implied by (B)", "(A) holds if (B) holds", "(A) follows from (B)", "(A) is necessary for (B)".

(ii) "(A) $\Rightarrow$ (B)" means "(A) implies (B)", "(A) holds only if (B) holds", "(A) leads to (B)", "(A) is sufficient for (B)".

(iii) "(A) $\Leftrightarrow$ (B)" stands for "(A) holds if and only if (B) holds", "(A) is equivalent to (B)", "(A) is necessary and sufficient for (B)", "(A) is a criterion for (B)".

For example, let $x \in \mathbb{R}$. Then the two implications

$$(x = 1) \Rightarrow (x > 0) \quad \text{and} \quad (x > 0) \Leftarrow (x = 1)$$

are both true statements because $x = 1$ implies $x > 0$ (this is a *proof*, even though trivial). On the other hand, the two implications

$$(x = 1) \Leftarrow (x > 0) \quad \text{and} \quad (x > 0) \Rightarrow (x = 1)$$

are both false statements, and it is sufficient to give a *counterexample* to prove that: for $2 \in \mathbb{R}$, we have $2 > 0$ but $2 \neq 1$. So the negated implications

$$(x = 1) \not\Leftarrow (x > 0) \quad \text{and} \quad (x > 0) \not\Rightarrow (x = 1)$$

are both true statements.

The double-sided arrow ("$\Leftrightarrow$") means a forward implications and a backward one at the same time, and the *negation* thereof ("$\not\Leftrightarrow$") means that at least one of the two one-sided implications is negated (claimed to be false).

Note that the context of an implication matters. In the example definition of h given above, we noted that

$$\big((h(n) - n)\big(h(n) + 1\big) = 0\big) \Leftrightarrow \big(h(n) = n\big)$$

is true in the context of $n, h(n) \in \mathbb{N}$ but false in the context of $n, h(n) \in \mathbb{Z}$ (the backward implication remains true, but the forward one does not).

§ I.1.17. Injective function

A function $\phi \colon X \to Y$ is called *injective* if and only if

$$(x' \neq x) \Rightarrow \big(\phi(x') \neq \phi(x)\big)$$

for all $x, x' \in X$, which is a formal way to express that "distinct arguments are mapped into distinct values".

Injectivity is preserved under composition (§ I.1.11). In the notations of § I.1.11, let us assume that both ϕ and ψ are injective. For any $x, x' \in X$, the equality $\psi(\phi(x)) = \psi(\phi(x'))$ implies $\phi(x) = \phi(x')$ due to the injectivity of ψ, and the latter equality implies $x = x'$ due to the injectivity of ϕ. So we have $(\psi \circ \phi)(x) \neq (\psi \circ \phi)(x')$ for any $x, x' \in X$, so that the composition $\psi \circ \phi$ is an injective function.

§ I.1.18. Surjective function

A function $\phi\colon X \to Y$ is called *surjective* if and only if $\phi(X) = Y$. Since we always have $\phi(X) \subseteq Y$, this effectively means that $Y \subseteq \phi(X)$, that is, that for every $y \in Y$ there exists $x \in X$ such that y is the value of ϕ at x.

Surjectivity is preserved under composition (§ I.1.11) if the domain of the outer function is equal to the co-domain of the inner function. In the notations of § I.1.11, that condition means $\tilde{Y} = Y$. How ψ maps $y \in \tilde{Y} \setminus Y$ is irrelevant for the composition, and we can always redefine ψ by restricting it (see § I.1.30 below) to Y. The surjectivity of the restriction is, however, essential for that of the composition.

In the notations of § I.1.11, let us assume that $\tilde{Y} = Y$ and that both ϕ and ψ are surjective. Consider $z \in Z$. Due to the surjectivity of ψ, there exists $y \in \tilde{Y}$ such that $\psi(y) = z$. Since $\tilde{Y} = Y$, we have $y \in Y$. Then, due to the surjectivity of ϕ, there exists $x \in X$ such that $\phi(x) = y$. Substituting this expression for y into ψ, we obtain $\psi(\phi(x)) = z$. So we have shown that, for any $z \in Z$, there exists $x \in X$ such that $(\psi \circ \phi)(x) = z$, i.e., that the composition $\psi \circ \phi$ is a surjective function.

§ I.1.19. Bijective function

A function $\phi\colon X \to Y$ is called *bijective* if it is both injective and surjective.

Combining what we showed in §§ I.1.17 and I.1.18, we conclude that bijectivity is preserved under composition (§ I.1.11) if the domain of the outer function is equal to the co-domain of the inner function. Indeed, in the notations of § I.1.11, let us assume that $\tilde{Y} = Y$ and that both ϕ and ψ are bijective. Then $\psi \circ \phi$ is both injective (§ I.1.17) and surjective (§ I.1.18), i.e., is bijective.

§ I.1.20. Bijectivity is equivalent to invertibility

Note that

$$(\phi \text{ is bijective}) \Leftrightarrow (\phi \text{ is invertible})$$

for any $\phi\colon X \to Y$. This claim breaks into two implications, which are to be proven independently.

(a) ϕ is bijective $\Rightarrow$ ϕ is invertible (that is, “bijectivity is sufficient for invertibility”, or “invertibility is necessary for bijectivity”)

Proof. Let us assume that ϕ is bijective and define a function $\psi\colon Y \to X$ as follows.

Consider arbitrary $y \in Y$. By the surjectivity of ϕ, there exists $x \in X$ such that $\phi(x) = y$. The injectivity of ϕ gives $\phi(x') \neq \phi(x) = y$ for any $x' \in X$ such that $x' \neq x$, so there is actually exactly one such $x \in X$ that $\phi(x) = y$. In other words, $x \in X$ such that $\phi(x) = y$ is a well-defined element of X, and we can set $\psi(y) = x$.

Let us verify that ψ is the inverse of ϕ.

On the one hand, for every $y \in Y$, we considered $x = \psi(y)$ such that $\phi(x) = y$, i.e., $(\phi \circ \psi)(y) = \phi(\psi(y)) = \phi(x) = y$. On the other hand, for any $x \in X$, we can set $y = \phi(x) \in Y$ and $x' = \psi(y)$. The latter implies, by the definition of ψ, that $\phi(x') = y$, and then the injectivity of ϕ gives $x' = x$. From that we obtain $(\psi \circ \phi)(x) = \psi(y) = x' = x$. We have showed that

$(\phi \circ \psi)(y) = \phi(x) = y$ for all $y \in Y$ and $(\psi \circ \phi)(x) = x$ for all $x \in X$. So ϕ has an inverse and is therefore invertible.

(b) ϕ is invertible $\Rightarrow$ ϕ is bijective (that is, "invertibility is sufficient for bijectivity", or "bijectivity is necessary for invertibility")

Proof. Let us assume that ϕ is invertible.

To show injectivity, consider $x, x' \in X$ and assume that $\phi(x) = \phi(x')$. Since ϕ is invertible, we may apply ϕ^{-1} to both the sides and obtain $\phi^{-1}(\phi(x')) = \phi^{-1}(\phi(x))$ by the definition of the inverse function, which entails $x' = x$. So whenever $x, x' \in X$ are such that $x' \neq x$, it necessarily holds that $\phi(x') \neq \phi(x)$. This means that ϕ is injective.

To show surjectivity, for any $y \in Y$ we consider $x = \phi^{-1}(y)$; by the definition of the inverse function, these satisfy $\phi(x) = \phi(\phi^{-1}(y)) = y$. So ϕ is surjective.

§ I.1.21. Bijectivity as a one-to-one correspondence

Invertibility (equivalently, bijectivity) establishes a *one-to-one correspondence* between X and Y. First, for every $x \in X$, there is a *unique* corresponding element $y = \phi(x) \in Y$, where uniqueness is due to that ϕ is a function. Second, for every $y \in Y$, there is a *unique* corresponding element $x = \phi^{-1}(y) \in X$, where uniqueness is due to that ϕ^{-1} is a function.

§ I.1.22. Finite set

A set is called *finite* if it can be bijectively mapped into $\{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}_0$. Then n is the number of elements in the set, or its *cardinality*. We denote the cardinality of a set X by $\#X$. In particular, $\#\varnothing = 0$ and $(\#X = 0 \Leftrightarrow X = \varnothing)$.

The bijective mapping mentioned above is nothing more than a way to enumerate the elements of the set: every element has to be counted once (hence the enumeration process induces an injective function) and every number from $\{1, 2, \ldots, n\}$ should be used for some element (hence the function is surjective).

§ I.1.23. Infinite set. Countable set. Uncountable set

A set is called *infinite* if it is not finite; the cardinality of an infinite set is defined as infinity: $\#X = \infty$ for any infinite set X.

$\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are all infinite sets (this requires a proof, even if simple).

Any finite set is called *countable*. An infinite set is called *countable* if it can be bijectively mapped into $\mathbb{N}$. A set that is not countable is called *uncountable*.

$\mathbb{Z}$ is a countable set. This is proven by constructing a bijection $\phi : \mathbb{Z} \to \mathbb{N}$; for example, we may define one as follows:

$$\phi(n) = \begin{cases} 1 & \text{if } n = 0, \\ 2n & \text{if } n > 0, \\ 2(-n) + 1 & \text{if } n < 0 \end{cases} \qquad \text{for all} \quad n \in \mathbb{Z}\,.$$

$\mathbb{Q}$ is a countable infinite set. This is proven by constructing a bijection $\phi : \mathbb{Q} \to \mathbb{N}$; for example, see a scheme for the so-called *diagonal argument* in Wikipedia.

§ I.1.24. Bijectivity of transformations

Let ϕ be a transformation of a *finite* set. Then ϕ is bijective if it is injective *or* surjective. This statement is equivalent to a combination of the following two implications.

(a) ϕ is injective $\Rightarrow$ ϕ is surjective

Proof. Let ϕ be an injective transformation of a set $X = \{x_k\}_{k=1}^n$ with $n \in \mathbb{N}$ distinct elements $x_1, \dots, x_n$.

Due to the injectivity of ϕ, the elements $\phi(x_1), \dots, \phi(x_n) \in \phi(X)$ are distinct: indeed, if they were not so, there would exist $i, j \in \{1, \dots, n\}$ such that $\phi(x_i) = \phi(x_j)$. Since we would necessarily have $x_i \neq x_j$, that would contradict the injectivity of ϕ.

Consider $x \in X$. Assuming that $x \notin \phi(X) = \{\phi(x_k)\}_{k=1}^n$, we conclude that X contains at least $n+1$ distinct elements, which contradicts the assumption. So $x \in X$ implies $x \in \phi(X)$, i.e., $X \subseteq \phi(X)$, and so ϕ is surjective.

(b) ϕ is surjective $\Rightarrow$ ϕ is injective

Proof. Let ϕ be a surjective transformation of a set $X = \{x_k\}_{k=1}^n$ with $n \in \mathbb{N}$ distinct elements $x_1, \dots, x_n$.

If ϕ were not injective, there would exist $i, j \in \{1, \dots, n\}$ such that $\phi(x_i) = \phi(x_j)$, and $\phi(X) = \{\phi(x_k)\}_{k=1}^n \subseteq X$ would therefore contain at most $n-1$ elements. Then it would be a proper subset of X, i.e., $\phi(X) \neq X$, which would contradict the surjectivity of ϕ.

Note again that the context of an implication matters. In the above short proofs, we explicitly use that the cardinality of the set is finite when we derive contradictions by increasing or decreasing it by one. For infinite sets, the statement is not generally true. For example, consider $\phi, \psi \colon \mathbb{N} \to \mathbb{N}$ given by

$$\phi(k) = 2k \quad \text{and} \quad \psi(2k) = \psi(2k-1) = k \quad \text{for each} \quad k \in \mathbb{N}\,.$$

It is easy to see that ϕ is injective but not surjective since $\phi(X)$ is the set of even natural numbers and that ψ is surjective but not injective since, for example, $\psi(1) = \psi(2)$.

§ I.1.25. Set partitions

Consider a nonempty set X and a set $\mathcal{Y}$ of nonempty subsets of X (the latter means that $Y \neq \varnothing$ and $Y \subseteq X$ for every $Y \in \mathcal{Y}$). If the elements of $\mathcal{Y}$ are pairwise disjoint (see § I.1.2), i.e., $Y \cap Y' = \varnothing$ for any $Y, Y' \in \mathcal{Y}$ such that $Y \neq Y'$, and the union of all elements of $\mathcal{Y}$ equals X, i.e., $\cup_{Y \in \mathcal{Y}} Y = X$ (see § I.1.2), then $\mathcal{Y}$ is referred to as a *partition* of X. Any partition of a nonempty set clearly has to be a nonempty set.

For example, the sets of odd and even integers form a partition of all integers: for $\mathbb{Z}_{\text{odd}} = \{2k+1 \colon k \in \mathbb{Z}\}$ and $\mathbb{Z}_{\text{even}} = \{2k \colon k \in \mathbb{Z}\}$, we have $\mathbb{Z}_{\text{even}} \cap \mathbb{Z}_{\text{odd}} = \varnothing$ and $\mathbb{Z} = \mathbb{Z}_{\text{odd}} \cup \mathbb{Z}_{\text{even}}$. So $\mathbb{Z}_{\text{odd}}$ and $\mathbb{Z}_{\text{even}}$ form a two-element partition $\mathcal{Z} = \{\mathbb{Z}_{\text{odd}}, \mathbb{Z}_{\text{even}}\}$ of $\mathbb{Z}$.

A set partition is called *ordered* if it is an *ordered set*; we will only use ordered set partitions indexed by sets of natural numbers. In that narrow sense, an ordered partition of a nonempty set X is an injective function (§ I.1.17) defined on a nonempty set $\mathcal{I} \subseteq \mathbb{N}$ that maps every $i \in \mathcal{I}$ into a subset Y_i of Y such that $\mathcal{Y} = \{Y_i\}_{i \in \mathcal{I}}$ is a partition of X. We require injectivity so as to have every element of the partition indexed exactly once.

The above partition $\mathcal{Z} = \{\mathbb{Z}_{\text{odd}}, \mathbb{Z}_{\text{even}}\}$ of $\mathbb{Z}$ can be ordered in many ways; for example, we can choose $\mathcal{I} = \{1, 2\} \subset \mathbb{N}$ and set $Z_1 = \mathbb{Z}_{\text{odd}}$ and $Z_2 = \mathbb{Z}_{\text{even}}$.

§ I.1.26. Real numbers

The set of real numbers is denoted by $\mathbb{R}$. The following chain of inclusions holds:

$$\{0, 1\} \subset \mathbb{N}_0 \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}\,.$$

The margin between $\mathbb{Q}$ and $\mathbb{R}$ includes *algebraic irrational numbers*, (such as $\sqrt{2}$) and *transcendental numbers* (such as π). The set $\mathbb{R}$ is *uncountable*. Loosely speaking, $\mathbb{R}$ is packed with elements so densely that they cannot be enumerated.

There are many constructions that allow to transition from $\mathbb{Q}$ to $\mathbb{R}$; see, for example, en.wikipedia.org/wiki/Construction_of_the_real_numbers. One of the most intuitive ways is to define $\mathbb{R}$ as the set of all $\mathbb{Q}$-valued *Cauchy sequences* identifying any two such sequences (as real numbers, not as sequences) the difference of which converges to zero. This approach, however, requires the basic theory of infinite sequences and of the convergence of such sequences in a form applicable to $\mathbb{Q}$-valued sequences.

§ I.1.27. Kronecker symbol

For notational convenience, we will casually use the so-called *Kronecker symbol*, which is often denoted by δ. This is nothing else than a function of two arguments (usually integers) that compares the arguments and evaluates to 1 if they are equal and to 0 otherwise. So, in the present course, by the Kronecker symbol we will mean the function $\delta\colon \mathbb{Z} \to \{0,1\}$ given by

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

We will use the symbol δ for other functions and constant numbers, and it will be clear in each case — from the context or additionally provided comments — what is meant.

§ I.1.28. Permutations and exchanges

Permutations as tuples and functions. For $n \in \mathbb{N}$, a *permutation* of $\{1,\ldots,n\}$ is a tuple $\sigma = (\sigma_1,\ldots,\sigma_n)$ such that $\sigma_k \in \{1,\ldots,n\}$ for each $k \in \{1,\ldots,n\}$ and $\sigma_i \neq \sigma_j$ for any $i,j \in \{1,\ldots,n\}$ such that $i \neq j$. As an n-tuple with values in $\{1,\ldots,n\}$, a permutation of $\{1,\ldots,n\}$ is naturally identified with the transformation (see § I.1.10) of $\{1,\ldots,n\}$ mapping every $k \in \{1,\ldots,n\}$ into $\sigma_k \in \{1,\ldots,n\}$.

A trivial example is the identity permutation $\imath = (1,2,\ldots,n-1,n)$ of $\{1,\ldots,n\}$, which corresponds to mapping k into $\imath_k = k$ for every $k \in \{1,\ldots,n\}$.

What moves where. A permutation σ of $\{1,\ldots,n\}$, as a function, maps $1,\ldots,n$ into $\sigma_1,\ldots,\sigma_n$, respectively. The term *permutation* refers to the re-ordering of the former list performed by such a function. The latter list is then understood as the result of permuting the former list. In other words, for every $k \in \{1,\ldots,n\}$, the element σ_k of the tuple σ is the element of $\{1,\ldots,n\}$ moved into position k by the permutation.

Bijectivity. Since the components of the permutation tuple are distinct, the associated transformation is injective (see § I.1.17). Then, using § I.1.24, we conclude that it is also surjective and therefore bijective.

Compositions of permutations. The abstract function composition, discussed in § I.1.11, fully applies to permutations considered as functions. In this sense, the composition of any two permutations of $\{1,\ldots,n\}$, which are necessarily injective transformations of $\{1,\ldots,n\}$, is also an injective transformation of $\{1,\ldots,n\}$ and is therefore a permutation.

It is, however, instructive to specialize the abstract notion of function composition to permutations in order to see what a composition of two permutations does in the sense of actually permuting $1,\ldots,n$. Consider two permutations π and σ of $\{1,\ldots,n\}$.

(i) The composition $\pi \circ \sigma$ (we use the standard prefix notation) is given by $(\pi \circ \sigma)_j = \pi_{\sigma_j}$ for every $j \in \{1,\ldots,n\}$.

(ii) As we discussed above, π_k with every $k \in \{1,\ldots,n\}$ is the element of $\{1,\ldots,n\}$ moved into position k by π.

(iii) Substituting σ_j with $j \in \{1,\dots,n\}$ for k in the previous Statements, we obtain that $(\pi\circ\sigma)_j = \pi_{\sigma_j}$ with every $j \in \{1,\dots,n\}$ is the element of $\{1,\dots,n\}$ moved into position σ_j by π.

(iv) On the other hand, σ_j with every $j \in \{1,\dots,n\}$ is the element of $\{1,\dots,n\}$ moved into position j by σ.

As a result, in the description of "what moves where" under the composition $\pi\circ\sigma$, the order of action reverses: for every $j \in \{1,\dots,n\}$, first π moves the element π_{σ_j} into position σ_j, and then σ moves that element into position j. Confusing indeed! This inconsistency stems from that, for a function that is a bijective transformation of $\{1,\dots,n\}$, every *function value represents* what *element* is moved into the respective position *and not* into what *position* the respective element is moved. In other words, the function values represent the old positions of the elements enumerated according to the new arrangement.

Inversion. Applying § I.1.20, we obtain that any permutation π of $\{1,\dots,n\}$ is invertible: there exists a unique bijective transformation π^{-1} of $\{1,\dots,n\}$ such that $\pi^{-1}\circ\pi$ and $\pi\circ\pi^{-1}$ are both equal to the identity permutation of the same set. Since π^{-1} is bijective, it is also injective; the images $\pi^{-1}_1,\dots,\pi^{-1}_n$ of $1,\dots,n$ are therefore distinct, so that π^{-1} is also a permutation.

Let us return to the composition of permutations. If π is a permutation of $\{1,\dots,n\}$, then π_k with every $k \in \{1,\dots,n\}$ is the element of $\{1,\dots,n\}$ moved into position k by π $(\star)$.

For each $j \in \{1,\dots,n\}$, let us denote by σ_j the position into which the element j is moved by π. Since π moves any two elements into distinct positions, $\sigma_1,\dots,\sigma_n$ are all pairwise distinct. Then $\sigma = (\sigma_1,\dots,\sigma_n)$ is a permutation of $\{1,\dots,n\}$. Furthermore, substituting π_k with $k \in \{1,\dots,n\}$ for j in the definition of σ_j, we obtain that σ_{π_k} is the position into which the element π_k is moved by π, which is k by $(\star)$. So we have that $\sigma\circ\pi$ is the identity transformation of $\{1,\dots,n\}$. Similarly, by $(\star)$, π_{σ_j} with every $j \in \{1,\dots,n\}$ is the element moved into position σ_j by π, which is j by the definition of σ_j. So we have that $\pi\circ\sigma$ is the identity transformation of $\{1,\dots,n\}$. As a result, $\sigma = \pi^{-1}$, i.e., σ is the inverse of π, the existence of which we justified in abstract terms above.

To summarize, a permutation π of $\mathcal{I} = \{1,\dots,n\}$ is a bijective function $\pi\colon \mathcal{I} \to \mathcal{I}$. It has a unique inverse $\pi^{-1}\colon \mathcal{I} \to \mathcal{I}$, which is also a bijection. The function values $\pi_1,\dots,\pi_n$ of π are the *elements* that move into positions $1,\dots,n$ under π, whereas the function values $\pi^{-1}_1,\dots,\pi^{-1}_n$ of π^{-1} are the *positions* into which elements $1,\dots,n$ move under π.

Exchanges. Let σ be a permutation of $\{1,\dots,n\}$ that permutes *at most* two elements, i.e., let σ be different from $(1,\dots,n)$ in *at most* two elements. This means that there exist $k, j \in \{1,\dots,n\}$ such that $\sigma_i = i$ for each $i \in \{1,\dots,n\} \setminus \{k,j\}$. Then, since $\sigma_1,\dots,\sigma_n$ are distinct, we have that neither of σ_k and σ_j can coincide with any of $\sigma_i = i$ with $i \in \{1,\dots,n\} \setminus \{k,j\}$. This implies that $\sigma_k, \sigma_j \in \{k,j\}$. Since $\sigma_1,\dots,\sigma_n$ are distinct, this leaves only the following possibilities: $\sigma_k = k$ and $\sigma_j = j$ or $\sigma_k = j$ and $\sigma_j = k$. In either case, such a permutation is called the (k,j)*-exchange permutation of* $\{1,\dots,n\}$ and also the (j,k)*-exchange permutation of* $\{1,\dots,n\}$. The case of $k = j$ is a special but perfectly valid example. Exchanges are permutations that are very simple and nice, not just in how they are defined or parametrized but also in that $\sigma^{-1} = \sigma$ evidently holds for any exchange σ.

§ I.1.29. Cartesian product

For any sets $X_1,\dots,X_n$ with $n \in \mathbb{N}$, one can consider the set of n-tuples with elements form the respective sets. This set is called the *Cartesian product* of $X_1,\dots,X_n$ and is denoted by $X_1\times\cdots\times X_n$:

$$X_1\times\cdots\times X_n = \big\{(x_1,\dots,x_n)\colon x_k \in X_k \text{ for each } k \in \{1,\dots,n\}\big\}\,.$$

For fastidious readers and, in particular, those with previous exposure to mathematical logic and computer programming, let us note the following. In mathematics, it is a widespread convention that $(X_1 \times X_2) \times X_3$ is identified with $X_1 \times X_2 \times X_3$, i.e., every tuple $(x_1, x_2) \in X_1 \times X_2$ is broken into $x_1 \in X_1$ and $x_2 \in X_2$ before being combined into all possible tuples (x_1, x_2, x_3) with $x_3 \in X_3$. This convention trivially implies that the binary operation of Cartesian multiplication is *associative*: $(X_1 \times X_2) \times X_3 = X_1 \times X_2 \times X_3 = X_1 \times (X_2 \times X_3)$. This convention is a useful notational device that, however, may introduce undesired ambiguity. In programming languages designed for carefully keeping track of the *types* of *objects* (these terms can be understood here in a broad sense or, for object-oriented programming languages, also in the special sense), the convention is not used, and a *splatting operator* has to be explicitly applied to any tuple (x_1, x_2) when it is used to construct a tuple (x_1, x_2, x_3) with some object x_3.

§ I.1.30. Function restriction

For sets U, V, W such that $U \subseteq V$, any function $\phi \colon V \to W$ trivially induces another function, $\phi|_U \colon U \to W$ given by $\phi|_U(u) = \phi(u)$ for all $u \in U$ and called the *restriction of* ϕ *to* U.

Essentially, the restriction of a function to a subset of its domain is obtained from the function by discarding the information about how the elements from the remainder of the domain are mapped.

A restriction of a function can have different properties than the original function. For example, the function $\phi \colon [-1, 1] \to [0, 2]$ given by $\phi(x) = 2x^2$ for all $x \in [-1, 1]$ is not invertible (it is not injective), whereas $\phi|_{[0,1]}$, given by the same formula but with the domain $[0, 1]$, is invertible.

As we noted in § I.1.30, by *function* we understand not just a formula, so many different functions can be defined using one formula and one function can be defined using many different formulae. In this sense, ϕ and $\phi|_U$ in the above context are different functions unless $U = V$.

§ I.1.31. Unary and binary operations

The following examples of *operations* with numbers should be familiar to the reader:

(i) addition, denoted by “$\cdot + \cdot$”: $(x, y) \mapsto x + y \in \mathbb{R}$ for $x, y \in \mathbb{R}$;

(ii) multiplication, denoted by “$\cdot \times \cdot$”: $(x, y) \mapsto x \times y \in \mathbb{R}$ for $x, y \in \mathbb{R}$;

(iii) additive inversion, denoted by “$- \cdot$”: $x \mapsto -x \in \mathbb{R}$ for $x \in \mathbb{R}$;

(iv) multiplicative inversion, denoted by “$\cdot^{-1}$” or “$\frac{1}{\cdot}$”: $x \mapsto x^{-1} \in \mathbb{R}$ for $x \in \mathbb{R}$ such that $x \neq 0$;

(v) doubling: $x \mapsto 2 \times x \in \mathbb{R}$ for $x \in \mathbb{R}$;

(vi) division, denoted by “$\frac{\cdot}{\cdot}$” (fraction) or by the signs “$\div$”, “$/$”: $(p, q) \mapsto \frac{p}{q} = x \div y = p/q \in \mathbb{Q}$ for $p \in \mathbb{Z}$ and $q \in \mathbb{N}$;

(vii) squaring, denoted by “$\cdot^2$”: $x \mapsto x \times x \in \mathbb{R}$ for $x \in \mathbb{R}$;

(viii) absolute value, denoted by $|\cdot|$: $x \mapsto x$ for $x \in \mathbb{R}$ such that $x \geq 0$ and $x \mapsto -x \in \mathbb{R}$ for $x \in \mathbb{R}$ such that $x < 0$;

(ix) extraction of the square root of a natural number, denoted by “$\sqrt{\cdot}$”: $n \mapsto \sqrt{n} \in \mathbb{R}$ for $n \in \mathbb{N}$.

In general, *operation* is a essentially a synonym for *function* (see § I.1.7), but the former term is often used with the connotation of simplicity or fundamentality. For example, the idea that two numbers can be added or multiplied ("operated on") is usually presented long ahead of the formal notion of function. Essentially, the term *operation*, being somewhat less formal, is not so strictly associated with specific sets as its domain and co-domain and is useful in avoiding the notions of domain and co-domain. For example, it is useful to think about function composition "$\circ$" (see § I.1.11) and Cartesian product "$\times$" of sets (see § I.1.29) even before considering any sets of functions or any sets of sets as possible domains and co-domains of these operations as functions.

As we see from the above list of examples, very diverse notational conventions have been adopted for the most "fundamental" operations.

An interpunct ("$\cdot$") is used as a placeholder for arguments when it is desirable for notational reasons to indicate (i) that arguments can be specified, i.e., that the symbol meant to denote an operation actually refers to an operation, or (ii) where exactly arguments can be specified to denote the evaluation of the operation. For example, "$|\cdot|$", "$\sqrt{\cdot}$" or "$\cdot^2$" could look rather ambiguous without such a placeholder. Another example is the functions that we will introduce as *inner products* in this course. We will use the short notation "$\langle\cdot\,,\cdot\rangle$" to denote such a function, meaning nothing more than a function with two arguments (mapping *any* pair of u and v from certain sets U and V into some well-defined value $\langle u, v\rangle$) and not just a pair of angular brackets, *without implying that the arguments are equal.* For inner products, the arguments will be chosen from the same set, but the notation "$\langle\cdot\,,\cdot\rangle$" itself is not understood so as to imply that.

Additionally, the modifiers *unary*, *binary*, *ternary* and so on are used to indicate how many single arguments (of a type that is known from the context) the operation under consideration takes. So, if X, Y, Z are some sets, then a function $f\colon X \to Z$ is called a *unary operation on X with values in Z*. Similarly, a function g mapping every possible pair of $x \in X$ and $y \in Y$ into $g(x, y) \in Z$ is called a *binary operation on X and Y with values in Z*. For the sake of brevity, when $Y = X$, such an operation is often called a *binary operation on X with values in Z*.

Note that an arbitrary pair of $x \in X$ and $y \in Y$ can be identified with the tuple $(x, y) \in X \times Y$, and vice versa (refer to § I.1.29 regarding the Cartesian product "$\times$" of sets). In this sense, a binary operation g on X and Y can be considered as a unary operation on $X \times Y$ and, for example, a unary operation on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ can be considered as a binary operation on $\mathbb{R}$.

§ I.1.32. Quantifiers

"$\forall$", "$\exists$", "!" are so-called quantifiers that are often used as abbreviations for "for all", "there exist(s)" and "unique". For example, the definition of that $\phi\colon X \to Y$ is injective could be expressed in terms of quantifiers as follows:

$$\phi(x') \neq \phi(x) \quad \forall x, x' \in X : x' \neq x$$

or, equivalently,

$$(\phi(x') = \phi(x) \Rightarrow x' = x) \quad \forall x, x' \in X$$

or, equivalently,

$$\forall y \in \phi(X)\ \exists!\, x \in X : \phi(x) = y\,.$$

In the present notes, words will be preferred to the quantifiers "$\forall$", "$\exists$" and "!". The quantifiers will, however, be occasionally used at the lectures in the interest of time. Both at the lectures and in the notes, we will work with natural-language statements and proofs. Formal proofs, such as the proof given at us.metamath.org/mpeuni/binomlem.html, are not very useful for the purpose of the course and will be practiced *neither* at the lectures *nor* in the notes.

§ I.2. Complex numbers

This section is a summary of everything that we may need to know about complex numbers for this course.

§ I.2.1. Motivation

The set $\mathbb{R}$ is richer than $\mathbb{Q}$ as it includes all points to which $\mathbb{Q}$-valued sequences can converge. Actually, $\mathbb{R}$ is much richer than $\mathbb{Q}$ in the sense that the former is uncountable whereas the latter is countable. Still, $\mathbb{R}$ is not sufficiently rich in another way: not all algebraic equations with coefficients in $\mathbb{R}$, such as

$$a_0 + \sum_{k=1}^{n} a_k x^k = 0 \quad \text{with respect to } x$$

with $n \in \mathbb{N}$ and $a_0, \ldots, a_n \in \mathbb{R}$, have solutions in $\mathbb{R}$, and among those that do, not all have as many solutions as one may need. Even the simple equation

$$x^2 + 1 = 0 \quad \text{with respect to } x\,,$$

which is an algebraic equation of degree two with real (in fact, integral) coefficients, has no solution in $\mathbb{R}$. This can be considered as a motivation for transcending the notion of real number.

§ I.2.2. Complex numbers as pairs of reals

Let us consider the set of pairs (two-tuples) of real numbers, denoted by $\mathbb{R}^2$:

$$\mathbb{R}^2 = \{(x,y)\colon x, y \in \mathbb{R}\} = \mathbb{R} \times \mathbb{R}\,.$$

It is natural to add such pairs and to multiply them by real numbers elementwise:

$$(x,y) + (x',y') = (x+x', y+y') \quad \text{and} \quad \alpha \cdot (x,y) = (\alpha \cdot x, \alpha \cdot y) \quad \text{for all} \quad x, x', y, y', \alpha \in \mathbb{R}\,. \tag{I.2.2.1}$$

These equalities, in fact, use the operations of addition and multiplication ($+$ and “ $\cdot$ ”) on $\mathbb{R}$ to define two new operations: of adding two elements of $\mathbb{R}^2$ (also denoted by $+$) and of multiplying an element of $\mathbb{R}^2$ by an element of $\mathbb{R}$ (also denoted by “ $\cdot$ ”).

> **Remark I.2.2.1.** Note that the operation of addition that is denoted by $+$ on the right-hand side of the first of equalities (I.2.2.1) is similar to but is *different from* the operation of addition that is denoted by the same symbol on the left-hand side. We use the former, assumed to be well defined, to define the latter. A similar remark holds for multiplication by a scalar.

Let us select two elements of $\mathbb{R}^2$: $\vec{1} = (1,0)$ and $\mathsf{i} = (0,1)$. Then, using the above operations, we obtain

$$(x,y) = x \cdot \vec{1} + y \cdot \mathsf{i} \quad \text{for all} \quad x, y \in \mathbb{R}\,. \tag{I.2.2.2}$$

The elements $\vec{1}$ and i are called the *real unit* and *imaginary unit.*

§ I.2.3. Complex multiplication. Complex numbers

Further, one can introduce another operation, which will work for the elements of $\mathbb{R}^2$ very similarly to how the usual multiplication works for the elements of $\mathbb{R}$. Let us denote it also by “ $\cdot$ ” and define it as follows:

$$(x,y) \cdot (x',y') = (x \cdot x' - y \cdot y', x \cdot y' + x' \cdot y) \quad \text{for all} \quad x, x', y, y' \in \mathbb{R}\,. \tag{I.2.3.1}$$

This binary operation is *commutative*: $(x,y) \cdot (x',y') = (x',y') \cdot (x,y)$ holds trivially for all $x, x', y, y' \in \mathbb{R}$ (inspect the right-hand side of the defining formula) because the real multiplication and addition are commutative. What is interesting is that the above definition implies

$$\vec{1} \cdot (x,y) = (x,y) \quad \text{and} \quad \mathsf{i} \cdot (x,y) = (-y,x) \quad \text{for all} \quad x, y \in \mathbb{R} \tag{I.2.3.2}$$

and, in particular,

$$\vec{1}^2 = \vec{1} \quad \text{and} \quad \mathsf{i}^2 = -\vec{1}\,. \tag{I.2.3.3}$$

This shows that $\vec{1}$ behaves in $\mathbb{R}^2$ with respect to the newly introduced multiplication just as 1 behaves in $\mathbb{R}$ with respect to the real multiplication. The imaginary unit, on the other hand, exhibits a different behavior.

The function $\phi\colon \mathbb{R} \to \mathbb{R} \times \{0\}$ given by

$$\phi(x) = (x, 0) \quad \text{for all} \quad x \in \mathbb{R} \tag{I.2.3.4}$$

is a bijection; it allows to “safely” (in a one-to-one fashion) identify $\mathbb{R}$ with

$$\mathbb{R} \times \{0\} = \{(x, 0)\colon x \in \mathbb{R}\} \subset \mathbb{R}^2\,.$$

Specifically, any $x \in \mathbb{R}$ this bijection allows to understand as $\phi(x) = (x, 0) \in \mathbb{R}^2$ whenever we need to interpret it as a pair of real numbers. That gives meaning to certain statements; for example, $(x, y) = 3$ for $x, y \in \mathbb{R}$: indeed, we understand 3 as $\phi(3) = (3, 0) \in \mathbb{R}^2$, and precisely in this sense is the statement $(x, y) = 3$ equivalent, for any $x, y \in \mathbb{R}$, to that $x = 3$ and $y = 0$. This identification is just a notational device. One often refers to mappings like (I.2.3.4) as *identity embeddings* in the sense that they allow to identify their domains and co-domains. The standard notation for such identification is “$\simeq$”; the identification of $\mathbb{R} \simeq \mathbb{R} \times \{0\}$ is expressed by $\mathbb{R} \simeq \mathbb{R} \times \{0\}$, where the identity embedding is supposed to have been specified or to be obvious. Other examples are the identity embeddings of $\mathbb{N}$ into $\mathbb{Z}$, of $\mathbb{Z}$ into $\mathbb{Q}$ and of $\mathbb{Q}$ into $\mathbb{R}$, which are used in the construction of integers, rational numbers and real numbers.

Identifying $1 \in \mathbb{R}$ and $0 \in \mathbb{R}$ with $\vec{1} \in \mathbb{R}^2$ and $\vec{0} = (0, 0) \in \mathbb{R}^2$ and placing the imaginary unit as a fore factor (which we can always do thanks to the commutativity noted above), we transform the above relations into

$$(x, y) = x + \mathsf{i}y \quad \text{for all} \quad x, y \in \mathbb{R}\,, \tag{I.2.3.5}$$

$$(x + \mathsf{i}y) + (x' + \mathsf{i}y') = (x + x') + \mathsf{i}(y + y') \quad \text{for all} \quad x, x', y, y' \in \mathbb{R} \tag{I.2.3.6}$$

and

$$(x + \mathsf{i}y) \cdot (x' + \mathsf{i}y') = (xx' - yy') + \mathsf{i}(xy' + x'y) \quad \text{for all} \quad x, x', y, y' \in \mathbb{R}\,. \tag{I.2.3.7}$$

Expressions of the form $x + \mathsf{i}y$ with $x, y \in \mathbb{R}$, understood as elements of $\mathbb{R}^2$ in the sense of (I.2.3.5), with the addition and multiplication given by (I.2.3.6) and (I.2.3.7) are called *complex numbers*. The set of such expressions, which can be seen as elements of $\mathbb{R}^2$, is called the set of complex numbers and is denoted, as a set, by $\mathbb{C}$:

$$\mathbb{C} = \{x + \mathsf{i}y\colon x, y \in \mathbb{R}\}\,.$$

We will formally introduce the notion of *field* below. At this point, we only note that $\mathbb{C}$ and the addition and multiplication given by (I.2.3.6) and (I.2.3.7) are such that $\mathbb{C}$ is a field with respect to these operations. Specifically, $\mathbb{C}$ with addition and multiplication defined according to (I.2.3.6) and (I.2.3.7) is called the *field of complex numbers*, and the notation $\mathbb{C}$ is understood to refer to the specific operations of addition and multiplication of complex numbers *as well as* the set itself.

§ I.2.4. Real and imaginary parts. Absolute value. Complex conjugation and multiplicative inverse

For any complex number $z \in \mathbb{C}$, the real numbers $x, y \in \mathbb{R}$ such that $z = x + \mathsf{i}y$ are well defined (i.e., exist and are unique) and are called the *real part* and *imaginary part* of z respectively; they are denoted by $\operatorname{Re} z$ and $\operatorname{Im} z$: $\operatorname{Re}(x + \mathsf{i}y) = x$ and $\operatorname{Im}(x + \mathsf{i}y) = y$ for any $x, y \in \mathbb{R}$.

For any complex number $z \in \mathbb{C}$, the nonnegative real number $|z| = \sqrt{(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2}$ is called the *absolute value* of z. Note $(z = 0 \Leftrightarrow |z| = 0)$ for every $z \in \mathbb{C}$.

For any complex number $z \in \mathbb{C}$, the complex number $\overline{z} = \operatorname{Re} z - \mathrm{i} \operatorname{Im} z$ is called the *complex conjugate* of z. For $z \in \mathbb{C}$ with $x = \operatorname{Re} z \in \mathbb{R}$ and $y = \operatorname{Im} z \in \mathbb{R}$, we therefore have

$$\overline{z} = \overline{\operatorname{Re} z + \mathrm{i} \operatorname{Im} z} = \overline{x + \mathrm{i}y} = x - \mathrm{i}y = \operatorname{Re} z - \mathrm{i} \operatorname{Im} z\,. \tag{I.2.4.1}$$

We see immediately from (I.2.3.7) that $\overline{z_1 z_2} = \overline{z_1}\;\overline{z_2}$ for any $z_1, z_2 \in \mathbb{C}$; furthermore, $|\overline{z}| = |z|$ and $z\overline{z} = |z|^2$ for any $z \in \mathbb{C}$. Then $|z_1 z_2|^2 = z_1 z_2 \overline{z_1 z_2} = z_1 \overline{z_1} z_2 \overline{z_2} = |z_1|^2\,|z_2|^2$ and hence $|z_1 z_2| = |z_1| \cdot |z_2|$ for any $z_1, z_2 \in \mathbb{C}$. Finally, for any $z \in \mathbb{C}$ such that $z \neq 0$, we can define the "reciprocal" (*multiplicative inverse*) of z as

$$\frac{1}{z} = \frac{\overline{z}}{|z|^2}\,. \tag{I.2.4.2}$$

§ I.2.5. Algebraic form of complex numbers

For any $z \in \mathbb{C}$, let us consider $x = \operatorname{Re} z \in \mathbb{R}$ and $y = \operatorname{Im} z \in \mathbb{R}$. Then the representation

$$z = x + \mathrm{i}y = \operatorname{Re} z + \mathrm{i} \operatorname{Im} z \tag{I.2.5.1}$$

is the so-called *algebraic form* of z as a complex number.

In §§ I.2.6 and I.2.8, we will explore two other standard ways of representing complex numbers. The three standard forms of complex numbers will be useful for particular tasks in the present course.

§ I.2.6. Trigonometric form of complex numbers

Let us consider $z \in \mathbb{C}$ such that $z \neq 0$ and set $x = \operatorname{Re} z$, $y = \operatorname{Im} z$ and $\rho = |z|$. Then, in particular, we have $|z| = \sqrt{x^2 + y^2} \neq 0$.

Note that

$$\frac{z}{\rho} = \frac{x}{\sqrt{x^2 + y^2}} + \mathrm{i}\,\frac{y}{\sqrt{x^2 + y^2}}\,.$$

As we know from trigonometry, there exists a unique number $\theta \in [0, 2\pi)$ such that

$$\frac{x}{\sqrt{x^2 + y^2}} = \cos\theta \quad \text{and} \quad \frac{y}{\sqrt{x^2 + y^2}} = \sin\theta\,;$$

for example, one way to express this number is as follows:

$$\theta = \begin{cases} \arccos \dfrac{x}{\sqrt{x^2 + y^2}} & \text{if } y \geq 0\,, \\ 2\pi - \arccos \dfrac{x}{\sqrt{x^2 + y^2}} & \text{if } y < 0\,. \end{cases}$$

Clearly, θ could be replaced with $\theta + 2\pi k$ with any $k \in \mathbb{Z}$ if we had not required that $\theta \in [0, 2\pi)$. However, it is convenient to consider θ "modulo 2π", i.e., to choose the only one from $[0, 2\pi)$. This unique θ is called the *normalized argument* of the nonzero complex number z and is denoted by $\operatorname{Arg} z$, so that $\operatorname{Arg} z = \theta$ in terms of our current notations. This leads us to the *trigonometric form* of z as a complex number:

$$z = \rho\cos\theta + \mathrm{i}\rho\sin\theta = |z|\cos\operatorname{Arg} z + \mathrm{i}|z|\sin\operatorname{Arg} z\,. \tag{I.2.6.1}$$

Recall that $z = (x, y) = (\rho\cos\theta, \rho\sin\theta) \in \mathbb{R}^2$ to notice that $\operatorname{Arg} z = \theta$ is actually the angle between the primary (horizontal) axis and vector $(x, y) \in \mathbb{R}^2$ measured counterclockwise, whereas $|z| = \rho$ is the length of this vector on the plane.

As clearly follows from (I.2.4.1) and from our decision to measure arguments of complex numbers in $[0, 2\pi)$, conjugation of a complex number in the trigonometric form amounts to replacing the argument by its complement to 2π (unless $\mathsf{Arg}\, z = 0$): for any $z \in \mathbb{C}$, we have

$$\overline{z} = \rho\cos\theta - \mathrm{i}\rho\sin\theta = \rho\cos\theta + \mathrm{i}\rho\sin(-\theta) = |z|\cos\mathsf{Arg}\, z + \mathrm{i}|z|\sin(-\mathsf{Arg}\, z) \tag{I.2.6.2}$$

and, in particular, $\mathsf{Arg}\,\overline{z} = 2\pi - \mathsf{Arg}\, z$ if $\mathsf{Arg}\, z \neq 0$. This allows to rewrite (I.2.4.2) as follows:

$$\frac{1}{z} = \frac{1}{\rho(\cos\theta + \mathrm{i}\sin\theta)} = \rho^{-1}(\cos\theta - \mathrm{i}\sin\theta) = |z|^{-1}\big(\cos\mathsf{Arg}\, z + \mathrm{i}\sin(-\mathsf{Arg}\, z)\big)\,. \tag{I.2.6.3}$$

§ I.2.7. The complex exponential function

The complex exponential function can be defined by setting

$$\exp(x + \mathrm{i}y) = \exp(x)\cdot\cos(y) + \mathrm{i}\exp(x)\cdot\sin(y) \quad \text{for all} \quad x, y \in \mathbb{R}\,. \tag{I.2.7.1}$$

This definition is given in terms of the real trigonometric functions cos and sin and of the real exponential function exp, which are assumed to have been defined. Within this construction, strictly speaking, $\exp(x)$ on the right-hand side and $\exp(x + \mathrm{i}y)$ on the left-hand side of (I.2.7.1) are values of different functions, with domains $\mathbb{R}$ and $\mathbb{C}$, at arguments that are different even for $y = 0$. The former domain, $\mathbb{R}$, is in a one-to-one correspondence (I.2.3.4) with $\mathbb{R} \times \{0\} = \{x + \mathrm{i}0 \colon x \in \mathbb{R}\} \subset \mathbb{C}$, the real line of the complex plane. This correspondence immediately yields another function $\widetilde{\exp} = \phi \circ \exp \circ\, \phi^{-1} \colon \mathbb{R} \times \{0\} \to \mathbb{C}$, so that $\widetilde{\exp}(x + \mathrm{i}0) = \exp(x) + \mathrm{i}0 \in \mathbb{C}$ for any $x \in \mathbb{R}$. There is, of course, little substance in this definition: it says no more than that $\widetilde{\exp}(\,x + \mathrm{i}0)$ is defined by converting the real complex number $x + \mathrm{i}0$ into a real number, applying the real exponential function exp and converting the real result into a real complex number. Finally, we extend $\widetilde{\exp}$ to a third function, $\widehat{\exp} \colon \mathbb{C} \to \mathbb{C}$, given by $\widehat{\exp}(x + \mathrm{i}y) = \widetilde{\exp}(x + \mathrm{i}0)\cos(y) + \mathrm{i}\widetilde{\exp}(x + \mathrm{i}0)\sin(y)$ for all $x, y \in \mathbb{R}$. This is an *extension* of $\widetilde{\exp}$ from $\mathbb{R} \times \{0\}$ to $\mathbb{C}$, which means that the domain of $\widehat{\exp}$ is $\mathbb{C}$ and that $\widetilde{\exp}$ is the restriction of $\widehat{\exp}$ to $\mathbb{R} \times \{0\}$.

If we use the identity embedding (I.2.3.4) $\mathbb{R}$ into $\mathbb{R} \times \{0\}$, the functions $\widetilde{\exp}$ and exp coincide ("up to the identity embedding", which is convenient not to write explicitly) and $\widehat{\exp}$ becomes the extension of exp from $\mathbb{R}$ to $\mathbb{C}$. On the other hand, exp is then just a restriction of $\widehat{\exp}$ to $\mathbb{R}$. Typically there is no need to have these three distinct functions denoted by symbols. So "exp" is used to denote the complex exponential, which can be evaluated, in particular, on $\mathbb{R} \times \{0\} \simeq \mathbb{R}$.

It can be shown that the definition (I.2.7.1) implies, with respect to operations on complex numbers, many of the properties that the real exponential function has with respect to operations on real numbers. For example, $\exp(z)\cdot\exp(z_0) = \exp(z + z_0)$ for all $z, z_0 \in \mathbb{C}$, $\exp(0 + \mathrm{i}0) = 1 + \mathrm{i}0$ and

$$\lim_{z \to z_0} \frac{\exp(z) - \exp(z_0)}{z - z_0} = \exp(z_0) \quad \text{for each} \quad z_0 \in \mathbb{C}\,.$$

These properties are consequences of (I.2.7.1) and of properties of the real trigonometric and exponential functions. Euler's formula

$$e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta\,, \tag{I.2.7.2}$$

valid for every $\theta \in \mathbb{R}$, trivially follows from (I.2.7.1).

An alternative but equivalent definition is based on extending the real trigonometric and exponential functions to the complex plane using their Taylor series:

$$\cos z = \sum_{k=0}^{\infty}(-1)^k\frac{z^{2k}}{(2k)!}\,, \quad \sin z = \sum_{k=0}^{\infty}(-1)^k\frac{z^{2k+1}}{(2k+1)!} \quad \text{and} \quad \exp(z) = \sum_{k=0}^{\infty}\frac{z^k}{k!} \tag{I.2.7.3}$$

for all $z \in \mathbb{C}$. This definition of exp implies (I.2.7.1) and the aforementioned properties. This definition requires certain care as it involves infinite functional series. In Analysis courses, one learns that the infinite summation is not an issue in (I.2.7.3): the series behave extremely well on the whole of the complex plane (they exhibit *uniform absolute convergence on any bounded subset of* $\mathbb{C}$).

§ I.2.8. Exponential form of complex numbers

In the context of § I.2.6, Euler's formula (I.2.7.2) for $\theta \in [0, 2\pi)$, leads us to the so-called *exponential form* of z as a complex number:

$$z = \rho e^{\mathsf{i}\theta} = |z|\, e^{\mathsf{i}\,\mathsf{Arg}\,z}\,. \tag{I.2.8.1}$$

Rewriting (I.2.6.2) in the exponential form, we obtain

$$\overline{z} = \overline{\rho e^{\mathsf{i}\theta}} = \rho e^{\mathsf{i}(-\theta)} = |z|\, e^{-\mathsf{i}\,\mathsf{Arg}\,z}\,. \tag{I.2.8.2}$$

Finally, recasting (I.2.6.3) in exponential form, we arrive at

$$\frac{1}{z} = \frac{1}{\rho e^{\mathsf{i}\theta}} = \rho^{-1} e^{-\mathsf{i}\theta} = |z|^{-1}\, e^{-\mathsf{i}\,\mathsf{Arg}\,z}\,, \tag{I.2.8.3}$$

which we could have also derived from basic properties of the exponential function.

§ I.2.9. Roots of unity

The exponential form of complex numbers is very helpful in solving certain equations, which will appear in the present course in relation to the *eigenvalue decomposition* of matrices. So let us consider, for $a \in \mathbb{C}$ nonzero and $n \in \mathbb{N}$, the equation

$$z^n = a \tag{I.2.9.1}$$

with respect to $z \in \mathbb{C}$. Representing a in the exponential form $a = \rho e^{\mathsf{i}\theta}$ with $\rho = |a|$ and $\theta = \mathsf{Arg}\,a$ and seeking z in the exponential form $z = re^{\mathsf{i}\varphi}$ with $r = |z|$ and $\varphi = \mathsf{Arg}\,z$, we transform the original equation into the *equivalent* form $r^n e^{\mathsf{i}n\varphi} = \rho e^{\mathsf{i}\theta}$, which is, in turn, equivalent to

$$r^n e^{\mathsf{i}(n\varphi - \theta)} = \rho\,. \tag{I.2.9.2}$$

Since the absolute values of the two sides have to be equal for every solution, we immediately conclude that $r = \sqrt[n]{\rho}$. Now it remains to find the set of all suitable arguments: with the found value of r, equation (I.2.9.2) is equivalent to $e^{\mathsf{i}(n\phi-\theta)} = 1$, which is, by Euler's formula, equivalent to that $\cos(n\phi - \theta) = 1$ and $\sin(n\phi - \theta) = 0$. These two conditions are equivalent to $n\phi = \theta + 2\pi k$ with $k \in \mathbb{Z}$. So the set of all solutions to equation (I.2.9.1) over $\mathbb{C}$ is

$$\Big\{ z_k = \sqrt[n]{\rho}\, \exp\Big(\frac{\mathsf{i}}{n}(\theta + 2\pi k)\Big) \colon k \in \mathbb{Z} \Big\} = \Big\{ z_k = \sqrt[n]{\rho}\, \exp\Big(\frac{\mathsf{i}}{n}(\theta + 2\pi k)\Big) \Big\}_{k=0}^{n-1} \tag{I.2.9.3}$$

and contains *exactly* n *distinct elements*. This result, in fact, expresses the complex nth root, possibly multivalued, in terms of the algebraic nth root (the single-valued nth root of a nonnegative real number): for any $\rho \in \mathbb{R}$ nonnegative, $\theta \in \mathbb{R}$ and $n \in \mathbb{N}$,

$$\big\{ z \in \mathbb{C} \colon z^n = \rho e^{\mathsf{i}\theta} \big\} = \Big\{ \sqrt[n]{\rho}\, e^{\mathsf{i}(\theta + 2\pi k)/n} \colon k \in \{0, \ldots, n-1\} \Big\}\,.$$

In the particular case of $a = 1$ (which corresponds to $\rho = 1$ and $\theta = 0$), the numbers $z_0, \ldots, z_{n-1}$, defined in (I.2.9.3), are referred to as the *roots of unity of degree* n.

For example, the equation $z^3 = 1$ with respect to $z \in \mathbb{C}$ has *three* solutions: one is $1 + \mathsf{i} \cdot 0 = 1$, lying on the real line, and the other two are

$$\exp\frac{2\pi\mathsf{i}}{3} = -\frac{1}{2} + \mathsf{i}\frac{\sqrt{3}}{2} \quad \text{and} \quad \exp\frac{4\pi\mathsf{i}}{3} = -\frac{1}{2} - \mathsf{i}\frac{\sqrt{3}}{2}\,,$$

both lying in the complex plane off the real line!

Chapter II. Columns and matrices. Gaussian elimination

§ II.1. Field

§ II.1.1. Definition of a field

Definition II.1.1.1 (field)**.** Consider a set $\mathbb{F}$ and functions $\boxplus\colon \mathbb{F}\times\mathbb{F}\to\mathbb{F}$ and $\boxtimes\colon \mathbb{F}\times\mathbb{F}\to\mathbb{F}$. The set $\mathbb{F}$ is called a *field with respect to operations* $\boxplus$ *and* $\boxtimes$ *of addition and multiplication* if the following conditions are satisfied with some distinct elements $0, 1 \in \mathbb{F}$. These elements are then called the *additive* and *multiplicative identity elements* of $\mathbb{F}$.

- **(a)** Commutativity: for any $\alpha, \beta \in \mathbb{F}$, we have $\alpha \boxplus \beta = \beta \boxplus \alpha$ and $\alpha \boxtimes \beta = \beta \boxtimes \alpha$.
- **(b)** Associativity: for any $\alpha, \beta, \gamma \in \mathbb{F}$, we have $(\alpha\boxplus\beta)\boxplus\gamma = \alpha\boxplus(\beta\boxplus\gamma)$ and $(\alpha\boxtimes\beta)\boxtimes\gamma = \alpha\boxtimes(\beta\boxtimes\gamma)$.
- **(c)** Additive and multiplicative identity elements: $\alpha\boxplus 0 = \alpha$ and $1\boxtimes\alpha = \alpha$ for any $\alpha \in \mathbb{F}$;
- **(d)** Additive inverse: for every $\alpha \in \mathbb{F}$, there exists $\beta \in \mathbb{F}$ such that $\alpha \boxplus \beta = 0$.
- **(e)** Multiplicative inverse: for every $\alpha \in \mathbb{F}$ such that $\alpha \neq 0$, there exists $\beta \in \mathbb{F}$ such that $\alpha \boxtimes \beta = 1$.
- **(f)** Distributivity: for any α, β, γ, we have $\alpha \boxtimes (\beta \boxplus \gamma) = (\alpha \boxtimes \beta) \boxplus (\alpha \boxtimes \gamma)$.

The above conditions are known as *field axioms*.

The elements of a field are called *scalars*. Actually, by saying that something is a scalar, one states nothing else than that it is an element of a field, the latter usually being clear from the context.

Loosely speaking, a field is a pool of *scalars*.

It is hard to argue that the properties required of a field and field operations in definition II.1.1.1 are anything unusual for what we know as *numbers*, or *scalars*. These conditions are everything we need from $\mathbb{F}$, from the identity elements $0, 1$ of $\mathbb{F}$ and from the operations of multiplication and addition for all definitions and results of this chapter II to be correctly formulated and valid.

§ II.1.2. Examples of fields

Example II.1.2.1 ($\mathbb{Q}$ is a field)**.** The set $\mathbb{Q}$ of real rational numbers (see § I.1.6) is a field with respect to the operations of rational addition and multiplication.

Example II.1.2.2 ($\mathbb{R}$ is a field)**.** The set $\mathbb{R}$ of real numbers is a field with respect to the operations of real addition and multiplication.

Proof. The reader is referred to the *axioms of real numbers*, which define an abstract *field of real numbers* as a field with additional properties (to be precise, the field of real numbers is a unique, up to *isomorphism*, *Dedekind-complete ordered* field). Any construction of real numbers (see § I.1.26) has to be justified as such by proving that it satisfies the axioms of real numbers.

Example II.1.2.3 ($\mathbb{C}$ is a field)**.** The set $\mathbb{C}$ of complex numbers is a field with respect to the operations of complex addition and multiplication.

Proof. The proof is left to the reader as an exercise. The claim follows from that $\mathbb{R}$ is a field and from the construction of the complex numbers presented in §§ I.2.2 and I.2.3. In particular, the definitions (I.2.3.6) and (I.2.3.7) of complex addition and multiplication, given in terms of their real counterparts, translate the field structure of $\mathbb{R}$ into that of $\mathbb{C}$.

To illustrate that fields $\mathbb{R}$ and $\mathbb{C}$ are special and that quite different a set can be a field with respect to rather unusual operations of addition and multiplication, let us consider the following example.

Example II.1.2.4 ($\mathbb{F}_2$ is a finite field)**.** Consider $\mathbb{F}_2 = \{0, 1\}$ with just two distinct elements, denoted by 0 and 1, and the functions $\boxplus, \boxtimes \colon \mathbb{F}_2 \times \mathbb{F}_2 \to \mathbb{F}_2$ defined as follows:

$$0 \boxplus 0 = 1 \boxplus 1 = 0 \quad \text{and} \quad 1 \boxplus 0 = 0 \boxplus 1 = 1\,,$$

$$0 \boxtimes 0 = 1 \boxtimes 0 = 0 \boxtimes 1 = 0 \quad \text{and} \quad 1 \boxtimes 1 = 1\,.$$

Then $\mathbb{F}_2$ is a field with respect to $\boxplus$ and $\boxtimes$ as operations of addition and multiplication, and 0 and 1 are its additive and multiplicative identity elements respectively.

The field $\mathbb{F}_2$, considered in example II.1.2.4, underlies modern digital signal processing, computer science and cryptography. The use of the symbols “0” and “1” for denoting the elements of $\mathbb{F}_2$ is a pure convention, and we could as well have denoted them by “☹” and “☺” or “`false`” and “`true`” instead. In the latter case, the addition and multiplication of $\mathbb{F}_2$ are the logical operations known as `XOR` (exclusive `OR`) and `AND`. For $\mathbb{F}_2 = \{0, 1\}$ to be a field, the only thing that matters is that 0 and 1 are distinct and behave under $\boxplus$ and $\boxtimes$ according to the playbook specified in example II.1.2.4. In fact, how they behave under the two operations is the only interpretation that 0 and 1 have as elements of $\mathbb{F}_2$.

Another thing to note is that $\mathbb{F}_2$ is the smallest field: it contains two elements while any field has to contain at least two elements, and for any set with two elements to be a field with respect to some operations $\boxplus, \boxtimes \colon \mathbb{F}_2 \times \mathbb{F}_2 \to \mathbb{F}_2$, it is required that the elements behave under $\boxplus$ and $\boxtimes$ as 0 and 1 do in example II.1.2.4.

§ II.1.3. Scope of chapter II in regard to the notion of a field

The theoretical construction of this chapter II is entirely applicable to any field $\mathbb{F}$. The reader is invited to keep in mind that all notions and results presented in chapter II are correct and valid for any abstract field $\mathbb{F}$, and, in particular, for the field $\mathbb{F}_2$ considered in example II.1.2.4 as well as for $\mathbb{R}$ and $\mathbb{C}$, notwithstanding that the exposition of the material is motivated and illustrated by examples specific to $\mathbb{R}$ and $\mathbb{C}$.

§ II.2. Column vector. Vector space

§ II.2.1. Definitions and linear properties of column vectors

Definition II.2.1.1 (column vector)**.** Let $n \in \mathbb{N}$. By a column vector with n components from $\mathbb{F}$ (alternatively, “with n elements”, “with n entries” or “of size n”), we mean an n-element $\mathbb{F}$-valued tuple,

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \equiv [x_j]_{j=1}^{n}\,,$$

where $x_1, \ldots, x_n \in \mathbb{F}$ are called *components*, *elements* or *entries*.

According to (II.2.1.1), column vectors are essentially nothing else than tuples of scalars from $\mathbb{F}$ except for that we use the square brackets to represent them as columns for future conformity with matrices. In particular, two tuples are equal if and only if all the respective components are equal. The equivalence sign (“$\equiv$”) is used rather indiscriminately in mathematics to express “strong” equalities; in definition II.2.1.1, its use between two expressions that are equal but are given in slightly different notations serves to indicate that the equality is somewhat universal (notational in this case) and holds independently from the context (in this case, from what the definition is about).

Remark II.2.1.2 (column vectors as functions; entry notation)**.** As any tuple (see § I.1.10), a column vector of size $n \in \mathbb{N}$ can be considered as a function mapping elements of $\{1, \dots, n\}$ into elements of $\mathbb{F}$. Arguments of such functions (indices) will be attached in subscript to symbols and expressions denoting column vectors (placed in parentheses when needed for clarity) to evaluate them and obtain components (cf. § I.1.7). In this sense, depending on the context, x_1 may denote the first component of a column vector $x = [x_j]_{j=1}^n \in \mathbb{F}^n$ with $n \in \mathbb{N}$ or a column vector from $\mathbb{F}^n$ with $n \in \mathbb{N}$, the first component of which could then be referenced with the notation $(x_1)_1$.

Definition II.2.1.3 ($\mathbb{F}^n$, the set of n-component $\mathbb{F}$-valued tuples)**.** Let $n \in \mathbb{N}$. By $\mathbb{F}^n$ we denote the set of n-component column vectors with components from $\mathbb{F}$:

$$\mathbb{F}^n = \left\{ [x_j]_{j=1}^n : x_1, \dots, x_n \in \mathbb{F} \right\}.$$

As it is often done with tuples, we can now use the addition and multiplication defined on $\mathbb{F}$ to define the addition of column vectors and the multiplication of column vectors by scalars *componentwise.*

Definition II.2.1.4 (addition on $\mathbb{F}^n$)**.** Let $n \in \mathbb{N}$. For any $x = [x_j]_{j=1}^n \in \mathbb{F}^n$ and $y = [y_j]_{j=1}^n \in \mathbb{F}^n$, we define $x + y \in \mathbb{F}^n$ by setting

$$(x + y)_j = x_j + y_j \quad \text{for each} \quad j \in \{1, \dots, n\} \tag{II.2.1.1a}$$

or, equivalently, by

$$x + y = [x_j + y_j]_{j=1}^n . \tag{II.2.1.1b}$$

Definition II.2.1.5 (multiplication by scalars on $\mathbb{F}^n$)**.** Let $n \in \mathbb{N}$. For any $x = [x_j]_{j=1}^n \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$, we define $\alpha \cdot x \in \mathbb{F}^n$ by setting

$$(\alpha \cdot x)_j = \alpha \cdot x_j \quad \text{for each} \quad j \in \{1, \dots, n\} \tag{II.2.1.2a}$$

or, equivalently, by

$$\alpha \cdot x = [\alpha \cdot x_j]_{j=1}^n . \tag{II.2.1.2b}$$

Definition II.2.1.6 (zero column vector)**.** For any $n \in \mathbb{N}$, by the *zero column vector of size* n we mean the column vector

$$[0]_{j=1}^{n} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{F}^n .$$

Here, 0 denotes the zero from $\mathbb{F}$.

Remark II.2.1.7. Zero column vectors are often denoted by 0, but this notation, obviously, does not specify the size of the column vector and therefore does not uniquely identify any particular zero column vector. Whenever this notation is used, the size should be clear from the context.

For any $n \in \mathbb{N}$, one can verify that due to the definition of the zero column vector (definition II.2.1.6) and to the definitions of the linear operations on $\mathbb{F}^n$ (definitions II.2.1.4 and II.2.1.5), the field axioms (the conditions required in definition II.1.1.1) translate into the following properties of $\mathbb{F}^n$.

Proposition II.2.1.8 (linear structure of $\mathbb{F}^n$)**.** Let $n \in \mathbb{N}$ and 0 denote the zero column vector of $\mathbb{F}^n$. The addition of column vectors and the multiplication of column vectors by scalars introduced above satisfy the following properties.

(a) Commutativity: for any $x, y \in \mathbb{F}^n$, we have $x + y = y + x$.

(b) Associativity: for any $x, y, z \in \mathbb{F}^n$, we have $(x + y) + z = x + (y + z)$. Also, for any $\alpha, \beta \in \mathbb{F}$ and $x \in \mathbb{F}^n$, we have $(\alpha \cdot \beta) \cdot x = \alpha \cdot (\beta \cdot x)$.

(c) Additive identity: the zero column vector 0 satisfies $0 + x = x$ for any $x \in \mathbb{F}^n$, and there is no other element in $\mathbb{F}^n$ with this property.

(d) Additive inverse: for every $x \in \mathbb{F}^n$, the element $y = [-x_j]_{j=1}^{n} \in \mathbb{F}^n$ satisfies $x + y = 0$ and there is no other element in $\mathbb{F}^n$ satisfying this property in place of y.

(e) Neutrality of the multiplicative identity element of $\mathbb{F}$: the element $1 \in \mathbb{F}$ satisfies $1 \cdot x = x$ for every $x \in \mathbb{F}^n$.

(f) Distributivity: for all $x, y \in \mathbb{F}^n$ and $\alpha, \beta \in \mathbb{F}$,
  (i) $(\alpha + \beta) \cdot x = (\alpha \cdot x) + (\beta \cdot x)$;
  (ii) $\alpha \cdot (x + y) = (\alpha \cdot x) + (\alpha \cdot y)$.

Proof. These facts are easy to show using definitions II.1.1.1 and II.2.1.4 to II.2.1.6. The proof is left to the reader as an exercise.

§ II.2.2. Vector space

In this section, we consider an *arbitrary* set V and *arbitrary* functions $\oplus \colon V \times V \to V$ and $\odot \colon \mathbb{F} \times V \to V$. Certain conditions, called *vector-space axioms* and given under (a) to (f) in definition II.2.2.1 below, mean that the functions $\oplus$ and $\odot$ operate on the elements of $\mathbb{F}$ and V similarly to as the corresponding operations of addition and multiplication by scalars, defined in definitions II.2.1.4 and II.2.1.5 operate on $\mathbb{F}$ and $\mathbb{F}^n$ with $n \in \mathbb{N}$. We can use these conditions to select a class of sets with structure, which includes $\mathbb{F}^n$ with any $n \in \mathbb{N}$ as a particular case but also any other set V with two operations that *resemble*, in how they operate on the elements of $\mathbb{F}$ and V, that *specific example* of addition and multiplication by scalars.

Definition II.2.2.1 (vector space)**.** Let $\mathbb{F}$ be a field with respect to operations $\boxplus$ and $\boxtimes$ and 1 denote its multiplicative identity element. Consider a set V and functions $\oplus\colon V \times V \to V$ and $\odot\colon \mathbb{F} \times V \to V$. The set V is called a *vector space over the field* $\mathbb{F}$ *with respect to operations* $\oplus$ *and* $\odot$ *of addition and multiplication by a scalar* if the following conditions are satisfied with some $\vec{0} \in V$.

- **(a)** Commutativity: for any $u, v \in V$, we have $u \oplus v = v \oplus u$.
- **(b)** Associativity: for any $u, v, w \in V$, we have $(u \oplus v) \oplus w = u \oplus (v \oplus w)$. Also, for any $\alpha, \beta \in \mathbb{F}$ and $u \in V$, we have $(\alpha \boxtimes \beta) \odot u = \alpha \odot (\beta \odot u)$.
- **(c)** Additive identity: $\vec{0} \oplus u = u$ for all $u \in V$
- **(d)** Additive inverse: for every $u \in V$, there exists an element $v \in V$ such that $u \oplus v = \vec{0}$.
- **(e)** Neutrality of the multiplicative identity element of the field: $1 \odot u = u$ for every $u \in V$.
- **(f)** Distributivity: for all $u, v \in V$ and $\alpha, \beta \in \mathbb{F}$, we have:
  - (i) $(\alpha \boxplus \beta) \odot u = (\alpha \odot u) \oplus (\beta \odot u)$;
  - (ii) $\alpha \odot (u \oplus v) = (\alpha \odot u) \oplus (\alpha \odot v)$.

A vector space is called *trivial* if it consists of a single element and *nontrivial*, otherwise.

A vector space over a field $\mathbb{F}$ is called a *real vector space* when $\mathbb{F} = \mathbb{R}$ and a *complex vectro space* when $\mathbb{F} = \mathbb{C}$.

The elements of a vector space are called *vectors*. Actually, by saying that something is a vector, one states nothing else than that it is an element of a vector space, the latter usually being clear from the context. Loosely speaking, a vector space is a pool of *vectors*.

Remark II.2.2.2. Proving that a set V is a vector space over the field $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$ of addition and multiplication by a scalar means showing the following. First, one needs to verify that $\oplus\colon V \times V \to V$ and $\odot\colon \mathbb{F} \times V \to V$, i.e., that V is closed under these operations. This means that $u \oplus v$ and $\alpha \odot u$ are *well defined as elements of* V for all $u, v \in V$ and $\alpha \in \mathbb{F}$. Many candidates for vector spaces (sets together with two operations) that we can dream up fall short of this closedness. Second, one needs to find a candidate $\vec{0}$ for an additive identity. This is usually easy to do by taking $0 \in \mathbb{F}$ and looking at what $\vec{0} = 0 \odot u \in V$ is for some $u \in V$. In fact, $0 \odot u$ *has to* be equal to *the* additive identity of V for any $u \in V$ *if* V is a vector space over $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$ (we may soon choose to prove this simple fact, as well as the uniqueness of an additive identity, to get the taste and meaning of the vector-space abstraction). Third, one needs to check all the axioms stated in definition II.2.2.1. These steps lead to the conclusion that V is a vector space over the field $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$.

The associativity condition (b) listed in definition II.2.2.1 allows to define the sum of any number r of vectors from a vector space V recursively, applying the addition operation to add one vector to the sum of the other $r - 1$ vectors. The commutativity condition (a) ensures that the result does not depend on the way in which that one vector is selected among the r vectors. The convenience of introducing the abstract notion of a vector space is that many properties, such as this one, hold in the same abstract setting and follow from the abstract definition alone, independently from what exactly the scalars and vectors are and how exactly the two operations are defined in any specific example of the notion.

Remark II.2.2.3. Let V be a trivial vector space over a field $\mathbb{F}$ and v be its only element. Then the operations $\oplus\colon V\times V\to V$ and $\odot\colon \mathbb{F}\times V\to V$ of V necessarily satisfy

$$v\oplus v=v \quad \text{and} \quad \alpha\odot v=v \quad \text{for each} \quad \alpha\in\mathbb{F}\,.$$

The only element of a trivial vector space necessarily takes the role of additive identity element. Trivial vector spaces are rather useless as they do not provide any structure, i.e., relations between their elements — for the very lack of even two distinct elements. Nevertheless, within any nontrivial vector space, the additive identity element forms the corresponding trivial vector space.

The direct comparison of definition II.2.2.1 and proposition II.2.1.8 yields the following.

Corollary II.2.2.4. For any $n\in\mathbb{N}$, consider the set $\mathbb{F}^n$ and the zero column vector $0\in\mathbb{F}^n$. The set $\mathbb{F}^n$ is a vector space over the field $\mathbb{F}$ in the sense of definition II.2.2.1 with additive identity 0 with respect to the operations of addition and multiplication by a scalar defined in definitions II.2.1.4 and II.2.1.5.

Example II.2.2.5 ($\mathbb{C}^n$ as a real vector space)**.** For any $n\in\mathbb{N}$, consider the set $\mathbb{C}^n$ and the zero column vector $0\in\mathbb{C}^n$. The set $\mathbb{C}^n$ is a vector space over the field $\mathbb{R}$ in the sense of definition II.2.2.1 with additive identity 0 with respect to the operations "$+$" and "$\cdot$" of addition and multiplication by a scalar given by (II.2.1.1) and (II.2.1.2) for all $x,y\in\mathbb{C}^n$ and $\alpha\in\mathbb{R}$.

Proof. The proof is left to the reader as an exercise.

Example II.2.2.6 (function space)**.** Consider a field $\mathbb{F}$ with operations $+$ and "$\cdot$" of addition and multiplication and with the respective identity elements 0 and 1. Let V denote the set $\mathcal{F}(\mathcal{D},\mathbb{F})$ of $\mathbb{F}$-valued functions defined on a set $\mathcal{D}$:

$$V=\mathcal{F}(\mathcal{D},\mathbb{F})=\{(f\colon\mathcal{D}\to\mathbb{F})\}\,.$$

The equality of the elements of V is understood in the sense of function equality (see § I.1.8), i.e., pointwise. This setting covers, in particular, the set $\mathbb{F}$-valued sequences (with $\mathcal{D}=\mathbb{N}$) and, somewhat artificially, also the set $\mathbb{F}$ (with a one-element set $\mathcal{D}$), which we considered in Problem 1 (a) of Assignment 1 in the particular case of $\mathbb{F}=\mathbb{C}$.

One can verify that the set V is a vector space over the field $\mathbb{F}$ with respect to the *pointwise operations* $\oplus\colon V\times V\to V$ *and* $\odot\colon\mathbb{F}\times V\to V$ *of addition and multiplication by a scalar*, which are given by

$$(f\oplus g)(t)=f(t)+g(t) \quad \text{and} \quad (\alpha\odot f)(t)=\alpha\cdot f(t) \quad \text{at every} \quad t\in\mathcal{D}$$

for all $f,g\in V$ and $\alpha\in\mathbb{F}$. The zero function $\vec{0}$, given by

$$\vec{0}(t)=0 \quad \text{for all} \quad t\in\mathcal{D}\,,$$

is a unique additive identity. Indeed, every $f\in V$ satisfies $\vec{0}(t)+f(t)=f(t)$ for all $t\in\mathcal{D}$ by conditions (a) and (c) of definition II.1.1.1, which is equivalent to $\vec{0}\oplus f=f$ due to the definitions of function equality and addition (both pointwise). Second, if $g\in V$ is such that $g \oplus f = f$ for any $f \in V$, then the definitions of function equality and addition imply $g(t) + \alpha = \alpha$ for all $\alpha \in \mathbb{F}$ and $t \in \mathcal{D}$ (for every $\alpha \in \mathbb{F}$, consider the function $f \in V$ given by $f(t) = \alpha$ for all $t \in \mathcal{D}$), which implies $g(t) = 0$ for all $t \in \mathcal{D}$ due to conditions (a) and (c) of definition II.1.1.1, so that $g = \vec{0}$.

Note that, even though the uniqueness of the additive identity can be derived from the vector-space axioms in general, it can also be established explicitly in any specific setting as a consequence of the specific definitions of addition and equality, as we claimed in part (c) of proposition II.2.1.8 for vector spaces of column vectors, showed in example II.2.2.6 for vector spaces of functions and will claim in part (c) of proposition II.3.1.7 for vector spaces of matrices. The beauty of an abstract definition, such as definition II.2.2.1, is in its generality and minimality: it uses a minimal set of requirements to pinpoint a general structure exhibited in specific settings, so that various properties that may be observed completely trivially in those specific settings turn out to be corollaries of those minimal requirements.

Definition II.2.2.7 (linear combination)**.** Let V be a vector space over the field $\mathbb{F}$ with respect to operations $+$ and $\cdot$ of addition and multiplication by a scalar. Consider $r \in \mathbb{N}$, vectors $v_1, \ldots, v_r \in V$ and scalars $\alpha_1, \ldots, \alpha_r \in \mathbb{F}$. Then the *linear combination* of $v_1, \ldots, v_r$ with *coefficients* $\alpha_1, \ldots, \alpha_r$ is the vector $v \in V$ defined by

$$v = \alpha_1 \cdot v_1 + \cdots + \alpha_r \cdot v_r \equiv \sum_{k=1}^{r} \alpha_k \cdot v_k \, . \tag{II.2.2.1}$$

The coefficient tuple $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{F}^r$ is often referred to as a *coefficient* of v *with respect to* $v_1, \ldots, v_r$. This coefficient is called *trivial* if it is a zero tuple and *nontrivial* otherwise. The linear combination is called *trivial* if it equals to the additive identity element of V and *nontrivial* otherwise.

Remark II.2.2.8 (linear combination and vector space)**.** Definition II.2.2.7 is correct due to definition II.2.2.1. Indeed, the vector space is closed under its linear operations. Further, due to the commutativity of vector addition (condition (a) of definition II.2.2.1), the order of summation in (II.2.2.1) has no effect on the value of the sum. Furthermore, the vectors $v_1, \ldots, v_r$ and coefficients $\alpha_1, \ldots, \alpha_r$ can be reordered consistently in any way with no effect on the value of the corresponding linear combination.

Remark II.2.2.9 (nonuniqueness of representation in the form of a linear combination)**.** If V is a vector space over a field $\mathbb{F}$, then the linear combination of vectors $v_1, \ldots, v_r \in V$ with coefficients $\alpha_1, \ldots, \alpha_r \in \mathbb{F}$ is uniquely defined by (II.2.2.1) in definition II.2.2.7 once $v_1, \ldots, v_r \in V$ and $\alpha_1, \ldots, \alpha_r \in \mathbb{F}$ have been fixed. At the same time, it may be the case that one vector can be represented as a linear combination of the same vectors with many different coefficients, so the representation of v given by (II.2.2.1) may be nonunique.

Example II.2.2.10. Let us see how definition II.2.2.7 specializes in the case of $V = \mathbb{F}^n$ with $n \in \mathbb{N}$, in which case the linear operations take the explicit form given by definitions II.2.1.4 and II.2.1.5. Consider $r \in \mathbb{N}$. For any $x_k = [x_{jk}]_{j=1}^{n} \in \mathbb{F}^n$ and $\alpha_k \in \mathbb{F}$ with $k \in \{1, \ldots, r\}$, we have

$$\alpha_1 \cdot v_1 + \cdots + \alpha_r \cdot v_r = \sum_{k=1}^{r} \alpha_k \cdot v_k = [\alpha_1 \cdot x_{j1} + \cdots + \alpha_r \cdot x_{jr}]_{j=1}^{n} = \begin{bmatrix} \alpha_1 \cdot x_{11} + \cdots + \alpha_r \cdot x_{1r} \\ \vdots \\ \alpha_1 \cdot x_{n1} + \cdots + \alpha_r \cdot x_{nr} \end{bmatrix} .$$

The first and third equalities merely relate equivalent notations. The second equality is more substantial; it is obtained by applying definitions II.2.1.4 and II.2.1.5. Given those componentwise definitions, it should not be surprising that linear combinations in $\mathbb{F}^n$ turn out to be defined componentwise.

Notice that, while introducing the column vectors in this example, we immediately set for their entries a notation more convenient in this example than the generic one presented in remark II.2.1.2.

Example II.2.2.11 (uniqueness and nonuniqueness of representation for $\mathbb{R}^2$)**.** In the case of $\mathbb{F} = \mathbb{R}$ and $V = \mathbb{R}^2$, let us consider $e_1 = (1, 0) \in V$, $e_2 = (0, 1) \in V$ and $u = (u_1, u_2) \in V$.

Any column vector $v = (v_1, v_2) \in V$ can be represented in terms of e_1 and e_2 as follows:

$$v = \alpha_1 \cdot e_1 + \alpha_2 \cdot e_2 \tag{II.2.2.2}$$

with $\alpha_1 = v_1$ and $\alpha_2 = v_2$. In fact, such a representation is unique. Indeed, expressing the column-vector equation (II.2.2.2) componentwise, we obtain an equivalent system of scalar equations with respect to $\alpha_1, \alpha_2 \in \mathbb{R}$:

$$\begin{cases} 1 \cdot \alpha_1 + 0 \cdot \alpha_2 = & v_1 , \\ 0 \cdot \alpha_1 + 1 \cdot \alpha_2 = & v_2 . \end{cases} \tag{II.2.2.3}$$

This obviously has a unique solution, which is (v_1, v_2). On the other hand, we can also represent the same column vector v as a a linear combination of e_1, e_2 and u:

$$v = \alpha_1 \cdot e_1 + \alpha_2 \cdot e_2 + \lambda \cdot u \tag{II.2.2.4}$$

with $\alpha_1 = v_1$, $\alpha_2 = v_2$ and $\lambda = 0$ follows from (II.2.2.2).

The fact that we now use a larger (and therefore redundant) set of column vectors (e_1, e_2 and u instead of e_1 and e_2) to represent v is closely related to that the representation (II.2.2.4) is not unique. Indeed, expressing the column-vector equation (II.2.2.4) componentwise, we obtain the following equivalent system of scalar equations with respect to $\alpha_1, \alpha_2, \lambda \in \mathbb{R}$:

$$\begin{cases} 1 \cdot \alpha_1 + 0 \cdot \alpha_2 + u_1 \cdot \lambda = & v_1 , \\ 0 \cdot \alpha_1 + 1 \cdot \alpha_2 + u_2 \cdot \lambda = & v_2 . \end{cases} \tag{II.2.2.5}$$

The unknown λ can be chosen arbitrary, with which α_1 and α_2 are determined uniquely, just as in (II.2.2.3). The set of all possible coefficients is therefore

$$\{(v_1 - \lambda u_1, v_2 - \lambda u_2, \lambda) \colon \lambda \in \mathbb{R}\} \subset \mathbb{R}^3 . \tag{II.2.2.6}$$

Actually, this result is closely related to the fact that u is a linear combination of e_1 and e_2: since $u = u_1 \cdot e_1 + u_2 \cdot e_2$, we can transform (II.2.2.2) into (II.2.2.4) as follows:

$$v = (v_1 \cdot e_1 + v_2 \cdot e_2 - \lambda \cdot u) + \lambda \cdot u = (v_1 - \lambda u_1) \cdot e_1 + v_2 - \lambda u_2 \cdot e_2 + \lambda \cdot u$$

for any $\lambda \in \mathbb{R}$, and the transformation of (II.2.2.4) with any $\lambda \in \mathbb{R}$ into (II.2.2.2) is even more straightforward.

§ II.3. Matrices: linear structure

§ II.3.1. Definitions and properties of matrices with regard to linear operations

Definition II.3.1.1 (matrix over $\mathbb{F}$)**.** Let $m, n \in \mathbb{N}$. By a *matrix* of size $m \times n$ over the field $\mathbb{F}$, we mean a function $\{1, \ldots, m\} \times \{1, \ldots, n\} \to \mathbb{F}$. Any such a function A is uniquely defined by its mn values at all possible index pairs, $A_{ij} \in \mathbb{F}$ with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, which are called the *entries* of the matrix A.

For any $a_{ij} \in \mathbb{F}$ with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, by any of

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \text{and} \quad [a_{ij}]_{i=1,\, j=1}^{m\quad n} \tag{II.3.1.1}$$

we denote the unique matrix A of size $m \times n$ over the field $\mathbb{F}$ such that $A_{ij} = a_{ij}$ for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$.

Let A be a matrix of size $m \times n$ over the field $\mathbb{F}$.

(a) For each $i \in \{1, \ldots, m\}$, the ith *row* of A is the matrix $[A_{kj}]_{k=i,\, j=1}^{i\quad n} \in \mathbb{F}^{1\times n}$.

(b) For each $j \in \{1, \ldots, n\}$, the jth *column* of A is the matrix $[A_{ik}]_{i=1,\, k=j}^{m\quad j} \in \mathbb{F}^{m\times 1}$.

Definition II.3.1.2 ($\mathbb{F}^{m\times n}$)**.** Let $m, n \in \mathbb{N}$. By $\mathbb{F}^{m\times n}$ we denote the set of matrices of size $m \times n$ with entries from $\mathbb{F}$:

$$\mathbb{F}^{m\times n} = \left\{ [a_{ij}]_{i=1,\, j=1}^{m\quad n} \colon \; a_{ij} \in \mathbb{F} \text{ for all } i \in \{1, \ldots, m\} \text{ and } j \in \{1, \ldots, n\} \right\}.$$

Definition II.3.1.1 introduces standard notations for $m \times n$ matrices over $\mathbb{F}$ based on the identification of such matrices with "tables", "two-dimensional arrays" or "two-index tuples" of scalars with m rows and n columns, and it is only due to the lack of a simpler formal definition of the quoted terms that we start with defining matrices as functions (cf. § I.1.10 for the case of tuples). The identification with tables of scalars — explicit, graphic and otherwise convenient — is, of course, the reason for defining rows and columns as in definition II.3.1.1 and for referring to the indices i and j in the context definition II.3.1.1 as, respectively, *row* and *column indices*.

Alternatively, one can see a matrix as a tuple of rows or as a tuple of columns. This helps in not mentioning functions in the definition of *matrix* (by hiding the notion of function in the definition of a tuple), and $\mathbb{F}^{m\times n}$ can then be introduced as

$$\bigtimes_{i=1}^{m} \mathbb{F}^n \quad \text{or} \quad \bigtimes_{j=1}^{n} \mathbb{F}^m .$$

Representing a matrix as a tuple of its rows or as a tuple of its columns in practice, at the level of syntax in programming languages, is rather unwieldy, as `Python` demonstrates.

Remark II.3.1.3 (entry notation and possible ambiguity)**.** It is standard to refer to the entries of a matrix $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ by A_{ij} with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, but one has to take care of avoiding ambiguity when using this notation.

The reference to an entry of a matrix can be interpreted as the evaluation of the matrix as a function at the respective index pair. Referencing matrix entries with attaching index pairs in subscript is standard, even though different from the usual notation for the evaluation of functions (see § I.1.7). A safe way of using that convention is to enclose the symbol denoting the matrix in a parenthesis whenever the omission of the parenthesis would lead to ambiguity and to omit it otherwise for convenience.

The only possible source of ambiguity is that the symbol obtained by omitting the parenthesis is already in use. For example, if A_{11} has been defined as a matrix of size 2×2 and A has been introduced as another matrix, then the upper left entry of $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ is clearly distinct from the matrix already denoted by A_{11} and may be referred to by $(A)_{11}$. A better solution, of course, is to use the natural language to introduce a notation for that entry (or for all the entries at once).

Finally, it is a standard convention that entry evaluation has *higher priority* than any of the matrix operations we define below. For that reason, matrix-valued expressions involving matrix operations should be enclosed in parentheses before entry evaluation. For example, for some $A \in \mathbb{F}^{n\times n}$ with $n \in \mathbb{N}$, we will define below a matrix $A^{-1} \in \mathbb{F}^{n\times n}$. Then the notation $\left(A^{-1}\right)_{ij}$ with $i, j \in \{1, \dots, n\}$ will refer to entry (i, j) of the matrix obtained by evaluating the matrix-valued expression A^{-1} (however we define this expression) and then evaluating the resulting matrix at (i, j). Clearly, this does not have to be the same as A_{ij}^{-1}, which is understood as the result of taking entry (i, j) of A (evaluating A at (i, j)) and inverting that number from $\mathbb{F}$ according to the rules of $\mathbb{F}$.

Exactly as for tuples and column vectors, we can use the zero of $\mathbb{F}$ and the addition and multiplication defined on $\mathbb{F}$ to introduce the zero matrix, the addition of matrices and the multiplication of matrices by scalars.

Definition II.3.1.4 (addition on $\mathbb{F}^{m\times n}$)**.** Let $m, n \in \mathbb{N}$. For any $A, B \in \mathbb{F}^{m\times n}$, we define $A + B \in \mathbb{F}^{m\times n}$ by setting $(A + B)_{ij} = A_{ij} + B_{ij}$ for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ or, equivalently, by $A + B = [A_{ij} + B_{ij}]_{i=1,\, j=1}^{m \quad n}$.

Definition II.3.1.5 (multiplication by scalars on $\mathbb{F}^{m\times n}$)**.** Let $m, n \in \mathbb{N}$. For any $A \in \mathbb{F}^{m\times n}$ and $\alpha \in \mathbb{F}$, we define $\alpha \cdot A \in \mathbb{F}^{m\times n}$ by setting $(\alpha \cdot A)_{ij} = \alpha \cdot A_{ij}$ for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ or, equivalently, by $\alpha \cdot A = [\alpha \cdot A_{ij}]_{i=1,\, j=1}^{m \quad n}$.

Definition II.3.1.6 (zero matrix)**.** For any $m, n \in \mathbb{N}$, by the *zero matrix of size* $m \times n$ we mean the matrix

$$[0]_{i=1,\, j=1}^{m \quad n} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \in \mathbb{F}^{m\times n}\,.$$

Here, 0 denotes the zero from $\mathbb{F}$.

Zero matrices are often denoted by O, and so remark II.2.1.7 equally applies to them.

With these definitions, just as for column vectors, we can verify that properties of real and complex numbers imply the following properties of the newly introduced operations on $\mathbb{F}^{m\times n}$.

Proposition II.3.1.7 (linear structure of $\mathbb{F}^{m\times n}$)**.** Let $m, n \in \mathbb{N}$ and $O \in \mathbb{F}^{m\times n}$ be the zero matrix given by definition II.3.1.6. The addition of matrices and multiplication of matrices by scalars introduced above satisfy the following properties.

(a) Commutativity: for any $A, B \in \mathbb{F}^{m\times n}$, we have $A + B = B + A$.

(b) Associativity: for any $A, B, C \in \mathbb{F}^{m\times n}$, we have $(A + B) + C = A + (B + C)$. Also, for any $\alpha, \beta \in \mathbb{F}$ and $A \in \mathbb{F}^{m\times n}$, we have $(\alpha \cdot \beta) \cdot A = \alpha \cdot (\beta \cdot A)$.

(c) Additive identity: there exists a matrix $O \in \mathbb{F}^{m\times n}$ such that $O + A = A$ for all $A \in \mathbb{F}^{m\times n}$. Specifically, the zero matrix given by definition II.3.1.6 is such an element, and there is no other.

(d) Additive inverse: for every $A \in \mathbb{F}^{m\times n}$, there exists a matrix $B \in \mathbb{F}^{m\times n}$ such that $A + B = O$. Specifically, for any $A \in \mathbb{F}^{m\times n}$, $B = [-A_{ij}]_{i=1,\,j=1}^{m\quad n} \in \mathbb{F}^{m\times n}$ is such a matrix, and there is no other.

(e) Neutrality of the multiplicative identity element of $\mathbb{F}$: the element $1 \in \mathbb{F}$ satisfies $1 \cdot A = A$ for every $A \in \mathbb{F}^{m\times n}$.

(f) Distributivity: for all $A, B \in \mathbb{F}^{m\times n}$ and $\alpha, \beta \in \mathbb{F}$,
  (i) $(\alpha + \beta) \cdot A = (\alpha \cdot A) + (\beta \cdot A)$;
  (ii) $\alpha \cdot (A + B) = (\alpha \cdot A) + (\alpha \cdot B)$.

Proof. Similarly to as in proposition II.2.1.8, these facts follow from the field axioms (definition II.1.1.1) due to definitions II.3.1.4 to II.3.1.6. The proof is left to the reader as an exercise.

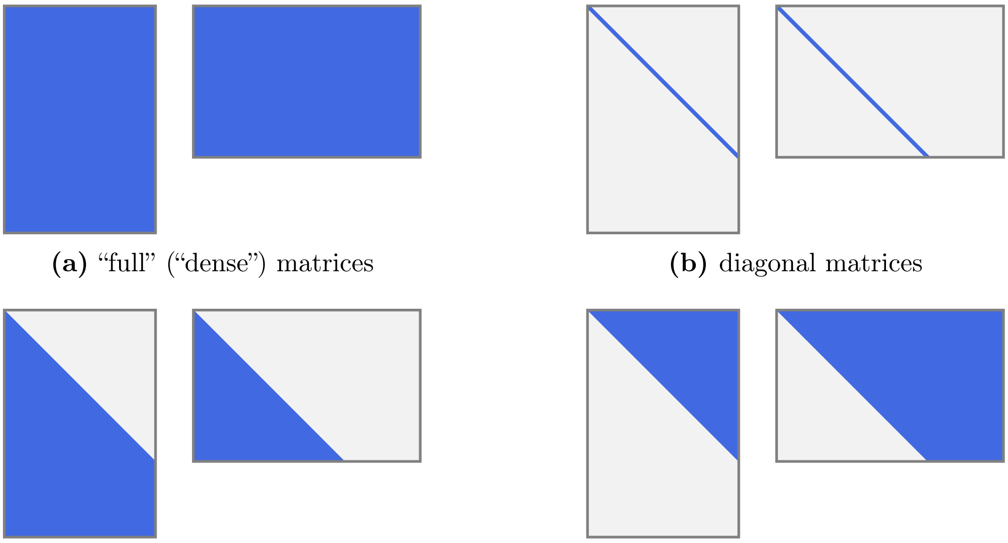
It is no coincidence that definitions II.3.1.4 and II.3.1.5 and proposition II.3.1.7 follow definitions II.2.1.4 and II.2.1.5 and proposition II.2.1.8 almost word by word. This similarity of $\mathbb{F}^n$ with $n \in \mathbb{N}$ and $\mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ (and of many other sets with linear structure) is an expression of that all those sets are vector spaces. The direct comparison of definition II.2.2.1 and proposition II.3.1.7 yields the following.

Corollary II.3.1.8. For any $m, n \in \mathbb{N}$, consider the set $\mathbb{F}^{m\times n}$ and the zero matrix $O \in \mathbb{F}^{m\times n}$. The set $\mathbb{F}^{m\times n}$ is a vector space in the sense of definition II.2.2.1 with additive identity O with respect to the operations of addition and multiplication by a scalar introduced in definitions II.3.1.4 and II.3.1.5.

Referring to the product of a scalar and a matrix, it is often convenient to omit the multiplication sign (the dot) or to place the scalar factor on the right of the matrix. To that end, we *define* the expressions αA and $A \cdot \alpha$ *to be equal to* $\alpha \cdot A$ for any $\alpha \cdot A$ with $\alpha \in \mathbb{F}$ and $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$. In this sense, one could possibly think that α and A commute, but we shall not do so. The reason is that the equality of $A \cdot \alpha$ to $\alpha \cdot A$ is just a notational definition. For example, in definition II.1.1.1, for any $\alpha, \beta \in \mathbb{F}$, the products $\alpha \boxtimes \beta \in \mathbb{F}$ and $\beta \boxtimes \alpha \in \mathbb{F}$ are assumed to be defined in some way and the equality of the two is not a notational device defining one expression in terms of the other but an essential condition, imposed in condition (a) of definition II.1.1.1 on two expressions defined independently, with the aim of requiring the commutativity of the multiplication operation.

§ II.3.2. Sparsity patterns and patterned matrices

Let us now introduce terms for certain parts of matrices and related patterned matrices, which will often appear throughout the course.



(a) “full” (“dense”) matrices

(b) diagonal matrices

(c) lower-trapezoid matrices

(d) upper-trapezoid matrices

Figure II.3.2.1. Illustration for definition II.3.2.1: sparsity patterns in tall ($m > n$) and wide ($m < n$) matrices of size $m \times n$. All possibly nonzero entries are depicted in blue.

By saying a matrix of size $m \times n$ is *full* or *dense*, we convey that we do *not assume* any of its mn entries to be zero (but in no way exclude such a possibility). By saying that a matrix is *sparse*, we convey that we assume the matrix to have *many* zero entries. In general, these terms are not exactly precise for a reason: in a specific context, what matters (and has an exact context-specific meaning) is the *degree of sparsity*.

For example, with regard to matrices of size $n \times n$ and for some positive constant C that is not much larger than one, we may be interested in matrices with at most $C \cdot n$ nonzero entries or with at most $C \log_2 n$ nonzero entries, while in some contexts even Cn^2 with C approximately equal to $\frac{1}{2}$ is a number of nonzero entries that is sufficiently small to result in useful theoretical properties or computational algorithms. The notion of sparsity based on the number of zero or nonzero entries is oblivious to the localization of the entries assumed to be zero in the matrix. That localization, associated with patterns formed by zeros or nonzeros, can be crucial, and we now turn to several basic examples of patterned matrices.

Definition II.3.2.1 (some patterns in matrices and some patterned matrices)**.** Let $m, n \in \mathbb{N}$, $r = \min\{m, n\}$ and $A = [a_{ij}]_{i=1,\, j=1}^{m \quad n} \in \mathbb{F}^{m \times n}$.

(a) Matrix A is called *square*, *tall* or *wide* if $m = n$, $m > n$ or $m < n$, respectively. By saying that matrix A is *rectangular*, we emphasize that we do *not assume* matrix A to be square but in no way rule out that particular case.

(b) The scalars $a_{11}, \ldots, a_{rr}$ (this notation lists the r elements of the tuple $(a_{kk})_{k=1}^{r}$) are called the *diagonal entries* of A. These scalars, as well as the positions they occupy, are collectively referred to as the *diagonal*, or the *main diagonal*, of A.

(c) A is said to have a *unit diagonal* if $a_{11} = \cdots = a_{rr} = 1$.

(d) The scalars $a_{21}, \ldots, a_{p,p-1}$, where $p = \min\{m, n+1\}$, as well as the positions they occupy, are collectively referred to as the *first subdiagonal* of A.

(e) The scalars $a_{12}, \dots, a_{q-1,q}$, where $q = \min\{m+1, n\}$, as well as the positions they occupy, are collectively referred to as the *first superdiagonal* of A.

(f) The matrices $L, D, U \in \mathbb{F}^{m \times n}$ given by

$$L_{ij} = \begin{cases} a_{ij} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases} \qquad D_{ij} = \begin{cases} a_{ij} & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \qquad U_{ij} = \begin{cases} a_{ij} & \text{if } i < j, \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ are called, respectively, the *strictly lower-trapezoid*, *diagonal* and *strictly upper-trapezoid* parts of A. The matrices $L + D$, $A - D$ and $U + D$ are called, respectively, the *lower-trapezoid*, *offdiagonal* and *upper-trapezoid* parts of A.

For any $d_1, \dots, d_n \in \mathbb{F}$, the square diagonal matrix with a diagonal $d_1, \dots, d_n$ is often denoted by $\operatorname{diag}(d_1, \dots, d_n)$:

$$\operatorname{diag}(d_1, \dots, d_n) = [\delta_{ij} d_i]_{i=1,\, j=1}^{n \quad n} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix} \in \mathbb{F}^{n \times n} . \tag{II.3.2.1}$$

Here, δ is the Kronecker symbol; see § I.1.27 for details. The diagonal part of a matrix A is often denoted by $\operatorname{diag} A$.

(g) A is called *diagonal*, *strictly lower trapezoid*, *strictly upper trapezoid*, *lower trapezoid* or *upper trapezoid* if and only if A is equal to its respective part defined as above (in other words, if and only if all the entries of A not included in the respective part are zero).

(h) A is called *unit lower trapezoid* (*unit upper trapezoid*) if A is lower trapezoid (upper trapezoid) with unit diagonal.

(i) A is called *tridiagonal* if $a_{ij} = 0$ for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ such that $|i - j| > 1$ (i.e., all entries except on the main diagonal, first subdiagonal and first superdiagonal are filled with zeros).

The modifier "‘trapezoid” in definition II.3.2.1 refers to the shape formed by all possibly nonzero entries of the trapezoid parts. On the other hand, the modifiers “upper”, “diagonal”, “off-diagonal” and “lower” refer to the location of all possibly nonzero entries with respect to the diagonal. Four sparsity patterns related to definition II.3.2.1 are illustrated in figure II.3.2.1. By adding “strictly” we mean that the diagonal entries are required to be zero.

Remark II.3.2.2 (“triangular trapezoid”)**.** Regarding the terminology presented in definition II.3.2.1, note that the terms *lower triangular* and *upper triangular* are often used in place of *lower trapezoid* and *upper trapezoid* for *any* matrices: tall, square and wide. The use of the term *triangular* is literal only for the respective cases of square or wide matrices and of square or tall matrices, when the respective trapezoid part degenerates into a triangle (see figure II.3.2.1).

Example II.3.2.3 (triangular, trapezoid and diagonal matrices)**.** The matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 7 & 7 \\ 6 & 18 & 22 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & 2 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & 2 \end{bmatrix} \qquad \text{(II.3.2.2)}$$

are all real (even integer-valued) and square, of size 3×3. Matrix L is unit lower triangular (also unit lower trapezoid). Matrix U is upper triangular (also upper trapezoid). Further, let us consider

$$\widehat{L} = \begin{bmatrix} 1 & \\ 2 & 1 \\ 6 & 2 \end{bmatrix}, \qquad \tilde{D} = \begin{bmatrix} 1 & & \sqrt{2} \\ & 1 & \mathsf{i} \end{bmatrix} \quad \text{and} \quad \tilde{U} = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & 0 \end{bmatrix}. \qquad \text{(II.3.2.3)}$$

Matrices $\widehat{L}$ (tall, of size 3×2) and $\tilde{U}$ (square, of size 3×3) are also real (even integer-valued). With regard to sparsity, $\tilde{U}$ is upper trapezoid and $\widehat{L}$ is unit lower trapezoid (or upper triangular and unit lower triangular, respectively; see remark II.3.2.2). Matrix $\tilde{D}$ (wide, of size 2×3) is complex and not real; with regard to sparsity, it is upper trapezoid (or upper triangular; see remark II.3.2.2) and not diagonal due to the presence of nonzeros ($\sqrt{2}$, i) above the diagonal.

Remark II.3.2.4 (all those blank spaces)**.** In (II.3.2.1) to (II.3.2.3) and throughout the course, we use *blank spaces* instead of zeros to emphasize sparsity patterns.

§ II.3.3. Transposition and symmetry

Definition II.3.3.1 (transposition)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. The matrix $[A_{ji}]_{j=1,\, i=1}^{n \quad m} \in \mathbb{F}^{n \times m}$ is called the *transpose* of A and is denoted by A^T. In other terms, we set

$$A^\mathsf{T} = [A_{ji}]_{j=1,\, i=1}^{n \quad m} \in \mathbb{F}^{n \times m}.$$

Proposition II.3.3.2 (reflexivity of transposition)**.** The operation of transposition is *reflexive*: for any $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$, we have $(A^\mathsf{T})^\mathsf{T} = A$.

Proof. Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. First, we note that definition II.3.3.1 defines A^T as an element of $\mathbb{F}^{n \times m}$ and $(A^\mathsf{T})^\mathsf{T}$, as an element of $\mathbb{F}^{m \times n}$. The claim for the m, n and A considered is therefore a matrix equality in $\mathbb{F}^{m \times n}$. Two matrices from $\mathbb{F}^{m \times n}$ are equal if and only if all their respective entries are equal (this notion of equality for matrices is a result of defining matrices as functions). Proving the claim therefore consists in showing $\big((A^\mathsf{T})^\mathsf{T}\big)_{ij} = A_{ij}$ for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. At this point, the only thing we know about transposition is the definition of the operation. So if we want to evaluate an entry of a matrix obtained by transposition, there is clearly no other way for us to proceed than to apply the definition of the transpose to *resolve*, or *evaluate*, the transposition.

Applying definition II.3.3.1 twice, first to resolve the outer transposition and then to resolve the inner transposition, we obtain

$$\left((A^\mathsf{T})^\mathsf{T}\right)_{ij} = (A^\mathsf{T})_{ji} = A_{ij}$$

for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. That this equality is valid for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ proves the claim.

Proposition II.3.3.3 (transposition under linear operations)**.** Let $m, n \in \mathbb{N}$, $\alpha \in \mathbb{F}$ and $A, B \in \mathbb{F}^{m\times n}$. Then $(A+B)^\mathsf{T} = A^\mathsf{T} + B^\mathsf{T}$ and $(\alpha \cdot A)^\mathsf{T} = \alpha \cdot A^\mathsf{T}$.

Proof. First, we note that definitions II.3.1.4 and II.3.1.5 define $A+B$ and $\alpha \cdot A$ as elements of $\mathbb{F}^{m\times n}$, so we are to prove two matrix equalities in $\mathbb{F}^{m\times n}$. Consider $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ arbitrary.

Applying the definition of transposition (definition II.3.3.1) and the definition of matrix addition (definition II.3.1.4) and marking the resulting equalities with $*$ and $**$, we obtain

$$\left((A+B)^\mathsf{T}\right)_{ji} \overset{*}{=} (A+B)_{ij} \overset{**}{=} A_{ij} + B_{ij} \overset{*}{=} (A^\mathsf{T})_{ji} + (B^\mathsf{T})_{ji} \overset{**}{=} (A^\mathsf{T} + B^\mathsf{T})_{ji}\,.$$

Since we have proved these equalities for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ at once, we have thereby proved the first claim.

Similarly, applying the definition of transposition (definition II.3.3.1) and the definition of the multiplication of a matrix by a scalar (definition II.3.1.5) and marking the resulting equalities with $*$ and $**$, we obtain

$$\left((\alpha \cdot A)^\mathsf{T}\right)_{ji} \overset{*}{=} (\alpha \cdot A)_{ij} \overset{**}{=} \alpha \cdot A_{ij} \overset{*}{=} \alpha \cdot (A^\mathsf{T})_{ji} \overset{**}{=} (\alpha \cdot A^\mathsf{T})_{ji}\,.$$

Again, we have obtained these equalities for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$ at once, so we have thereby proved the second claim.

Definition II.3.3.4 (symmetric and skew-symmetric matrices)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Matrix A is called *symmetric* if $A^\mathsf{T} = A$ and *skew-symmetric (antisymmetric)* if $A^\mathsf{T} = -A$.

Recall that “if” in definitions effectively means “if and only if” — as long as there is no additional definition relaxing the conditions of the original one and thereby generalizing it.

Remark II.3.3.5. For any $m, n \in \mathbb{N}$ and $A, B \in \mathbb{F}^{m\times n}$, the expressions $-A$ and $B - A$ refer to *the* additive inverse of A and to $B + (-A)$. The additive inverse is uniquely defined due to proposition II.3.1.7, whereas the sum is (uniquely) defined by definition II.3.1.4. From definitions II.3.1.4 and II.3.1.5, we immediately see that $-A = (-1) \cdot A$, where -1 is an additive inverse of 1 in $\mathbb{F}$ (which is also unique, as one can derive from definition II.1.1.1). We will see later that such an equality holds in any vector space, not just in $\mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ with respect to the specific operations introduced in definitions II.3.1.4 and II.3.1.5.

Proposition II.3.3.6 (symmetrization and antisymmetrization)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Matrices $A + A^\mathsf{T}$ and $A - A^\mathsf{T}$ are symmetric and skew-symmetric respectively.

Proof. From propositions II.3.3.2, II.3.3.3 and II.3.1.7, we deduce the following:

$$(A + A^\mathsf{T})^\mathsf{T} = A^\mathsf{T} + (A^\mathsf{T})^\mathsf{T} = A^\mathsf{T} + A = A + A^\mathsf{T}$$

and

$$(A - A^\mathsf{T})^\mathsf{T} = A^\mathsf{T} - (A^\mathsf{T})^\mathsf{T} = A^\mathsf{T} - A = -A + A^\mathsf{T} = -(A - A^\mathsf{T})\,.$$

Then definition II.3.3.4 yields the claimed statements. Note that the above equalities are all meant in $\mathbb{F}^{m\times n}$, i.e., we completely bypassed the entrywise comparison of matrices in proving the equality of those matrices. This fortunate circumstance is thanks to that the results of propositions II.3.3.2 and II.3.3.3 are formulated as matrix equalities.

Following the references given at the beginning of the proof and justifying every single one of the above equalities with a suitable individual property of matrices may be a very useful exercise for the reader.

Definition II.3.3.7 (symmetric and skew-symmetric parts of a matrix)**.** Assume that $\mathbb{F}$ is such that $1 + 1 \neq 0$ and let $\frac{1}{2}$ denote the multiplicative inverse of $1 + 1$ in $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Then the matrices $\frac{1}{2}A + \frac{1}{2}A^\mathsf{T}$ and $\frac{1}{2}A - \frac{1}{2}A^\mathsf{T}$ are called the symmetric and skew-symmetric parts of the matrix A.

Definition II.3.3.7 is motivated by proposition II.3.3.6, which shows that the symmetric and skew-symmetric parts of a matrix are indeed symmetric and skew-symmetric matrices themselves. Under the assumptions of definition II.3.3.7 (ruling out the case of $\mathbb{F} = \mathbb{F}_2$ from example II.1.2.4), we have $\frac{1}{2} + \frac{1}{2} = 1$, so that the sum of the symmetric and antisymmetric parts of A is equal to A.

§ II.4. Matrices: algebraic structure

§ II.4.1. Definitions and algebraic properties. Inverse matrix

Addition and multiplication by scalars are often called *linear operations*, and proposition II.3.1.7 shows that these operations, as defined above, endow $\mathbb{F}^{m\times n}$ with what is referred to as *linear structure*. This expresses the fact that the two operations relate some matrices to other matrices in specific ways. However, we can — and it *is* helpful to — do more with matrices than just adding and multiplying by scalars. Specifically, we will also multiply matrices by other matrices, which is an *algebraic operation*. We will now motivate and gradually introduce matrix multiplication, which additionally endows $\mathbb{F}^{m\times n}$ with what is called *algebraic structure*.

Think of systems of linear equations, which are central objects of the course. They express linear relations between vectors, and we are going to use a specific rule for multiplying matrices and column vectors that is suitable for concisely expressing such relations.

Let $m, n \in \mathbb{N}$. Consider a linear relation between vectors $x = [x_j]_{j=1}^n \in \mathbb{F}^n$ and $b = [b_i]_{i=1}^m \in \mathbb{F}^m$ expressed in the form of m equations with mn coefficients that form a matrix $A = [a_{ij}]_{i=1,\,j=1}^{m \quad n} \in \mathbb{F}^{m\times n}$:

$$\left\{\begin{array}{ccccc} a_{11} \cdot x_1 + \cdots + a_{1n} \cdot x_n & = & b_1 & , \\ \vdots \qquad\qquad\quad \vdots & & \vdots & \\ a_{m1} \cdot x_1 + \cdots + a_{mn} \cdot x_n & = & b_m & . \end{array}\right. \tag{II.4.1.1}$$

The same linear relation can be equivalently expressed as follows:

$$\sum_{j=1}^{n} a_{ij} \cdot x_j = b_i \quad \text{for each} \quad i \in \{1, \ldots, m\}\,. \tag{II.4.1.2}$$

In (II.4.1.2), equations, corresponding to the components of b, are indexed by i, whereas the components of x are indexed by j. We see from (II.4.1.2) that (II.4.1.1) and (II.4.1.2) can be recast in the equivalent form

$$Ax = b \tag{II.4.1.3}$$

provided that we define Ax as follows.

Definition II.4.1.1 (matrix-vector product)**.** Let $m, n \in \mathbb{N}$. For any $A = [a_{ij}]_{i=1,\, j=1}^{m \quad n} \in \mathbb{F}^{m\times n}$ and $x \in \mathbb{F}^n$, the matrix-vector product $Ax \in \mathbb{F}^m$ of A and x is defined by

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} \cdot x_j \quad \text{for each} \quad i \in \{1, \dots, m\} \tag{II.4.1.4}$$

or, equivalently,

$$Ax = \left[\sum_{j=1}^{n} a_{ij} \cdot x_j\right]_{i=1}^{m} . \tag{II.4.1.5}$$

Remark II.4.1.2. The product of a matrix and a vector is a linear combination of the columns of the matrix with the vector as coefficient. Indeed, in the context of definition II.4.1.1, let $a_1, \dots, a_n$ be the columns of A (see definition II.3.1.1). Then (II.4.1.4) and (II.4.1.5) can be recast as follows:

$$(Ax)_i = \sum_{j=1}^{n} (a_j)_i \cdot x_j = (Ax)_i = \left(\sum_{j=1}^{n} x_j \cdot a_j\right)_i \quad \text{for each} \quad i \in \{1, \dots, m\} ,$$

where the third equality relies on definitions II.3.1.4 and II.3.1.5. The above equality is the componentwise representation of the column-vector equality

$$Ax = \sum_{j=1}^{n} x_j \cdot a_j .$$

Let us now consider a chain of two linear relations, between $x = [x_j]_{j=1}^{n} \in \mathbb{F}^n$, $y = [y_k]_{k=1}^{r} \in \mathbb{F}^r$ and $z = [z_i]_{i=1}^{m} \in \mathbb{F}^m$, where $m, n, r \in \mathbb{N}$. Assuming that

$$Ay = z \quad \text{and} \quad Bx = y \tag{II.4.1.6}$$

for some matrices $A = [a_{ik}]_{i=1,\, k=1}^{m \quad r} \in \mathbb{F}^{m\times r}$ and $B = [b_{kj}]_{k=1,\, j=1}^{r \quad n} \in \mathbb{F}^{r\times n}$, we would like to establish a linear relation between x and z and to express it in a form similar to that of (II.4.1.3):

$$Cx = z \tag{II.4.1.7}$$

with some $C = [c_{ij}]_{i=1,\, j=1}^{m \quad n} \in \mathbb{F}^{m\times n}$. We will see now that both of these goals can be achieved.

Applying definition II.4.1.1 to translate equalities (II.4.1.6) in $\mathbb{F}^m$ and $\mathbb{F}^r$ back into componentwise equalities in $\mathbb{F}$, we rewrite the two linear relations as

$$\sum_{k=1}^{r} a_{ik} \cdot y_k = z_i \quad \text{for all} \quad i \in \{1, \dots, m\} \quad \text{and} \quad \sum_{j=1}^{n} b_{kj} \cdot x_j = y_k \quad \text{for all} \quad k \in \{1, \dots, r\} .$$

Substituting every value of y_k with $k \in \{1, \dots, r\}$ given by the second relation into the first relation for every $i \in \{1, \dots, m\}$, we obtain

$$\sum_{k=1}^{r} a_{ik} \cdot \left(\sum_{j=1}^{n} b_{kj} \cdot x_j \right) = z_i \quad \text{for all} \quad i \in \{1, \dots, m\} \, .$$

Changing the order of summation and multiplication (which is justified by conditions (a), (b) and (f) of definition II.1.1.1), we arrive at

$$\sum_{j=1}^{n} \left(\sum_{k=1}^{r} a_{ik} \cdot b_{kj} \right) \cdot x_j = z_i \quad \text{for all} \quad i \in \{1, \dots, m\} \, ,$$

which is equivalent to (II.4.1.7) with $C = AB$ defined as follows.

Definition II.4.1.3 (matrix-matrix multiplication)**.** Let $m, n, r \in \mathbb{N}$ and $A = [a_{ik}]_{i=1,\, k=1}^{m \quad r} \in \mathbb{F}^{m \times r}$ and $B = [b_{kj}]_{k=1,\, j=1}^{r \quad n} \in \mathbb{F}^{r \times n}$. The *matrix product* $AB \in \mathbb{F}^{m \times n}$ of A and B is defined by

$$(AB)_{ij} = \sum_{k=1}^{r} a_{ik} \cdot b_{kj} \quad \text{for all} \quad i \in \{1, \dots, m\} \quad \text{and} \quad j \in \{1, \dots, n\} \tag{II.4.1.8}$$

or, equivalently,

$$AB = \left[\sum_{k=1}^{r} a_{ik} \cdot b_{kj} \right]_{i=1,\, j=1}^{m \quad n} . \tag{II.4.1.9}$$

Remark II.4.1.4 (column vectors as matrices)**.** For every $n \in \mathbb{N}$, we can identify $\mathbb{F}^n$ with $\mathbb{F}^{n \times 1}$ by considering column vectors as one-column matrices. In this respect, definitions II.4.1.3 to II.3.1.5 are *consistent* with their preceding counterparts definitions II.4.1.1, II.2.1.4 and II.2.1.5: column vectors and one-column matrices are added, multiplied by scalars and multiplied by matrices on the left according to the respective definitions in the same way.

Remark II.4.1.5. In definition II.4.1.3, we define the product of two matrices componentwise. In calculating such products analytically ("on paper"), it is convenient not to look up all the relevant entries of the factors every time a single entry of the product is evaluated but rather to evaluate the product row by row or column by column. To elaborate on this, let us use the assumptions and notations of definition II.4.1.3 and set notations for the entries of AB as follows: $AB = [c_{ij}]_{i=1,\, j=1}^{m \quad n}$.

(a) **Calculating a product of matrices column-wise**. For any $j \in \{1, \dots, n\}$, the jth column of AB (see definition II.3.1.1) is formed by c_{ij} with $i \in \{1, \dots, m\}$. Considering (II.4.1.8) with the same fixed j, we see that the jth column of AB is determined by only the jth column of B and by all columns of A. Specifically, *the jth column of AB is the linear combination* (see example II.2.2.10) *of the columns of A with the coefficients forming the jth column of B* (there are r such columns and r such coefficients).

(b) **Calculating a product of matrices row-wise**. For any $i \in \{1, \dots, m\}$, the ith row of AB (see definition II.3.1.1) is formed by c_{ij} with $j \in \{1, \dots, n\}$. Considering (II.4.1.8) with the same fixed i, we see that the ith row of AB is determined by only the ith row of A and by all rows of B. Specifically, *the ith row of AB is the linear combination* (see example II.2.2.10) *of the rows of B with the coefficients forming the ith row of A* (there are r such rows and r such coefficients).

These trivial observations allow for the calculation of AB column by column or row by row, which means looking up (“caching”) and applying every tuple of r coefficients to perform all operations that involve that tuple before proceeding to the next such a tuple. With “all operations involving that tuple” we refer to the evaluation of a single linear combination (of columns or rows), which is done componentwise, which eventually reduces the calculation to (II.4.1.8).

Note that in § II.3.1 we introduced matrices essentially as collections of mn scalars that could, in principle, be re-enumerated in any way, possibly even using a single index instead of two. In §§ II.3.2 and II.3.3, we considered sparsity patterns and the operation of transposition in relation to the two-index enumeration of the entries, but with no clear purpose in sight. What really gives clear meaning to the row-and-column indexing is the role of matrices in expressing single linear relations between tuples (“column vectors”) and compositions of such relations, i.e., how they interact with tuples and with each other by means of the matrix-vector and matrix-matrix multiplication introduced in this section. That role is actually the motivation behind introducing the table notation for matrices in definition II.3.1.1 and even the column notation, together with the term *column vector*, for tuples in definition II.2.1.1.

Example II.4.1.6 (multiplication of matrices and LU decomposition)**.** Let us revisit the matrices given in example II.3.2.3. First, we are interested in the product LU of

$$L = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & 2 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & 2 \end{bmatrix}. \tag{II.4.1.10}$$

Let us calculate the product LU column by column (i.e., using rule (a) of remark II.4.1.5). The third column of LU is the product of L and of $(3, 1, 2) \in \mathbb{R}^3$, the third column of U, which is a linear combination (see definition II.2.2.7 and example II.2.2.10) of the columns of L with the coefficients 3, 1 and 2:

$$\begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & 2 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = 3 \cdot \begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + 2 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \cdot 1 + 1 \cdot 0 + 2 \cdot 0 \\ 3 \cdot 2 + 1 \cdot 1 + 2 \cdot 0 \\ 3 \cdot 6 + 1 \cdot 2 + 2 \cdot 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \\ 22 \end{bmatrix}.$$

We started with the third column because the linear combinations representing the other two happen to contain zeros (due to that matrix U is upper triangular), which makes it even easier to evaluate the the first and second columns:

$$\begin{bmatrix} 1 & & \\ 2 & \cancel{1} & \\ 6 & \cancel{2} & \cancel{1} \end{bmatrix} \begin{bmatrix} 1 \\ \cancel{0} \\ \cancel{0} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 6 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & 2 & \cancel{1} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ \cancel{0} \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 18 \end{bmatrix}.$$

In each of these two calculations, we strike out the zeros in the column of U involved in the calculation. Consider any of those zeros and note its row index k. That zero enters into the linear combination that we are calculating as the coefficient of the kth column of L. As a result, that kth column of L does not contribute to the result and, for the purpose of this calculation, can be treated as zero. We strike out (using the same color) all the entries of such irrelevant columns (except for the zeros associated with the lower-triangular sparsity pattern, which we do not even write in accordance with remark II.3.2.4). The product can now be composed from its columns:

$$LU = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 7 & 7 \\ 6 & 18 & 22 \end{bmatrix}, \tag{II.4.1.11}$$

so that we actually have $A = LU$ for the matrix A given in (II.3.2.2).

Let us now calculate the same product LU row by row (i.e., using rule (b) of remark II.4.1.5). We start again with the third row:

$$\begin{bmatrix} 6 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & 2 \end{bmatrix} = 6 \cdot \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} + 2 \cdot \begin{bmatrix} 0 & 3 & 1 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 6 & 18 & 22 \end{bmatrix}.$$

The calculation of the first and second rows is even a bit simpler due to the presence of zeros in L:

$$\begin{bmatrix} 1 & \cancel{0} & \cancel{0} \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ & \cancel{3} & \cancel{1} \\ & & \cancel{2} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 1 & \cancel{0} \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & \cancel{2} \end{bmatrix} = \begin{bmatrix} 2 & 7 & 7 \end{bmatrix}.$$

We see that all the three rows we obtained agree with the result (II.4.1.11) of the column-wise calculation.

Finally, we revisit the matrices $\tilde{D}$, $\tilde{U}$ and $\widehat{L}$ given in the same example II.3.2.3 and, using definition II.4.1.3 directly or in the form of the rules described in remark II.4.1.5, calculate the products $\widehat{U} = \tilde{D}\tilde{U}$ and $\widehat{A} = \widehat{L}\,\widehat{U}$:

$$\widehat{U} = \begin{bmatrix} 1 & & \cancel{\sqrt{2}} \\ & 1 & \cancel{\mathrm{i}} \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & \cancel{0} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \end{bmatrix} \quad \text{and} \quad \widehat{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 7 & 7 \\ 6 & 18 & \mathbf{20} \end{bmatrix}. \tag{II.4.1.12}$$

Note that the matrices $\widehat{L}$ and $\widehat{U}$ are obtained from the above L and U by removing the last column and the last row, respectively. The error introduced in the product is not dramatic: $\widehat{A}$ differs from A only in the bottom right entry (boldface) and only by 2.

The factorized representations $A = LU$ and $\widehat{A} = \widehat{L}\widehat{U}$ of A and $\widehat{A}$ are examples of the *LU decomposition (factorization)* of matrices, which we study in detail in § II.5 below.

Remark II.4.1.7. For $n \in \mathbb{N}$, consider the factorization $A = LU$ of $A \in \mathbb{F}^{n\times n}$ with a lower-triangular matrix $L \in \mathbb{F}^{n\times n}$ and an upper-triangular matrix $U \in \mathbb{F}^{n\times n}$. One such a factorization, with $n = 3$, is given in example II.4.1.6.

Similarly to what (II.4.1.6) expresses, L and U may represent a chain of linear relations between some $x, y, z \in \mathbb{F}^n$:

$$z = Ly \quad \text{and} \quad y = Ux\,.$$

Let us express these two relations in a form that may be more familiar to the reader:

$$\begin{cases} L_{11}\cdot y_1 & = z_1\,, \\ L_{21}\cdot y_1 + L_{22}\cdot y_2 & = z_2\,, \\ \vdots \quad \vdots \qquad \vdots\ \vdots \quad \ddots & \vdots \quad \vdots \\ L_{n1}\cdot y_1 + L_{n2}\cdot y_2 + \cdots + L_{nn}\cdot y_n & = z_n \end{cases} \tag{II.4.1.13a}$$

and

$$\begin{cases} U_{11}\cdot x_1 + U_{12}\cdot x_2 + \cdots + U_{1n}\cdot x_n & = y_1\,, \\ U_{22}\cdot x_2 + \cdots + U_{2n}\cdot x_n & = y_2\,, \\ \ddots \quad \vdots \quad \vdots & \vdots \quad \vdots \\ U_{nn}\cdot x_n & = y_n\,. \end{cases} \tag{II.4.1.13b}$$

In particular, for the matrices L and U from definition II.4.1.3, these relations take the following form:

$$\begin{cases} y_1 & = z_1\,, \\ 2\cdot y_1 + \quad y_2 & = z_2\,, \\ 6\cdot y_1 + 2\cdot y_2 + \ y_3 & = z_3 \end{cases} \quad \text{and} \quad \begin{cases} 1\cdot x_1 + 2\cdot x_2 + 3\cdot x_3 & = y_1\,, \\ 3\cdot x_2 + 1\cdot x_3 & = y_2\,, \\ 2\cdot x_3 & = y_3\,. \end{cases} \tag{II.4.1.14}$$

As we discussed in relation to (II.4.1.6), we can forgo the intermediate column vector y and consider the composition of the two linear relations: $Ax = z$. On the other hand, given a factorization $A = LU$, we can convert that linear relation into a chain of two, involving an intermediate column vector y.

That equivalent conversion is useful — and you must have already used this many times, even if implicitly, — for solving linear systems of equations using Gaussian elimination. For a system $Ax = z$ with A and z given and x unknown, one may exclude the unknown x_1 by (i) using the first equation to express x_1 in terms of $x_2, \ldots, x_n$ and z_1 and (ii) substituting the resulting expression into all the other equations. Since the system is linear, that is equivalent to subtracting, from all the other equations, the first equation multiplied by suitable scalar coefficients chosen so that none of the resulting equations involves x_1. This procedure, often called *forward substitution*, is then repeated sequentially, for x_2 and for all the other unknowns. If the method does not break down (we will soon come to understand that precisely), $n-1$ steps of this procedure produce $y_1, \ldots, y_n$ and transform the original system to a system of the form (II.4.1.13b), or $Ux = y$, with an upper-triangular matrix U in our setting. In the transformed system, one of the equations involves only one unknown (x_n in our setting) and can be immediately solved with respect to that unknown. That is followed by a second stage of the calculation, known as *backward substitution*. In our notations, it consists in substituting the found value of x_n into the other equations, finding x_{n-1} and repeating this step sequentially for all the other unknowns.

For now, we intend to merely allude to that form of Gaussian elimination and to claim that, for A and z given, (i) the forward substitution is equivalent to calculating U and solving $Ly = z$ with respect to y, whereas (ii) the backward substitution is equivalent to solving $Ux = y$ with respect to x. We study the exact mechanics of such calculations starting from § II.5

Matrix multiplication is drastically different form the linear operations we consider for matrices (see § II.3) and also from the multiplication we consider for scalars (see definition II.1.1.1). First of all, it is not commutative.

Remark II.4.1.8 (matrix multiplication is not commutative)**.** For any $n \in \mathbb{N}$ such that $n \geq 2$, matrix multiplication on $\mathbb{F}^{n\times n}$ is not commutative. Indeed, let $J \in \mathbb{F}^{n\times n}$ be the matrix the only nonzero entry of which is equal to 1 and is located in the upper-right corner:

$$J = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{bmatrix} .$$

Then

$$J J^{\mathsf{T}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \neq \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \vdots & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} = J^{\mathsf{T}} J .$$

Note that remark II.4.1.8 states that matrix multiplication is noncommutative for any matrix order larger than one, and the counterexample is therefore given for an arbitrary order. This particular counterexample, however, is very fortunate in the sense that it is easily obtained by generalizing its particular case corresponding to $n = 2$.

On the other hand, there is an obvious candidate for a *matrix multiplicative identity*, which should be a matrix behaving similarly to 1 in part (e) of proposition II.3.1.7 and part (e) of proposition II.2.1.8.

Definition II.4.1.9 (identity matrix)**.** For any $n \in \mathbb{N}$, by the *identity matrix of order n* we mean the matrix

$$[\delta_{ij}]_{i=1,\, j=1}^{n\quad n} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{bmatrix} = \operatorname{diag}(1, \ldots, 1) \in \mathbb{F}^{n\times n} .$$

Here, δ is the Kronecker symbol (see § I.1.27 for details) and diag denotes the construction of a diagonal matrix in the sense of part (g) of definition II.3.2.1.

Identity matrices are often denoted by the symbol "I" and its various derivatives, so that remark II.2.1.7 equally applies to them.

In spite of the seemingly unpleasant observation made in remark II.4.1.8, definition II.4.1.3, together with definitions II.3.1.4 and II.4.1.9, has certain important consequences for how we can work with matrices, which we will now summarize.

Proposition II.4.1.10 (algebraic structure of the set of matrices over $\mathbb{F}$)**.** The addition and multiplication of matrices introduced above satisfy the following properties, where, for any $s \in \mathbb{N}$, I_s denotes the identity matrix of order s.

- **(a)** Associativity: for all $A \in \mathbb{F}^{m\times n}, B \in \mathbb{F}^{n\times p}, C \in \mathbb{F}^{p\times q}$ with any $m, n, p, q \in \mathbb{N}$, we have $(AB)C = A(BC)$.
- **(b)** Multiplicative identity elements:
  - (i) $I_m A = A$ for every $A \in \mathbb{F}^{m\times n}$ and all $m, n \in \mathbb{N}$;
  - (ii) $A I_n = A$ for every $A \in \mathbb{F}^{m\times n}$ and all $m, n \in \mathbb{N}$.
- **(c)** Distributivity: for all $A, B \in \mathbb{F}^{m\times n}$ and $P \in \mathbb{F}^{p\times m}, Q \in \mathbb{F}^{n\times q}$ with any $m, n, p, q \in \mathbb{N}$,
  - (i) $P(A + B) = PA + PB$;
  - (ii) $(A + B)Q = AQ + BQ$.
- **(d)** Uniqueness of multiplicative identity elements:
  - (i) if $I'_n \in \mathbb{F}^{n\times n}$ with $n \in \mathbb{N}$ satisfies $I'_n A = A$ for every $A \in \mathbb{F}^{n\times n}$, then $I'_n = I_n$;
  - (ii) if $I'_n \in \mathbb{F}^{n\times n}$ with $n \in \mathbb{N}$ satisfies $A I'_n = A$ for every $A \in \mathbb{F}^{n\times n}$, then $I'_n = I_n$.

Note that part (b) of proposition II.4.1.10 explains the term "identity matrix" introduced in definition II.4.1.9: it shows that the identity matrix of any order $r \in \mathbb{N}$ is *a left identity element* of $\mathbb{F}^{r\times r}$ and *a right identity element* of $\mathbb{F}^{r\times r}$ with respect to matrix multiplication. Part (d) then establishes the fact that there are no other such identity elements. Of course, it is important for us how the identity matrices interact with rectangular matrices, not just with square ones.

Proof. Each of these properties are easy to verify. First, we can reduce the statement from matrices to scalars (real or complex) using definitions II.4.1.3 and II.3.1.4. Second, using properties of operations with scalars, we can conclude that all entries of the matrices claimed to be equal are indeed so. For demonstration, we give a proof by entrywise inspection for part (a) here. Parts (b) and (c) are shown similarly; the proof is left to the reader as an exercise. Part (d) relies upon parts (a) to (c), and we prove it here to show that the statement is not a tautology.

To prove part (a), let us consider arbitrary $i \in \{1, \ldots, m\}$ and $\ell \in \{1, \ldots, q\}$ and show that all corresponding entries of the two products are equal. Using definition II.4.1.3 twice, we obtain that

$$((AB)C)_{i\ell} = \sum_{k=1}^{p} (AB)_{ik}\, C_{k\ell} = \sum_{k=1}^{p} \Big(\sum_{j=1}^{n} A_{ij}\, B_{jk} \Big)\, C_{k\ell}$$

and

$$(A(BC))_{i\ell} = \sum_{j=1}^{n} A_{ij}\, (BC)_{j\ell} = \sum_{j=1}^{n} A_{ij} \left(\sum_{k=1}^{p} B_{jk}\, C_{k\ell} \right).$$

Now, using field axioms (the commutativity of addition, the associativity of multiplication and the distributivity of multiplication over addition), we can change the order of summation and multiplication and arrive at

$$((AB)C)_{i\ell} = \sum_{k=1}^{p} \sum_{j=1}^{n} A_{ij}\, B_{jk}\, C_{k\ell} = (A(BC))_{i\ell}\,.$$

We have been considering arbitrary i and ℓ from the respective ranges so far, so what we have just proved is that $A(BC) = (AB)C$.

To prove statement (i) of part (d), let $I'_n \in \mathbb{F}^{n\times n}$ be a matrix such that $I'_n A = A$ for every $A \in \mathbb{F}^{n\times n}$. Then we immediately notice that

$$I'_n = I'_n I_n = I_n \,,$$

where the first equality is due to statement (ii) of part (b) applied with $A = I'_n$ and the second equality is $I'_n A = A$ (assumed above) applied with $A = I_n$.

Statement (ii) of part (d) is proven analogously.

Proposition II.4.1.11 (transposition under matrix multiplication)**.** Let $m, n, r \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times r}, B \in \mathbb{F}^{r\times n}$. Then $(AB)^\mathsf{T} = B^\mathsf{T} A^\mathsf{T}$.

Proof. For all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, combining definitions II.3.3.1 and II.4.1.3, we obtain

$$\big((AB)^\mathsf{T}\big)_{ji} = (AB)_{ij} = \sum_{k=1}^{r} A_{ik}\, B_{kj} = \sum_{k=1}^{r} (A^\mathsf{T})_{ki}\, (B^\mathsf{T})_{jk} = \sum_{k=1}^{r} (B^\mathsf{T})_{jk}\, (A^\mathsf{T})_{ki} = \big(B^\mathsf{T} A^\mathsf{T}\big)_{ji} \,.$$

This gives $(AB)^\mathsf{T} = B^\mathsf{T} A^\mathsf{T}$ and therefore proves the claim.

Example II.4.1.12. In proposition II.4.1.11, reversing the order of the transposes is essential. Indeed, $A^\mathsf{T} B^\mathsf{T}$ does not need to be equal to $(AB)^\mathsf{T}$ even when the product $A^\mathsf{T} B^\mathsf{T}$ is defined, which is the case if and only if $m = n$, and even if the matrix $A^\mathsf{T} B^\mathsf{T}$ is of the same size as $(AB)^\mathsf{T}$, which is the case if and only if $m = n = r$.

For example, consider $n \in \mathbb{N}$ and the matrix $J \in \mathbb{F}^{n\times n}$ defined in remark II.4.1.8. For $A = J$ and $B = J^\mathsf{T}$, we have

$$(AB)^\mathsf{T} = AB \neq BA = (BA)^\mathsf{T} \,,$$

where the inequality is the one pointed out in remark II.4.1.8 and the equalities are evident from the explicit form of JJ^T and $J^\mathsf{T} J$ given in remark II.4.1.8. On the other hand, we also have $(AB)^\mathsf{T} = (JJ^\mathsf{T})^\mathsf{T} = JJ^\mathsf{T} = B^\mathsf{T} A^\mathsf{T}$ and $(BA)^\mathsf{T} = (J^\mathsf{T} J)^\mathsf{T} = J^\mathsf{T} J = A^\mathsf{T} B^\mathsf{T}$, consistently with proposition II.4.1.11.

Proposition II.4.1.13. Let $n \in \mathbb{N}$ and I be the identity matrix of order n. Consider $A, B, C \in \mathbb{F}^{n\times n}$ such that $AB = I = BA$ and $AC = I = CA$. Then $B = C$.

Proof. Using parts (a) and (b) of proposition II.4.1.10, we obtain $B = BI = B(AC) = (BA)C = IC = C$.

Definition II.4.1.14 (inverse matrix)**.** Let $n \in \mathbb{N}$ and I be the identity matrix of order n. Consider $A \in \mathbb{F}^{n\times n}$. If there exists a matrix $B \in \mathbb{F}^{n\times n}$ such that $AB = I = BA$, then A is called *invertible* and B is referred to as the *inverse* of A and is denoted by A^{-1}.

Remark II.4.1.15 (inverse matrix is well defined)**.** Definition II.4.1.14 defines *the* inverse matrix A^{-1} of a matrix $A \in \mathbb{F}^{n\times n}$ with $n \in \mathbb{N}$ as *any* $B \in \mathbb{F}^{n\times n}$ satisfying $AB = I = BA$, and such a definition alone should immediately raise a suspicion. This definition is, however, correct: due to the uniqueness result of proposition II.4.1.13, there is at most one such a matrix. This unique element of $\mathbb{F}^{n\times n}$ is therefore well defined and can therefore be correctly referred to as *the* inverse of A.

Example II.4.1.16. Let $n \in \mathbb{N}$ and $O, I \in \mathbb{F}^{n\times n}$ be the respective zero and identity matrices. By definition II.4.1.14, the zero matrix O is not invertible: indeed, for any $B \in \mathbb{F}^{n\times n}$, we obtain $OB = BO = O \neq I$ from definitions II.4.1.3, II.3.1.6 and II.4.1.9, where the inequality holds because the identity elements $0, 1$ of the field are distinct (by definition II.1.1.1). On the other hand, by definition II.4.1.14, the identity matrix I is invertible and equals its own inverse: $I \cdot I = I$ follows from part (b) of proposition II.4.1.10 or, alternatively, immediately from definitions II.4.1.3 and II.4.1.9).

Remark II.4.1.17. Let $n \in \mathbb{N}$ and $e_1, \ldots, e_n \in \mathbb{F}^{n\times 1}$ be the columns of the identity matrix I of order n, so that $e_1^\mathsf{T}, \ldots, e_n^\mathsf{T} \in \mathbb{F}^{1\times n}$ are the rows of I. Consider an invertible matrix $A \in \mathbb{F}^{n\times n}$. Using rules (a) and (b) of remark II.4.1.5 to interpret matrix multiplication in definition II.4.1.14, we obtain the following:

(a) $Av_j = e_j$ for each $j \in \{1, \ldots, n\}$, where $v_1, \ldots, v_n \in \mathbb{F}^{n\times 1}$ are the columns of A^{-1};

(b) $u_i A = e_i^\mathsf{T}$ for each $i \in \{1, \ldots, n\}$, where $u_1, \ldots, u_n \in \mathbb{F}^{1\times n}$ are the rows of A^{-1}.

The rows and columns of a matrix are defined in definition II.3.1.1. We can equivalently recast statement (b), transposing its equalities with the use of propositions II.3.3.2 and II.4.1.11: $A^\mathsf{T} w_i = e_i$ for each $i \in \{1, \ldots, n\}$, where $w_1, \ldots, w_n$ are the columns of $(A^{-1})^\mathsf{T}$.

Proposition II.4.1.18 (reflexivity of matrix inversion)**.** The operation of matrix inversion is *reflexive*: for any $n \in \mathbb{N}$ and each invertible matrix $A \in \mathbb{F}^{n\times n}$, we have that A^{-1} is invertible and

$$\left(A^{-1}\right)^{-1} = A\,.$$

Proof. Since matrix A is invertible, it has a unique inverse, denoted by A^{-1}. By definition II.4.1.14, we then have $AA^{-1} = I = A^{-1}A$, where I is the identity matrix of order n. Interpreting these two equalities in the context of definition II.4.1.14 with A and A^{-1} interchanged, we conclude that A^{-1} is invertible and that A is the inverse of A^{-1}.

Proposition II.4.1.19. Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$, $\alpha \in \mathbb{F}$. Then the following holds:

(a) $\alpha \cdot A = (\alpha \cdot I_m)A$, where I_m is the identity matrix of order m;

(b) $\alpha \cdot A = A\,(\alpha \cdot I_n)$, where I_n is the identity matrix of order n.

Proof. Let us prove part (a). We have $\alpha \cdot A \in \mathbb{F}^{m\times n}$ by definition II.3.1.5, $\alpha \cdot I_m \in \mathbb{F}^{m\times m}$ by definitions II.3.1.5 and II.4.1.9 and therefore $(\alpha \cdot I_m)A \in \mathbb{F}^{m\times n}$ by definition II.4.1.3, so our goal is proving the equality of two matrices from $\mathbb{F}^{m\times n}$. For arbitrary $i \in \{1, \ldots, m\}$

and $j \in \{1, \dots, n\}$, we obtain

$$\begin{aligned}
\big((\alpha \cdot I_m)A\big)_{ij} &= \sum_{k=1}^{m} (\alpha \cdot I_m)_{ik} \cdot A_{kj} = \sum_{k=1}^{m} (\alpha \cdot (I_m)_{ik}) \cdot A_{kj} \\
&= \sum_{k=1}^{m} (\alpha \cdot \delta_{ik}) \cdot A_{kj} = \sum_{k=1}^{m} (\delta_{ik} \cdot \alpha) \cdot A_{kj} = \sum_{k=1}^{m} \delta_{ik} \cdot (\alpha \cdot A_{kj}) \\
&= \sum_{k=1}^{i-1} 0 \cdot (\alpha \cdot A_{kj}) \quad + \; 1 \cdot (\alpha \cdot A_{ij}) \; + \quad \sum_{k=i+1}^{m} 0 \cdot (\alpha \cdot A_{kj}) \\
&\qquad\qquad = 1 \cdot (\alpha \cdot A_{ij}) = \alpha \cdot A_{ij} = (\alpha \cdot A)_{ij} \,,
\end{aligned}$$

so $(\alpha \cdot I_m)A = \alpha \cdot A$.

First, let us note that any sum of more than two scalars is understood in the sense of recursive reduction to the binary operation of addition. That any such a sum is correctly defined regardless of the order (to be precise, of the *binary tree*) according to which the binary operation of addition is performed on neighboring terms is entirely due to the associativity of the addition of scalars (condition (b) of definition II.1.1.1).

In the above chain of equalities, the first is due to the definition of matrix multiplication (definition II.4.1.3). The second holds by the definition of the product of a scalar and a matrix (definition II.3.1.5). The third follows from the definition of the identity matrix of a given order (definition II.4.1.9). The fourth we obtain using the the commutativity of the multiplication of scalars (condition (a) of definition II.1.1.1). The fifth is due to the associativity of the multiplication of scalars (condition (b) of definition II.1.1.1). The sixth invokes the definition of the Kronecker symbol (§ I.1.27). The seventh relies on the neutrality of the additive identity element of the field with respect to addition (condition (c) of definition II.1.1.1). The eighth is by the neutrality of the multiplicative identity element of the field with respect to addition (condition (c) of definition II.1.1.1). The ninth is, again, due to the definition of the product of a scalar and a matrix (definition II.3.1.5).

Such a depth of justification may seem a bit excessive at this point and was not originally intended. It serves to address questions raised at the lecture and to emphasize yet another time that the field axioms still underlie those transformations that we gradually come to consider trivial and somewhat elementary.

Part (b) is proven analogously. Writing a complete proof is left to the reader as an exercise.

Proposition II.4.1.20 (matrix inversion under multiplication by a scalar)**.** Let $n \in \mathbb{N}$, $\alpha \in \mathbb{F}$ and $A \in \mathbb{F}^{n\times n}$ be invertible. Then $\alpha \cdot A$ is invertible if and only if $\alpha \neq 0$. When $\alpha \neq 0$, we have $(\alpha \cdot A)^{-1} = \alpha^{-1} \cdot A^{-1}$.

Proof. Let I denote the identity matrix of order n.

When $\alpha = 0$, matrix $\alpha \cdot A$ is zero and hence not invertible: for any $B \in \mathbb{F}^{n\times n}$, matrix $(\alpha \cdot A)\, B$ is also zero and cannot be equal to I. So let us focus on the case of $\alpha \neq 0$ in the remainder of the proof.

Since A is invertible, it has a unique inverse, denoted by A^{-1}. By definition II.4.1.14, we then have $AA^{-1} = I = A^{-1}A$. Then, using the above proposition II.4.1.19 and algebraic properties of matrices (parts (a) and (b) of proposition II.4.1.10), we obtain

$$\begin{aligned}(\alpha\cdot A)\,(\alpha^{-1}\cdot A^{-1}) &= \big(A\,(\alpha\cdot I)\big)\,\big((\alpha^{-1}\cdot I)\,A^{-1}\big)\\ &= A\,\big((\alpha\cdot I)\,\big((\alpha^{-1}\cdot I)\,A^{-1}\big)\big) = A\,\big((\alpha\cdot I)\,(\alpha^{-1}\cdot I)\big)\,A^{-1}\\ &= A\,\big((\alpha^{-1}\alpha)\cdot I\big)\,A^{-1} = A\,(1\cdot I)\,A^{-1} = A\,I\,A^{-1} = A\,A^{-1} = I\end{aligned}$$

and, similarly, $(\alpha^{-1}\cdot A^{-1})\,(\alpha\cdot A) = I$. Applying definition II.4.1.14, we conclude that $\alpha\cdot A$ is invertible and that its inverse is $\alpha^{-1}\cdot A^{-1}$.

Remark II.4.1.21. For matrices, as well as for scalars, inversion does not distribute with respect to addition: $A, B \in \mathbb{F}^{n\times n}$ are invertible matrices of order $n \in \mathbb{N}$, then matrices $A^{-1} + B^{-1}$ and $(A+B)^{-1}$, when the latter exists, may be very different.

For example, if $\mathbb{F}$ is such that $1+1 \neq 0$, we know from proposition II.4.1.20 that $(A+A)^{-1} = (1+1)^{-1}\cdot A^{-1}$, while the matrix $A+(-A)$ is not even invertible (example II.4.1.16).

Proposition II.4.1.22 (matrix inversion under matrix multiplication)**.** Let $n \in \mathbb{N}$ and $A, B \in \mathbb{F}^{n\times n}$ be invertible matrices. Then AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$.

Proof. Let I denote the identity matrix of order n. By definition II.4.1.14, we have $AA^{-1} = I = A^{-1}A$ and $BB^{-1} = I = B^{-1}B$. Combining these equalities with parts (a) and (b) of proposition II.4.1.10, we obtain

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$$

and

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}IB = B^{-1}B = I$$

so that AB is invertible and $(AB)^{-1} = B^{-1}A^{-1}$ by definition II.4.1.14.

Proposition II.4.1.23 (matrix inversion under transposition)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$ be an invertible matrix. Then $A^\intercal$ is invertible and $(A^\intercal)^{-1} = (A^{-1})^\intercal$.

Proof. Let I denote the identity matrix of order n. By definition II.4.1.14, we have $AA^{-1} = I = A^{-1}A$. Combining these equalities with proposition II.4.1.11, we obtain

$$A^\intercal(A^{-1})^\intercal = (A^{-1}A)^\intercal = I^\intercal = I \quad\text{and}\quad (A^{-1})^\intercal A^\intercal = (AA^{-1})^\intercal = I^\intercal = I\,,$$

so that $A^\intercal$ is invertible and $(A^\intercal)^{-1} = (A^{-1})^\intercal$ by definition II.4.1.14. This proves the claim.

§ II.4.2. Subvectors and submatrices

For any $n, n_1 \in \mathbb{N}$ such that $n_1 \leq n$, we consider the selection of a n_1-element ordered subset of the set $\{1,\ldots,n\}$. This will be useful for selecting the respective components (or, equivalently, removing all the others) in a tuple (column vector) from $\mathbb{F}^n$ and forming a tuple (column vector) from the selected components in the prescribed order. To keep track of the ordering, we will represent every such a subset by an index tuple $\sigma = (\sigma_\beta)_{\beta=1}^{n_1} \in \{1,\ldots,n\}^{n_1}$ (see §§ I.1.9 and I.1.10) of indices $\sigma_1,\ldots,\sigma_{n_1} \in \{1,\ldots,n\}$.

We call any index tuple $\sigma = (\sigma_\beta)_{\beta=1}^{n_1}$ with $n_1 \in \mathbb{N}$ *contiguous* if

$$\{\sigma_\beta\}_{\beta=1}^{n_1} = \{j \in \mathbb{N} \colon \sigma_{\min} \leq j \leq \sigma_{\max}\} \tag{II.4.2.1}$$

for $\sigma_{\min} = \min\{\sigma_\beta\}_{\beta=1}^{n_1}$ and $\sigma_{\max} = \max\{\sigma_\beta\}_{\beta=1}^{n_1}$ and *order-preserving* if

$$\sigma_{\beta-1} < \sigma_\beta \quad \text{for each} \quad \beta \in \{2, \ldots, n_1\}\,. \tag{II.4.2.2}$$

Clearly, any index tuple $(\sigma_\beta)_{\beta=1}^{n_1}$ with $n_1 \in \mathbb{N}$ that is both contiguous and order-preserving satisfies $\sigma_\beta = \sigma_1 + \beta - 1$ for every $\beta \in \{1, \ldots, n_1\}$.

Definition II.4.2.1 (subcolumn). Let $n, n_1 \in \mathbb{N}$ be such that $n_1 \leq n$ and $x \in \mathbb{F}^n$.

(a) Consider distinct $\sigma_1, \ldots, \sigma_{n_1} \in \{1, \ldots, n\}$. Then the column vector

$$[x_{\sigma_\beta}]_{\beta=1}^{n_1} \in \mathbb{F}^{n_1}$$

is called the *subcolumn* of x formed by components $\sigma_1, \ldots, \sigma_{n_1}$.

(b) If $\widehat{x} \in \mathbb{F}^{n_1}$ is the subcolumn of x formed by components $\sigma_1, \ldots, \sigma_{n_1}$, where $\sigma = (\sigma_\beta)_{\beta=1}^{n_1}$ is a tuple with distinct components, then $\widehat{x}$ is said to be *a subcolumn* of x. If additionally

(i) σ is a contiguous index tuple with distinct components,

(ii) σ is an order-preserving index tuple,

then $\widehat{x}$ is said to be, respectively,

(i) *a contiguous subcolumn* of x,

(ii) *a subcolumn* of x *with the original ordering of the components.*

Remark II.4.2.2. A contiguous subcolumn $\widehat{x}$ of a column vector x with the original ordering of the components is often called a *block* of x.

We can define a submatrix in a similar (and consistent, in view of remark II.4.1.4) fashion, using two tuples of distinct indices encoding the rows and the columns to be selected together with their ordering.

Definition II.4.2.3 (submatrix). Let $m, n, m_1, n_1 \in \mathbb{N}$ be such that $m_1 \leq m$ and $n_1 \leq n$ and $A \in \mathbb{F}^{m \times n}$.

(a) Consider distinct $\pi_1, \ldots, \pi_{m_1} \in \{1, \ldots, m\}$ and distinct $\sigma_1, \ldots, \sigma_{n_1} \in \{1, \ldots, n\}$. Then the matrix

$$[A_{\pi_\alpha \sigma_\beta}]_{\alpha=1,\,\beta=1}^{m_1 \quad n_1} \in \mathbb{F}^{m_1 \times n_1}$$

is called the *submatrix* of A formed by rows $\pi_1, \ldots, \pi_{m_1}$ and columns $\sigma_1, \ldots, \sigma_{n_1}$.

(b) If $\widehat{A} \in \mathbb{F}^{m_1 \times n_1}$ is the submatrix of A formed by rows $\pi_1, \ldots, \pi_{m_1}$ and columns $\sigma_1, \ldots, \sigma_{n_1}$, where $\pi = (\pi_\alpha)_{\alpha=1}^{m_1}$ and $\sigma = (\sigma_\beta)_{\beta=1}^{n_1}$ are tuples with distinct components, then $\widehat{A}$ is said to be *a submatrix* of A. If additionally

(i) π is a contiguous index tuple with distinct components,

(ii) σ a contiguous index tuple with distinct components,

(iii) π and σ are contiguous index tuples with distinct components,

(iv) π is an order-preserving index tuple,

(v) σ is an order-preserving index tuple,

(vi) π and σ are order-preserving index tuples,

then $\widehat{A}$ is said to be, respectively,

(i) *a row-contiguous submatrix* of A,

(ii) *a column-contiguous submatrix* of A,
(iii) *a contiguous submatrix* of A,
(iv) *a submatrix* of A *with the original ordering of the rows*,
(v) *a submatrix* of A *with the original ordering of the columns*,
(vi) *a submatrix* of A *with the original ordering of the entries.*

Example II.4.2.4. Let $m, n, m_1, n_1 \in \mathbb{N}$ be such that $m_1 \leq m$ and $n_1 \leq n$ and $O, \widehat{O}$ denote the zero matrices from $\mathbb{F}^{m\times n}$ and $\mathbb{F}^{m_1\times n_1}$. Then $\widehat{O}$ is a contiguous submatrix of O with the original ordering of the entries. Indeed, $\widehat{O}$ is the submatrix of O corresponding to any contiguous order-preserving (or just any) index tuples $\pi \in \{1, \ldots, m\}^{m_1}$ and $\sigma \in \{1, \ldots, n\}^{n_1}$ (with distinct entries, at least). This example is, however, trivial and of little interest.

Example II.4.2.5. Consider the matrix $A \in \mathbb{R}^{3\times 3}$ from example II.3.2.3. Selecting a 2 × 2 submatrix, or a submatrix formed by two rows and two columns, in A corresponds to considering definition II.4.2.3 with $m_1 = 2$ and $n_1 = 2$. For example, we can choose to form such a submatrix from rows 1 and 3 and from columns 2 and 3.

Having decided upon the size of the submatrix to be selected and on the *sets* of rows and columns that should form it, we still get to choose the order of the selected rows and columns. In this example, that means choosing which of the two rows selected in A should form the first row of the submatrix and which of the two columns selected in A should form the first column of the submatrix.

To have the third row and the second column of A form the first row and the first column of the submatrix, we set $\pi_1 = 3$ and $\sigma_1 = 2$. This choice exhausts the freedom in ordering the rows and columns of the submatrix we have in this example: it necessarily results in $\pi_2 = 1$ and $\sigma_2 = 3$ and yields the submatrix

$$\widehat{A} = \begin{bmatrix} \mathbf{18} & \mathbf{22} \\ \mathbf{2} & \mathbf{3} \end{bmatrix} \quad \text{of} \quad A = \begin{bmatrix} 1 & \mathbf{2} & \mathbf{3} \\ 2 & 7 & 7 \\ 6 & \mathbf{18} & \mathbf{22} \end{bmatrix}. \tag{II.4.2.3}$$

Notice that $\widehat{A}$ is formed from two contiguous columns of A (one of the selected indices follows the other) taken in the original order and from two noncontiguous rows of A (the set of selected rows contains 1 and 3 but not 2). In fact, there is no way to obtain $\widehat{A}$ as a submatrix of A by selecting contiguous rows or by selecting rows in the original order. So $\widehat{A}$ is a column-contiguous submatrix of A with the original ordering of the columns but not a row-contiguous submatrix of A and not a submatrix of A with the original ordering of the rows.

Remark II.4.2.6. In the context of definition II.4.2.3, a contiguous submatrix $\widehat{A}$ of A with the original ordering of the entries is often called a *block* of A. Such a block is called a *block row* of A if $n_1 = n$ and a *block column* of A if $m_1 = m$. Any block of A is the submatrix of A formed by the intersection of the respective block row and of the respective block column.

For example, for the matrix A given in (II.4.2.3), the submatrices

$$\begin{bmatrix} 1 & \mathbf{2} & \mathbf{3} \\ 2 & 7 & 7 \end{bmatrix}, \quad \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ 7 & 7 \\ \mathbf{18} & \mathbf{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ 7 & 7 \end{bmatrix}$$

are, respectively, a block row, a block column and the block formed by the intersection of that block row and of that block column.

Definition II.4.2.7 (principal submatrix)**.** In the context of definition II.4.2.3, assume $m_1 = n_1$ and let $\pi_1, \dots, \pi_{n_1}$ coincide with $\sigma_1, \dots, \sigma_{n_1}$. Then $\widehat{A}$ is referred to as the *principal submatrix* of A formed by rows and columns $\pi_1, \dots, \pi_{n_1}$.

Definition II.4.2.8 (leading submatrix)**.** In the context of definition II.4.2.3, assume $\pi_i = i$ for each $i \in \{1, \dots, m_1\}$ and $\sigma_j = j$ for each $j \in \{1, \dots, n_1\}$. Then $\widehat{A}$ is referred to as the *leading submatrix* of size $m_1 \times n_1$ of A.

In other words, the leading submatrix of size $m_1 \times n_1$ of a matrix with suitable $m_1, n_1 \in \mathbb{N}$ is the submatrix located in the original matrix at the intersection of the first m_1 rows and of the first n_1 columns.

For any $m, n, r \in \mathbb{N}$ such that $r \leq m, n$, the *leading principal submatrix* of order $r \in \mathbb{N}$ of any $A \in \mathbb{F}^{m \times n}$ is the leading submatrix of size $r \times r$ of A. Clearly, it is also a principal submatrix of A.

In definitions II.4.2.1 and II.4.2.3, we extract a single subcolumn $\widehat{x} \in \mathbb{F}^{n_1}$ of a column vector $x \in \mathbb{F}^n$ and a single submatrix $\widehat{A} \in \mathbb{F}^{m_1 \times n_1}$ of a matrix $A \in \mathbb{F}^{m \times n}$. Except in the trivial case of $m_1 = m$ and $n_1 = n$, that means discarding some of the entries. Now we consider *partitions* of a column vector and of a matrix into several subcolumns and submatrices, respectively, the goal of which is to include every entry in exactly one of the subcolumns and submatrices, respectively. We start with introducing ordered partitions (see § I.1.25) of index sets into ordered index subsets and define the associated partitions of column vectors and matrices.

Definition II.4.2.9 (index-set partition)**.** For $n, q \in \mathbb{N}$ such that $q \leq n$, assume that $\sigma_j = (\sigma_{j\beta})_{\beta=1}^{n_j} \in \{1, \dots, n\}^{n_j}$ is an index tuple with $n_j \in \mathbb{N}$ distinct components for each $j \in \{1, \dots, q\}$ and that the respective index sets $\{\sigma_{j\beta}\}_{\beta=1}^{n_j}$ with $j \in \{1, \dots, q\}$ form a partition of $\{1, \dots, n\}$, so that $n = n_1 + \dots + n_q$. Then we say that the index tuples $\sigma_1, \dots, \sigma_q$ define *an ordered partition of the index set* $\{1, \dots, n\}$ *into ordered subsets.* We refer to q as *the length of the partition* and to $n_1, \dots, n_q$, as *the sizes of the subsets.*

Definition II.4.2.10 (column-vector partition)**.** For $n \in \mathbb{N}$, consider the ordered partition of $\{1, \dots, n\}$, of length $q \in \mathbb{N}$, into ordered subsets of sizes $n_1, \dots, n_q \in \mathbb{N}$ and defined by index tuples $\sigma_j = (\sigma_{j\beta})_{\beta=1}^{n_j}$ with $j \in \{1, \dots, q\}$. For any $x \in \mathbb{F}^n$, that partition induces the subcolumns

$$\widehat{x}_j = [x_{\sigma_{j\beta}}]_{\beta=1}^{n_j} \in \mathbb{F}^{n_j}$$

with $j \in \{1, \dots, q\}$, which are said to form *the partition of the column vector x associated with the index tuples $\sigma_1, \dots, \sigma_q$.*

Definition II.4.2.11 (matrix partition)**.** For $m, n \in \mathbb{N}$, consider ordered partitions of $\{1, \dots, m\}$ and $\{1, \dots, n\}$, of lengths $p \in \mathbb{N}$ and $q \in \mathbb{N}$, into ordered subsets of sizes $m_1, \dots, m_p \in \mathbb{N}$ and $n_1, \dots, n_q \in \mathbb{N}$, defined by index tuples $\pi_i = (\pi_{i\alpha})_{\alpha=1}^{m_i}$ with $i \in \{1, \dots, p\}$ and $\sigma_j = (\sigma_{j\beta})_{\beta=1}^{n_j}$ with $j \in \{1, \dots, q\}$, respectively. Then the index sets $\{(\pi_{i\alpha}, \sigma_{j\beta})\}_{\alpha=1,\, \beta=1}^{m_i \quad n_j}$ with $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, q\}$ form a partition of $\{1, \dots, m\} \times \{1, \dots, n\}$. For any $A \in \mathbb{F}^{m \times n}$, that partition induces the submatrices

$$\widehat{A}_{ij} = [A_{\pi_{i\alpha}\, \sigma_{j\beta}}]_{\alpha=1,\, \beta=1}^{m_i \quad n_j} \in \mathbb{F}^{m_i \times n_j}$$

with $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, q\}$, which are said to form *the partition of the matrix A associated with the row-index tuples $\pi_1, \dots, \pi_p$ and with the column-index tuples $\sigma_1, \dots, \sigma_q$.* The partition is said to have *size* $p \times q$, and $m_1, \dots, m_p$ and $n_1, \dots, n_q$ are referred to as the *row and column sizes of the submatrices* of the partition.

Remark II.4.2.12 (block matrix partition)**.** When all the index tuples considered in definition II.4.2.11 are contiguous and order-preserving (as we define in (II.4.2.1) and (II.4.2.2) above), the submatrices of A defined in definition II.4.2.11 are blocks of A in the sense of definition II.4.2.3: they all are contiguous and inherit the order of entries (of rows and columns) from A. If, in addition, the index tuples satisfy $\pi_{i+1,1} = \pi_{im_i} + 1$ for every $i \in \{1, \dots, p-1\}$ and $\sigma_{j+1,1} = \sigma_{jn_j} + 1$ for every $j \in \{1, \dots, q-1\}$, then the partition introduced in definition II.4.2.11 is referred to as a *block partition* of A. Such partitions are often invoked using the following block-matrix notation, cf. the original entrywise notation (II.3.1.1):

$$\begin{bmatrix} \widehat{A}_{11} & \cdots & \widehat{A}_{1q} \\ \vdots & \ddots & \vdots \\ \widehat{A}_{p1} & \cdots & \widehat{A}_{pq} \end{bmatrix}, \tag{II.4.2.4}$$

which denotes the matrix A partitioned into the given submatrices as described in remark II.4.2.12.

Let us consider how the product of two matrices can be expressed in terms of consistent partitions of the two factors and of the product.

Proposition II.4.2.13 (matrix product under matrix partition)**.** For $m, n, r \in \mathbb{N}$, consider ordered partitions of $\{1, \dots, m\}$, $\{1, \dots, n\}$ and $\{1, \dots, r\}$ into ordered subsets, of lengths $p \in \mathbb{N}$, $q \in \mathbb{N}$ and $s \in \mathbb{N}$ and of sizes $m_1, \dots, m_p \in \mathbb{N}$, $n_1, \dots, n_q$ and $r_1, \dots, r_s$, defined by index tuples $\pi_i = (\pi_{i\alpha})_{\alpha=1}^{m_i}$ with $i \in \{1, \dots, p\}$, $\sigma_j = (\sigma_{j\beta})_{\beta=1}^{n_j}$ with $j \in \{1, \dots, q\}$ and $\tau_k = (\tau_{k\gamma})_{\gamma=1}^{r_k}$ with $k \in \{1, \dots, s\}$, respectively.

Let $A \in \mathbb{F}^{m \times r}$, $B \in \mathbb{F}^{r \times n}$ and $C \in \mathbb{F}^{m \times n}$ and consider the following:

(i) the submatrices $\widehat{A}_{ik} = [A_{\pi_{i\alpha}\, \tau_{k\gamma}}]_{\alpha=1,\, \gamma=1}^{m_i \quad r_k} \in \mathbb{F}^{m_i \times r_k}$ with $i \in \{1, \dots, p\}$ and $k \in \{1, \dots, s\}$ forming the partition of A associated with the row-index tuples $\pi_1, \dots, \pi_p$ and with the column-index tuples $\tau_1, \dots, \tau_s$;

(ii) the submatrices $\widehat{B}_{kj} = [B_{\tau_{k\gamma}\,\sigma_{j\beta}}]_{\gamma=1,\,\beta=1}^{r_k\quad n_j} \in \mathbb{F}^{r_k \times n_j}$ with $k \in \{1,\ldots,s\}$ and $j \in \{1,\ldots,q\}$ forming the partition of B associated with the row-index tuples $\tau_1,\ldots,\tau_s$ and with the column-index tuples $\sigma_1,\ldots,\sigma_q$;

(iii) the submatrices $\widehat{C}_{ij} = [C_{\pi_{i\alpha}\,\sigma_{j\beta}}]_{\alpha=1,\,\beta=1}^{m_i\quad n_j} \in \mathbb{F}^{m_i \times n_j}$ with $i \in \{1,\ldots,p\}$ and $j \in \{1,\ldots,q\}$ forming the partition of C associated with the row-index tuples $\pi_1,\ldots,\pi_p$ and with the column-index tuples $\sigma_1,\ldots,\sigma_q$.

Then $C = AB$ holds if and only if

$$\widehat{C}_{ij} = \sum_{k=1}^{s} \widehat{A}_{ik}\,\widehat{B}_{kj} \qquad \text{(II.4.2.5)}$$

holds for all $i \in \{1,\ldots,p\}$ and $j \in \{1,\ldots,q\}$.

The above proposition II.4.2.13, which we prove below, shows that the matrix product of consistently partitioned matrices can be evaluated according to a formula similar to (II.4.1.9), in terms of the submatrices forming the partitions.

One crucial difference of (II.4.2.5) from the formula (II.4.1.9), defining the matrix multiplication in terms of the individual entries, is that the equality of (II.4.2.5) with any $i \in \{1,\ldots,p\}$ and $j \in \{1,\ldots,q\}$ is an equality in $\mathbb{F}^{m_i \times n_i}$ and not in $\mathbb{F}$, as the equality of (II.4.1.9) with any $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$. Another important difference is that the product involved on the right-hand side of (II.4.2.5) is the matrix product (which itself is defined in (II.4.1.9)) and not the product of scalars in $\mathbb{F}$, as in (II.4.1.9). Since the matrix multiplication is not commutative, the submatrices of A and B need to be multiplied in (II.4.2.5) in the specified order. On the upside, multiplying such submatrices in the wrong order often happens to be impossible due to size inconsistency.

Before we proceed to the proof, let us consider *block matrix partitions*, a special and very useful type of matrix partitions.

Remark II.4.2.14. Consider proposition II.4.2.13 in the case of $p = q = s = 2$ and when the partitions of A, B and C are block partitions in the sense of remark II.4.2.6. Then C, A and B can be expressed using the block notation (II.4.2.4) as follows:

$$C = \begin{bmatrix} \widehat{C}_{11} & \widehat{C}_{12} \\ \widehat{C}_{21} & \widehat{C}_{22} \end{bmatrix}, \qquad A = \begin{bmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ \widehat{A}_{21} & \widehat{A}_{22} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} \widehat{B}_{11} & \widehat{B}_{12} \\ \widehat{B}_{21} & \widehat{B}_{22} \end{bmatrix}. \qquad \text{(II.4.2.6)}$$

Then (II.4.2.5) with any $i \in \{1,\ldots,p\}$ and $j \in \{1,\ldots,q\}$ can be interpreted as the evaluation of submatrix (i,j) of $C = AB$ according to rules analogous to the two ones given in remark II.4.1.5. For example, the first block row and the second block column of C can be evaluated as follows:

$$\begin{aligned} \begin{bmatrix} \widehat{C}_{11} & \widehat{C}_{12} \end{bmatrix} &= \begin{bmatrix} \widehat{A}_{11} & \widehat{A}_{12} \end{bmatrix} \cdot B = \widehat{A}_{11} \cdot \begin{bmatrix} \widehat{B}_{11} & \widehat{B}_{12} \end{bmatrix} + \widehat{A}_{12} \cdot \begin{bmatrix} \widehat{B}_{21} & \widehat{B}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \widehat{A}_{11}\widehat{B}_{11} + \widehat{A}_{12}\widehat{B}_{21} & \widehat{A}_{11}\widehat{B}_{12} + \widehat{A}_{12}\widehat{B}_{22} \end{bmatrix} \qquad \text{(II.4.2.7)} \end{aligned}$$

and

$$\begin{bmatrix} \widehat{C}_{12} \\ \widehat{C}_{22} \end{bmatrix} = A \cdot \begin{bmatrix} \widehat{B}_{12} \\ \widehat{B}_{22} \end{bmatrix} = \begin{bmatrix} \widehat{A}_{11} \\ \widehat{A}_{21} \end{bmatrix} \cdot \widehat{B}_{12} + \begin{bmatrix} \widehat{A}_{12} \\ \widehat{A}_{22} \end{bmatrix} \cdot \widehat{B}_{22} = \begin{bmatrix} \widehat{A}_{11}\widehat{B}_{12} + \widehat{A}_{12}\widehat{B}_{22} \\ \widehat{A}_{21}\widehat{B}_{12} + \widehat{A}_{22}\widehat{B}_{22} \end{bmatrix} \qquad \text{(II.4.2.8)}$$

In place of scalar coefficients of the linear combinations considered in remark II.4.1.5 and example II.4.1.6, above we have matrix coefficients (highlighted in color). In every product of submatrices of A and B appearing in (II.4.2.7) and (II.4.2.8), the order of the submatrices of A and B is always the same as the order of A and B in the product.

Example II.4.2.15. Let us revisit the matrices we considered in examples II.3.2.3 and II.4.1.6. Let $\ell_1, \ell_2, \ell_3 \in \mathbb{R}^{3\times 1}$ be the columns of L and $u_1, u_2, u_3 \in \mathbb{R}^{3\times 1}$ be the columns of U^T, so that $u_1^\mathsf{T}, u_2^\mathsf{T}, u_3^\mathsf{T}$ are the rows of U. It happens that ℓ_1, ℓ_2 are exactly the columns of $\widehat{L}$ and $u_1^\mathsf{T}, u_2^\mathsf{T}$ are exactly the rows of $\widehat{U}$. So the matrices $\widehat{L}$, $\widehat{U}$, L and U can be obtained by "concatenation" as described in remark II.4.2.12:

$$\widehat{L} = \begin{bmatrix} \ell_1 \ \ell_2 \end{bmatrix}, \quad \widehat{U} = \begin{bmatrix} u_1^\mathsf{T} \\ u_2^\mathsf{T} \end{bmatrix}, \quad L = \begin{bmatrix} \ell_1 \ \ell_2 \ \ell_3 \end{bmatrix} = \begin{bmatrix} \widehat{L} \ \ell_3 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_1^\mathsf{T} \\ u_2^\mathsf{T} \\ u_3^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \widehat{U} \\ u_3^\mathsf{T} \end{bmatrix}.$$

The definition of matrix multiplication (definition II.4.1.3) yields

$$\widehat{A} = \widehat{L}\,\widehat{U} = \sum_{k=1}^{2} \ell_k u_k^\mathsf{T} \quad \text{and} \quad A = LU = \sum_{k=1}^{3} \ell_k u_k^\mathsf{T} = \widehat{A} + \ell_3 u_3^\mathsf{T}\,. \tag{II.4.2.9}$$

This shows, in particular, that the effect of removing a column from the left factor of a product and the corresponding row from the second factor consists in omitting the corresponding one of the three terms forming the product. This relation between A and $\widehat{A}$ can be deduced from the above block structure, using proposition II.4.2.13 and remark II.4.2.14 instead of the definition of the matrix multiplication (definition II.4.1.3):

$$LU = \begin{bmatrix} \widehat{L} \ \ell_3 \end{bmatrix} \begin{bmatrix} \widehat{U} \\ u_3^\mathsf{T} \end{bmatrix} = \widehat{L}\widehat{U} + \ell_3 u_3^\mathsf{T}\,. \tag{II.4.2.10}$$

Proof of proposition II.4.2.13. For arbitrary indices $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$, the two sides of (II.4.2.5) are elements of $\mathbb{F}^{m_i \times n_j}$. Consider $\alpha \in \{1, \ldots, m_i\}$ and $\beta \in \{1, \ldots, n_j\}$. First, the definition of the submatrices of C yields

$$(\widehat{C}_{ij})_{\alpha\beta} = C_{\pi_{i\alpha}\,\sigma_{j\beta}}\,. \tag{II.4.2.11}$$

Second, the definition of matrix multiplication (definition II.4.1.3) applied to A and B results in the first of the following equalities:

$$\begin{aligned}(AB)_{\pi_{i\alpha}\,\sigma_{j\beta}} &= \sum_{\ell=1}^{r} A_{\pi_{i\alpha}\,\ell} \cdot B_{\ell\,\sigma_{j\beta}} = \sum_{k=1}^{s}\sum_{\gamma=1}^{r_k} A_{\pi_{i\alpha}\,\tau_{k\gamma}} \cdot B_{\tau_{k\gamma}\,\sigma_{j\beta}} \\ &= \sum_{k=1}^{s}\sum_{\gamma=1}^{r_k} (\widehat{A}_{ik})_{\alpha\gamma} \cdot (\widehat{B}_{kj})_{\gamma\beta} = \sum_{k=1}^{s} (\widehat{A}_{ik}\widehat{B}_{kj})_{\alpha\beta}\,. \end{aligned} \tag{II.4.2.12}$$

Let us justify the other equalities of (II.4.2.12). By assumption, the tuples $\tau_k = (\tau_{k\gamma})_{\gamma=1}^{r_k}$ with $k \in \{1, \ldots, s\}$ define an ordered partition of $\{1, \ldots, r\}$ into ordered subsets. That is meant in the sense of definition II.4.2.9, so that the indices $\tau_{k\gamma}$ with $k \in \{1, \ldots, s\}$ and

$\gamma \in \{1, \dots, r_k\}$ are all distinct and constitute the set $\{1, \dots, r\}$. The summation with respect to ℓ can therefore be replaced with the one with respect to k and γ provided that $\tau_{k\gamma}$ is substituted for ℓ in the summand. That leads to the second equality of (II.4.2.12). Further, applying the definition of the submatrices of A and B involved in the resulting expression gives the third equality of (II.4.2.12). Finally, the definition of matrix multiplication (definition II.4.1.3) applied to those submatrices results in the fourth equality of (II.4.2.12).

First, let us assume $C = AB$ and consider $i \in \{1, \dots, p\}$, $j \in \{1, \dots, q\}$ and $\alpha \in \{1, \dots, m_i\}$, $\beta \in \{1, \dots, n_j\}$. The *right-hand-side* expression of (II.4.2.11) and the *first* expression of (II.4.2.12) are equal by assumption. Then (II.4.2.11) and (II.4.2.12) imply that the *left-hand-side* expression of (II.4.2.11) and the *last* expression of (II.4.2.12) are equal. That leads to (II.4.2.5) for the indices $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, q\}$, which we assumed to be arbitrary.

To prove the converse, let us assume (II.4.2.5) for all $i \in \{1, \dots, p\}$ and $j \in \{1, \dots, q\}$ and prove $C = AB$ By assumption, $\pi_i = (\pi_{i\alpha})_{\alpha=1}^{p_i}$ with $i \in \{1, \dots, p\}$ and $\sigma_j = (\sigma_{j\beta})_{\beta=1}^{q_j}$ with $j \in \{1, \dots, q\}$ define ordered partitions of $\{1, \dots, p\}$ and $\{1, \dots, q\}$ into ordered subsets. That is meant, again, in the sense of definition II.4.2.9, so that the indices $\pi_{i\alpha}$ with $i \in \{1, \dots, p\}$ and $\alpha \in \{1, \dots, p_i\}$ are all distinct and constitute the set $\{1, \dots, m\}$ and the indices $\sigma_{j\beta}$ with $j \in \{1, \dots, q\}$ and $\beta \in \{1, \dots, q_j\}$ are all distinct and constitute the set $\{1, \dots, n\}$. As a result, we can use these indices to enumerate all rows and columns of C and AB, and it therefore suffices to show the *right-hand side* expression of (II.4.2.11) to be equal to the *first* expression of (II.4.2.12) for all $i \in \{1, \dots, p\}$, $j \in \{1, \dots, q\}$ and $\alpha \in \{1, \dots, m_i\}$, $\beta \in \{1, \dots, n_j\}$. That follows by (II.4.2.11) and (II.4.2.12) since the *left-hand side* expression of (II.4.2.11) is equal to the *last* expression of (II.4.2.12) by assumption.

§ II.4.3. Permutation matrices

For a definition of permutations and exchanges, see § I.1.28.

Definition II.4.3.1 (permutation matrix)**.** For $n \in \mathbb{N}$, a matrix $\mathit{\Pi} \in \mathbb{F}^{n \times n}$ is called a *permutation matrix of order n* if $\mathit{\Pi}_{ij} \in \{0, 1\}$ for all $i, j \in \{1, \dots, n\}$ and the following two conditions are satisfied:

(i) for every $i \in \{1, \dots, n\}$, there exists a unique $j \in \{1, \dots, n\}$ such that $\mathit{\Pi}_{ij} = 1$;
(ii) for every $j \in \{1, \dots, n\}$, there exists a unique $i \in \{1, \dots, n\}$ such that $\mathit{\Pi}_{ij} = 1$.

In definition II.4.3.1, conditions (i) and (ii) mean that there is exactly one unit entry in every row and in every column.

Example II.4.3.2. The following two matrices are permutation matrices of order three:

$$\mathit{\Pi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathit{\Sigma} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

The multiplication of any $x \in \mathbb{F}^3$ by these matrices yields

$$\varPi x = \begin{bmatrix} x_1 \\ x_3 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_{\pi_1} \\ x_{\pi_2} \\ x_{\pi_3} \end{bmatrix}, \quad \varSigma x = \begin{bmatrix} x_3 \\ x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} x_{\sigma_1} \\ x_{\sigma_2} \\ x_{\sigma_3} \end{bmatrix} \quad \text{and} \quad Tx = \begin{bmatrix} x_2 \\ x_3 \\ x_1 \end{bmatrix} = \begin{bmatrix} x_{\tau_1} \\ x_{\tau_2} \\ x_{\tau_3} \end{bmatrix},$$

where $\pi = (1,3,2)$, $\sigma = (3,2,1)$ and $\tau = (2,3,1)$ are permutations of $\{1,2,3\}$ (see § I.1.28), so that the multiplication of a column vector by any of the above permutation matrices amounts to applying the respective permutation to the components of the column vector.

Proposition II.4.3.3. For $n \in \mathbb{N}$, let π be a permutation of $\{1,\dots,n\}$. Then there exists a unique matrix $\varPi \in \mathbb{F}^{n\times n}$ satisfying $\varPi x = [x_{\pi_i}]_{i=1}^n$ for any $x \in \mathbb{F}^n$. That matrix $\varPi$ is uniquely defined by $\varPi_{ij} = \delta_{j\pi_i}$ for all $i,j \in \{1,\dots,n\}$ and is a permutation matrix.

For a definition of the Kronecker symbol, see § I.1.27.

Proof of proposition II.4.3.3. Assume that $\varPi \in \mathbb{F}^{n\times n}$ is such that $\varPi x = [x_{\pi_i}]_{i=1}^n$ for any $x \in \mathbb{F}^n$. In particular, denoting by e_j the jth column of the identity matrix of order n for $j \in \{1,\dots,n\}$, so that $e_j = [\delta_{ji}]_{i=1}^n \in \mathbb{F}^n$, we obtain

$$(\varPi e_j)_i = (e_j)_{\pi_i} = \delta_{j\pi_i}$$

for every $i \in \{1,\dots,n\}$, i.e., $\varPi e_j = [\delta_{j\pi_i}]_{i=1}^n$. On the other hand, $\varPi e_j$ is the jth column of $\varPi$ by rule (a) of remark II.4.1.5. So the condition $\varPi x = [x_{\pi_i}]_{i=1}^n$ for any $x \in \mathbb{F}^n$ imposed on a matrix $\varPi \in \mathbb{F}^{n\times n}$ implies $\varPi_{ij} = \delta_{j\pi_i}$ for all $i,j \in \{1,\dots,n\}$, i.e., $\varPi = [\delta_{j\pi_i}]_{i=1,\,j=1}^{n\quad n}$. We have now identified precisely what $\varPi$ *has* to be:

$$\varPi = [\delta_{j\pi_i}]_{i=1,\,j=1}^{n\quad n} \tag{II.4.3.1}$$

is a well-defined element of $\mathbb{F}^{n\times n}$. Still, we have not proved the existence claim yet: it remains to show that the only possible value of $\varPi$ *is* a suitable one. That is the case since

$$(\varPi x)_i = \sum_{j=1}^n \varPi_{ij}\, x_j = \sum_{j=1}^n \delta_{j\pi_i}\, x_j = x_{\pi_i} \quad \text{for each} \quad i \in \{1,\dots,n\} \tag{II.4.3.2}$$

for any $x \in \mathbb{F}^n$. So we conclude that there is a unique matrix $\varPi \in \mathbb{F}^{n\times n}$ satisfying $\varPi x = [x_{\pi_i}]_{i=1}^n$ for any $x \in \mathbb{F}^n$, which is given by (II.4.3.1). It remains to prove that $\varPi$ is a permutation matrix.

First, we note that $\varPi_{ij} \in \{0,1\}$ for all $i,j \in \{1,\dots,n\}$ by (II.4.3.1). Second, $\varPi$ satisfies condition (i) of definition II.4.3.1: for every $i \in \{1,\dots,n\}$, we have $\pi_i \in \{1,\dots,n\}$ since π is a permutation of $\{1,\dots,n\}$ and, by (II.4.3.1), there is a unique $j = \pi_i \in \{1,\dots,n\}$ such that $\varPi_{ij} = 1$. Third, the matrix $\varPi$ satisfies condition (ii) of definition II.4.3.1. Indeed, let $i,k,j \in \{1,\dots,n\}$ be such that $\varPi_{ij} = 1 = \varPi_{kj}$. By the above definition of $\varPi$, that implies $\pi_k = j = \pi_i$, which implies $k = i$ since π is a permutation.

The calculation presented in (II.4.3.2) can be expanded into a sequence of elementary steps, see the proof of proposition II.4.1.19 for an example.

Definition II.4.3.4 (permutation matrix associated with a permutation)**.** Let $n \in \mathbb{N}$ and π be a permutation of $\{1,\dots,n\}$. The matrix $\varPi \in \mathbb{F}^{n\times n}$ defined and proven to be a permutation matrix in proposition II.4.3.3 is referred to as *the permutation matrix associated with* π and as *the permutation matrix realizing* π.

The following proposition shows that the the association between a permutation and the respective permutation matrix established in proposition II.4.3.3 and definition II.4.3.4 is a one-to-one correspondence.

Proposition II.4.3.5. Let $\varPi$ be a permutation matrix of order $n \in \mathbb{N}$. Then there exists a unique permutation π of $\{1, \ldots, n\}$ such that $\varPi$ is the permutation matrix associated with π, and that permutation π is uniquely defined by $\varPi_{i\pi_i} = 1$ for every $i \in \{1, \ldots, n\}$.

Proof. Assume that σ is a permutation of $\{1, \ldots, n\}$ realized by $\varPi$. Then we have $\varPi_{ij} = \delta_{j\sigma_i}$ for all $i, j \in \{1, \ldots, n\}$; in particular, we have $\varPi_{i\sigma_i} = 1$ for all $i \in \{1, \ldots, n\}$. So far we have proven that any permutation realized by $\varPi$ satisfies that condition. Whether that condition actually defines a permutation and whether that permutation is realized by $\varPi$ are different matters, which remain to be addressed.

By condition (i) of definition II.4.3.1, for every $i \in \{1, \ldots, n\}$, there exists a unique column index $\pi_i \in \{1, \ldots, n\}$ such that $\varPi_{i\pi_i} = 1$. This renders $\pi \in \{1, \ldots, n\}^n$ uniquely defined.

By the same condition (i) of definition II.4.3.1, the matrix $\varPi$ then satisfies (II.4.3.1). It remains to prove that π is a permutation.

Let us assume that $\pi_i = \pi_k$ for some $i, k \in \{1, \ldots, n\}$ such that $i \neq k$. Then we have $\varPi_{i\pi_i} = 1 = \varPi_{k\pi_i}$, i.e., column π_i of $\varPi$ contains two unit entries, which contradicts condition (ii) of definition II.4.3.1. We therefore conclude that the indices $\pi_1, \ldots, \pi_n$ are distinct and π is therefore a permutation. By definition II.4.3.4 and proposition II.4.3.3, the permutation π is realized by $\varPi$ due to (II.4.3.1).

Proposition II.4.3.6. Consider permutation matrices $\varPi$ and $\varSigma$ of order $n \in \mathbb{N}$ associated with permutations π and σ. Then $\varPi\varSigma$ is the permutation matrix of order n associated with the permutation $\sigma \circ \pi$.

Proof. We have $\varPi, \varSigma \in \mathbb{F}^{n \times n}$ by definition II.4.3.1 and therefore $\varPi\varSigma \in \mathbb{F}^{n \times n}$ by definition II.4.1.3. The composition $\tau = \sigma \circ \pi$ is the permutation given by $\tau_i = \sigma_{\pi_i}$ for each $i \in \{1, \ldots, n\}$, see § I.1.28.

By definition II.4.3.4, we have $\varPi_{ik} = \delta_{k\pi_i}$ and $\varSigma_{kj} = \delta_{j\sigma_k}$ for all $i, j, k \in \{1, \ldots, n\}$. Then

$$(\varPi\varSigma)_{ij} = \sum_{k=1}^{n} \varPi_{ik} \cdot \varSigma_{kj} = \sum_{k=1}^{n} \delta_{k\pi_i} \cdot \delta_{j\sigma_k} = \delta_{j\sigma_{\pi_i}} = \delta_{j\tau_i}$$

for all $i, j \in \{1, \ldots, n\}$. Applying proposition II.4.3.3 again, we find $\varPi\varSigma$ to be the permutation matrix associated with the permutation τ.

Notice the order reversal in proposition II.4.3.6: the product of the permutation matrices $\varPi$ and $\varSigma$, associated with the respective permutations π and σ, corresponds to the composition $\sigma \circ \pi$. This order reversal is due to the fact that, for a given permutation π of $\{1, \ldots, n\}$, the permutation matrices $\mathit{\Pi} = [\delta_{j\pi_i}]_{i=1,\,j=1}^{n\quad n} = [\delta_{i\pi_j^{-1}}]_{i=1,\,j=1}^{n\quad n}$ and $\mathit{\Pi}^\mathsf{T} = [\delta_{i\pi_j}]_{i=1,\,j=1}^{n\quad n}$ have historically been associated with π and π^{-1} (respectively), which are the permutations applied to the components of a column vector under the action of the matrices. For example, in proposition II.4.3.6, if instead we associated the same matrices $\mathit{\Pi} = [\delta_{j\pi_i}]_{i=1,\,j=1}^{n\quad n} = [\delta_{i\pi_j^{-1}}]_{i=1,\,j=1}^{n\quad n}$ and $\Sigma = [\delta_{j\sigma_i}]_{i=1,\,j=1}^{n\quad n} = [\delta_{i\sigma_j^{-1}}]_{i=1,\,j=1}^{n\quad n}$ with π^{-1} and σ^{-1}, then we would find $\mathit{\Pi}\mathit{\Sigma}$ to be associated with $\tau^{-1} = \pi^{-1} \circ \sigma^{-1}$.

Proposition II.4.3.7. Let $\mathit{\Pi}$ be a permutation matrix of order $n \in \mathbb{N}$. Then $\mathit{\Pi}^\mathsf{T}$ is a permutation matrix, matrix $\mathit{\Pi}$ is invertible and $\mathit{\Pi}^{-1} = \mathit{\Pi}^\mathsf{T}$. Further, the permutations π and σ realized by $\mathit{\Pi}$ and $\mathit{\Pi}^\mathsf{T}$ are mutually inverse.

Proof. By definition II.4.3.1, $\mathit{\Pi}^\mathsf{T}$ is a permutation matrix since such is $\mathit{\Pi}$: conditions (i) and (ii) swap under a swap of the row and column indices.

Consider $i, j \in \{1, \dots, n\}$. Then

$$(\mathit{\Pi}^\mathsf{T}\mathit{\Pi})_{ij} = \sum_{k=1}^{n} (\mathit{\Pi}^\mathsf{T})_{ik}\, \mathit{\Pi}_{kj} = \sum_{k=1}^{n} \mathit{\Pi}_{ki}\, \mathit{\Pi}_{kj} = \delta_{ij}\,.$$

To justify the last equality, we consider two cases. If $i \neq j$, then each term of the latter sum equals zero: by condition (i) of definition II.4.3.1, for any $k \in \{1, \dots, n\}$, row k contains $n-1$ zero entries and one unit entry, so at least one of $\mathit{\Pi}_{ki}$ and $\mathit{\Pi}_{kj}$ is zero. If $i = j$, by condition (ii) of definition II.4.3.1, column i contains $n-1$ zeros entries and one unit entry, which results in the value 1 for the sum. So we indeed have $\mathit{\Pi}^\mathsf{T}\mathit{\Pi} = I$.

Further, we have also $\mathit{\Pi}\mathit{\Pi}^\mathsf{T} = I$. That can be shown by an argument analogous to the one right above or, better, by applying the result for $\mathit{\Pi}^\mathsf{T}$ in place of $\mathit{\Pi}$ and by using the reflexivity of transposition (proposition II.3.3.2).

Finally, applying the definition of matrix inversion (definition II.4.1.14), we conclude that matrix $\mathit{\Pi}$ is invertible and $\mathit{\Pi}^{-1} = \mathit{\Pi}^\mathsf{T}$.

By proposition II.4.3.6, $\mathit{\Pi}^\mathsf{T}\mathit{\Pi}$ and $\mathit{\Pi}\mathit{\Pi}^\mathsf{T}$ are associated with the permutations $\sigma \circ \pi$ and $\pi \circ \sigma$. Since those products are both equal to the identity matrix of order n, both the compositions are identity permutations, so σ and π are indeed mutually inverse.

Remark II.4.3.8. Let us also note that the multiplication of a matrix by a permutation matrix on the left or on the right amounts to permuting the rows or to the columns of the matrix. Specifically, for any $m, n \in \mathbb{N}$, $A \in \mathbb{F}^{m\times n}$ and permutation matrices $\mathit{\Pi}$ and $\mathit{\Sigma}$ of orders m and n associated with permutations π and σ, we have

$$(\mathit{\Pi} A)_{ij} = \sum_{k=1}^{m} \delta_{k\pi_i} \cdot A_{kj} = A_{\pi_i\, j}$$

and

$$(A\mathit{\Sigma}^\mathsf{T})_{ij} = \sum_{k=1}^{n} A_{ik} \cdot \delta_{j\sigma_k^{-1}} = \sum_{k=1}^{n} A_{i\sigma_k} \cdot \delta_{j\sigma_{\sigma_k}^{-1}} = \sum_{k=1}^{n} A_{i\sigma_k} \cdot \delta_{jk} = A_{i\sigma_j}$$

for all $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. In the second chain of equalities, we use the fact that $\mathit{\Sigma}^\mathsf{T}$ is the permutation matrix associated with the permutation σ^{-1}, the inverse of the permutation σ (due to proposition II.4.3.7), and re-enumerate the terms using σ, which allows to cancel σ^{-1}.

Using permutation matrices, we can reduce a general matrix partition (see definition II.4.2.11) to a block partition (see remark II.4.2.12). This is based on that, in the context of definition II.4.2.9, the tuple

$$(\sigma_{11}, \ldots, \sigma_{1n_1}, \ldots, \ldots, \sigma_{q1}, \ldots, \sigma_{qn_q}),$$

obtained by concatenating the tuples $\sigma_1, \ldots, \sigma_q$, represents a permutation of $\{1, \ldots, n\}$.

Proposition II.4.3.9. In the context of definition II.4.2.11, let $\mathit{\Pi}$ and $\mathit{\Sigma}$ be the permutation matrices of orders m and n associated with the permutations

$$(\pi_{11}, \ldots, \pi_{1m_1}, \ldots, \ldots, \pi_{p1}, \ldots, \pi_{pm_p}) \quad \text{and} \quad (\sigma_{11}, \ldots, \sigma_{1n_1}, \ldots, \ldots, \sigma_{q1}, \ldots, \sigma_{qn_q}).$$

Then the submatrices $\widehat{A}_{ij}$ of A with $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$ are also the submatrices of $\widehat{A} = \mathit{\Pi} A \mathit{\Sigma}^\mathsf{T}$ and form a block partition of $\widetilde{A}$ in the sense of remark II.4.2.12:

$$\mathit{\Pi} A \mathit{\Sigma}^\mathsf{T} = \widehat{A} = \begin{bmatrix} \widehat{A}_{11} & \cdots & \widehat{A}_{1q} \\ \vdots & \ddots & \vdots \\ \widehat{A}_{p1} & \cdots & \widehat{A}_{pq} \end{bmatrix}. \tag{II.4.3.3}$$

In what follows, we will be most interested in *exchanges* (see § I.1.28), which can be considered elementary permutations. We now define the respective class of *exchange matrices*, which can be seen as elementary permutation matrices.

Definition II.4.3.10 (exchange matrix)**.** Let $n \in \mathbb{N}$ and $\pi = (\pi_1, \ldots, \pi_n)$ be the (k, ℓ)-exchange permutation of $\{1, \ldots, n\}$ for $k, \ell \in \{1, \ldots, n\}$. Then the permutation matrix associated with the permutation π is called the (k, ℓ)-exchange matrix of order n.

For any $n \in \mathbb{N}$ and $k, \ell \in \{1, \ldots, n\}$, the (k, ℓ)-exchange matrix $\mathit{\Pi}$ of order n is given by

$$\mathit{\Pi}_{ij} = \delta_{ij} - \delta_{ik}\delta_{jk} - \delta_{i\ell}\delta_{j\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} = \begin{cases} 1 & \text{if } i = k \text{ and } j = \ell, \\ 1 & \text{if } i = \ell \text{ and } j = k, \\ 0 & \text{if } i = j = k \neq \ell, \\ 0 & \text{if } i = j = \ell \neq k, \\ \delta_{ij} & \text{otherwise.} \end{cases} \tag{II.4.3.4}$$

Using the block-matrix notation (see remark II.4.2.12), we can express the same exchange matrix $\mathit{\Pi}$ as follows when $k < \ell$:

$$\mathit{\Pi} = \begin{bmatrix} I_1 & & & & \\ & 0 & & 1 & \\ & & I_2 & & \\ & 1 & & 0 & \\ & & & & I_3 \end{bmatrix}, \tag{II.4.3.5}$$

where I_1, I_2 and I_3 are the identity matrices of orders $k-1$, $\ell-k-1$ and $n-\ell$ (when $k-1=0$, $\ell-k-1=0$ or $n-\ell=0$, the respective block row and column vanish).

§ II.4.4. Matrix rank

Definition II.4.4.1 (matrix rank)**.** Let $m, n \in \mathbb{N}$, $A \in \mathbb{F}^{m\times n}$. The *matrix rank* of A, denoted by $\operatorname{rank} A$, is defined as the *least* number $r \in \mathbb{N}_0$ such that there exist $u_1, \dots, u_r \in \mathbb{F}^m$ and $v_1, \dots, v_r \in \mathbb{F}^n$ satisfying

$$A = \sum_{k=1}^{r} u_k v_k^\mathsf{T} . \tag{II.4.4.1}$$

We will use $\mathbb{F}_r^{m\times n}$ to denote the set of matrices from $\mathbb{F}^{m\times n}$ of rank *not exceeding* $r \in \mathbb{N}_0$:

$$\mathbb{F}_r^{m\times n} = \left\{A \in \mathbb{F}^{m\times n} \colon \operatorname{rank} A \le r\right\} = \left\{ \sum_{k=1}^{r} u_k v_k^\mathsf{T} \colon u_1, \dots, u_r \in \mathbb{F}^m,\ v_1, \dots, v_r \in \mathbb{F}^n \right\} . \tag{II.4.4.2}$$

Note that $\operatorname{rank} A = 0$ holds if and only if A is the zero matrix of $\mathbb{F}^{m\times n}$ (the set of all $k \in \mathbb{N}$ such that $1 \le k \le 0$ is empty, and sums over empty index sets are zero). So $\mathbb{F}_0^{m\times n}$ contains only one element, which is very important (see proposition II.3.1.7) but also so simple that we already know everything about it.

Any $u \in \mathbb{F}^m$ and $v \in \mathbb{F}^n$ with $m, n \in \mathbb{N}$ induce the matrix $uv^\mathsf{T} \in \mathbb{F}^{m\times n}$, which is called *the outer product* of u and v:

$$(uv^\mathsf{T})_{ij} = u_i \cdot v_j \quad \text{for all} \quad i \in \{1, \dots, m\} \quad \text{and} \quad j \in \{1, \dots, n\} . \tag{II.4.4.3}$$

Note that the rank of uv^T is either zero or one; to be precise, $\operatorname{rank} uv^\mathsf{T} = 0$ if any of u and v is zero and $\operatorname{rank} uv^\mathsf{T} = 1$ otherwise. The condition (II.4.4.3) is also said to express the separation of variables in $A = uv^\mathsf{T}$ in the sense that A is a bivariate function represented by (II.4.4.3) as a product of two factors, each depending only on the respective one of the two variables.

The decomposition (II.4.4.1), representing A as a sum of r outer products, is often referred to as *r-term separation of variables*. Indeed, the equality is equivalent to

$$A_{ij} = \sum_{k=1}^{r} (u_k)_i\, (v_k)_j \quad \text{for all} \quad i \in \{1, \dots, m\} \quad \text{and} \quad j \in \{1, \dots, n\} , \tag{II.4.4.4}$$

where the indices i and j (discrete variables) separate in each term: each of the two enters only in the corresponding factor.

Let us verify that the notion of matrix rank is introduced by definition II.4.4.1 correctly and, additionally, obtain a trivial but important bound on the rank of a matrix in terms of its size. The reason why we are concerned with correctness is that, in terms of the notation introduced in (II.4.4.2), we effectively define the rank of $A \in \mathbb{F}^{m\times n}$ as follows:

$$\operatorname{rank} A = \min\{r \in \mathbb{N}_0 \colon A \in \mathbb{F}_r^{m\times n}\} .$$

Since the empty set has no minimum, we need to check that the set of which we take the minimum ("the least") element is not empty.

Lemma II.4.4.2 (matrix rank is defined correctly)**.** For any $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$, $\operatorname{rank} A$ is defined by definition II.4.4.1 correctly and does not exceed n.

Proof. To prove that $\operatorname{rank} A$ is well defined, we need to show that (II.4.4.1) holds for some $u_1, \dots, u_r \in \mathbb{F}^m$ and $v_1, \dots, v_r \in \mathbb{F}^n$ with *at least one* $r \in \mathbb{N}_0$, which is to ensure that the set of candidates $r \in \mathbb{N}_0$, among which $\operatorname{rank} A$ is defined as *the least*, is not empty. We will now easily see that a decomposition of the form (II.4.4.1) is always possible with $r = n$.

Indeed, let $a_1, \dots, a_n$ and $e_1, \dots, e_n$ denote the columns of A and of the identity matrix of order n respectively. Then we trivially have

$$A = \sum_{k=1}^{n} a_k e_k^{\mathsf{T}} .$$

So $\operatorname{rank} A$ is defined by definition II.4.4.1 correctly and satisfies $\operatorname{rank} A \leq n$.

Definition II.4.4.3 (rank-r factorization of a matrix)**.** A representation of $A \in \mathbb{F}^{m \times n}$ with $m, n \in \mathbb{N}$ of the form $A = UV^{\mathsf{T}}$ with $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ for some $r \in \mathbb{N}$ is often called a *rank-r* factorization of A.

Note that the rank of A does not have to be *equal* to r in the context of definition II.4.4.3. In fact, the notions of rank and rank-r factorization are related as follows: the existence of a rank-r factorization of the matrix is equivalent to that its rank does not exceed r.

Proposition II.4.4.4 (rank-r factorization and matrix rank)**.** Let $m, n, r \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. Then $\operatorname{rank} A \leq r$ holds if and only if there exist matrices $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ such that $A = UV^{\mathsf{T}}$.

Proof. The proof follows from definition II.4.1.3.

Indeed, if $\operatorname{rank} A \leq r$, then the decomposition (II.4.4.2) holds with some column vectors $u_1, \dots, u_r \in \mathbb{F}^m$ and $v_1, \dots, v_r \in \mathbb{F}^n$. Composing the block matrices

$$U = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \in \mathbb{F}^{m \times r} \quad \text{and} \quad V = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix} \in \mathbb{F}^{n \times r} ,$$

see remark II.4.2.12, we obtain $A = UV^{\mathsf{T}}$. That follows from our general result on the multiplication of partitioned matrices, see proposition II.4.2.13 and remark II.4.2.14; alternatively, one can easily deduce the equality immediately from the definition of matrix multiplication (definition II.4.1.3) since the above block partitions readily translate into trivial expressions for the entries of U and V in terms of those of $u_1, \dots, u_r$ and $v_1, \dots, v_r$. In any case, the transposition is resolved using definition II.3.3.1.

On the other hand, if $A = UV^{\mathsf{T}}$ with $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$, we can denote the columns of U and V by $u_1, \dots, u_r \in \mathbb{F}^m$ and $v_1, \dots, v_r \in \mathbb{F}^n$. Then the decomposition (II.4.4.2) holds due to definition II.4.1.3.

Remark II.4.4.5 (rank-r factorization and matrix rank, again)**.** In the context of definition II.4.4.3, the representation $A = UV^{\mathsf{T}}$ is often referred to as a *rank-r factorization* of A regardless of whether $\operatorname{rank} A = r$ or $\operatorname{rank} A < r$.

Remark II.4.4.6 (nonuniqueness of rank-r factorizations)**.** There is much freedom in choosing or modifying a rank-r factorization; for example, for any $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ with $m, n, r \in \mathbb{N}$, we have $UV^{\mathsf{T}} = \tilde{U}\tilde{V}^{\mathsf{T}}$ with $\tilde{U} = US \in \mathbb{F}^{m \times r}$ and $\tilde{V} = VS^{-\mathsf{T}} \in \mathbb{F}^{n \times r}$ for any invertible matrix $S \in \mathbb{F}^{r \times r}$.

Note that remark II.4.4.6 equivalently applies to the decompositions of the form (II.4.4.1) with $r \in \mathbb{N}$, but expressing the result in terms of rows and columns would require dealing with inverse matrices *entrywise*, which is less pleasant and completely unnecessary.

Proposition II.4.4.7 (matrix rank under transposition)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. Then $\operatorname{rank} A^\mathsf{T} = \operatorname{rank} A$.

Proof. When A is a zero matrix, the statement is trivial. Otherwise, the claim follows immediately from proposition II.4.4.4. Indeed, $(UV^\mathsf{T})^\mathsf{T} = VU^\mathsf{T}$ for any $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ with any $r \in \mathbb{N}$. This means that $\operatorname{rank} A \leq r$ implies $\operatorname{rank} A^\mathsf{T} \leq r$ for any $r \in \mathbb{N}$ and that $\operatorname{rank} A^\mathsf{T} \leq r$ implies $\operatorname{rank} A = \operatorname{rank}(A^\mathsf{T})^\mathsf{T} \leq r$ for any $r \in \mathbb{N}$.

Corollary II.4.4.8 (matrix rank does not exceed any of the sizes)**.** Let $m, n \in \mathbb{N}$. Then $\operatorname{rank} A \leq \min\{m, n\}$ for any $A \in \mathbb{F}^{m \times n}$.

Proof. The statement is an immediate corollary of propositions II.3.3.2, II.4.4.4 and II.4.4.7

Proposition II.4.4.9 (rank of a submatrix)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. Then the rank of any submatrix of A does not exceed the rank of A.

Proof. If $\widehat{A}$ is a submatrix of size $p \times q$ of A, then $p \in \{1, \ldots, m\}$, $q \in \{1, \ldots, n\}$ and there exist distinct row indices $\pi_1, \ldots, \pi_p \in \{1, \ldots, m\}$ and distinct column indices $\sigma_1, \ldots, \sigma_q \in \{1, \ldots, n\}$ such that $\widehat{A}_{ij} = A_{\pi_i \sigma_j}$ for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$.

Let us set $r = \operatorname{rank} A$. By proposition II.4.4.4, there exist matrices $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ such that $A = UV^\mathsf{T}$. In particular, we have then $A_{\pi_i \sigma_j} = \sum_{k=1}^r U_{\pi_i k} V_{\sigma_j k}$ for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$. This yields $\widehat{A} = \widehat{U} \widehat{V}^\mathsf{T}$ with

$$\widehat{U} = [U_{\pi_i k}]_{i=1,\, k=1}^{p \quad r} \in \mathbb{F}^{p \times r} \quad \text{and} \quad \widehat{V} = [V_{\sigma_j k}]_{j=1,\, k=1}^{q \quad r} \in \mathbb{F}^{p \times r} ,$$

so that $\operatorname{rank} \widehat{A} \leq r$ by proposition II.4.4.4.

Proposition II.4.4.10 (subadditivity of matrix rank)**.** For $m, n \in \mathbb{N}$, consider $A \in \mathbb{F}^{m \times n}$ and $B \in \mathbb{F}^{m \times n}$. Then $\operatorname{rank}(A + B) \leq \operatorname{rank} A + \operatorname{rank} B$.

Proof. The statement follows trivially in the case when A or B is zero. So let us assume for the remainder of the proof that both A and B are nonzero. By proposition II.4.4.4, there exist matrices $U \in \mathbb{F}^{m \times p}$, $V \in \mathbb{F}^{n \times p}$, $X \in \mathbb{F}^{m \times q}$ and $Y \in \mathbb{F}^{n \times q}$, where $p = \operatorname{rank} A$ and $q = \operatorname{rank} B$, such that $A = UV^\mathsf{T}$ and $B = XY^\mathsf{T}$. Then

$$A + B = UV^\mathsf{T} + XY^\mathsf{T} = \begin{bmatrix} U & X \end{bmatrix} \begin{bmatrix} V & Y \end{bmatrix}^\mathsf{T} ,$$

and we therefore have $\operatorname{rank}(A + B) \leq p + q$ by proposition II.4.4.4.

Proposition II.4.4.11 (matrix rank under multiplication)**.** Let $m, n, k \in \mathbb{N}$. Consider $A \in \mathbb{F}^{m\times k}$ and $B \in \mathbb{F}^{k\times n}$. Then $\operatorname{rank} AB \leq \min\{p, q\}$.

Proof. Let $p = \operatorname{rank} A$ and $q = \operatorname{rank} B$. The statement follows trivially in the case when any of A and B is the respective zero matrix. So let us assume for the remainder of the proof that both A and B are nonzero. By proposition II.4.4.4, there exist matrices $U \in \mathbb{F}^{m\times p}$, $V \in \mathbb{F}^{k\times p}$, $X \in \mathbb{F}^{k\times q}$ and $Y \in \mathbb{F}^{n\times q}$ such that $A = UV^\mathsf{T}$ and $B = XY^\mathsf{T}$. These representations lead to

$$AB = UV^\mathsf{T}XY^\mathsf{T} = U(YX^\mathsf{T}V)^\mathsf{T} \quad \text{and} \quad AB = UV^\mathsf{T}XY^\mathsf{T} = (UV^\mathsf{T}X)Y^\mathsf{T}\,,$$

and so $\operatorname{rank} AB \leq p$ and $\operatorname{rank} AB \leq q$ by proposition II.4.4.4.

The following is a trivial corollary of definition II.4.4.1, propositions II.4.4.4 and II.4.4.11 and corollary II.4.4.8, citing which will be very convenient in the sequel.

Corollary II.4.4.12 (matrix rank and minimal factorization)**.** For $m, n \in \mathbb{N}$, consider $A \in \mathbb{F}^{m\times n}$ nonzero and let $r = \operatorname{rank} A$. Then $r \in \mathbb{N}$, $r \leq \min\{m, n\}$ and there exist matrices $U \in \mathbb{F}^{m\times r}$ and $V \in \mathbb{F}^{n\times r}$ such that $\operatorname{rank} U = r = \operatorname{rank} V$ and $A = UV^\mathsf{T}$.

Note that the selection of a submatrix in a matrix can be represented as the multiplication of the matrix by suitable matrices on the left and on the right. So we could have obtained proposition II.4.4.9 as a corollary of proposition II.4.4.11. The proof of proposition II.4.4.9 given above is, however, more intuitive than this argument.

Corollary II.4.4.13 (matrix rank under multiplication by invertible matrices)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Then $\operatorname{rank} PAQ^{-1} = \operatorname{rank} A$ for any invertible matrices $P \in \mathbb{F}^{m\times m}$ and $Q \in \mathbb{F}^{n\times n}$.

Proof. By corollary II.4.4.8, we have $\operatorname{rank} P \leq m$, $\operatorname{rank} P^{-1} \leq m$, $\operatorname{rank} Q \leq n$, $\operatorname{rank} Q^{-1} \leq n$, $\operatorname{rank} A \leq \min\{m, n\}$ and $\operatorname{rank} PAQ^{-1} \leq \min\{m, n\}$. Applying proposition II.4.4.11 twice, we obtain

$$\operatorname{rank} PAQ^{-1} \leq \operatorname{rank} PA \leq \operatorname{rank} A\,.$$

On the other hand, in the very same way, we arrive at

$$\operatorname{rank} A = \operatorname{rank} P^{-1}PAQ^{-1}Q \leq \operatorname{rank} P^{-1}PAQ^{-1} \leq \operatorname{rank} PAQ^{-1}\,.$$

Remark II.4.4.14 (matrix invertibility and rank)**.** Let $n \in \mathbb{N}$. Combining definition II.4.1.14 and proposition II.4.4.11, we immediately obtain for any invertible matrix $A \in \mathbb{F}^{n\times n}$ that $\operatorname{rank} A \geq \operatorname{rank} I$ and $\operatorname{rank} A^{-1} \geq \operatorname{rank} I$, where I is the identity matrix of order n.

We can only notice at this point that it *does not seem possible* to obtain a representation of the form (II.4.4.1) with less than n terms for I. At the end of chapter II, we will have developed a universal technique allowing to reveal the rank of any rectangular matrix by constructing a representation of the form (II.4.4.1) with the minimal number of terms. In particular, it will follow from lemma II.5.9.2 that $\operatorname{rank} A = \operatorname{rank} A^{-1} = \operatorname{rank} I = n$.

Example II.4.4.15 (rank-one matrix)**.** Let $u = [1, 2, 3]^\mathsf{T} \in \mathbb{C}^3$ and

$$v = \big[2, \frac{1}{3}, -1, 7, 3\mathrm{i}, \pi, \sqrt{2}, e, e^{\mathrm{i}\pi}\big]^\mathsf{T} \in \mathbb{C}^9 .$$

Then

$$uv^\mathsf{T} = \begin{bmatrix} 2 & \frac{1}{3} & -1 & 7 & 3\mathrm{i} & \pi & \sqrt{2} & e & -1 \\ 4 & \frac{2}{3} & -2 & 14 & 6\mathrm{i} & 2\pi & 2\sqrt{2} & 2e & -2 \\ 6 & 1 & -3 & 21 & 9\mathrm{i} & 3\pi & 3\sqrt{2} & 3e & -3 \end{bmatrix} \in \mathbb{C}^{3\times 9}$$

is a rank-one matrix. Note that it is defined in terms of two column vectors and can therefore be parametrized by the $3+9 = 12$ entries of these column vectors instead of its own $3\cdot 9 = 27$ entries. This is an example of *low-parametric representation*: an element of $\mathbb{C}^{3\times 9}_1 \subset \mathbb{C}^{3\times 9}$ is represented in terms of 12 parameters from $\mathbb{C}$, whereas representing an arbitrary element of $\mathbb{C}^{3\times 9}$ requires 27 parameters. Thus far, our discussion has been focused on the cost of *representation* (or *storage*).

Exploiting the rank-one representation can also be beneficial in terms of *computation* costs. For example, consider the multiplication of uv^T by a column vector $w \in \mathbb{C}^9$. We can naively use uv^T in the entrywise representation and compute $(uv^\mathsf{T}) \cdot w$ in $N_\times = 3 \cdot 9 = 27$ multiplications over $\mathbb{C}$ and $N_+ = 3\cdot(9-1) = 24$ additions over $\mathbb{C}$ (assuming that the product uv^T has been evaluated or that this product is given and the factorization is not known). Alternatively, we can use the factorized representation to first evaluate $v^\mathsf{T} w$ using $1 \cdot 9 \cdot 1$ multiplications and $1 \cdot (9-1) \cdot 1 = 8$ additions over $\mathbb{C}$ and to then evaluate $u \cdot (v^\mathsf{T} w)$ using $3 \cdot 1 \cdot 1$ of multiplications over $\mathbb{C}$, which would bring us to the total numbers $N_\times = 12$ and $N_+ = 11$ multiplications and additions over $\mathbb{C}$.

Example II.4.4.16 (low-rank matrices and complexity)**.** For any $U \in \mathbb{F}^{m\times r}$ and $V \in \mathbb{F}^{n\times r}$ with $m, n, r \in \mathbb{N}$, the product $A = UV^\mathsf{T}$ can be stored as a pair of factors instead of being stored as a matrix of size $m \times n$. The storage cost then becomes $(m + n)r$ instead of mn, which may feature a significant reduction.

On the other hand, access cost per entry rises from a single (scalar) memory-access operation to $2r$ such operations followed by r scalar multiplications and $r-1$ scalar additions.

Finally, we note that the cost of computing the product $U(V^\mathsf{T} w)$ with a column vector $w \in \mathbb{F}^n$ in two steps, in the order indicated by the parenthesis, is $N_\times = r \cdot n \cdot 1 + m \cdot r \cdot 1 = (m+n)r$ scalar multiplications and $N_+ = r \cdot (n-1) \cdot 1 + m \cdot (r-1) \cdot 1 = (m+n)r - r - m$ scalar additions instead of the $N_\times = mn$ scalar multiplications and $N_+ = m(n-1)$ scalar additions involved in the evaluation of the product Aw according to definition II.4.1.1, with A represented entrywise.

Example II.4.4.17 (low rank vs. simple formulae)**.** For $m, n \in \mathbb{N}$, let $u = [1]_{i=1}^m \in \mathbb{F}^m$ and $v = [1]_{i=1}^n \in \mathbb{F}^n$. Then $A = uv^\mathsf{T} = [1]_{i=1,\,j=1}^{m\ \ \ n} \in \mathbb{F}^{m\times n}$, the matrix of ones of size $m \times n$, is a rank-one matrix. It can be represented as an element of $\mathbb{F}^{m\times n}$ (a full matrix with its mn entries as parameters), as an element of $\mathbb{F}^{m\times n}_1$ (a matrix of rank at most one and the $m+n$ entries of u and v as parameters) or as an element of $\{[\alpha]_{i=1,\,j=1}^{m\ \ \ n} \colon \alpha \in \mathbb{F}\}$ (a matrix with all entries equal, in terms of the value of all entries as a single parameter). For $m = n = 10^9$, these complexities are, respectively, 10^{18}, $2 \cdot 10^9$ and 1.

> **Remark II.4.4.18** (classes of data: the tradeoff between complexity and expressive power)**.** Matrices, just as column vectors, are data. As such, they need to be stored and manipulated as elements of a certain problem-dependent class (set) to which they, as data, are sensibly expected to belong. When we are *interested in all matrices* of size $m \times n$, all mn parameters are necessary and sufficient to represent an arbitrary *matrix of interest*. However, this class may be extravagantly broad for the nature of the data. On the other hand, the set $\big\{[\alpha]_{i=1,\,j=1}^{m\quad n} \colon \alpha \in \mathbb{F}\big\}$, which is a one-parametric family of matrices for fixed $m, n \in \mathbb{N}$, is rather useless: such matrices are essentially scalars ($\alpha \in \mathbb{F}$) and should not be considered, stored or even thought of as matrices.
>
> Finding a suitable intermediate class of data (with intermediate complexity) is an important goal for the mathematical study of many applied problems. Such classes are intended to realize a suitable tradeoff between accuracy and complexity, allowing thereby to reclaim the intrinsic structure of the data involved in the problem and to exploit this structure for efficient storage and computation. The sets $\mathbb{F}_r^{m\times n}$ with $m, n, r \in \mathbb{N}$ such that r is much less than $\min\{m, n\}$ are important examples of such classes. As we will establish in chapter III, the notion of matrix rank is immediately related to the notion of *dimension*, and having low-rank structure in a matrix allows to efficiently operate on a low-dimensional *subspace* of a vast *space*, which is vital when the dimension of the latter is prohibitively large for computational purposes.

§ II.5. Gaussian elimination and LU decomposition

§ II.5.1. Introduction

Matrices are important for the present course as a tool for the representation and analysis of linear relations between vectors. In § II.4.1, we focused on the representation part: we considered a vector $b \in \mathbb{F}^m$ defined in terms of a vector $x \in \mathbb{F}^n$ using a linear relation expressed in the equivalent forms (II.4.1.1) and (II.4.1.2). Using matrix multiplication, we recast this dependence in the form of (II.4.1.3).

In the present section, we will make the most straightforward attempt to analyze the same linear relation. Specifically, we are now interested in the opposite problem: we would like, assuming that $b \in \mathbb{F}^m$ is known, to understand what can be said about $x \in \mathbb{F}^n$ if it satisfies the same linear relation as before. In particular, we will try to understand whether this relation defines a unique such x. This must be well known to the reader as solving the system of linear equations expressed equivalently by (II.4.1.1) to (II.4.1.3) with respect to unknown $x \in \mathbb{F}^n$. Matrix A is called the matrix of this system, and column vector b is called the *right-hand side*. So we consider the following problem.

$$\text{Given } A \in \mathbb{F}^{m\times n} \text{ and } b \in \mathbb{F}^m\text{, find } \mathcal{X} = \big\{x \in \mathbb{F}^n \colon Ax = b\big\}\ . \tag{II.5.1.1}$$

The set $\mathcal{X} \subseteq \mathbb{F}^n$ is called the *set of solutions* of (II.4.1.3) considered as a linear system.

§ II.5.2. Motivation: the ABS system

At one of the lectures, we considered the following example of a linear system over $\mathbb{F} = \mathbb{R}$:

$$\begin{cases} x_1 = 2x_2 \\ x_2 = \dfrac{x_1}{2} \end{cases} \Leftrightarrow \left\{ \begin{aligned} 1 \cdot x_1 + (-2) \cdot x_2 &= \ 0 \\ \tfrac{1}{2} \cdot x_1 + (-1) \cdot x_2 &= \ 0 \end{aligned} \right. \Leftrightarrow x_1 = 2x_2 \tag{II.5.2.ABS}$$

It may seem unfortunate that we could not do more than realize that the second equation is useless and finding (x_1, x_2) as *the solution* is impossible. Indeed, the set of solutions of

(II.5.2.ABS) reads

$$\mathcal{X} = \left\{ \begin{bmatrix} 2\alpha & \alpha \end{bmatrix}^{\mathsf{T}} : \alpha \in \mathbb{R} \right\} \subset \mathbb{R}^2 \,, \tag{II.5.2.1}$$

where, for notational convenience, we use the identification of $\mathbb{R}^{2\times 1}$ with $\mathbb{R}^2$ (see remark II.4.1.4). There is an apparent bijection between $\mathcal{X}$ and $\mathbb{R}$, and so $\mathcal{X}$ is an uncountable set (since $\mathbb{R}$ is uncountable, see § I.1.23). Without additional conditions (which would change the problem) we have no reason to prefer any element of $\mathcal{X}$ to the others.

Let us recast the solution process expressed by (II.5.2.ABS) in terms of matrices and column vectors $x = [x_1 \ \ x_2]^{\mathsf{T}}$ and $b = 0 = [0 \ \ 0]^{\mathsf{T}}$:

$$\begin{bmatrix} 1 & -2 \\ \frac{1}{2} & -1 \end{bmatrix} x = 0 \Leftrightarrow \begin{bmatrix} 1 & -2 \\ 0 & 0 \end{bmatrix} x = 0 \Leftrightarrow \begin{bmatrix} 0 & 0 \\ \frac{1}{2} & -1 \end{bmatrix} x = 0 \,, \tag{II.5.2.ABS-M}$$

where we can express the fact that the two equations are equivalent by replacing one of them with a trivial equation (the one represented by a zero row in the matrix and a zero component in the right-hand side). We can also omit one of the equations in the above matrix notation by removing the corresponding rows; for example,

$$\begin{bmatrix} 1 & -2 \\ \frac{1}{2} & -1 \end{bmatrix} x = 0 \Leftrightarrow \begin{bmatrix} 1 & -2 \end{bmatrix} x = 0 \,. \tag{II.5.2.ABS-MM}$$

Clearly, the set $\mathcal{X}$ of solutions of (II.4.1.3) is determined by the *data of the problem*, which are the matrix and the right-hand side (see in (II.5.1.1) what is “given”).

Remark II.5.2.1 (all possible scenarios for systems with two equations and two unknowns)**.**

- **(a)** For A equal to any of the matrices involved in (II.5.2.ABS-M) and $b = [0 \ \ 0]^{\mathsf{T}} \in \mathbb{R}^2$, we have already obtained (II.5.2.1), i.e., (II.4.1.3) has a one-parametric family of solutions.
- **(b)** For A equal to any of the matrices involved in (II.5.2.ABS-M) and $b = [0 \ \ 1]^{\mathsf{T}} \in \mathbb{R}^2$, we would immediately find that
$$\mathcal{X} = \varnothing \,, \tag{II.5.2.2}$$
i.e., (II.4.1.3) would have *no* solution. The reason is that (in a sense to be made precise later) the right-hand side we consider here is *inconsistent* with the matrix, whereas the original one was *consistent*.
- **(c)** For $A = \operatorname{diag}(0,0)$ and $b = [0 \ \ 0]^{\mathsf{T}} \in \mathbb{R}^2$, we would arrive at
$$\mathcal{X} = \mathbb{R}^2 \,, \tag{II.5.2.3}$$
i.e., (II.4.1.3) *any* column vector with two components would be a solution.
- **(d)** For $A = \operatorname{diag}(1,1)$ and any $b \in \mathbb{R}^2$, we would obtain that
$$\mathcal{X} = \{b\} \,, \tag{II.5.2.4}$$
i.e., (II.4.1.3) would have a *unique* solution independently of b, and this unique solution would be completely determined by b.

The example of (II.5.2.ABS-M) and remark II.5.2.1 show that the structure of the solution set depends on the structure of the matrix and on its interaction with the right-hand side, so we need to study matrices in order to analyze (II.5.1.1).

§ II.5.3. Solving linear systems with square, tall and wide matrices

In general, for the problem (II.5.1.1), which involves $m \in \mathbb{N}$ scalar equations and $n \in \mathbb{N}$ scalar unknowns, one may rightfully wonder, even before trying to solve the system, if the set of solutions is empty, is a single point, is the whole of $\mathbb{F}^n$ or is a subset thereof with a number of parameters that can be chosen independently from each other.

As (II.5.2.ABS-MM) shows, a linear system with equal numbers of scalar equations and scalar unknowns may be apparently redundant and therefore admit equivalent representation with a smaller number of equations. Similarly, if all coefficients in front of one scalar unknown are zero, then this unknown should be allowed to take any value but, at the same time, can be excluded from the system with the effect of reducing the number of scalar unknowns.

Let us consider the problem (II.5.1.1) with a general matrix (square, tall or wide; see part (a) of definition II.3.2.1). At this point, using the notions, notations and results from §§ II.3 and II.4, we can completely solve a linear system with such a matrix assuming that its leading principal submatrix of maximum size is invertible.

Lemma II.5.3.1 (solving linear systems with "good" leading submatrices)**.**

(a) Assume that $n \in \mathbb{N}$, $A \in \mathbb{F}^{n\times n}$ is an invertible matrix and $x, y \in \mathbb{F}^n$. Then $Ax = y$ if and only if $x = A^{-1}y$.

(b) Assume that $m, r \in \mathbb{N}$ are such that $r < m$, $z \in \mathbb{F}^r$ and

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

with an invertible matrix $U_1 \in \mathbb{F}^{r\times r}$ and arbitrary $U_2 \in \mathbb{F}^{(m-r)\times r}$, $y_1 \in \mathbb{F}^r$ and $y_2 \in \mathbb{F}^{m-r}$. Then $Uz = y$ if and only if $z = U_1^{-1}y_1$ and $U_2\,U_1^{-1}y_1 = y_2$.

(c) Assume that $n, r \in \mathbb{N}$ are such that $r < n$, $z \in \mathbb{F}^r$ and

$$W = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

with an invertible matrix $W_1 \in \mathbb{F}^{r\times r}$ and arbitrary $W_2 \in \mathbb{F}^{r\times(n-r)}$, $x_1 \in \mathbb{F}^r$ and $x_2 \in \mathbb{F}^{n-r}$. Then $Wx = z$ if and only if $x_1 = W_1^{-1}z - W_1^{-1}W_2x_2$.

Proof. The proof of part (a) is trivial. Indeed, $Ax = y$ implies $A^{-1}Ax = A^{-1}y$, which leads to $x = A^{-1}y$ by definition II.4.1.14 and proposition II.4.1.10. On the other hand, $x = A^{-1}y$ implies $Ax = AA^{-1}y$, which leads to $Ax = y$ by definition II.4.1.14 and proposition II.4.1.10.

For part (b), we split the system into two by considering separately the first r rows and all the remaining rows (using proposition II.4.2.13, see also remark II.4.2.14). Together with part (a), this yields the claim as follows.

$$Uz = y \Leftrightarrow \begin{cases} U_1 z = y_1 \\ U_2 z = y_2 \end{cases} \Leftrightarrow \begin{cases} z = U_1^{-1}y_1 \\ U_2 z = y_2 \end{cases} \Leftrightarrow \begin{cases} z = U_1^{-1}y_1 \\ U_2\,U_1^{-1}y_1 = y_2 \end{cases}$$

To prove part (c), we note that, according to proposition II.4.2.13 (see also remark II.4.2.14), $Wx = z$ if and only if $W_1\,x_1 + W_2\,x_2 = z$, i.e., if and only if $W_1\,x_1 = y - W_2\,x_2$. By part (a), this is the case if and only if $x_1 = W_1^{-1}y - W_1^{-1}W_2\,x_2$.

The assumption that the leading principal submatrix of order r is invertible is related to that, in a sense, the first r equations are not redundant and the first r unknowns are also not redundant. We will give a precise meaning to such statements in this § II.5.

As we observed in § II.5.2, we may encounter drastically different sets $\mathcal{X}$ of solutions even for linear systems with equal numbers of scalar equations and scalar unknowns. Matrices that are not *full rank* ($\operatorname{rank} A < \min\{m, n\}$ for $A \in \mathbb{F}^{m\times n}$) are called *rank-deficient*. Each such a matrix can be represented as a product of a tall matrix and a wide matrix (or, to put it better in the present context, of a narrow matrix and of a low matrix) using proposition II.4.4.4. Let us now use lemma II.5.3.1 to obtain a characterization of the set of solutions to a linear system with a rank-deficient matrix.

Lemma II.5.3.2 (solving linear systems with “rank-deficient” matrices)**.** Let $m, n, r \in \mathbb{N}$ be such that $r < \min\{m, n\}$. Consider

$$U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

with invertible matrices $U_1, W_1 \in \mathbb{F}^{r\times r}$ and arbitrary $U_2 \in \mathbb{F}^{(m-r)\times r}$, $W_2 \in \mathbb{F}^{r\times(n-r)}$, $x_1 \in \mathbb{F}^r$, $x_2 \in \mathbb{F}^{n-r}$, $y_1 \in \mathbb{F}^r$ and $y_2 \in \mathbb{F}^{m-r}$. Then $UWx = y$ if and only if $x_1 = W_1^{-1}U_1^{-1}y_1 - W_1^{-1}W_2\,x_2$ and $U_2\,U_1^{-1}y_1 = y_2$.

Example II.5.3.3. Recall remark II.5.2.1, presenting several examples with the rank-one matrix

$$A = \begin{bmatrix} 1 & -2 \\ \frac{1}{2} & -1 \end{bmatrix} = UW\,, \quad \text{where} \quad U = \begin{bmatrix} 1 \\ \frac{1}{2} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} 1 & -2 \end{bmatrix}.$$

Applying lemma II.5.3.2 with $\mathbb{F} = \mathbb{R}$, $r = 1$, the above U, V and the respective right-hand sides, we obtain the conclusions of parts (a) and (b) of remark II.5.2.1. The difference between those two cases stems from the fact that the *consistency condition*, which takes the form of $\frac{1}{2}\cdot 1^{-1}\cdot b_1 = b_2$, i.e., $b_1 = 2b_2$, is satisfied by $b = (0, 0)$ and is violated by $b = (0, 1)$. The *solution formula* $x_1 = W_1^{-1}U_1^{-1}y_1 - W_1^{-1}W_2\,x_2$ evaluates to $x_1 = 1^{-1}\cdot 1^{-1}\cdot b_1 - 1^{-1}\cdot(-2)\cdot x_2 = b_1 + 2x_2$, where $x_2 \in \mathbb{R}^1$ is a free parameter, the only entry of which is denoted by α in (II.5.2.1). The solution set is therefore

$$\mathcal{X} = \left\{\begin{bmatrix} b_1 + 2\alpha & \alpha \end{bmatrix}^{\mathsf{T}} : \alpha \in \mathbb{R}\right\} \subset \mathbb{R}^2$$

provided that $b_1 = 2b_2$ (the consistency condition) holds. As in (II.5.2.1), we use here the identification of $\mathbb{R}^{2\times 1}$ with $\mathbb{R}^2$ (see remark II.4.1.4) for notational convenience. For $b = (0, 0)$, this solution set coincides with the one given in (II.5.2.1).

Proof of lemma II.5.3.2. The proof follows from lemma II.5.3.1. Indeed, by part (b), we have $UWx = y$ if and only if $Wx = U_1^{-1}y_1$ and $U_2\,U_1^{-1}y_1 = y_2$. By part (c), the first condition is equivalent to $x_1 = W_1^{-1}U_1^{-1}y_1 - W_1^{-1}W_2\,x_2$.

The remainder of chapter II is dedicated to developing the *pivoted LU decomposition* of matrices, which is particular but quite standard a version of Gaussian elimination, as a *universal* technique for exploring the structure of any matrix A and for finding the set $\mathcal{X}$ that solves (II.5.1.1) with an arbitrary right-hand side b. We will extensively use in that pursuit the notions, notations and results from §§ II.3 and II.4.

As lemma II.5.3.2 hints, this task is closely related to the notions of rank and rank-r factorization, introduced in § II.4.4. In fact, these notions will turn out to be the protagonists of our entire storyline at the end of chapter II. The technique of the *pivoted LU decomposition*, which we will start developing as a way of solving (II.5.1.1), will, in fact, produce a minimal low-rank factorization of the matrix. On the other hand, we will understand along the way how any such a factorization allows to solve (II.5.1.1) even if the invertibility assumptions of lemma II.5.3.2 do not hold.

Finally, the technique of *pivoted LU decomposition*, which we will develop, is *constructive*, which means that our exposition will allow immediate translation into *some* successful computational algorithms. The remainder of the present course (including the second semester), the diversity of possible applications and many important details being left aside, *may* be interpreted as the exposition and analysis of other successful computational algorithms for exploring the structure of matrices.

§ II.5.4. Gaussian elimination and LU decomposition: a detailed example

Let us revisit matrices A, L and U from example II.3.2.3. As we noted in example II.4.1.6, they satisfy the relation $A = LU$, i.e. A can be represented in a factorized form. We will now see how such a factorization can be obtained without knowing the factors beforehand.

We will follow the course of § II.5.2 and consider the problem (II.5.1.1) with A as given in example II.3.2.3, $b = [8\ 23\ 64]^{\mathsf{T}} \in \mathbb{R}^3$ and the matrix $B = [A\,|\,b] \in \mathbb{R}^{3\times 4}$. Then, by definition II.4.1.3, we have

$$Ax = b \Leftrightarrow B \begin{bmatrix} x \\ -1 \end{bmatrix} = 0$$

for all $x \in \mathbb{R}^3$. This equivalence can be seen as a motivation for composing A and b (all the data of the problem) into a single matrix, namely B.

One of the most straightforward approaches for finding $\mathcal{X}$, which the reader might know from high school, is based on choosing one equation and one unknown and subtracting the chosen equation from all other equations with suitable coefficients. The suitability of the coefficients is understood with respect to the purpose of this manipulation, which is to ensure that the modified equations *no longer involve* the chosen unknown. This first step can be repeated for the reduced system, which is obtained by excluding the chosen equation and therefore does not involve the chosen unknown.

For the present example, we will apply this procedure to the three equations encapsulated by $Ax = b$, expressing all linear systems in matrix form, as we did in (II.5.2.ABS-M). We denote the original matrix by B_0 (so that $B_0 = B$) and the transformed matrices by B_1 and B_2 (two transformation steps happen to be sufficient in this example). We will use vertical lines to remind ourselves that the last columns represent right-hand side vectors.

$$\left\{\begin{aligned} \textcircled{1}\cdot x_1 + {}\ 2\cdot x_2 + {}\ 3\cdot x_3 &= \ 8 \\ \mathbf{2}\ \cdot x_1 + {}\ 7\cdot x_2 + {}\ 7\cdot x_3 &= 23 \\ \mathbf{6}\ \cdot x_1 + 18\cdot x_2 + 22\cdot x_3 &= 64 \end{aligned}\right. \qquad \sim B_0 = \left[\begin{array}{ccc|c} \textcircled{1} & 2 & 3 & 8 \\ \mathbf{2} & 7 & 7 & 23 \\ \mathbf{6} & 18 & 22 & 64 \end{array}\right] \tag{II.5.4.1}$$

We start by choosing the first equation and the first unknown. Notice that the corresponding coefficient is circled in (II.5.4.1). Then we subtract the first equation from the second and third with coefficients $\ell_{21} = {}^{\mathbf{2}}\!/\textcircled{1} = 2$ and $\ell_{31} = {}^{\mathbf{6}}\!/\textcircled{1} = 6$.

$$\left\{\begin{aligned} 1\cdot x_1 + \ 2\ \cdot x_2 + 3\cdot x_3 &= \ 8\\ \textcircled{3}\cdot x_2 + 1\cdot x_3 &= \ 7\\ \mathbf{6}\ \cdot x_2 + 4\cdot x_3 &= 16\end{aligned}\right. \qquad \sim B_1 = \left[\begin{array}{ccc|c} 1 & 2 & 3 & 8\\ & \textcircled{3} & 1 & 7\\ & \mathbf{6} & 4 & 16\end{array}\right] \tag{II.5.4.2}$$

The result of this transformation is that the second and third equations in (II.5.4.2) no longer involve x_1. In terms of matrices, this corresponds to transforming B_0 into B_1, eliminating the entries that were shown in boldface in (II.5.4.1). The first equation can be put aside for a while, and so the problem is reduced to one with two equations and two unknowns.

At the second step, we choose the second equation and the second unknown. Notice that the corresponding entry is circled in (II.5.4.2). Then we subtract the second equation from the third with coefficient $\ell_{32} = {}^{\mathbf{6}}\!/\textcircled{3} = 2$.

$$\left\{\begin{aligned} 1\cdot x_1 + 2\cdot x_2 + \ 3\ \cdot x_3 &= 8\\ 3\cdot x_2 + \ 1\ \cdot x_3 &= 7\\ \textcircled{2}\cdot x_3 &= 2\end{aligned}\right. \qquad \sim B_2 = \left[\begin{array}{ccc|c} 1 & 2 & 3 & 8\\ & 3 & 1 & 7\\ & & \textcircled{2} & 2\end{array}\right] \tag{II.5.4.3}$$

The result of this transformation is that the second and third equations in (II.5.4.3) still do not involve x_1 and the third equation no longer involves x_2. In terms of matrices, this means that the corresponding matrix B_2 inherits all zeros below the diagonal in the first column from B_1 and, additionally, is zero below the diagonal in the second column.

Transformations (II.5.4.1) to (II.5.4.3) show that all intermediate data of the entire procedure can be conveniently expressed in terms of matrices. Indeed, we did not transform the unknowns in any way, so it is preferable to avoid writing them and summation signs all over. The equivalence of the matrix notation for linear systems is afforded by definition II.4.1.1 (see also the motivation that precedes it).

In fact, the transformations from B_0 to B_1 and from B_1 to B_2 can also be expressed in terms of matrix multiplication, which is thanks to definition II.4.1.3. Since the transformations amount to subtracting (with certain coefficients) the first and the second rows of B_0 and B_1 from all the rows underneath to obtain B_1 and B_2, they can be represented as multiplication by suitable matrices on the left. Specifically, we have

$$B_1 = M_1 B_0 = M_1 B \quad \text{and} \quad B_2 = M_2 B_1 = M_2 M_1 B_0 = M_2 M_1 B \tag{II.5.4.4}$$

with

$$M_1 = \begin{bmatrix} 1 & & \\ -\ell_{21} & 1 & \\ -\ell_{31} & & 1\end{bmatrix} = \begin{bmatrix} 1 & & \\ -2 & 1 & \\ -6 & & 1\end{bmatrix} \quad \text{and} \quad M_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & -\ell_{32} & 1\end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ & -2 & 1\end{bmatrix}, \tag{II.5.4.5}$$

which the reader can verify using definition II.4.1.3.

Note that these matrices are invertible: applying again definition II.4.1.3, we can verify that $M_1\tilde{M}_1$, $\tilde{M}_1 M_1$, $M_2\tilde{M}_2$ and $\tilde{M}_2 M_2$ are all equal to the identity matrix of order three (definition II.4.1.9) for

$$\tilde{M}_1 = \begin{bmatrix} 1 & & \\ \ell_{21} & 1 & \\ \ell_{31} & & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & & 1 \end{bmatrix} \quad \text{and} \quad \tilde{M}_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & \ell_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ & 2 & 1 \end{bmatrix} . \tag{II.5.4.6}$$

Recalling definition II.4.1.14, we then conclude that M_1 and M_2 are invertible and that $M_1^{-1} = \tilde{M}_1$ and $M_2^{-1} = \tilde{M}_2$.

Remark II.5.4.1 (inverse matrices represent inverse row transformations, or *what we are doing here*)**.** Just above, we established that the matrices given in (II.5.4.6) are the inverses of the matrices introduced in (II.5.4.5). We made the final conclusion in this regard *relatively formally*, using our *formal* definitions (definitions II.4.1.3, II.4.1.9 and II.4.1.14). The fact itself, however, follows in a *more intuitive way* from the nature of the two transformations that we performed in (II.5.4.1) to (II.5.4.3) and denoted by M_1 and M_2.

Recall that, at every step $k \in \{1, \ldots, 2\} = \{1, 2\}$, we subtracted equation (row) k from equations (rows) $k{+}1, \ldots, 3$ with coefficients $\ell_{k+1,k}, \ldots, \ell_{3,k}$. Then the inverse transformation amounts to adding the same equation (row) back, with the same coefficient and to the same equations (rows). This corresponds to multiplying on the left — another time — by almost the same matrix as M_k, obtained from M_k by inverting the sign of the strictly lower triangular part. This matrix is $\tilde{M}_k$.

The composition of these two steps in both possible orders amounts to doing nothing (to an identity transformation of equations and rows), and this does not depend on the matrix B_{k-1}. Indeed, for any $\ell_{k+1,k}, \ldots, \ell_{3,k} \in \mathbb{R}$, only the fact whether we eliminate column k below the diagonal (and not our observations on the inverse transformation) depends on B_{k-1}. Then, applying the two mutually inverse transformations to the rows of the identity matrix I of order three in both possible orders, we arrive at $\tilde{M}_k M_k I = I$ and $M_k \tilde{M}_k I = I$, and these equalities immediately give $\tilde{M}_k M_k = I$ and $M_k \tilde{M}_k = I$.

Remark II.5.4.2 (*applied* vs. *theoretical* and the common sense connecting the two)**.** As the reader may have noticed, the explanation offered in remark II.5.4.1, as compared to the formal argument given in the preceding text, is much more intuitive. Indeed, it does not *explicitly* refer to the formula (II.4.1.9) from definition II.4.1.3, which may seem rather formal, and instead appeals to the meaning of the transformations the matrices represent. On the other hand, this explanation may seem verbose and more costly to follow for anyone who is confident about definition II.4.1.3.

One of the goals of the present course and lecture notes is that the reader become fluent in the language of matrix transformations and, more generally, of linear mappings (which we will introduce at a later point). The study of such mappings allows to concisely and efficiently express and analyze various specific problems, to work only with the relevant algebraic structures and properties and, eventually, to advance further in the development, analysis and application of suitable algorithms. It is nevertheless crucial to maintain, whenever possible, a parallel interpretation of abstract reasoning in terms of the specific problem at hand and in its context. Indeed, problem-specific intuition can indicate errors in formal reasoning, expose the inadequacy of the abstract formulation of the problem or suggest how the theoretical construction used may be extended.

> For the example considered in § II.5.4, the solution of a system of three equations plays the role of an *applied problem* (even though any such a system is usually an abstract model of something more applied), and the body of notions and properties developed in §§ II.2 to II.4 plays the role of *abstract theory*. In general, the terms *applied* and *theoretical* are wildly relative! What is important for applications is whether the *abstract theory* is good for the *applied problem* under consideration, that is, has anything to offer for solving the problem.

In (II.5.4.1) to (II.5.4.3), we were transforming the coefficients of the equations (matrix A) and the right-hand side (column vector b) at the same time. Let us now use the partitioning indicated for B_0, B_1, B_2 in (II.5.4.3) and consider the intermediate matrices and right-hand sides: we set $B_k = [\,U_k \,|\, b_k\,]$ with $U_k \in \mathbb{R}^{3\times 3}$ and $b_k \in \mathbb{R}^3$ for $k \in \{0,1,2\}$. Of course, we have $U_0 = A$ and $b_0 = b$ because we set $B_0 = B = [\,A \,|\, b\,]$ earlier.

Notice that in (II.5.4.1) to (II.5.4.3) we actually computed our old acquaintances L and U from example II.3.2.3 in the course of Gaussian elimination in (II.5.4.1) to (II.5.4.3):

$$L = \begin{bmatrix} 1 & & \\ 2 & 1 & \\ 6 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ \ell_{21} & 1 & \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 1 \\ & & 2 \end{bmatrix} = U_2 \,. \tag{II.5.4.7}$$

Further, even though the equality $A = LU$ was verified in example II.4.1.6, our above calculation actually ensures it independently. Indeed, using definition II.4.1.3, it is easy to verify that $\tilde{M}_1 \tilde{M}_2 = L$. The relation $B_2 = M_2 M_1 B$, which we noted in (II.5.4.4), gives, in particular, $U = M_2 M_1 A$, which is equivalent to $\tilde{M}_1 \tilde{M}_2 U = A$, i.e., $LU = A$.

Finally, let us set $y = b_2$. The transformations we performed in (II.5.4.1) to (II.5.4.3) collectively amount to multiplying $b = b_0$ by L^{-1} on the left, so that $y = L^{-1} b$. We therefore have

$$Ax = b \Leftrightarrow Ux = y$$

for all $x \in \mathbb{R}^3$.

To summarize, by transforming the composed matrix B instead of A in (II.5.4.3), we effectively calculated the very same matrices L and U as those defined in example II.3.2.3 and, at the same time, solved the linear system $Ly = b$ with respect to $y \in \mathbb{R}^3$. So the solution of problem (II.5.1.1) with the chosen data can be expressed as follows:

$$\mathcal{X} = \left\{ x \in \mathbb{R}^3 \colon Ux = y \right\},$$

and solving the above system with matrix U is the only thing that remains to be done.

Note that, by construction, U is an upper-triangular matrix and that this very detail — seemingly technical — makes it remarkably easy to solve linear systems with matrix U. In particular, for every $x \in \mathbb{R}^3$, we have

$$Ux = y \Leftrightarrow \begin{bmatrix} 1 & 2 & 3 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \\ 2 \end{bmatrix} \Leftrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix},$$

where we perform the so-called *backward substitution* by finding first x_3 from the third equation, then x_2 from the second and finally x_1 from the first. The original system is therefore uniquely solvable, and the set of solutions consists of a single element:

$$\mathcal{X} = \left\{ [1 \;\; 2 \;\; 1]^{\mathsf{T}} \right\}.$$

In contrast, what we were doing first, in order to transform the equation $Ax = b$ into $Ux = y$, is called *forward substitution*. Indeed, at the first step ($B_0 \rightsquigarrow B_1$), we effectively used the first equation to express x_1 in terms of x_2 and x_3 and substituted the resulting expressions into the second and third equations. At the second step ($B_1 \rightsquigarrow B_2$), we dealt with only two equations (the second and third) involving only two unknowns (x_2 and x_3). Then we effectively used the (new) second equation to express x_2 in terms of x_3 and substituted the resulting expression into the (new) third equation. This process did not give us a single component of the solution but was nevertheless crucial for solving the original system $Ax = b$ by transforming it into the equivalent but much simpler system $Ux = y$ with a triangular matrix U. That simpler system we then easily solved by backward substitution.

§ II.5.5. Properties of triangular matrices

As we can conclude from § II.5.4, Gaussian elimination may be a way to factorize a matrix $A \in \mathbb{F}^{n\times n}$ in the form $A = LU$ with a unit lower-triangular matrix $L \in \mathbb{F}^{n\times n}$ and an upper-triangular matrix $U \in \mathbb{F}^{n\times n}$. Such a factorization can, for example, dramatically simplify the solution of linear systems with A, but we will see very soon that it can, generally speaking, provide us with important insights into intrinsic properties of A.

Taking this as a motivation for considering triangular matrices important, we will now focus on their properties. For the definition of triangular matrices, refer to definition II.3.2.1 and remark II.3.2.2.

Lemma II.5.5.1 (triangular structure under matrix multiplication)**.** For $m \in \mathbb{N}$, consider matrices $A \in \mathbb{F}^{m\times m}$ and $B \in \mathbb{F}^{m\times m}$. Then the following properties hold.

(a) If matrices A and B are lower-triangular, then so is AB and $(AB)_{ii} = A_{ii}\, B_{ii}$ for every $i \in \{1, \dots, m\}$.

(b) If matrices A and B are unit lower-triangular, then so is AB.

(c) If matrices A and B are strictly lower-triangular, then so is AB.

(d) If matrices A and B are upper-triangular, then so is AB and $(AB)_{ii} = A_{ii}\, B_{ii}$ for every $i \in \{1, \dots, m\}$.

(e) If matrices A and B are unit upper-triangular, then so is AB.

(f) If matrices A and B are strictly upper-triangular, then so is AB.

Proof. Let us first prove part (a). Consider $i, j \in \{1, \dots, m\}$ such that $i \le j$. Applying definition II.4.1.3, we obtain that

$$(AB)_{ij} = \sum_{k=1}^{m} A_{ik}\, B_{kj} = \sum_{k=1}^{i-1} A_{ik}\, B_{kj} + \sum_{k=i}^{j} A_{ik}\, B_{kj} + \sum_{k=j+1}^{m} A_{ik}\, B_{kj}$$

Since matrix B is lower triangular, we have $B_{kj} = 0$ for each $k \in \{1, \dots, j-1\}$. Similarly, we have $A_{ik} = 0$ for each $k \in \{i+1, \dots, m\}$ because A is lower triangular. Due to that $i \le j$, these equalities result in

$$(AB)_{ij} = \sum_{k=1}^{i-1} A_{ik} \cdot 0 + \sum_{k=i}^{j} A_{ik}\, B_{kj} + \sum_{k=j+1}^{m} 0 \cdot B_{kj} = \sum_{k=i}^{j} A_{ik}\, B_{kj}\,.$$

Further, in the case of $i < j$, we obtain

$$(AB)_{ij} = \sum_{k=i}^{j} A_{ik}\, B_{kj} = A_{ii}\, B_{ij} + \sum_{k=i+1}^{j} A_{ik}\, B_{kj} = A_{ii} \cdot 0 + \sum_{k=i+1}^{j} 0 \cdot B_{kj} = 0\,,$$

and, in the case of $i = j$, we arrive at

$$(AB)_{ij} = \sum_{k=i}^{j} A_{ik}\, B_{kj} = A_{ii}\, B_{ii}\,,$$

which competes the proof of part (a).

Parts (b) and (c) follow immediately from part (a). Indeed, if both A and B are lower-triangular, and so is AB by part (a). If, additionally, both A and B have unitary (as in part (b)) or zero (as in part (c)) diagonals, then, by part (a), the product AB exhibits the same property.

Part (d) follows from part (a) by transposition. Indeed, using proposition II.4.1.11, we obtain $(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}}$. If A and B are both upper-triangular matrices, then A^{T} and B^{T} are both lower-triangular. Then, by part (a), so is $B^{\mathsf{T}} A^{\mathsf{T}} = (AB)^{\mathsf{T}}$ and $(B^{\mathsf{T}} A^{\mathsf{T}})_{ii} = (B^{\mathsf{T}})_{ii}\, (A^{\mathsf{T}})_{ii}$ for every $i \in \{1, \ldots, m\}$. So AB is an upper-triangular matrix and $(AB)_{ii} = \big((AB)^{\mathsf{T}}\big)_{ii} = (B^{\mathsf{T}} A^{\mathsf{T}})_{ii} = (B^{\mathsf{T}})_{ii}\, (A^{\mathsf{T}})_{ii} = A_{ii}\, B_{ii}$ for every $i \in \{1, \ldots, m\}$.

Parts (e) and (f) are derived from part (d) in the same way as parts (b) and (c) from part (a).

Our goal is to be able to explicitly invert any triangular matrix. We start with a simple class of unit lower-triangular matrices generalizing matrices M_1 and M_2 from (II.5.4.5). Let us recall that each of these two matrices appeared in the course of Gaussian elimination in § II.5.4 as a matrix representation of a single elimination step.

Lemma II.5.5.2 (inversion of a single step of Gaussian elemination)**.** For $m \in \mathbb{N}$ and $k \in \{1, \ldots, m\}$, let I denote the identity matrix of order $m \in \mathbb{N}$, $e_k \in \mathbb{F}^m$ be column k of I and $\ell \in \mathbb{F}^m$ have zero components 1 to k. Then the matrix $I - \ell e_k^{\mathsf{T}}$ is invertible and

$$\big(I - \ell e_k^{\mathsf{T}}\big)^{-1} = I + \ell e_k^{\mathsf{T}}\,.$$

Proof. Multiplying $I - \ell e_k^{\mathsf{T}}$ and $I + \ell e_k^{\mathsf{T}}$ in both possible orders and using the associativity, multiplicative identity and distributivity of matrix multiplication (parts (a) to (c) of proposition II.4.1.10), we obtain

$$(I \mp \ell e_k^{\mathsf{T}})(I \pm \ell e_k^{\mathsf{T}}) = I + \ell e_k^{\mathsf{T}} - \ell e_k^{\mathsf{T}} - \ell (e_k^{\mathsf{T}} \ell) e_k^{\mathsf{T}}\,.$$

By the assumption of the lemma, all components of e_k except component k are zeros and component k of ℓ is zero. This leads to $e_k^{\mathsf{T}} \ell = 0$ and hence

$$(I \mp \ell e_k^{\mathsf{T}})(I \pm \ell e_k^{\mathsf{T}}) = I\,.$$

Applying definition II.4.1.14, we obtain the claim of the lemma.

Further, we will now see that lower-triangular matrices (such as L from (II.5.4.7), which also was produced by Gaussian elimination in § II.5.4) can be decomposed into the matrices of the form considered in lemma II.5.5.2: first, additively; second, multiplicatively.

For $m \in \mathbb{N}$ and $r \in \{1, \ldots, m-1\}$, let us consider a unit lower-triangular matrix $L = [\ell_{ij}]_{i=1,\, j=1}^{m \quad m} \in \mathbb{F}^{m \times m}$ the strictly lower triangular part of which is zero starting from column

$r+1$:

$$L = \left[\begin{array}{cccc|ccc} 1 & & & & & & \\ \ell_{21} & \ddots & & & & & \\ \vdots & \ddots & \ddots & & & & \\ \ell_{r,1} & \cdots & \ell_{r,r-1} & 1 & & & \\ \hline \ell_{r+1,1} & \cdots & \ell_{r+1,r-1} & \ell_{r+1,r} & 1 & & \\ \vdots & \cdots & \vdots & \vdots & & \ddots & \\ \ell_{m1} & \cdots & \ell_{m,r-1} & \ell_{m,r} & & & 1 \end{array}\right] . \tag{II.5.5.1}$$

In what follows, we will often encounter such matrices and, for convenience, will partition them as follows in the case of $r < m$:

$$L = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} , \tag{II.5.5.2}$$

where $L_1 \in \mathbb{F}^{r \times r}$ is a unit lower-triangular matrix, $L_2 \in \mathbb{F}^{(m-r) \times r}$ and I_2 is the identity matrix of order $m-r$. Let I denote the identity matrix of order m and $e_1, \dots, e_m$ be its columns. For every $k \in \{1, \dots, r\}$, let ℓ_k be the kth column of $L - I$:

$$\ell_k = [\ell_{ik}]_{i=1}^m = \begin{bmatrix} 0 & \cdots & 0 & \ell_{k+1\,k} & \cdots & \ell_{mk} \end{bmatrix}^\mathsf{T} \in \mathbb{F}^m . \tag{II.5.5.3}$$

Since columns $r+1$ to m of $L - I$ are zeros, we have $L - I = \sum_{k=1}^r \ell_k e_k^\mathsf{T}$, which implies

$$L = I + \sum_{k=1}^{r} \ell_k e_k^\mathsf{T} . \tag{II.5.5.4}$$

In particular, any unit lower-triangular matrix $L \in \mathbb{F}^{m \times m}$ with $m \in \mathbb{N}$ can be represented in the form of (II.5.5.4), with $r = m-1$ in the most general case. We will exploit (II.5.5.4) as a way to decompose an arbitrary unit lower-triangular matrix L into elementary unit lower-triangular matrices of the form considered in lemma II.5.5.2. In the general implementation and rigorous analysis of Gaussian elimination, such matrices as L in (II.5.5.1) with $r \in \{1, \dots, m\}$ will appear as representations of the first r steps of Gaussian elimination.

Lemma II.5.5.3 (representation of r steps of Gaussian elimination)**.** For $m \in \mathbb{N}$ and $r \in \{1, \dots, m-1\}$, let I denote the identity matrix of order $m \in \mathbb{N}$, $e_1, \dots, e_m \in \mathbb{F}^m$ be its columns and $\ell_1, \dots, \ell_r \in \mathbb{F}^m$ be such that, for each $k \in \{1, \dots, r\}$, components 1 to k of ℓ_k are all zero. Then

$$(I + \ell_1 e_1^\mathsf{T}) \cdots (I + \ell_r e_r^\mathsf{T}) = I + \sum_{k=1}^{r} \ell_k e_k^\mathsf{T} .$$

Proof. Let us prove the claim by *induction* with respect to r. For $r = 1$, the claimed statement is trivial and therefore holds (*induction base*). Let us consider $n \in \{1, \dots, m-2\}$, assume that the claim holds for $r = n$ (*induction assumption*) and prove the claim for $r = n+1$ (*induction step*). That will complete the proof since the induction step can be applied iteratively, starting from the base case and yielding thereby the claimed equality for every $r \in \{1, \dots, m-1\}$.

By the induction assumption, we have

$$(I+\ell_1 e_1^{\mathsf{T}})\cdots(I+\ell_n e_n^{\mathsf{T}}) = I + \sum_{k=1}^{n} \ell_k e_k^{\mathsf{T}} \,.$$

Multiplying both sides by $I+\ell_{n+1}e_{n+1}^{\mathsf{T}}$ on the right and applying the associativity, multiplicative identity and distributivity of matrix multiplication (parts (a) to (c) of proposition II.4.1.10), we arrive at

$$\begin{aligned}(I+\ell_1 e_1^{\mathsf{T}})\cdots(I+\ell_{n+1} e_{n+1}^{\mathsf{T}}) &= \Big(I + \sum_{k=1}^{n} \ell_k e_k^{\mathsf{T}}\Big)(I+\ell_{n+1}e_{n+1}^{\mathsf{T}}) \\ &= I + \sum_{k=1}^{n} \ell_k e_k^{\mathsf{T}} + \ell_{n+1}e_{n+1}^{\mathsf{T}} + \sum_{k=1}^{n} \ell_k e_k^{\mathsf{T}} \ell_{n+1} e_{n+1}^{\mathsf{T}} \\ &= I + \sum_{k=1}^{n+1} \ell_k e_k^{\mathsf{T}} + \sum_{k=1}^{n} \ell_k (e_k^{\mathsf{T}}\ell_{n+1}) e_{n+1}^{\mathsf{T}} \,.\end{aligned}$$

Showing that the last sum is zero would be sufficient (and is, in fact, necessary) for completing the induction step (and therefore the entire proof). Indeed, all its terms are equal to zero: for every $k \in \{1,\ldots,n\}$, all components of e_k except component k are zeros, whereas component k of ℓ_{n+1} is zero by the assumption of the lemma. This implies $e_k^{\mathsf{T}}\ell_{n+1} = 0$ for each $k \in \{1,\ldots,n\}$.

Let us now use the results of lemmata II.5.5.2 and II.5.5.3 to derive properties of triangular matrices with respect to matrix inversion.

Lemma II.5.5.4 (lower-triangular structure under matrix inversion)**.** For $m \in \mathbb{N}$, consider a lower-triangular matrix $L \in \mathbb{F}^{m\times m}$. Then the following properties hold.

(a) If matrix L is unit lower triangular, then it is invertible and matrix L^{-1} is unit lower triangular.

(b) If matrix L has no zeros on the diagonal, then it is invertible and matrix L^{-1} is lower triangular.

(c) If L is invertible, then it has no zeros on the diagonal.

Proof. Let I be the identity matrix of order m throughout the proof.

Let us first prove part (a). Denoting the columns of the strictly lower-triangular matrix $L-I$ by $\ell_1,\ldots,\ell_m$, we obtain $L = I + \sum_{k=1}^{m-1} \ell_k e_k^{\mathsf{T}}$, cf. (II.5.5.4). Clearly, columns $\ell_1,\ldots,\ell_{m-1} \in \mathbb{F}^m$ are of the form (II.5.5.3). Using the multiplicative representation given by lemma II.5.5.3, we obtain

$$L = (I+\ell_1 e_1^{\mathsf{T}})\cdots(I+\ell_{m-1}e_{m-1}^{\mathsf{T}}) \,.$$

For each $k \in \{1,\ldots,m-1\}$, lemma II.5.5.2 yields that matrix $I+\ell_k e_k^{\mathsf{T}}$ is invertible and that its inverse equals $I-\ell_k e_k^{\mathsf{T}}$. So L is a product of invertible matrices and is therefore invertible by proposition II.4.1.22; furthermore, by the same proposition,

$$L^{-1} = (I-\ell_{m-1}e_{m-1}^{\mathsf{T}})\cdots(I-\ell_1 e_1^{\mathsf{T}}) \,.$$

Each of the right-hand factors is a unit lower-triangular matrix, and hence L^{-1} is also such by part (b) of lemma II.5.5.1.

Now we will prove part (b). Let $D \in \mathbb{F}^{m\times m}$ be the diagonal part of L (see definition II.3.2.1). Since L has no zeros on the diagonal, matrix D is invertible and matrix $D^{-1}L$ is unit lower-triangular. Applying part (a), we deduce that matrix $D^{-1}L$ is invertible and that matrix $(D^{-1}L)^{-1}$ is unit lower triangular. In particular, this means that $L = D\,(D^{-1}L)$ is a product of two invertible matrices, and we obtain by proposition II.4.1.22 that L is invertible and $L^{-1} = (D^{-1}L)^{-1}D^{-1}$. Both the right-hand factors are lower-triangular matrices, and hence L^{-1} is also such by part (a) of lemma II.5.5.1.

Finally, let us prove part (c) by contradiction, assuming that L is invertible but has a zero on the diagonal. Let $k \in \{1,\dots,m\}$ be such that $L_{kk} = 0$ and $L_{ii} \neq 0$ for each $i \in \{1,\dots,k-1\}$ (in other words, the kth diagonal entry is the first of the zero diagonal entries).

If $k = 1$, then we have $L_{11} = 0$, so that the first row of L is zero. Then so is the first row of LL^{-1}, which contradicts that, by definition II.4.1.14, $LL^{-1} = I$.

If $k = m$, then we have $L_{mm} = 0$, so that column m of L is zero. Then so is column m of $L^{-1}L$, which contradicts that, by definition II.4.1.14, $L^{-1}L = I$.

If $1 < k < m$, matrix L is of the form

$$L = \begin{bmatrix} L_{11} & & \\ u^{\mathsf{T}} & 0 & \\ L_{21} & v & L_{22} \end{bmatrix},$$

where $L_{11} \in \mathbb{F}^{(k-1)\times(k-1)}$, $L_{22} \in \mathbb{F}^{(m-k)\times(m-k)}$, $L_{21} \in \mathbb{F}^{(m-k)\times(k-1)}$, $u \in \mathbb{F}^{k-1}$ and $v \in \mathbb{F}^{m-k}$. Due to our definition of k, matrix L_{11} has no zeros on the diagonal. It is therefore invertible by part (b). Denoting by I_1 and I_2 the identity matrices of orders $k-1$ and $m-k$, we define the matrices

$$X = \begin{bmatrix} I_1 & & \\ -u^{\mathsf{T}}L_{11}^{-1} & 1 & \\ & & I_2 \end{bmatrix} \quad\text{and}\quad Y = \begin{bmatrix} I_1 & & \\ u^{\mathsf{T}}L_{11}^{-1} & 1 & \\ & & I_2 \end{bmatrix},$$

cf. (II.5.6.7) Direct multiplication in block form shows that $XY = I = YX$, so that X and Y are invertible and mutually inverse by part (b). Direct multiplication in block form also leads to

$$XL = \begin{bmatrix} L_{11} & & \\ & 0 & \\ L_{21} & v & L_{22} \end{bmatrix}, \tag{II.5.5.5}$$

so that row k of XL is zero. Note that the matrices X and $Y = X^{-1}$ are analogues of the matrices M and $L = M^{-1}$ appearing in part (a) of lemma II.5.6.3. The difference is that a single step of the standard LU decomposition (as presented in lemma II.5.6.3) is aimed at the elimination a one-column block of maximum height located below the diagonal and uses a single entry as a pivot, whereas the transformation (II.5.5.5) serves to eliminate a one-row block of maximum width located below the diagonal and uses a square submatrix of the suitable order as a pivot.

Since both L and X are invertible, XL is invertible by proposition II.4.1.22. Then column k of $(XL)^{-1}(XL)$ is zero, which contradicts that, by definition II.4.1.14, $(XL)^{-1}(XL) = I$.

So we have proved that it is not possible for L to be invertible and, at the same time, to have a zero on the diagonal. This completes the proof of part (c).

Combining the results of lemma II.5.5.4 and proposition II.4.1.23, we immediately obtain analogous properties for upper-triangular matrices.

Corollary II.5.5.5 (upper-triangular structure under matrix inversion)**.** For $m \in \mathbb{N}$, consider an upper-triangular matrix $U \in \mathbb{F}^{m\times m}$. Then the following properties hold.

(a) If matrix U is unit upper triangular, then it is invertible and matrix U^{-1} is unit upper triangular.

(b) If matrix U has no zeros on the diagonal, then it is invertible and matrix U^{-1} is upper triangular.

(c) If U is invertible, then it has no zeros on the diagonal.

Proof. The proof is left to the reader as an exercise.

§ II.5.6. LU decomposition

In this section, we consider $m, n \in \mathbb{N}$ and turn to generalizing the calculation from § II.5.4 to an algorithm applicable to a matrix of size $m \times n$, as arbitrary as possible. To be precise, for a matrix $A \in \mathbb{F}^{m\times n}$, we are interested in finding a unit lower-triangular matrix $L \in \mathbb{F}^{m\times m}$ and an upper-trapezoid matrix $U \in \mathbb{F}^{m\times n}$ such that $A = LU$.

As we observed in § II.5.4, this *may be* possible to accomplish by eliminating the entries below the diagonal in one column after another. To represent and analyze this process step by step, we will also consider *incomplete LU decompositions*. We will see later in this section that such a decomposition may be the best we can obtain for an arbitrary matrix by means of the procedure considered in § II.5.4.

Definition II.5.6.1 (r-step LU decomposition of a matrix)**.** For $m, n, r \in \mathbb{N}$ such that $r \le \min\{m, n\}$, let $L \in \mathbb{F}^{m\times m}$ and $U \in \mathbb{F}^{m\times n}$ satisfy the following conditions.

(a) Assume that matrix L is unit lower triangular and differs from the identity matrix of order m in no other column than the first r, i.e.:

(i) if $r < m$, then $L \in \mathbb{F}^{m\times m}$ is of the form

$$L = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \tag{II.5.6.1}$$

with a unit lower-triangular matrix $L_1 \in \mathbb{F}^{r\times r}$, an arbitrary matrix $L_2 \in \mathbb{F}^{(m-r)\times r}$ and the identity matrix I_2 of order $m - r$;

(ii) if $r = m$, then $L \in \mathbb{F}^{m\times m}$ is a unit lower-triangular matrix.

(b) Assume that matrix U is upper triangular in the first r columns with no zero among its first r diagonal entries:

(i) if $r < \min\{m,n\}$, then $U \in \mathbb{F}^{m\times n}$ is of the form

$$U = \begin{bmatrix} U_1 & U_2 \\ & S \end{bmatrix} \tag{II.5.6.2}$$

with an upper-triangular matrix $U_1 \in \mathbb{F}^{r\times r}$ *with no zeros on the diagonal*, and arbitrary matrices $U_2 \in \mathbb{F}^{r\times(n-r)}$ and $S \in \mathbb{F}^{(m-r)\times(n-r)}$;

(ii) if $r = \min\{m,n\}$, then $U \in \mathbb{F}^{m\times n}$ is an upper-triangular matrix *with no zeros on the diagonal.*

For $A = LU \in \mathbb{F}^{m\times n}$, the equality $A = LU$, as a representation of A in terms of L and U satisfying conditions (a) and (b), is called an *r-step LU decomposition.* Matrices L and U are referred to as *r-step LU factors.* In the case of $r < \min\{m,n\}$, the matrix S is called an *r-step Schur complement.*

When $r < \min\{m,n\}$ and S is nonzero, the above r-step LU decomposition is called *incomplete.*

When $r < \min\{m,n\}$ and S is zero or when $r = \min\{m,n\}$, the above r-step LU decomposition is called *complete.*

The assumptions on the LU factors made in definition II.5.6.1 immediately imply the following.

Proposition II.5.6.2 (the leading principal submatrices of the factors are invertible)**.** Let $m,n,r \in \mathbb{N}$ be such that $r \le \min\{m,n\}$ and $A \in \mathbb{F}^{m\times n}$. Consider an r-step LU decomposition $A = LU$. Then the leading principal submatrices of L and U of orders $1,\ldots,r$ are invertible.

Proof. Consider $k \in \{1,\ldots,r\}$. By definition II.5.6.1, the matrix L is unit lower triangular and the matrix U is upper triangular in columns 1 to r with no zeros among its first r diagonal entries. Let us denote by L_1 and U_1 the leading principal submatrices of order k of L and U. Then matrix L_1 is unit lower triangular and matrix U_1 is upper triangular with no zeros on the diagonal. Applying part (b) of lemma II.5.5.4 to matrix L_1 and part (b) of corollary II.5.5.5 to matrix U_1, we obtain that matrices L_1 and U_1 are both invertible.

For $m,n,r \in \mathbb{N}$ such that $r \le \min\{m,n\}$, let us consider an r-step LU decomposition $A = LU$ of a matrix $A \in \mathbb{F}^{m\times n}$ in the sense of definition II.5.6.1, with L and U partitioned as in conditions (a) and (b) of definition II.5.6.1. Let us partition the matrix in the same way:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{II.5.6.3}$$

with $A_{11} \in \mathbb{F}^{r\times r}$. So A_{11} is the leading principal submatrix of size $r \times r$ (see definitions II.4.2.7 and II.4.2.8), and the sizes of A_{21}, A_{12} and A_{22} unambiguously follow from those of A and A_{11}. For $r = m$, the second block row in the right-hand side of (II.5.6.3) should be omitted; similarly, for $r = n$, the second block column should be omitted. Rewriting the r-step LU decomposition $A = LU$ blockwise, we obtain the following expressions for the blocks of A:

$$\begin{aligned} &A_{11} = L_1 U_1\,, && A_{12} = L_1 U_2 && \text{if } r < n, \\ &A_{21} = L_2 U_1 \quad \text{if } r < m, && A_{22} = L_2 U_2 + S && \text{if } r < \min\{m,n\}. \end{aligned} \tag{II.5.6.4}$$

We will use the notations of (II.5.6.3) and (II.5.6.4) in several proofs below.

Definition II.5.6.1 introduces what we mean by *LU decomposition* but does not shed light on how such decompositions may be obtained. We have, however, explored a way to do that in § II.5.4. The calculation showcased in (II.5.4.1) and (II.5.4.2), which is an instance of good old Gaussian elimination, is *iterative* and *recursive* in its nature: all steps are similar to each other in how they are performed and in what they achieve. At every step k, the calculation does not involve the first $k-1$ columns and the first $k-1$ rows of the current matrix, focusing therefore on a submatrix. The result of step k is twofold. First, the current matrix becomes zero below the diagonal in column k and remains so in the first $k-1$ columns. Second, step $k+1$ of the calculation can ignore the first k columns and the first k rows of the current matrix, focusing therefore on a submatrix with one fewer rows and one fewer columns than step k.

In the following lemma, we will generalize the first and second steps of Gaussian elimination from (II.5.4.1) and (II.5.4.2). First, we obtain a one-step LU decomposition for every matrix under a certain condition. Second, we show that if a p-step LU decomposition is available for a matrix and a q-step LU decomposition for the corresponding Schur complement is also available, then the two decompositions can be explicitly combined to yield a $(p+q)$-step LU decomposition of the original matrix. These two results allow, for example, to start with a one-step LU decomposition and iterate with respect to the number of steps, increasing it by one at a time.

Lemma II.5.6.3 (iterated Gaussian elimination)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$.

(a) Let I, e_1 and $a_1 = [a_{i1}]_{i=1}^m$ be the identity matrix of order m, the first column of I and the first column of A. Assume that $a_{11} \neq 0$ and consider

$$\ell_1 = a_{11}^{-1} a_1 - e_1\,, \quad L = I + \ell_1 e_1^\top \quad \text{and} \quad U = (I - \ell_1 e_1^\top)\,A\,. \tag{II.5.6.5}$$

Then $A = LU$ is a one-step LU decomposition.

(b) Let $r, q \in \mathbb{N}$ be such that $r + q \leq \min\{m, n\}$. Assume that $A = LU$ is an r-step LU decomposition with

$$L = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} U_1 & U_2 \\ & S \end{bmatrix}\,, \tag{II.5.6.6}$$

where $L_1 \in \mathbb{F}^{r\times r}$, $L_2 \in \mathbb{F}^{(m-r)\times r}$, I_2 is the identity matrix of order $m-p$, $U_1 \in \mathbb{F}^{r\times r}$, $U_2 \in \mathbb{F}^{r\times(m-r)}$ and $S \in \mathbb{F}^{(m-r)\times(n-r)}$. Further, assume that $S = \tilde{L}\tilde{U}$ is a q-step LU decomposition. Then $A = L_\star U_\star$ with

$$L_\star = \begin{bmatrix} L_1 & \\ L_2 & \tilde{L} \end{bmatrix} \quad \text{and} \quad U_\star = \begin{bmatrix} U_1 & U_2 \\ & \tilde{U} \end{bmatrix}$$

is an $(r+q)$-step LU decomposition

Note that equation (II.5.6.5) in part (a) of lemma II.5.6.3 defines

$$\ell_1 = \frac{1}{a_{11}} \begin{bmatrix} 0 \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} \in \mathbb{F}^m \quad \text{and} \quad L = \begin{bmatrix} 1 & & & \\ a_{11}^{-1} a_{21} & 1 & & \\ \vdots & & \ddots & \\ a_{11}^{-1} a_{m1} & & & 1 \end{bmatrix} \in \mathbb{F}^{m\times m}. \tag{II.5.6.7}$$

Proof of lemma II.5.6.3.

(a) Since $a_{11} \neq 0$ by assumption, ℓ_1, L and U are well defined by (II.5.6.5). Furthermore, the first component of ℓ_1 is zero by (II.5.6.5). We note immediately that $L = I + \ell_1 e_1^\mathsf{T}$ satisfies condition (a) of definition II.5.6.1 with $r = 1$.

Let us set $M_1 = I - \ell_1 e_1^\mathsf{T}$. By lemma II.5.5.2, matrix M_1 is invertible and $M_1^{-1} = L$. The definition of U given in (II.5.6.5) then implies $LU = M_1^{-1}(M_1 A) = (M_1 M_1^{-1})A = I \cdot A = A$.

Note that the particular definition of ℓ_1 made in (II.5.6.5) ensures that U is zero below the diagonal in the first column. Indeed, the first column of A equals $a_1 = Ae_1 = a_{11}\, e_1 + a_{11}\, \ell_1$ due to the definition of ℓ_1 given in (II.5.6.5). Then, using that $e_1^\mathsf{T} e_1 = 1$ and $e_1^\mathsf{T} \ell_1 = 0$ (these equalities are verified using definition II.4.1.3), we find that the first column of U reads

$$\begin{aligned} Ue_1 &= (M_1 A)\, e_1 = M_1 (Ae_1) \\ &= \big(I - \ell_1 e_1^\mathsf{T}\big)(a_{11} e_1 + a_{11} \ell_1) \\ &= a_{11} e_1 + a_{11} \ell_1 - a_{11} \ell_1 e_1^\mathsf{T} e_1 - a_{11} \ell_1 e_1^\mathsf{T} \ell_1 \\ &= a_{11} e_1 + a_{11} \ell_1 - a_{11} \ell_1 \cdot 1 - a_{11} \ell_1 \cdot 0 \\ &= a_{11} e_1 \,. \end{aligned}$$

These transformations rely on the associativity, multiplicative identity and distributivity of matrix multiplication (parts (a) to (c) of proposition II.4.1.10) and on the associativity and commutativity of matrix addition (parts (a) and (b) of proposition II.3.1.7).

The conclusion is that the first column of U is zero in all components except the first, which is equal to a_{11} and is therefore nonzero. So U satisfies condition (b) of definition II.5.6.1 with $r = 1$.

We finally conclude that $A = LU$ is a one-step LU decomposition by definition II.5.6.1.

(b) By assumption, $S = \tilde{L}\tilde{U}$ is a q-step LU decomposition. Then condition (a) of definition II.5.6.1 gives that $\tilde{L}$ is a unit lower triangular matrix. Applying part (a) of lemma II.5.5.4, we deduce that matrix $\tilde{L}$ is invertible and that matrix $\tilde{L}^{-1}$ is unit lower triangular. Let us denote the identity matrix of order p by I_1 and consider

$$M = \begin{bmatrix} I_1 & \\ & \tilde{L}^{-1} \end{bmatrix} = \quad \text{with} \quad M^{-1} = \begin{bmatrix} I_1 & \\ & \tilde{L} \end{bmatrix} .$$

With these notations, the LU decomposition $A = LU$ implies $A = (LM^{-1})(MU)$.

Let us note that the first factor,

$$LM^{-1} = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \begin{bmatrix} I_1 & \\ & \tilde{L} \end{bmatrix} = \begin{bmatrix} L_1 & \\ L_2 & \tilde{L} \end{bmatrix} = L_\star \,,$$

is a unit lower-triangular matrix since L_1 and $\tilde{L}$ are such by condition (a) of definition II.5.6.1 with $r = p$ and $r = q$ respectively. Furthermore, since matrix L_1 is of size $r \times r$ and matrix $\tilde{L}$ can differ from I_2 only in the first q columns by condition (a) of definition II.5.6.1, we conclude that $L_\star$ can differ from I only in the first $r + q$ columns.

Further, the equality $\tilde{U} = \tilde{L}^{-1}S$ implies

$$MU = \begin{bmatrix} I_1 & \\ & \tilde{L}^{-1} \end{bmatrix} \begin{bmatrix} U_1 & U_2 \\ & S \end{bmatrix} = \begin{bmatrix} U_1 & U_2 \\ & \tilde{U} \end{bmatrix} = U_\star \, .$$

This block structure of the matrix $U_\star$ shows that it is upper triangular in the first $r+q$ columns since matrix U_1 is upper triangular by definition II.5.6.1 and is of size $r \times r$ and matrix $\tilde{U}$ is upper triangular in the first r columns by definition II.5.6.1. Furthermore, the diagonal entries of U_1 and the first r diagonal entries of $\tilde{U}$ are all nonzero by definition II.5.6.1, respectively. We therefore conclude that the first $r+q$ diagonal entries of $U_\star$ are all nonzero.

As a result, $A = L_\star U_\star$ is an $(r+q)$-step LU decomposition in the sense of definition II.5.6.1.

In fact, lemma II.5.6.3 presents a *computational scheme*, *an algorithm* for obtaining LU decompositions, which we summarize in pseudocode in algorithm II.5.6.4 below. The algorithm is a product of the same ideas that we used in the proof of lemma II.5.6.3; in this sense, understanding how and why exactly the algorithm works is quite the same as understanding how and why exactly the proof succeeds.

Algorithm II.5.6.4 (r-step LU decomposition of a matrix)**.**

Input: $m, n, r \in \mathbb{N}$ and $r \leq \min\{m, n\}$, $A = [a_{ij}]_{i=1,\,j=1}^{m\quad n} \in \mathbb{F}^{m \times n}$ with $a_{11} \neq 0$.

Output: matrices $L \in \mathbb{F}^{m \times m}$ and $U \in \mathbb{F}^{m \times n}$ such that $A = LU$ is a k-step LU decomposition with $k \in \{1, \ldots, r\}$ such that $U_{k+1\,k+1} = 0$ if $k < r$.

```
 1: set $U \leftarrow A$
 2: initialize $L$ as the identity matrix of order $m$
 3: for $k = 1, \ldots, r$ do
 4:    {perform step $k$ of Gaussian elimination}
 5:    for $i \in \{k+1, \ldots, r\}$ do
 6:       set $L_{ik} \leftarrow U_{ik}/U_{kk}$ {compute the elimination coefficient for row $i$}
 7:       set $U_{ik} \leftarrow 0$ {eliminate component $i$ in the pivot column in $U$}
 8:       {compute the step-$k$ Schur}
 9:       for $j \in \{k+1, \ldots, n\}$ do
10:          set $U_{ij} \leftarrow U_{ij} - L_{ik} U_{kj}$
11:       end for
12:    end for
13:    if $k < r$ and $U_{k+1\,k+1} = 0$ then
14:       return $L$ and $U$ {step $k+1$ cannot be performed due to a zero pivot}
15:    end if
16: end for
17: return $L$ and $U$ {step $r$ has been performed}
```

First, let us remark that algorithm II.5.6.4 may fail to produce an r-step LU decomposition of a given matrix, which case is handled by the conditional of line 13. Indeed, we cannot proceed with Gaussian elimination in the same way as we did in (II.5.4.1) to (II.5.4.3) when the entry by which we need to divide is zero.

Our goal is to exhaustively characterize such situations, entirely in terms of the input data and with no reference to any intermediate computational results (such as $U_{k+1\,k+1}$ on line 13). A closely related question is whether such a situation can be a deficiency of the algorithm or that of the decomposition, i.e. whether the failure of algorithm II.5.6.4 to produce an r-step LU decomposition means that no such a decomposition exists. To start addressing these questions, we will now establish a sufficient condition for the existence of an LU decomposition of a given matrix.

> **Lemma II.5.6.5** (sufficient condition for the existence of an LU decomposition)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m\times n}$. Assume that all leading principal submatrices of A of orders $1, \ldots, r$ are invertible. Then A has an r-step LU decomposition.

Proof. Throughout the proof, we will denote the identity matrix of order m and its columns by I and $e_1, \ldots, e_m$. We will show this result by induction with respect to the number $k \in \{1, \ldots, r\}$ of steps for m, n and $A = [a_{ij}]_{i=1,\,j=1}^{m\quad n}$ fixed.

The case of $k = 1$ will serve as the *base of induction*: the leading principal submatrix of order one contains only one entry, which is a_{11}, so the invertibility of that submatrix yields $a_{11} \neq 0$. Then A has a one-step LU decomposition by part (a) of lemma II.5.6.3.

In the remainder of the proof, we consider the case of $k \in \{2, \ldots, r\}$ and prove a single *step of induction*. We assume that the leading principal submatrices of A of orders $1, \ldots, k$ are all invertible. In addition, we make the following *induction assumption*: the statement of the lemma holds with $r = k - 1$. We will then prove that the statement holds for $r = k$. This induction step can be iterated $r - 1$ times, starting from the induction base ($k = 1$), to yield a proof of the lemma.

Applying the claim with $r = k - 1$, we conclude that there exists a $(k - 1)$-step LU decomposition $A = LU$ since the leading principal submatrices of A of orders $1, \ldots, k - 1$ are all invertible. We denote by S and u_{kk} the corresponding Schur complement and entry (k, k) of U, which is also entry $(1, 1)$ of S.

We are yet to use the invertibility of the leading principal submatrix of A of order k, from which we will obtain $u_{kk} \neq 0$. That inequality will complete the proof: indeed, we will be able to apply part (a) of lemma II.5.6.3 to obtain a one-step LU decomposition of S. Then we will be able to apply part (b) of lemma II.5.6.3 with $p = k - 1$ and $q = 1$ to combine the LU decompositions of A and S, which will result in a k-step LU decomposition of A.

Let us denote by $\tilde{L}_1$ and $\tilde{U}_1$ the leading principal submatrices of L and U of order k. Since matrix L is lower triangular, the leading principal submatrix A_{11} of A of order k satisfies $A_{11} = \tilde{L}_1\tilde{U}_1$. This equality follows from the mentioned triangular structure in the same way as the first of equalities (II.5.6.4).

By the assumption of condition (a) of definition II.5.6.1, matrix L is unit lower triangular. Matrix $\tilde{L}_1$ is also unit lower triangular and is therefore invertible by part (a) of lemma II.5.5.4. Multiplying A_{11} by the inverse of $\tilde{L}_1$ on the left, we arrive at $\tilde{L}_1^{-1}A_{11} = \tilde{U}_1$. Applying proposition II.4.1.22, we conclude that matrix $\tilde{U}_1$ is invertible as a product of two invertible matrices. Then part (c) of corollary II.5.5.5 yields $u_{kk} \neq 0$.

Note that the equality $A_{11} = \tilde{L}_1\tilde{U}_1$ can be quite intuitively interpreted in terms of Gaussian elimination. For every $i \in \{1, \ldots, k - 1\}$, step i modifies equations 1 to k only by subtracting from them multiples of row i, located above row k, and rows $k + 1$ to m are not at all involved in the modification of row k.

Lemma II.5.6.5 provides a sufficient condition for the existence of an r-step LU decomposition of a given matrix, and its proof is *constructive* in the sense that it follows a concrete computational procedure (lemma II.5.6.3 and algorithm II.5.6.4), which generalizes the observations made in § II.5.4, and shows that it succeeds in producing an r-step LU decomposition of the matrix (specifically, that the special case handled on line line 13 does not occur).

On the other hand, the condition on leading principal submatrices, stipulated in lemma II.5.6.5, is not so easy to verify. In fact, we are developing the LU decomposition of matrices as our basic tool for matrix inversion and for establishing matrix invertibility. So the second question one may ask about lemma II.5.6.3 and algorithm II.5.6.4 is whether there is a more convenient condition or, at least, whether the one we currently have is *necessary* for the existence of LU decompositions. If it were not necessary, then for some $m, n, r \in \mathbb{N}$ such that $r \leq \min\{m, n\}$ and some matrix $A \in \mathbb{F}^{m \times n}$, an r-step LU decomposition would exist but lemma II.5.6.5, in its present form, would be inapplicable and therefore useless. The inquisitive reader would then feel tempted to return to the proof of lemma II.5.6.5 to see how it could be adapted to the most general setting possible. This is, however, not needed, as the following result shows.

Lemma II.5.6.6 (necessary condition for the existence of an LU decomposition)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m \times n}$. Assume that there exists an r-step LU decomposition of A. Then the leading principal submatrices of A of orders $1, \dots, r$ are invertible.

Proof. Let us assume that $A = LU$ is an r-step LU decomposition of A and consider $k \in \{1, \dots, r\}$. By proposition II.5.6.2, the leading principal submatrices L_1 and U_1 of order k of L and U are invertible. By definition II.5.6.1, the matrix L is lower triangular, which implies $A_{11} = L_1 U_1$ for the leading principal submatrix A_{11} of order k of A in the same way as the first of equalities (II.5.6.4). Then, by proposition II.4.1.22, we obtain that A_{11} is invertible as a product of two invertible matrices.

Example II.5.6.7. Lemmata II.5.6.5 and II.5.6.6 provide a criterion (see § I.1.16) for the existence of an LU decomposition of a matrix. Even if this criterion is stated in terms of the invertibility of submatrices, which makes it difficult to verify in practice without, in effect, attempting to invert those submatrices, the criterion does not depend on any particular technique we could use for that purpose. It also exposes the limitations of the LU decomposition. For example, for any $n \in \mathbb{N}$ such that $n > 1$, the matrix

$$P = \begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix} = [\delta_{i \; n-j+1}]_{i=1,\, j=1}^{n \quad n} \in \mathbb{F}^{n \times n}$$

is invertible but has no r-step LU decomposition with any r. We address this deficiency by generalizing the LU decomposition in §§ II.5.7 and II.5.8.

Finally, there is a question of uniqueness. Definition II.5.6.1 introduces, for an arbitrary matrix, *an* r-step LU decomposition *of that matrix*, and it is logical to ask whether such a decomposition for any single matrix is unique whenever it exists. Having a positive answer to this question is important for the procedure devised in lemma II.5.6.3 and algorithm II.5.6.4: the uniqueness would mean that *all* existing r-step LU decompositions (of all matrices) could be constructed by algorithm II.5.6.4.

Lemma II.5.6.8 (uniqueness of an LU decomposition of a matrix)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m \times n}$. Then A has at most one r-step LU decomposition of A: if $A = LU$ and $A = \tilde{L}\tilde{U}$ are both r-step LU decompositions of A, then $\tilde{L} = L$ and $\tilde{U} = U$.

Proof. Let us use the partition (II.5.6.3) with $A_{11} \in \mathbb{F}^{r \times r}$ and partition $L, U, \tilde{L}, \tilde{U}$ as in definition II.5.6.1.

The first of equalities (II.5.6.4) gives $L_1 U_1 = \tilde{L}_1 \tilde{U}_1$. By definition II.5.6.1, lemma II.5.5.4 and corollary II.5.5.5, all the four matrices are invertible, which gives, in particular, $\tilde{L}_1^{-1} L_1 = \tilde{U}_1 U_1^{-1}$. Applying definition II.5.6.1, lemmata II.5.5.1 and II.5.5.4 and corollary II.5.5.5, we find that the left-hand side is a unit lower-triangular matrix and the right-hand side is an upper-triangular matrix. They both are therefore equal to the identity matrix of order r, which is the only matrix of size $r \times r$ that combines the two properties. This leads us to that $L_1 = \tilde{L}_1$ and $U_1 = \tilde{U}_1$.

For $r < m$, the second of equalities (II.5.6.4) and the above discussion yield $L_2 U_1 = \tilde{L}_2 U_1$ and hence $L_2 = \tilde{L}_2$ since U_1 is invertible.

Similarly, for $r < n$, the third of equalities (II.5.6.4) and the above discussion result in $L_1 U_2 = L_1 \tilde{U}_2$ and hence $U_2 = \tilde{U}_2$ because L_1 is invertible.

Finally, for $r < \min\{m, n\}$, the last of equalities (II.5.6.4) and the above argument lead to $S = \tilde{S}$.

We can combine the results of lemmata II.5.6.3, II.5.6.5, II.5.6.6 and II.5.6.8 into the following statement. In fact, the four lemmata could have been proven at once, in the course of same induction with respect to the number of steps of Gaussian elimination. Instead, we proved them separately here in order to better structure the analysis of the LU decomposition.

Theorem II.5.6.9. Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m \times n}$. Then an r-step LU decomposition of A exists if and only if all leading principal submatrices of A of orders $1, \ldots, r$ are invertible, in which case it is unique and is produced by algorithm II.5.6.4.

Let us emphasize the meaning and role and importance of the requirement of definition II.5.6.1 that the diagonal entries of U_1 are all nonzero.

Remark II.5.6.10 (not an r-step LU decomposition if step r cannot be explicitly performed)**.** In definition II.5.6.1, the first r diagonal entries of U are *required to be nonzero* for LU to be referred to as an r-step LU decomposition.

For example,

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

is a one-step LU decomposition, which is incomplete since the bottom right entry is nonzero, and is *not* a two-step LU decomposition (the second diagonal entry is zero). In fact, the matrix has no two-step LU decomposition by lemma II.5.6.6 and, since the one-step decomposition is incomplete, has therefore has no complete LU decomposition.

Let us now look at another example, obtained by removing the third row from the above matrix. In the sense of definition II.5.6.1, the representation

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

is still, as above, a (complete) one-step LU decomposition and is *not* a two-step LU decomposition. This matrix as well has no two-step LU decomposition, as lemma II.5.6.6 shows.

The requirement that the first r diagonal entries of U be nonzero and the associated terminological subtlety are mainly due to that, to keep terminology and analysis simple, we do not wish to call $A = LU$ with a unit lower-triangular matrix L and an upper-triangular matrix U an r-step LU decomposition of $A = LU$ in any case *when this decomposition cannot be produced by* r *steps of algorithm II.5.6.4*. Another reason is that by saying that matrices L and U form an r-step LU decomposition we wish to imply that the claim of proposition II.5.6.2 holds.

Consider a situation when $A = LU$ is an $(r-1)$-step LU decomposition and column r of U is zero below the diagonal (which means that the first column of the corresponding Schur complement S is zero below the diagonal). For step r of Gaussian elimination, the *pivot entry* is entry (r, r) of U, which is also entry $(1, 1)$ of S. The step can be performed if and only the pivot entry is nonzero. Such an elimination step, however, would not modify the matrix, so we can skip it altogether and proceed straight to the next column. When the pivot entry is zero, however, step r cannot be performed in the way we formally described. Skipping such a step would require a modification in algorithm II.5.6.4, on which we choose not to focus at this point. By requiring that the diagonal entries of U_1 be nonzero we omit such special cases from consideration — only to address them in a systematic way later.

When the theoretical construction is complete and the decomposition is fully understood in its more general form, this fine distinction may be not so important. For this reason, you can easily encounter in the literature the use of the term *LU decomposition* even for such representations as the above examples, but we will adhere to definition II.5.6.1 in the present course.

Definition II.5.6.11 (truncated r-step LU decomposition of a matrix)**.** In the context of definition II.5.6.1, matrices

$$\widehat{L} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \in \mathbb{F}^{m \times r} \quad \text{and} \quad \widehat{U} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{F}^{r \times n}\,, \tag{II.5.6.8}$$

which consist of the first r columns of L and of the first r rows of U respectively, are often called the *truncated* r*-step LU factors* of A. The matrix $\widehat{L}\,\widehat{U} \in \mathbb{F}^{m \times n}$ is often referred to as the r*-step LU approximation of* A.

Remark II.5.6.12 (*decomposition* vs. *factorization* vs. *representation*)**.** Note that the words *decomposition*, *factorization* and *representation* are typically used interchangeably in such collocations as *LU decomposition*, but some authors make various specific distinctions in this regard.

Remark II.5.6.13 (complete decomposition and zero approximation error)**.** In the context of definitions II.5.6.1 and II.5.6.11 with $r < \min\{m,n\}$, it is easy to see that an r-step LU decomposition $A = \widehat{L}\widehat{U}$ is complete if and only if the r-step LU *approximation* $\widehat{A} = \widehat{L}\widehat{U}$ of A introduces no error: $\widehat{A} = A$ is equivalent to that $r = \min\{m,n\}$ or $r < \min\{m,n\}$ but $S = 0$.

Example II.5.6.14 (incomplete LU decomposition and Schur complements)**.** In the context of examples II.3.2.3 and II.4.1.6 and (II.4.2.9), matrix $\widehat{A}$ is a two-step LU approximation of $A \in \mathbb{R}^{3\times 3}$. The error of that approximation is

$$A - \widehat{A} = \ell_3 u_3^\mathsf{T} = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 2 \end{bmatrix}, \tag{II.5.6.9}$$

which is the matrix of the same size as A and $\widehat{A}$ composed from entrywise errors. As we see, that error is concentrated in the two-step Schur complement $[\,2\,] \in \mathbb{R}^{1\times 1}$.

Let us now consider the LU decomposition in the context of the so-called *cross approximation*, which consists in using a certain number r of rows and columns of a matrix for constructing an approximation to the matrix that interpolates the matrix in those rows and columns and, hopefully, accurately approximates the matrix in the remaining entries.

Lemma II.5.6.15 (cross approximation and the LU decomposition)**.** Consider $m, n, r \in \mathbb{N}$ such that $r < \min\{m,n\}$, a matrix $A \in \mathbb{F}^{m\times n}$ partitioned as in (II.5.6.3) with $A_{11} \in \mathbb{F}^{r\times r}$ invertible and

$$C = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}, \quad R = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \quad \text{and} \quad G = A_{11}^{-1}. \tag{II.5.6.10}$$

Then

$$A - CGR = \begin{bmatrix} O & \\ & S \end{bmatrix} \quad \text{with} \quad S = A_{22} - A_{21} A_{11}^{-1} A_{12}, \tag{II.5.6.11}$$

where O is the zero square matrix of order r. In the cases of $r = m$ and $r = n$, the second block rows and second block columns vanish in (II.5.6.10) and (II.5.6.11).

Furthermore, if $A = LU$ is an r-step LU decomposition, then S is the r-step Schur complement of that decomposition and the truncated r-step LU factors $\widehat{L} \in \mathbb{F}^{m\times r}$ and $\widehat{U} \in \mathbb{F}^{r\times n}$ satisfy the following in terms of the partitions (II.5.6.8):

$$\widehat{L} = CU_1^{-1}, \quad \widehat{U} = L_1^{-1}R, \quad G = U_1^{-1}L_1^{-1} \quad \text{and} \quad \widehat{L}\widehat{U} = CGR. \tag{II.5.6.12}$$

Proof of lemma II.5.6.15. Let us first calculate the product CGR and compare it with A:

$$CGR = \begin{bmatrix} A_{11} A_{11}^{-1} A_{11} & A_{11} A_{11}^{-1} A_{12} \\ A_{21} A_{11}^{-1} A_{11} & A_{21} A_{11}^{-1} A_{12} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{21} A_{11}^{-1} A_{12} \end{bmatrix}.$$

In the cases of $r = m$ and $r = n$, the second block rows and second block columns vanish and $A = CGR$ holds. For $r < \min\{m, n\}$, subtracting CGR from A blockwise, we obtain (II.5.6.11).

If A has an r-step LU decomposition $A = LU$ with factors

$$L = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \in \mathbb{F}^{m\times m} \quad \text{and} \quad U = \begin{bmatrix} U_1 & U_2 \\ & S_\star \end{bmatrix} \in \mathbb{F}^{m\times n},$$

where L_1 and U_1 are the leading principal submatrices of L and U of order r. These leading principal submatrices are invertible by definition II.5.6.1 and parts (a) and (b). Equalities (II.5.6.4) then give $A_{11} = L_1 U_1$ and hence $G = U_1^{-1} L_1^{-1}$ and $A_{22} = L_2 U_2 + S_\star = (L_2 U_1)\, U_1^{-1} L_1^{-1} (L_1 U_2) + S_\star = A_{21} A_{11}^{-1} A_{12} + S_\star$, which implies $S_\star = S$ due to (II.5.6.11). Furthermore, equalities (II.5.6.4) yield also

$$C = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} L_1 U_1 \\ L_2 U_1 \end{bmatrix} = \widehat{L} U_1 \quad \text{and} \quad R = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} = \begin{bmatrix} L_1 U_1 & L_1 U_2 \end{bmatrix} = L_1 \widehat{U}\,.$$

Multiplying the factors, we immediately obtain $CGR = (\widehat{L} U_1)\, U_1^{-1} L_1^{-1} (L_1 \widehat{U}) = \widehat{L}\widehat{U}$.

The approximation CGR of A defined in lemma II.5.6.15 is often called the *cross approximation of A based on the first r rows and on the first r columns*. As the lemma shows, the r-step LU approximation of A, when it exists, actually coincides with that approximation, but the cross approximation may also be defined when A has no r-step LU decomposition.

§ II.5.7. Need for pivoting: detailed examples

As we see from algorithm II.5.6.4, each step of Gaussian elimination involves division by one entry of the current matrix. Such entries were carefully circled at each step of Gaussian elimination performed in (II.5.4.1) to (II.5.4.3) for the example considered in § II.5.4.

This special entry is often referred to as the *pivot entry* or *pivot element* of the respective step, and the row and column to which it belongs are called the *pivot row* and *pivot column* of the step. For example, in the proof of lemma II.5.6.5, the pivot entry of every step k is u_{kk}. In algorithm II.5.6.4 and in the proof of lemma II.5.6.5, row k, column k and entry (k, k) are always used as the pivot row, pivot column and pivot entry at each step k. As lemma II.5.6.6 shows, the resulting computational scheme is not universally applicable in the sense that a certain condition on leading principal submatrices of the matrix has to be satisfied. We will, however, see now that any nonzero entry of the current Schur complement can be used as a pivot entry even our usual candidate is zero.

Let us revisit the calculation that we discussed in detail in § II.5.4, where pivot entries were carefully circled.

Example II.5.7.1 (column pivoting)**.** The same calculation as (II.5.4.1)–(II.5.4.2) for a slightly modified matrix,

$$A = M_1^{-1} U_1 = \tilde{M}_1 U_1 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 7 \\ 6 & 18 & 22 \end{bmatrix} \tag{II.5.7.1}$$

produces

$$U_1 = M_1 U_0 = \begin{bmatrix} 1 & 2 & 3 \\ & \textcircled{0} & 1 \\ & 6 & 4 \end{bmatrix} \quad \text{with} \quad M_1 = \begin{bmatrix} 1 & & \\ -2 & 1 & \\ -6 & & 1 \end{bmatrix} . \tag{II.5.7.2}$$

For demonstration purposes, the matrix A considered here is such that the transformation M_1 performing the first step of Gaussian elimination is the same as the matrix M_1 considered in § II.5.4 and given by (II.5.4.5) and (II.5.4.6). This example shows another time that it may well be the case at the beginning of step k of Gaussian elimination that the kth diagonal entry is zero and the step therefore cannot be performed as usual. In fact, we know from theorem II.5.6.9 that matrix A has no two-step LU decomposition because if it had existed, it would be unique and we would be able to apply the procedure of lemma II.5.6.5 and algorithm II.5.6.4 to obtain it.

There is, however, a powerful remedy that allows to continue Gaussian elimination! We can find a nonzero entry in the same row and exchange columns in U_1. Let us set

$$\Sigma_2 = \begin{bmatrix} 1 & & \\ & & 1 \\ & 1 & \end{bmatrix} \tag{II.5.7.3}$$

and proceed for $U_1 \Sigma_2$ as we did for B_1 in (II.5.4.2), which we know now as the second step of algorithm II.5.6.4:

$$U_1 \Sigma_2 = \begin{bmatrix} 1 & 3 & 2 \\ & \textcircled{1} & 0 \\ & \mathbf{4} & 6 \end{bmatrix} = \tilde{M}_2 U_2 \quad \text{with} \quad \tilde{M}_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & \mathbf{4} & 1 \end{bmatrix} \quad \text{and} \quad U_2 = \begin{bmatrix} 1 & 3 & 2 \\ & 1 & 0 \\ & & 6 \end{bmatrix} .$$

It is easy to see that the column exchange we applied between the first and second steps of Gaussian elimination can be applied even before the first step. In fact, the first step can be performed in the very same way after such a column exchange as before:

$$A \Sigma_2 = \tilde{M}_1 U_1 \Sigma_2 = \tilde{M}_1 \tilde{M}_2 U_2 = L_2 U_2 \,, \tag{II.5.7.4}$$

where $L_2 = \tilde{M}_1 \tilde{M}_2$ is a unit lower-triangular matrix. Since Σ_2 is its own inverse and transpose, so we can recast (II.5.7.4) in the form

$$A \Sigma_2^{\mathsf{T}} = L_2 U_2 \,, \tag{II.5.7.5}$$

which obviously is a two-step LU decomposition.

Note that we *implicitly* relied on that the column exchange, realized by Σ_2, does not affect the first column, in which elimination occurred at the previous step, before we arrived at the idea of exchanging columns. This circumstance, seemingly minor, is important as it ensures that matrix $U_1 \Sigma_2$ is upper triangular (zero below the diagonal) in the first column. Then so is U_2, as we see in this example and in general (see the proof of lemma II.5.6.5).

Example II.5.7.2 (row pivoting)**.** The case of a zero diagonal entry, encountered in example II.5.7.1, could be handled differently. Indeed, for the same matrices A and U_1 as in (II.5.7.7), one could use a nonzero below the zero diagonal entry. Multiplying by $\Pi_2 = \Sigma_2$ on the left, where $\varSigma_2$ is given by (II.5.7.3), we interchange the second and third rows. Actually, this immediately produces a two-step LU decomposition for matrix $\varPi_2\,U_1$ with no need for any further computation:

$$\varPi_2\,U_1 = \begin{bmatrix} 1 & 2 & 3 \\ & \textcircled{6} & 4 \\ & \mathbf{0} & 1 \end{bmatrix} = \tilde{M}_2\,U_2 \quad \text{with} \quad \tilde{M}_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & \mathbf{0} & 1 \end{bmatrix} \quad \text{and} \quad U_2 = \begin{bmatrix} 1 & 2 & 3 \\ & 6 & 4 \\ & & 1 \end{bmatrix}.$$

Similarly to a column exchange, a row exchange can be put in front of Gaussian elimination, which means performing even *before* the first elimination step an exchange that we would otherwise select *in the course of* Gaussian elimination. There is, however, a caveat: every preceding step is a transformation of rows and needs therefore to be modified accordingly, so as to take into account the row exchange we wish to put in front. This is intuitively clear: the coefficients of Gaussian elimination from any step correspond to rows, and if two rows of the original matrix are exchanged ahead of the respective elimination step, then the coefficients should be exchanged in the same way. It is important that these coefficients only undergo an exchange but do not change otherwise, which is a direct consequence of that each row exchange is performed *only among the rows that were not used as pivot rows* at any of the preceding steps.

For our example, using the fact that matrix $\varPi_2$ is its own inverse and transpose, we obtain

$$\varPi_2\,A = \varPi_2\,\tilde{M}_1\,U_1 = (\varPi_2\,\tilde{M}_1\,\varPi_2^{\mathsf{T}})\,(\varPi_2\,U_1) = \tilde{M}_1^{\star}\tilde{M}_2\,U_2 = L_2\,U_2\,, \tag{II.5.7.6}$$

where we introduce $\tilde{M}_1^{\star} = \varPi_2\,\tilde{M}_1\,\varPi_2^{\mathsf{T}}$ and $L_2 = \tilde{M}_1^{\star}\tilde{M}_2 = \tilde{M}_1^{\star}$. Unlike column pivoting, discussed in example II.5.7.2, realizing that $\varPi_2\,A = L_2\,U_2$ has to be a two-step LU decomposition in *any* such a calculation requires a bit of thinking. The point is that matrix $\tilde{M}_1^{\star}$ is unit lower triangular.

Note that $\tilde{M}_1$ can be expressed in the form (II.5.5.4): $\tilde{M}_1 = I + \ell_1 e_1^{\mathsf{T}}$, where I is the identity matrix of order three, e_1 is its first column and ℓ_1 is the first column of strictly lower-triangular matrix $\tilde{M}_1 - I$. Considering the action of the exchange, we notice that it does not affect the first term of $\tilde{M}_1$: indeed, $\varPi_2 I \varPi_2^{\mathsf{T}} = \varPi_2\varPi_2^{\mathsf{T}} = I$. As for the second term, it is important to notice that it is nonzero only in the first column (which is due to the fact that $M_1 = \tilde{M}_1^{-1}$ is the elimination matrix of the *first* step). Since multiplication by $\varPi_2^{\mathsf{T}} = \varPi_2$ on the right exchanges the second and third columns and these are both zero in e_1^{T}, we immediately see that $e_1^{\mathsf{T}}\varPi_2^{\mathsf{T}} = e_1^{\mathsf{T}}$. So we have $\tilde{M}_1^{\star} = \varPi_2\tilde{M}_1\varPi_2^{\mathsf{T}} = I + (\varPi_2\ell_1)\,e_1^{\mathsf{T}}$. Similarly, multiplication by $\varPi_2$ on the left only exchanges the second and third rows, so the first component of $\varPi_2\ell_1$ is zero because that of ℓ_1 is so, and matrix $\tilde{M}_1^{\star}$ is therefore unit lower triangular. This leads us to the conclusion that $\varPi_2 A = L_2\,U_2$ is the two-step (also three-step and complete) LU decomposition of $\varPi_2 A$.

Example II.5.7.3 (row and column pivoting)**.** It is not difficult to imagine a situation requiring that both rows and columns be exchanged. For example, calculating as in (II.5.4.1), we obtain

$$U_1 = M_1 U_0 = \begin{bmatrix} 1 & 2 & 3 \\ & \enclose{circle}{0} & 0 \\ & 0 & 4 \end{bmatrix} \quad \text{for} \quad A = M_1^{-1} U_1 = \widetilde{M}_1 U_1 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 6 & 12 & 22 \end{bmatrix}, \tag{II.5.7.7}$$

where we use the matrices M_1 and $\widetilde{M}_1$ given by (II.5.4.5) and (II.5.4.6).

The first Schur complement (see definition II.5.6.1) has only one nonzero entry, which cannot be brought into position $(2, 2)$ by exchanging rows only or by exchanging columns only. On the other hand, if we allow both, that is possible. Setting $\varPi_2 = \varSigma_2$, where $\varSigma_2$ is given by (II.5.7.3), and applying row and column exchange, we immediately obtain a two-step LU decomposition for matrix $\varPi_2 \, U_1 \, \varSigma_2^{\mathsf{T}}$ with no need for any further computation:

$$\varPi_2 U_1 \varSigma_2^{\mathsf{T}} = \begin{bmatrix} 1 & 3 & 2 \\ & \enclose{circle}{4} & 0 \\ & \mathbf{0} & 0 \end{bmatrix} = \widetilde{M}_2 U_2 \quad \text{with} \quad \widetilde{M}_2 = \begin{bmatrix} 1 & & \\ & 1 & \\ & \mathbf{0} & 1 \end{bmatrix} \quad \text{and} \quad U_2 = \begin{bmatrix} 1 & 3 & 2 \\ & 4 & 0 \\ & & 0 \end{bmatrix}.$$

Then, in the same way as in examples II.5.7.1 and II.5.7.2, one discovers that

$$\varPi_2 A \varSigma_2^{\mathsf{T}} = L_2 U_2 \tag{II.5.7.8}$$

with $\widetilde{M}_1^\star = \varPi_2 \widetilde{M}_1 \varPi_2^{\mathsf{T}}$ and $L_2 = \widetilde{M}_1^\star \widetilde{M}_2 = \widetilde{M}_1^\star = I + (\varPi_2 \ell_1) \, e_1^{\mathsf{T}}$ is the two-step LU decomposition of the matrix $\varPi_2 A \, \varSigma_2^{\mathsf{T}}$. Here, as in example II.5.7.2, I is the identity matrix of order three, e_1 is its first column and ℓ_1 is the first column of strictly lower-triangular matrix $\widetilde{M}_1 - I$.

§ II.5.8. Pivoted LU decomposition

In this section, we consider a *pivoted LU decomposition* of a matrix, which is nothing else than an LU decomposition of the same matrix with rows and columns permuted.

Definition II.5.8.1 (r-step pivoted LU decomposition of a matrix)**.** Consider $m, n, r \in \mathbb{N}$ such that $r \leq \min\{m, n\}$. For each $k \in \{1, \ldots, r\}$, consider $\pi_k \in \{k, \ldots, m\}$ and $\sigma_k \in \{k, \ldots, n\}$ and let $\varPi_k$ be the (k, π_k)-exchange matrix of order m and $\varSigma_k$ be the (k, σ_k)-exchange matrix of order n. Let $P = \varPi_r \cdots \varPi_1$ and $Q = \varSigma_r \cdots \varSigma_1$.

For the matrix $\widetilde{A} = PAQ^{\mathsf{T}}$, assume that $\widetilde{A} = LU$ is an r-step LU decomposition in the sense of definition II.5.6.1. Then the equality $A = P^{\mathsf{T}} LUQ$, as a representation of A in terms of P, L, U and Q, is called an *r-step pivoted LU decomposition corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$.*

When the r-step LU decomposition $\widetilde{A} = LU$ is incomplete or complete, the r-step pivoted LU decomposition $A = P^{\mathsf{T}} LUQ$ is called *incomplete* or *complete*, respectively.

Clearly, a zero matrix has no pivoted LU decomposition: no matter how its rows and columns are permuted, the resulting matrix has no invertible leading principal submatrices of any order and hence has no LU decomposition by lemma II.5.6.6.

Definition II.5.8.1 defines *pivoted* LU decomposition in terms of an LU decomposition. So we can refer to lemmata II.5.6.5, II.5.6.6 and II.5.6.8 to establish the existence, nonexistence and uniqueness of pivoted LU decompositions for a specific matrix *once* we have fixed some column and row permutations (to which matrices P and Q in definition II.5.8.1 correspond).

Still, this approach is not constructive in the sense that it does not allow us to obtain pivoted LU decompositions in general settings and to make an educated choice of the permutations.

Remark II.5.8.2. Unlike the LU decomposition, the *pivoted* LU decomposition is essentially nonunique. For example, consider any $n, r \in \mathbb{N}$ such that $n \geq 2$ and $r \leq n$. Then the identity matrix I of order n satisfies $I = P^{\mathsf{T}}LUQ$ with $L = U = I$ and P and Q defined as in definition II.5.8.1 for any $\pi_k \in \{k, \ldots, n\}$ and $\sigma_k = \pi_k$ with $k \in \{1, \ldots, r\}$. Indeed, we then have $Q = P$ by definition II.5.8.1 and $P^{\mathsf{T}}P = I$ by proposition II.4.3.7. So $I = P^{\mathsf{T}}LUQ$ is an r-step pivoted LU decomposition of I.

Note that there are $n!/(n-r)!$ ways to choose $\pi_1, \ldots, \pi_r$. One can show that these choices result in the same number of distinct values for P, so that matrix I has exactly $n!/(n-r)!$ r-step pivoted LU decompositions with distinct permutation matrices P defined as in definition II.5.8.1.

However, for any matrix A, once row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$ have been fixed, an r-step pivoted LU decomposition of A corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$, as an r-step LU decomposition of PAQ^{T} with P and Q as specified in definition II.5.8.1, is unique by lemma II.5.6.8 if it exists, and a criterion of its existence is provided by lemmata II.5.6.5 and II.5.6.6.

Similarly to as in definition II.5.6.11, which followed the definition of *LU decomposition*, we now define *truncated pivoted LU decomposition*. By lemma II.5.6.8, it is unique for every matrix *and* every admissible choice of exchange indices.

Definition II.5.8.3 (truncated r-step pivoted LU decomposition of a matrix)**.** In the context of definition II.5.8.1, the *truncated r-step LU factors* $\widehat{L}$ and $\widehat{U}$ of PAQ^{T}, introduced in definition II.5.6.11, are called the *truncated r-step pivoted LU factors of A corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$.*

The matrix $\widehat{A} = P^{\mathsf{T}}\widehat{L}\widehat{U}Q$ is called the *r-step pivoted LU approximation of A corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$.*

Similarly to as in remark II.5.6.13, we note the following.

Remark II.5.8.4. In the context of definitions II.5.8.1 and II.5.8.3, it is easy to see from definition II.5.6.1 that an r-step pivoted LU decomposition $A = P^{\mathsf{T}}LUQ$ is complete if and only if the r-step pivoted LU *approximation* $\widehat{A} = P^{\mathsf{T}}\widehat{L}\widehat{U}Q$ of A is *exact*, i.e., $\widehat{A} = A$.

In examples II.5.7.1 to II.5.7.3, we observed that a combination of row and column pivoting *in the course of* Gaussian elimination allows to proceed further with Gaussian elimination. This solution immediately leads to a modification of the computational scheme formalized in lemma II.5.6.3 and algorithm II.5.6.4 in which elimination steps and pivoting steps are *interlaced*. Nevertheless, as we worked out in examples II.5.7.1 to II.5.7.3, this scheme can be equivalent to calculating an LU decomposition, complete or incomplete, of the original matrix with rows and columns permuted *before* elimination and according to the exchanges which are otherwise discovered *in the course of* elimination. This, in fact, means finding what we have just introduced in definition II.5.8.1 as a pivoted LU decomposition of the original matrix. As examples II.5.7.1 to II.5.7.3 show, putting each pivoting step in front of the entire procedure of Gaussian elimination requires a certain modification of the elimination steps preceding the pivoting step.

In the following lemma, we will generalize the illustration of pivoting that we considered in example II.5.7.3.

Lemma II.5.8.5 (iterated Gaussian elimination with pivoting)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$ be nonzero.

(a) Let I and e_1 denote the identity matrix of order m and its first column.

Assume that $\pi_1 \in \{1,\ldots,m\}$ and $\sigma_1 \in \{1,\ldots,n\}$ are such that $A_{\pi_1\sigma_1} \neq 0$ and let $\varPi_1$ be the $(1,\pi_1)$-exchange matrix of order m and $\varSigma_1$ be the $(1,\sigma_1)$-exchange matrix of order n. Then the matrix $\varPi_1 A \varSigma_1^\mathsf{T}$ has a one-step LU decomposition $\varPi_1 A \varSigma_1^\mathsf{T} = LU$ and $A = \varPi_1^\mathsf{T} LU \varSigma_1$ is a one-step pivoted LU decomposition corresponding to the row- and column-exchange indices π_1 and σ_1.

(b) Let $r, q \in \mathbb{N}$ be such that $r+q \leq \min\{m,n\}$. Assume that $A = P^\mathsf{T} LUQ$ is an r-step pivoted LU decomposition corresponding to row- and column-exchange indices $\pi_1,\ldots,\pi_r$ and $\sigma_1,\ldots,\sigma_r$. Consider a partitioning of L and U analogous to that of (II.5.6.6):

$$L = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} U_1 & U_2 \\ & S \end{bmatrix}, \tag{II.5.8.1}$$

where $L_1 \in \mathbb{F}^{r\times r}$, $L_2 \in \mathbb{F}^{(m-r)\times r}$, I_2 is the identity matrix of order $m-r$, $U_1 \in \mathbb{F}^{r\times r}$, $U_2 \in \mathbb{F}^{r\times(n-r)}$ and $S \in \mathbb{F}^{(m-r)\times(n-r)}$.

Further, assume that $S = \tilde{P}^\mathsf{T} \tilde{L}\tilde{U}\tilde{Q}$ is a q-step pivoted LU decomposition corresponding to row- and column-exchange indices $\tilde{\pi}_1,\ldots,\tilde{\pi}_q$ and $\tilde{\sigma}_1,\ldots,\tilde{\sigma}_q$. For each $k \in \{1,\ldots,q\}$, let

$$\pi_{r+k} = r + \tilde{\pi}_k \quad \text{and} \quad \sigma_{r+k} = r + \tilde{\sigma}_k\,. \tag{II.5.8.2}$$

Then A has an $(r+q)$-step pivoted LU decomposition $A = P_\star^\mathsf{T} L_\star U_\star Q_\star$ corresponding to the row- and column-exchange indices $\pi_1,\ldots,\pi_{r+q}$ and $\sigma_1,\ldots,\sigma_{r+q}$ with

$$L_\star = \begin{bmatrix} L_1 & \\ \tilde{P}L_2 & \tilde{L} \end{bmatrix} \quad \text{and} \quad U_\star = \begin{bmatrix} U_1 & U_2\tilde{Q}^\mathsf{T} \\ & \tilde{U} \end{bmatrix}. \tag{II.5.8.3}$$

Proof.

(a) Since $(\varPi_1 A \varSigma_1^\mathsf{T})_{11} = A_{\pi_1\sigma_1}$ and $A_{\pi_1\sigma_1} \neq 0$ by assumption, part (a) of lemma II.5.6.3 yields a one-step LU decomposition $\varPi_1 A \varSigma_1^\mathsf{T} = LU$. The claim follows by definition II.5.8.1 due to that $\varPi_1$ and $\varSigma_1$ are exchange matrices and are therefore their own inverses.

(b) To prove the claim, we need to show two points: that the equality $A = P_\star^\mathsf{T} L_\star U_\star Q_\star$ holds and that the factors satisfy the conditions imposed in the definition of an $(r+q)$-step pivoted LU decomposition (definition II.5.8.1).

Let I_1 denote the identity matrix of order p and consider

$$\bar{P} = \begin{bmatrix} I_1 & \\ & \tilde{P} \end{bmatrix} \quad \text{and} \quad \bar{Q} = \begin{bmatrix} I_1 & \\ & \tilde{Q} \end{bmatrix}. \tag{II.5.8.4}$$

By definition II.5.8.1, the given r-step pivoted LU decomposition implies the r-step LU decomposition $PAQ^\mathsf{T} = LU$. We can transform this decomposition as follows using the given r-step pivoted LU decomposition of the Schur complement S:

$$
\begin{aligned}
PAQ^{\mathsf{T}} = \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \\ & \tilde{P}^{\mathsf{T}} \tilde{L} \tilde{U} \tilde{Q} \end{bmatrix} &= \begin{bmatrix} L_1 & \\ L_2 & I_2 \end{bmatrix} \begin{bmatrix} I_1 & \\ & \tilde{P}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} U_1 & U_2 \tilde{Q}^{\mathsf{T}} \\ & \tilde{L} \tilde{U} \end{bmatrix} \begin{bmatrix} I_1 & \\ & \tilde{Q} \end{bmatrix} \\
= \begin{bmatrix} L_1 & \\ L_2 & \tilde{P}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} U_1 & U_2 \tilde{Q}^{\mathsf{T}} \\ & \tilde{L} \tilde{U} \end{bmatrix} \bar{Q} \overset{*}{=} \begin{bmatrix} I_1 & \\ & \tilde{P}^{\mathsf{T}} \end{bmatrix} & \begin{bmatrix} L_1 & \\ \tilde{P} L_2 & I_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \tilde{Q}^{\mathsf{T}} \\ & \tilde{L} \tilde{U} \end{bmatrix} \bar{Q} \\
&= \bar{P}^{\mathsf{T}} \begin{bmatrix} L_1 & \\ \tilde{P} L_2 & I_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \tilde{Q}^{\mathsf{T}} \\ & \tilde{L} \tilde{U} \end{bmatrix} \bar{Q} \, , \qquad \text{(II.5.8.5)}
\end{aligned}
$$

where the step marked with $*$ shows that the row permutations performed on S (after the first p steps performed on A) can be applied *a priori* and that those permutations have to be applied to the elimination coefficients corresponding to rows $r+1$ to $r+q$ obtained at the first r steps, which are collected in L_2. This is a generalization of how $\tilde{M}_1$ was replaced with $\tilde{M}_1^{\star}$ in example II.5.7.2.

Further, (II.5.8.5) yields

$$
(\bar{P}P)A(\bar{Q}Q)^{\mathsf{T}} = \begin{bmatrix} L_1 & \\ \tilde{P} L_2 & I_2 \end{bmatrix} \begin{bmatrix} U_1 & U_2 \tilde{Q}^{\mathsf{T}} \\ & \tilde{L} \tilde{U} \end{bmatrix} ,
$$

which is an r-step LU decomposition with the corresponding Schur complement given in a q-step LU decomposition. Applying part (b) of lemma II.5.6.3, we obtain the $(r+q)$-step LU decomposition

$$
(\bar{P}P)A(\bar{Q}Q)^{\mathsf{T}} = L_{\star} U_{\star} \, . \qquad \text{(II.5.8.6)}
$$

It therefore remains to show that $\bar{P}P = P_{\star}$ and $\bar{Q}Q = Q_{\star}$, i.e., that those matrices are products of the exchange matrices corresponding (in the sense of definition II.5.8.1) to the exchange indices specified in the statement.

For each $k \in \{1, \ldots, r+q\}$, let $\mathit{\Pi}_k$ be the (k, π_k)-exchange matrix of order m and $\mathit{\Sigma}_k$ be the (k, σ_k)-exchange matrix of order n. Similarly, for each $k \in \{1, \ldots, q\}$, let $\tilde{\mathit{\Pi}}_k$ be the $(k, \tilde{\pi}_k)$-exchange matrix of order $m-r$ and $\tilde{\mathit{\Sigma}}_k$ be the $(k, \tilde{\sigma}_k)$-exchange matrix of order $n-r$.

For the permutation matrices involved in the given and claimed pivoted LU decompositions, according to definition II.5.8.1, we have the following factorizations:

$$
P = \mathit{\Pi}_r \cdots \mathit{\Pi}_1 \quad \text{and} \quad Q = \mathit{\Sigma}_r \cdots \mathit{\Sigma}_1
$$

for the outer decomposition,

$$
\tilde{P} = \tilde{\mathit{\Pi}}_q \cdots \tilde{\mathit{\Pi}}_1 \quad \text{and} \quad \tilde{Q} = \tilde{\mathit{\Sigma}}_q \cdots \tilde{\mathit{\Sigma}}_1
$$

for the inner decomposition,

$$
P_{\star} = \mathit{\Pi}_{r+q} \cdots \mathit{\Pi}_1 \quad \text{and} \quad Q_{\star} = \mathit{\Sigma}_{r+q} \cdots \mathit{\Sigma}_1
$$

for the aggregate decomposition.

For each $k \in \{1, \ldots, q\}$, due to (II.5.8.2), we have

$$
\mathit{\Pi}_{r+k} = \begin{bmatrix} I_1 & \\ & \tilde{\mathit{\Pi}}_k \end{bmatrix} \quad \text{and} \quad \mathit{\Sigma}_{r+k} = \begin{bmatrix} I_1 & \\ & \tilde{\mathit{\Sigma}}_k \end{bmatrix} . \qquad \text{(II.5.8.7)}
$$

Using matrix multiplication in block form, we then obtain from (II.5.8.4) that

$$\bar{P} = \varPi_{r+q} \cdots \varPi_{r+1} \quad \text{and} \quad \bar{Q} = \varSigma_{r+q} \cdots \varSigma_{r+1} \,.$$

The above factorizations then imply $P_\star = \bar{P}P$ and $Q_\star = \bar{Q}Q$, so that the claimed pivoted LU decomposition follows from (II.5.8.6).

A logical sequel to lemma II.5.8.5 is the following result, obtained by iterating the single-step procedure.

Lemma II.5.8.6 (pivoted LU decomposition of a matrix)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$ be nonzero. Then the iterative application of lemma II.5.8.5 produces a complete pivoted LU decomposition of the matrix A. Specifically, the following statements hold with *some* number $r \in \mathbb{N}$ of steps such that $r \leq \min\{m, n\}$.

(a) Lemma II.5.8.5 applied with $k = 1$ produces row- and column-exchange indices $\pi_1 \in \{1, \ldots, m\}$ and $\sigma_1 \in \{1, \ldots, n\}$ and a one-step pivoted LU decomposition of A corresponding to the row- and column-exchange indices π_1 and σ_1.

(b) For every $k \in \{2, \ldots, r\}$, lemma II.5.8.5 produces row- and column-exchange indices $\pi_k \in \{k, \ldots, m\}$ and $\sigma_k \in \{k, \ldots, n\}$ and a k-step pivoted LU decomposition of A corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_k$ and $\sigma_1, \ldots, \sigma_k$ from the previously produced $(k-1)$-step pivoted LU decomposition of A corresponding to the previously produced row- and column-exchange indices $\pi_1, \ldots, \pi_{k-1}$ and $\sigma_1, \ldots, \sigma_{k-1}$.

(c) For every $k \in \{1, \ldots, r\}$, the k-step pivoted LU decomposition that is obtained as described in statements (a) and (b) is incomplete if $k < r$ and complete if $k = r$.

Proof. Since A is nonzero, the assumptions of lemma II.5.8.5 with $k = 1$ are satisfied, and A has an one-step pivoted LU decomposition corresponding to some row- and column-exchange indices π_1 and σ_1.

Consider $k \in \mathbb{N}$ such that $1 < k \leq \min\{m, n\}$ and assume that A has an $(k-1)$-step pivoted LU decomposition corresponding to some row- and column-exchange indices $\pi_1, \ldots, \pi_{k-1}$ and $\sigma_1, \ldots, \sigma_{k-1}$. If this decomposition is incomplete, the assumptions of lemma II.5.8.5 are satisfied with some $\pi_k \in \{k, \ldots, m\}$ and $\sigma_k \in \{k, \ldots, n\}$. Then the lemma gives a k-step pivoted LU decomposition of A corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_k$ and $\sigma_1, \ldots, \sigma_k$.

Applying this argument iteratively, we prove the claim.

§ II.5.9. Implications of the pivoted LU decomposition for matrix rank

As we will see in this section, the pivoted LU decomposition can serve as a universal technique for understanding, calculating and computing the rank of any matrix. To start with, let us note that a complete r-step pivoted LU decomposition of a matrix provides an *exact* rank-r factorization of the matrix.

Lemma II.5.9.1 (a complete r-step decomposition implies a rank-r factorization)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m \times n}$ be nonzero. Assume A has an r-step pivoted LU decomposition that is complete. Then $\operatorname{rank} A \leq r$.

Proof. Let $A = P^\mathsf{T} LUQ$ be a complete r-step pivoted LU decomposition. Then, as we mentioned in remark II.5.6.13, we have $PAQ^\mathsf{T} = \widehat{L}\widehat{U}$, where $\widehat{L} \in \mathbb{F}^{m\times r}$ and $\widehat{U} \in \mathbb{F}^{r\times n}$ are the truncated r-step LU factors of PAQ^T (see definition II.5.6.11). So we have $A = VW$ with a tall matrix $V = P^\mathsf{T}\widehat{L} \in \mathbb{F}^{m\times r}$ and a wide matrix $W = \widehat{U}Q \in \mathbb{F}^{r\times n}$, which implies that $\operatorname{rank} A \leq r$ by proposition II.4.4.4.

Let us now use our fundamental results on the pivoted LU decomposition of matrices to prove that a square matrix has full rank if an only if it is invertible. We split this statement into two implications and prove them separately as lemmata II.5.9.2 and II.5.9.3.

Lemma II.5.9.2 (any invertible matrix is full rank)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$ be invertible. Then $\operatorname{rank} A = n$.

Proof. Let us first prove the following: if $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ is such that $\operatorname{rank} A < n$, then there exists $x \in \mathbb{F}^n$ nonzero such that Ax is zero.

If A is zero, then any nonzero column vector $x \in \mathbb{F}^n$ fulfills the claim. So let us assume that A is nonzero. By proposition II.4.4.4, there exist matrices $V \in \mathbb{F}^{m\times(n-1)}$ and $W \in \mathbb{F}^{(n-1)\times n}$ such that $A = VW$. Note that W is nonzero since A is nonzero. By lemma II.5.8.6, matrix W has a complete r-step pivoted LU decomposition $W = P^\mathsf{T} LUQ$ for some $r \in \{1, \ldots, n-1\}$. Truncating that complete decomposition, we obtain $W = P^\mathsf{T}\widehat{L}\widehat{U}Q$. Due to definition II.5.8.1, the truncation yields a unit lower-trapezoid matrix $\widehat{L} \in \mathbb{F}^{(n-1)\times r}$ and an upper-trapezoid matrix $\widehat{U} \in \mathbb{F}^{r\times n}$ with no zeros on the diagonal. Note also that $P \in \mathbb{F}^{(n-1)\times(n-1)}$ and $Q \in \mathbb{F}^{n\times n}$ are permutation matrices (by the same definition II.5.8.1), so that they are both inverted by transposition. Then we have $PWQ^\mathsf{T} = \widehat{L}\widehat{U}$.

The matrix $\widehat{U}$ is wide since $r \leq n-1 < n$, and its leading principal submatrix U_1 of order r is square upper triangular with no zeros on the diagonal and is therefore invertible by part (b) of corollary II.5.5.5. Then part (c) of lemma II.5.3.1 guarantees the existence of a *nonzero* column vector $\widehat{x} \in \mathbb{F}^n$ such that $\widehat{U}\widehat{x} = 0$. As a consequence, $x = Q^\mathsf{T}\widehat{x} \in \mathbb{F}^n$ is also nonzero and satisfies $Wx = P^\mathsf{T}(PWQ^\mathsf{T})\widehat{x} = P^\mathsf{T}\widehat{L}\widehat{U}x = 0$ and hence $Ax = VW\widehat{x} = 0$.

As a consequence, when $m = n$, the matrix A cannot be invertible. Indeed, combining the above result with the definition of the inverse matrix (definition II.4.1.14), we conclude that $x = A^{-1}Ax$ is zero. This contradicts that x is nonzero. When A is invertible, the inequality $\operatorname{rank} A < n$ is therefore false, which leaves us with the only possible value: $\operatorname{rank} A = n$ since $\operatorname{rank} A \leq n$ by lemma II.4.4.2.

Lemma II.5.9.2 finally fills one particular lacuna in our understanding of matrix rank, which we noted in remark II.4.4.14: we can now conclude that every identity matrix has full (maximum) rank. Since we proved the same for any invertible matrix in lemma II.5.9.2, there is no use developing the argument initiated in remark II.4.4.14. There is a converse statement to be proven: a full-rank square matrix is invertible. This is another fact that we now, with the pivoted LU decomposition at our disposal, are in position to prove.

Lemma II.5.9.3 (any full-rank square matrix is invertible)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$ be such that $\operatorname{rank} A = n$. Then matrix A is invertible.

Proof. Since $\operatorname{rank} A = n \in \mathbb{N}$, matrix A is nonzero. By lemma II.5.8.6, matrix A has a complete pivoted LU decomposition. Specifically, this implies that, for some number $r \in \mathbb{N}$ of steps, there exist a unit lower-triangular matrix $\widehat{L} \in \mathbb{F}^{n\times r}$, an upper-triangular matrix $\widehat{U} \in \mathbb{F}^{r\times n}$ with no zeros on the diagonal and invertible matrices $P \in \mathbb{F}^{n\times n}$ and $Q \in \mathbb{F}^{n\times n}$ such that $r \le n$ and

$$A = P^{-1}\widehat{L}\widehat{U}Q = (P^{-1}\widehat{L})(\widehat{U}Q)\,.$$

By proposition II.4.4.4, this implies $\operatorname{rank} A \le r$, and so $r = n$. The matrices $\widehat{L}$ and $\widehat{U}$ are therefore square. Applying part (b) of lemma II.5.5.4 to $\widehat{L}$ and part (b) of corollary II.5.5.5 to $\widehat{U}$, we obtain that both the matrices are invertible. Then, using proposition II.4.1.22, we conclude that A is invertible as a product of invertible matrices.

In the definition of the inverse matrix (definition II.4.1.14), we required that both the possible products of a matrix and a candidate for its inverse be equal to the identity matrix. Using lemma II.5.9.2, we can now show that inspecting one of the two products is sufficient.

Lemma II.5.9.4. Let $n \in \mathbb{N}$, $A, B \in \mathbb{F}^{n\times n}$ and I be the identity matrix of order n. Then the following statements hold:

(a) if $AB = I$, then A is invertible and $A^{-1} = B$;

(b) if $BA = I$, then A is invertible and $A^{-1} = B$.

Proof. Let us prove part (a). Note that lemma II.4.4.2 and proposition II.4.4.11 yield, respectively, $\operatorname{rank} A \le n$ and $\operatorname{rank} I \le \operatorname{rank} A$. On the other hand, we have $\operatorname{rank} I = n$ due to lemma II.5.9.2, so that $\operatorname{rank} A = n$. By lemma II.5.9.3, A is invertible. Then $AB = I$ implies $A^{-1}AB = A^{-1}I$, i.e., $B = A^{-1}$.

The proof of part (b) is completely analogous and is left to the reader as an exercise.

Using that the rank of a submatrix of a matrix does not exceed the rank of the matrix, we can strengthen the statement of lemma II.5.9.2 as follows.

Corollary II.5.9.5 (rank is at least the order of an invertible submatrix)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \le \min\{m, n\}$ and assume that $A \in \mathbb{F}^{m\times n}$ has an invertible square submatrix of order r. Then $r \le \operatorname{rank} A$.

Proof. Let $\widehat{A}$ be an invertible square submatrix of A of order r. Then $\operatorname{rank} \widehat{A} = r$ by lemma II.5.9.2 and $\operatorname{rank} \widehat{A} \le \operatorname{rank} A$ by proposition II.4.4.9. So we conclude that $r \le \operatorname{rank} A$, which proves the claim.

Lemma II.5.9.6 (number of steps in a pivoted LU decomposition and “determinantal rank“)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \le \min\{m, n\}$ and $A \in \mathbb{F}^{m\times n}$ be nonzero. Assume that A has an r-step pivoted LU decomposition. Then A has an invertible square submatrix of order r.

Proof. Let $A = P^{\mathsf{T}} LUQ$ be an r-step pivoted LU decomposition. Since $PAQ^{\mathsf{T}} = LU$ is an r-step LU decomposition, lemma II.5.6.6 yields that the leading principal submatrix $\widetilde{A}_{11}$ of PAQ^{T} of order r is invertible. Since P and Q are permutation matrices, that submatrix is a submatrix of A, and so A has an invertible square submatrix of order r.

Specifically, since P and Q are permutation matrices, there exist unique distinct row indices $\pi_1, \ldots, \pi_r \in \{1, \ldots, m\}$ and unique distinct column indices $\sigma_1, \ldots, \sigma_r \in \{1, \ldots, n\}$ such that $P_{k\pi_k} = 1$ and $Q_{k\sigma_k} = 1$ for each $k \in \{1, \ldots, r\}$ (see proposition II.4.3.5). Then $(PAQ^{\mathsf{T}})_{ij} = A_{\pi_i \sigma_j}$ for all $i, j \in \{1, \ldots, r\}$. The indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$ can be extracted directly from P and Q whenever P and Q are available in entrywise form. Alternatively, if we have P and Q parametrized by exchange indices, as in definition II.5.8.1 (note that the indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$ have a different meaning here), then we can evaluate the row and column permutations $(\pi_1, \ldots, \pi_m)$ and $(\sigma_1, \ldots, \sigma_n)$ associated with P and Q by composing the respective exchanges. Selecting the first r components in each of the two tuples, representing the permutations, we obtain the row and column indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$.

Now we are ready to put the finishing touches on our result regarding the existence of complete pivoted LU decompositions. In fact, we can now prove that the number of steps in a complete decomposition is always equal to the rank of the matrix.

Theorem II.5.9.7 (pivoted LU decomposition and matrix rank)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m\times n}$ be a nonzero matrix. Then the following statements hold:

(a) A has an invertible square submatrix of order r if and only if $r \leq \operatorname{rank} A$;

(b) an r-step pivoted LU decomposition of A exists if and only if $r \leq \operatorname{rank} A$;

(c) an r-step pivoted LU decomposition of A is complete if and only if $r = \operatorname{rank} A$.

Proof. The proof consists in carefully putting together several results obtained in §§ II.5.8 and II.5.9. Splitting each part of the claim into two implications, we obtain the following six statements to verify:

(i) if A has an invertible square submatrix of order r, then $r \leq \operatorname{rank} A$;

(ii) if A has an r-step pivoted LU decomposition, then $r \leq \operatorname{rank} A$;

(iii) if an r-step pivoted LU decomposition of A is complete, then $r = \operatorname{rank} A$;

(iv) if $r \leq \operatorname{rank} A$, then A has an r-step pivoted LU decomposition;

(v) if $r = \operatorname{rank} A$, then any r-step pivoted LU decomposition of A is complete.

(vi) if $r \leq \operatorname{rank} A$, then A has an invertible square submatrix of order r;

First, statement (i) trivially follows from corollary II.5.9.5.

Second, lemma II.5.9.6 shows that if A has an r-step pivoted LU decomposition, then A has an invertible square submatrix of order r. Applying statement (i), we conclude that $r \leq \operatorname{rank} A$. This proves statement (ii).

Third, applying lemma II.5.9.1 to any complete r-step decomposition of A, we obtain $\operatorname{rank} A \leq r$. On the other hand, we have $r \leq \operatorname{rank} A$ by statement (ii). These two inequalities together result in $r = \operatorname{rank} A$. This proves statement (iii).

Fourth, lemma II.5.8.6 guarantees the existence of a complete p-step pivoted LU decomposition $A = P_\star^\mathsf{T} L_\star U_\star Q_\star$, where $p \in \mathbb{N}$ is such that $p \leq \min\{m, n\}$. Applying statement (iii), we conclude that $p = \operatorname{rank} A$. It therefore remains to prove that, for any $r \in \{1, \dots, p-1\}$, there exists an r-step pivoted LU decomposition of A.

Due to definition II.5.8.1, we have $P_\star = \mathit{\Pi}_p \cdots \mathit{\Pi}_1$ and $Q_\star = \mathit{\Sigma}_p \cdots \mathit{\Sigma}_1$, where, for each $k \in \{1, \dots, p\}$, the factors $\mathit{\Pi}_k$ and $\mathit{\Sigma}_k$ are the (k, π_k)- and (k, σ_k)-exchange matrices of orders m and n, respectively. Let us consider $r \in \{1, \dots, p-1\}$, the permutation matrices $P = \mathit{\Pi}_r \cdots \mathit{\Pi}_1$ and $Q = \mathit{\Sigma}_r \cdots \mathit{\Sigma}_1$, accumulating the exchanges from the first r steps, and the permutation matrices $\bar{P} = \mathit{\Pi}_p \cdots \mathit{\Pi}_{r+1}$ and $\bar{Q} = \mathit{\Sigma}_p \cdots \mathit{\Sigma}_{r+1}$, accumulating the exchanges from steps $r+1$ to p. Then we have $P_\star = \bar{P}P$ and $Q_\star = \bar{Q}Q$. Since $P_\star A Q_\star^\mathsf{T} = L_\star U_\star$ is a p-step LU decomposition (by definition II.5.8.1), lemma II.5.6.6 yields, in particular, that the leading principal submatrices of $P_\star A Q_\star^\mathsf{T}$ of orders $1, \dots, r$ are invertible. Note that, since $\pi_k \geq k > r$ and $\sigma_k \geq k > r$ for each $k \in \{r+1, \dots, p\}$, the leading principal submatrices of orders $1, \dots, r$ of $PAQ^\mathsf{T} = \bar{P}^\mathsf{T}(P_\star A Q_\star^\mathsf{T})\bar{Q}$ are exactly the same as those of $P_\star A Q_\star^\mathsf{T}$ and are therefore invertible. Then, by lemma II.5.6.5, the matrix PAQ^T has an r-step LU decomposition $PAQ^\mathsf{T} = LU$. Applying definition II.5.8.1, we see that $A = P^\mathsf{T} LUQ$ is an r-step LU decomposition of A. This argument proves statement (iv).

Fifth, if A has an r-step pivoted LU decomposition that is not complete, then it follows from lemma II.5.8.5 that A has an $(r+1)$-step pivoted LU decomposition. By statement (ii), this implies $r + 1 \leq \operatorname{rank} A$. So this is not possible when $r = \operatorname{rank} A$. This proves statement (v).

Sixth, if $r \leq \operatorname{rank} A$, then A has an r-step pivoted LU decomposition by statement (iv). Then lemma II.5.9.6 ensures the existence of an invertible square submatrix of order r in A. This proves statement (vi).

We can now revisit lemma II.5.8.6 and apply theorem II.5.9.7 to specify that the number $r \in \mathbb{N}$ stated to exist in the lemma is actually nothing else than the rank of the matrix.

Theorem II.5.9.8 (pivoted LU decomposition of a matrix, revisited)**.** The statement of lemma II.5.8.6 holds with and only with $r = \operatorname{rank} A$.

Theorems II.5.9.7 and II.5.9.8 are master theorems of chapter II. They show two important results.

First, the rank of a matrix, as it was defined in § II.4.4, can be found for *any* matrix using the *computational algorithm* presented in lemma II.5.8.5. In fact, it is not just the rank what it gives us: the algorithm produces a complete LU decomposition of the matrix with appropriately permuted rows and columns. This allows, for example, to solve a linear system, see (II.5.1.1), using lemma II.5.3.2. Doing so requires inverting certain leading principal submatrices, which are in this case triangular and can therefore be inverted with the help of lemma II.5.5.4 and corollary II.5.5.5.

Second, we have established that the rank of a matrix, as it was defined in § II.4.4, is nothing else than the order of its largest invertible square submatrix. In fact, this equivalent characteristic is often used to define matrix rank and is known as the *determinantal rank* of a matrix. Such a naming convention will make sense once we have studied determinants later in the course. Let us now use this interpretation of matrix rank to prove a useful auxiliary result.

Lemma II.5.9.9 (rank of a block-diagonal matrix)**.** For $m_1, m_2, n_1, n_2 \in \mathbb{N}$, let us consider matrices $A_1 \in \mathbb{F}^{m_1 \times n_1}$, $A_2 \in \mathbb{F}^{m_2 \times n_2}$ and

$$A = \begin{bmatrix} A_1 & \\ & A_2 \end{bmatrix} \in \mathbb{F}^{(m_1+m_2)\times(n_1+n_2)} .$$

Then $\operatorname{rank} A = \operatorname{rank} A_1 + \operatorname{rank} A_2$.

Proof. For convenience, let us set $p_1 = \operatorname{rank} A_1$, $p_2 = \operatorname{rank} A_2$, $m = m_1 + m_2$, $n = n_1 + n_2$ and $p = p_1 + p_2$. When A_1 or A_2 is zero, the claim follows trivially, so we assume for the remainder of the proof that both A_1 and A_2 are nonzero.

First, let us prove that $\operatorname{rank} A \leq p$. By corollary II.4.4.12, $p_1, p_2 \in \mathbb{N}$ and there exist matrices $U_1 \in \mathbb{F}^{m_1 \times p_1}$, $V_1 \in \mathbb{F}^{n_1 \times p_1}$, $U_2 \in \mathbb{F}^{m_2 \times p_2}$ and $V_2 \in \mathbb{F}^{n_2 \times p_2}$ such that $A_1 = U_1 V_1^{\mathsf{T}}$ and $A_2 = U_2 V_2^{\mathsf{T}}$. Then the matrices

$$U = \begin{bmatrix} U_1 & \\ & U_2 \end{bmatrix} \in \mathbb{F}^{m \times p} \quad \text{and} \quad V = \begin{bmatrix} V_1 & \\ & V_2 \end{bmatrix} \in \mathbb{F}^{n \times p}$$

satisfy $A = UV^{\mathsf{T}}$. Applying proposition II.4.4.4, we arrive at the bound $\operatorname{rank} A \leq p$.

Second, let us prove that $\operatorname{rank} A \geq p$. By part (a) of theorem II.5.9.7, the matrices A_1 and A_2 have invertible submatrices of orders p_1 and p_2 respectively. According to definition II.4.2.3, this means the following: for each $k \in \{1, 2\}$, there exist distinct row indices $\pi_{ki} \in \{1, \ldots, m_k\}$ with $i \in \{1, \ldots, p_k\}$ and distinct column indices $\sigma_{kj} \in \{1, \ldots, n_i\}$ with $j \in \{1, \ldots, p_k\}$ such that the corresponding submatrix

$$\widehat{A}_k = \left[(A_k)_{\pi_{ki}\sigma_{kj}}\right]_{i=1,\, j=1}^{p_k \quad p_k} \in \mathbb{F}^{p_k \times p_k}$$

of A_k is invertible. Let us set $\pi_i = \pi_{1i}$ for $i \in \{1, \ldots, p_1\}$ and $\pi_{p_1+i} = m_1 + \pi_{2i}$ for $i \in \{1, \ldots, p_2\}$. Similarly, let us set $\sigma_j = \sigma_{1j}$ for $j \in \{1, \ldots, p_1\}$ and $\sigma_{p_1+j} = n_1 + \sigma_{2j}$ for $j \in \{1, \ldots, p_2\}$. Then the row indices $\pi_i \in \{1, \ldots, m\}$ with $i \in \{1, \ldots, p\}$ are distinct and the column indices $\sigma_j \in \{1, \ldots, n\}$ with $j \in \{1, \ldots, p\}$ are also distinct. The corresponding submatrix

$$\widehat{A} = [A_{\pi_i \sigma_j}]_{i=1,\, j=1}^{p \quad p} \in \mathbb{F}^{p \times p}$$

of A satisfies

$$\widehat{A} = \begin{bmatrix} \widehat{A}_1 & \\ & \widehat{A}_2 \end{bmatrix} \quad \text{and hence} \quad \widehat{A}^{-1} = \begin{bmatrix} \widehat{A}_1^{-1} & \\ & \widehat{A}_2^{-1} \end{bmatrix} ,$$

i.e., is invertible. Then part (a) of theorem II.5.9.7 yields $\operatorname{rank} A \geq p = p_1 + p_2 = \operatorname{rank} A_1 + \operatorname{rank} A_2$.

Let us now revisit the cross approximation, upon which we touched in the context of the LU decomposition in § II.5.6, and cover its most general form. For the the cross approximation of a matrix A built upon the first r rows and the first r columns of A, we can augment our analysis, given in lemma II.5.6.15, with the following result.

Lemma II.5.9.10. In the context of lemma II.5.6.15, we have $\operatorname{rank} CGR = r$ and

$$\operatorname{rank}(A - CGR) = \operatorname{rank} A - r = \begin{cases} \operatorname{rank} S & \text{if } r < \min\{m, n\}, \\ 0 & \text{if } r = \min\{m, n\}. \end{cases}$$

Proof. On the one hand, $\operatorname{rank} CGR = \operatorname{rank}(CG)R \leq r$ by proposition II.4.4.4 since CG and R are the factors of a rank-r factorization of CGR. On the other hand, as (II.5.6.11) shows, the leading principal submatrix of CGR is A_{11}, which is invertible by the assumption of lemma II.5.6.15. Part (a) of theorem II.5.9.7 then yields $\operatorname{rank} CGR \geq r$. Combining the two inequalities, we obtain $\operatorname{rank} CGR = r$.

When $r = \min\{m, n\}$, the block S vanish in (II.5.6.11), so that $A = CGR$ and hence $\operatorname{rank}(A - CGR) = 0$. Let us assume $r < \min\{m, n\}$ for the remainder of the proof.

From (II.5.6.11), we deduce that $\operatorname{rank}(A - CGR) = \operatorname{rank} S$ by lemma II.5.9.9 (one can also argue in a simpler way: $\operatorname{rank} S \leq \operatorname{rank}(A - CGR)$ by proposition II.4.4.9, and the factors of any factorization of S can be extended with zeros so as to form a factorization of $A - CGR$ of the same rank).

It remains to prove $\operatorname{rank} S = \operatorname{rank} A - r$. To this end, let us denote the identity matrices of orders r, $m - r$ and $n - r$ by I_1, I_2 and I_2'. First, consider the block-triangular matrices

$$X = \begin{bmatrix} I_1 & \\ -A_{21}A_{11}^{-1} & I_2 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} I_1 & \\ A_{21}A_{11}^{-1} & I_2 \end{bmatrix} .$$

Direct multiplication according to definition II.4.1.3 shows YX is the identity matrix of order m. Then X and Y are mutually inverse by lemma II.5.9.4. In terms of the partitioning given in (II.5.6.3) with $A_{11} \in \mathbb{F}^{r\times r}$, we note that the multiplication of A by X on the left eliminates the block A_{21} in the lower-triangular part:

$$A = (YX)A = Y(XA) = Y \begin{bmatrix} A_{11} & A_{12} \\ & S \end{bmatrix} , \tag{II.5.9.1}$$

cf. (II.5.6.7). The block elimination performed in (II.5.9.1) is nothing else than a block analogue of the standard Gaussian elimination we considered in part (a) of lemma II.5.6.3 (and hence in algorithm II.5.6.4) and a multi-row analogue of the multi-column but single-row elimination performed in (II.5.5.5).

Let us now eliminate the block A_{12} in the upper-triangular part in a similar way. First, defining the block-triangular matrices

$$X' = \begin{bmatrix} I_1 & -A_{11}^{-1}A_{12} \\ & I_2' \end{bmatrix} \quad \text{and} \quad Y' = \begin{bmatrix} I_1 & A_{11}^{-1}A_{12} \\ & I_2' \end{bmatrix} ,$$

we observe that $X'Y'$ is the identity matrix of order n, so that X' and Y' are mutually inverse by lemma II.5.9.4. As in (II.5.9.1), we find that

$$A = Y \begin{bmatrix} A_{11} & A_{12} \\ & S \end{bmatrix} = Y \begin{bmatrix} A_{11} & A_{12} \\ & S \end{bmatrix} X'Y' = Y \begin{bmatrix} A_{11} & \\ & S \end{bmatrix} Y' . \tag{II.5.9.2}$$

By lemma II.5.9.9 and corollary II.4.4.13, this factorization implies $\operatorname{rank} A = \operatorname{rank} A_{11} + \operatorname{rank} S$, so that $\operatorname{rank} S = \operatorname{rank} A - r$.

The construction analyzed in § II.5.6 and lemma II.5.9.10, however, relies on the invertibility of the order-r leading principal submatrix A_{11} of A. In general, if $\operatorname{rank} A \geq r$, then A has an invertible square submatrix of order r due to lemma II.5.9.6. Any such a submatrix can be used to construct a cross approximation of A of a more general form than the one considered in lemma II.5.6.15. In fact, the first r steps of pivoted Gaussian elimination find such an invertible submatrix (see the proof of lemma II.5.9.6) and produce a pivoted LU decomposition of that submatrix (as we show in the following theorem).

Theorem II.5.9.11 (cross approximation and the pivoted LU decomposition)**.** Let $m, n, r \in \mathbb{N}$ be such that $r \leq \min\{m, n\}$ and $A \in \mathbb{F}^{m \times n}$ be such that $r \leq \operatorname{rank} A$. Consider permutation matrices P and Q of orders m and n and their partitions

$$P = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \tag{II.5.9.3}$$

with $P_1 \in \mathbb{F}^{r \times m}$ and $Q_1 \in \mathbb{F}^{r \times n}$, so that $\widetilde{A} = PAQ^{\mathsf{T}}$ satisfies

$$\widetilde{A} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix} \tag{II.5.9.4}$$

with $\widetilde{A}_{ij} = P_i A Q_j^{\mathsf{T}}$ for all $i, j \in \{1, 2\}$. Assume that $\widetilde{A}_{11} = P_1 A Q_1^{\mathsf{T}}$ is invertible and consider

$$C = AQ_1^{\mathsf{T}}, \quad R = P_1 A \quad \text{and} \quad G = \widetilde{A}_{11}^{-1} = (P_1 A Q_1^{\mathsf{T}})^{-1}\,. \tag{II.5.9.5}$$

Then $\operatorname{rank} CGR = r$,

$$A - CGR = P^{\mathsf{T}} \begin{bmatrix} O & \\ & S \end{bmatrix} Q \quad \text{with} \quad S = \widetilde{A}_{22} - \widetilde{A}_{21} \widetilde{A}_{11}^{-1} \widetilde{A}_{12}\,, \tag{II.5.9.6}$$

where O is the zero square matrix of order r, and $\operatorname{rank}(A - CGR) = \operatorname{rank} S = \operatorname{rank} A - r$. In the cases of $r = m$ and $r = n$, the second block rows and second block columns vanish in (II.5.9.3) to (II.5.9.5).

Furthermore, if $A = P^{\mathsf{T}} LUQ$ is an r-step pivoted LU decomposition, then S is the r-step Schur complement of the LU decomposition $PAQ^{\mathsf{T}} = LU$, and the truncated r-step LU factors $\widehat{L} \in \mathbb{F}^{m \times r}$ and $\widehat{U} \in \mathbb{F}^{r \times n}$ of PAQ^{T} satisfy the following in terms of the partitions (II.5.6.8):

$$\widehat{L} = PCU_1^{-1}\,, \quad \widehat{U} = L_1^{-1} R Q^{\mathsf{T}}\,, \quad G = U_1^{-1} L_1^{-1} \quad \text{and} \quad P^{\mathsf{T}} \widehat{L} \widehat{U} Q = CGR\,. \tag{II.5.9.7}$$

Proof. The claims follow from lemmata II.5.9.10 and II.5.6.15 applied to $\widetilde{A}$, which yields the cross approximation $\widetilde{C}\widetilde{G}\widetilde{R}$ with the factors

$$\widetilde{C} = \begin{bmatrix} \widetilde{A}_{11} \\ \widetilde{A}_{21} \end{bmatrix} = PC, \quad \widetilde{R} = \begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \end{bmatrix} = RQ^{\mathsf{T}} \quad \text{and} \quad G = \widetilde{G} = \widetilde{A}_{11}^{-1}\,.$$

The ranks of $CGR = P^{\mathsf{T}} \widetilde{C}\widetilde{G}\widetilde{R} Q$ and $A - CGR = P^{\mathsf{T}} (\widetilde{A} - \widetilde{C}\widetilde{G}\widetilde{R})\, Q$ are equal to those of $\widetilde{C}\widetilde{G}\widetilde{R}$ and $\widetilde{A} - \widetilde{C}\widetilde{G}\widetilde{R}$ due to corollary II.4.4.13 (and to proposition II.4.3.7, which we use throughout for applying and undoing the column and row pivoting).

The cross approximation CGR given by (II.5.9.6) is formed by the rows and columns of A selected by P_1 and Q_1 and collected in R and C. Those rows and columns are often called the *basis rows* and *basis columns* of the cross approximation. The term *cross* refers to the pattern formed by the basis rows and columns. The submatrix $\widetilde{A}_{11}$ of A, formed by the basis rows and columns (the order matters), is often called the *basis submatrix* of the cross approximation. For now, the word “basis” should be seen as a modifier in the terms “basis rows”, "basis columns" and "basis submatrix" merely reflecting that the objects to which they refer are involved in the construction of CGR. The error of the approximation is zero "at the cross", i.e., in the basis rows and columns, as the first equality of (II.5.9.6) demonstrates. The cross approximation is therefore called *interpolatory*: it interpolates the data upon which it is based, and the error concentrates in the part of the matrix to which the approximation is oblivious. Finally, theorem II.5.9.11 shows that the rank of the associated error is exactly $\operatorname{rank} A - r$, so that cross approximation depletes the rank: any increase of the order (and hence of the rank) of the basis submatrix leads to a commensurate decrease of the rank of the associated approximation error.

In many applications, we can expect the Schur complement to be small in one sense or another (we will study norms later in the course) if a sufficiently "good" submatrix is used as a basis submatrix. In that case, if the error is sufficiently small for CGR to reasonably accurately approximate A, we can store and work with only $mr + rn - r^2$ entries of A instead of all mn entries (see examples II.4.4.15 to II.4.4.17). Such a reduction is crucial in many applications, especially when such low-rank approximations are arranged in multilevel cascades (analogous to the currently so fashionable deep neural networks).

The proof of theorem II.5.9.11 given above is somewhat higher level in that it uses *block* row and column elimination to transform $\widetilde{A}$ to a block-diagonal matrix, involving the inverse of $\widetilde{A}_{11}$. Any specific implementation of this approach requires that the inverse be obtained in one way or another. The pivoted LU decomposition is a perfectly suitable tool for this task and can be used as a specific implementation of the block elimination performed in the proof of theorem II.5.9.11, even though it was not used explicitly in the proof itself.

> **Remark II.5.9.12** (pivoted LU decomposition as the depletion of approximation error)**.** Theorem II.5.9.11 has an important implication for the interpretation of the pivoted LU decomposition. Consider a matrix $A \in \mathbb{F}^{m\times n}$ with $m, n \in \mathbb{N}$ of rank $r = \operatorname{rank} A \geq 2$. By lemma II.5.8.6 and theorem II.5.9.8, matrix A has a complete r-step pivoted LU decomposition. Consider the corresponding truncated decomposition: $PAQ^\mathsf{T} = \widehat{L}\widehat{U}$ with $\widehat{L} \in \mathbb{F}^{m\times r}$ and $\widehat{U} \in \mathbb{F}^{r\times n}$. Let $u_1, \ldots, u_r \in \mathbb{F}^m$ and $v_1, \ldots, v_r \in \mathbb{F}^n$ be the columns of $\widehat{L}$ and $\widehat{U}^\mathsf{T}$ respectively. Then
>
> $$PAQ^\mathsf{T} = \sum_{k=1}^{r} u_k v_k^\mathsf{T} \,,$$
>
> and PAQ^T cannot be represented as a sum of fewer rank-one matrices because $\operatorname{rank} PAQ^\mathsf{T} = \operatorname{rank} A = r$. For each $k \in \{1, \ldots, r\}$, step k of the pivoted LU decomposition consists in discovering the kth term of this sum, and the effect of this discovery is that the rank of the pivoted LU approximation is increased by one while the rank of the associated error is decreased by one.

Apart from being fundamental for our construction of matrix rank, theorem II.5.9.7 has various useful consequences. In particular, we are now in position to consider the one-sided inversion of matrices.

> **Definition II.5.9.13** (one-sided inverses)**.** For $m, n \in \mathbb{N}$, consider $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times m}$ such that AB is an identity matrix. Then A is called the *left inverse* of B and B is called the *right inverse* of A.

Definitions II.4.1.9 and II.5.9.13 and proposition II.4.1.11 yield the following counterpart of proposition II.4.1.23 in the context of one-sided matrix inversion.

Proposition II.5.9.14 (one-sided matrix inversion under transposition)**.** For $m, n \in \mathbb{N}$, consider $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times m}$. Then the following statements are equivalent:

(a) A is a left inverse of B;

(b) A^T is a right inverse of B^T.

Comparing definitions II.5.9.13 and II.4.1.14 and applying proposition II.4.1.13 and lemma II.5.9.4, we immediately obtain the following proposition.

Proposition II.5.9.15 (one-sided inversion of square matrices)**.** Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Then the following statements hold:

(a) if A has a left inverse, then A is invertible;

(b) if A has a right inverse, then A is invertible;

(c) if A is invertible, then A^{-1} is a unique left inverse of A.

(d) if A is invertible, then A^{-1} is a unique right inverse of A;

Proof. The proof is left to the reader as an exercise.

Proposition II.5.9.15 shows that one-sided invertibility brings about nothing new in the case of a square matrix, being is equivalent to invertibility. The rationale for introducing the notion is therefore associated entirely with non-square matrices.

Example II.5.9.16 (one-sided inverses of nonsquare matrices may be nonunique)**.** For $n, r \in \mathbb{N}$ such that $r < n$, consider $U \in \mathbb{F}^{n\times r}$ and $V \in \mathbb{F}^{r\times n}$ partitioned as follows:

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix},$$

where $U_1, V_1 \in \mathbb{F}^{r\times r}$. Then $UV = U_1V_1 + U_2V_2$. If U_2 is zero, then U is a left inverse of V (and, at the same time, V is a right inverse of U) if and only if U_1 and V_1 are mutually inverse, regardless of V_2. Similarly, if V_2 is zero, then U is a left inverse of V (and, at the same time, V is a right inverse of U) if and only if U_1 and V_1 are mutually inverse, regardless of U_2.

In the following lemma, we establish, loosely speaking, that the existence and uniqueness of one-sided inverses is equivalent to the full-rank property.

Lemma II.5.9.17 (one-sided invertibility and the full-rank property)**.** For $m, n \in \mathbb{N}$, consider a matrix $A \in \mathbb{F}^{m\times n}$.

(a) A has a left inverse if and only if $\operatorname{rank} A = n \leq m$.

(b) A has several left inverses if $\operatorname{rank} A = n < m$.

(c) A has a right inverse if and only if $\operatorname{rank} A = m \leq n$.

(d) A has several right inverses if $\operatorname{rank} A = m < n$.

(e) A has left and right inverses if and only if $\operatorname{rank} A = m = n$.

Proof. We have $\operatorname{rank} A \in \mathbb{N}_0$ by definition II.4.4.1 and $\operatorname{rank} A \leq \min\{m, n\}$ by corollary II.4.4.8.

If A has a left inverse B, then $\operatorname{rank} BA = n$ by definition II.5.9.13 and lemma II.5.9.2. On the other hand, $\operatorname{rank} BA \leq \operatorname{rank} A$ by proposition II.4.4.11. So we have $n \leq \operatorname{rank} A \leq \min\{m, n\}$ and hence $\operatorname{rank} A = n \leq m$, which proves the "only if" implication of part (a).

Let us now prove the "if" implication of part (a). If $\operatorname{rank} A = m = n$, then A is invertible by lemma II.5.9.3 and therefore has a left inverse by statement (c) of proposition II.5.9.15. Assuming $\operatorname{rank} A = n < m$ and using part (a) of theorem II.5.9.7, we conclude that A has an invertible square submatrix of order n. By definition II.4.2.3, that means that there exists a permutation matrix $P \in \mathbb{F}^{m \times m}$ with a leading submatrix P_1 of size $n \times m$ such that the matrix $A_1 = P_1 A$ is invertible (the permutation of columns of an invertible matrix does not affect the invertibility due to proposition II.4.3.7 and proposition II.4.1.22). Consider the zero matrix $O \in \mathbb{F}^{n \times (m-n)}$ and let $B = [\, A_1^{-1} \;\; O \,] P$. Then $BA = (BP^\mathsf{T})(PA)$ is an identity matrix of order n. By definition II.5.9.13, the matrix B is a left inverse of A. This proves the "if" implication of part (a).

Part (c) follows from part (a) due to propositions II.4.4.7 and II.5.9.14.

Let us prove part (d). By part (c), A has a right inverse B. By lemma II.5.9.2, there exists a nonzero column vector $x \in \mathbb{F}^n$ such that Ax is zero. This implies that there exists a nonzero matrix $\varDelta \in \mathbb{F}^{n \times m}$ such that $A\varDelta$ is zero. Then $A(B + \varDelta) = AB$, so $B + \varDelta$ is a right inverse of A by definition II.5.9.13.

Part (b) follows from part (d) due to propositions II.4.4.7 and II.5.9.14.

The "if" implication of part (e) is a consequence of lemma II.5.9.3 and of statements (c) and (d) of proposition II.5.9.15. The "only if" implication of part (e) follows immediately from parts (a) and (c).

Part (a) of lemma II.5.9.17 immediately yields the following.

Lemma II.5.9.18 (relation between the minimal factorizations of a matrix)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$ be of rank $r \in \mathbb{N}$. Consider $U \in \mathbb{F}^{m \times r}$ and $V \in \mathbb{F}^{n \times r}$ such that $A = UV^\mathsf{T}$. Then the following holds.

(a) There exist matrices $\varPsi \in \mathbb{F}^{r \times m}$ and $\varOmega \in \mathbb{F}^{r \times n}$ such that $U = A\varOmega^\mathsf{T}$ and $V = A^\mathsf{T}\varPsi^\mathsf{T}$.

(b) For any $\tilde{U} \in \mathbb{F}^{m \times r}$ and $\tilde{V} \in \mathbb{F}^{n \times r}$ such that $A = \tilde{U}\tilde{V}^\mathsf{T}$, there exists an invertible matrix $T \in \mathbb{F}^{r \times r}$ such that $\tilde{U} = US$ and $\tilde{V} = VS^{-\mathsf{T}}$.

Proof. By part (a) of lemma II.5.9.17, U and V have left inverses $\varPsi$ and $\varOmega$. Let I denote the identity matrix of order r. By definition II.5.9.13, we have that $\varPsi U = I = \varOmega V$ and hence $U^\mathsf{T}\varPsi^\mathsf{T} = I$. Then $A = UV^\mathsf{T}$ implies $U = A\varOmega^\mathsf{T}$ and $V = A^\mathsf{T}\varPsi^\mathsf{T}$, which proves part (a).

Consider $\tilde{U} \in \mathbb{F}^{m \times r}$ and $\tilde{V} \in \mathbb{F}^{n \times r}$ such that $A = \tilde{U}\tilde{V}^\mathsf{T}$. By proposition II.4.4.11, $\operatorname{rank} \tilde{U} = r = \tilde{V}$, so part (a) yields $\tilde{U} = A\tilde{\varOmega}^\mathsf{T}$ and $\tilde{V} = A^\mathsf{T}\tilde{\varPsi}^\mathsf{T}$ with suitable matrices $\tilde{\varPsi} \in \mathbb{F}^{r \times m}$ and $\tilde{\varOmega} \in \mathbb{F}^{r \times n}$. Given the original factorization $A = UV^\mathsf{T}$, that implies $\tilde{U} = US$ with $S = V^\mathsf{T}\tilde{\varOmega}^\mathsf{T}$ and $\tilde{V} = VT$ with $T = U^\mathsf{T}\tilde{\varPsi}^\mathsf{T}$. Substituting these factors, we obtain $A = USTV^\mathsf{T}$ and hence $USTV^\mathsf{T} = A = UV^\mathsf{T}$. Multiplying this equality with $\varPsi$ on the left and with $\varOmega^\mathsf{T}$ on the right, we obtain $ST = I$. By proposition II.5.9.15, that leads to that the matrix S is invertible and $T = S^{-1}$, which completes the proof of part (b).

Lemma II.5.9.18 follows up on remark II.4.4.6 in an important way by showing that *any* two minimal (rank-r) factorizations of a matrix A (of rank r) are related by an invertible matrix.

The truncated LU decomposition and, more generally, the cross approximation (see lemma II.5.6.15 and theorem II.5.9.11) are specific techniques for approximating a matrix $A \in \mathbb{F}^{m\times n}$ with a product UCV^T, where $U \in \mathbb{F}^{m\times p}$ is a square or tall matrix, $V^\mathsf{T} \in \mathbb{F}^{q\times n}$ is a square or wide matrix and $C \in \mathbb{F}^{p\times q}$, where $m, n, p, q \in \mathbb{N}$. Such factorized approximations (or exact representations) are particularly informative and useful when both p and q are less (or even "much less") than both m and n. For both the truncated LU approximation and cross approximation, the columns of U and V, being obtained from those of A and A^T in specific ways, are naturally adapted to A, and it just suffices to construct the same number of columns and of rows ($p = q$). In addition, it is always possible to absorb the coefficient matrix C into the other two factors as long as at least one of them can be modified accordingly. In a different setting, we may be interested in whether a given matrix $A \in \mathbb{F}^{m\times n}$ has the following factorizations:

(i) $A = UC$ with a *given* matrix $U \in \mathbb{F}^{m\times p}$ and *any* $C \in \mathbb{F}^{n\times p}$,

(ii) $A = CV^\mathsf{T}$ with a *given* matrix $V \in \mathbb{F}^{n\times q}$ and *any* $C \in \mathbb{F}^{m\times q}$,

(iii) $A = UCV^\mathsf{T}$ with *given* matrices $V \in \mathbb{F}^{n\times q}$ and $U \in \mathbb{F}^{m\times q}$ and *any* $C \in \mathbb{F}^{p\times q}$.

The third factorization trivially covers the first two as particular cases (consider V equal to the identity matrix of order $q = n$ or U equal to the identity matrix of order $p = m$), and we will now address the question with respect to the third, most general, factorization. As we will now establish, the relation between the above three factorizations is, actually, deeper: the existence of the first two factorizations implies the existence of the third.

Lemma II.5.9.19. Let $m, n, p, q \in \mathbb{N}$, $U \in \mathbb{F}^{m\times p}$, $V \in \mathbb{F}^{n\times q}$ and $A \in \mathbb{F}^{m\times n}$. A matrix $C \in \mathbb{F}^{p\times q}$ such that $A = UCV^\mathsf{T}$ exists *if and only if* matrices $X \in \mathbb{F}^{m\times q}$ and $Y \in \mathbb{F}^{n\times p}$ such that $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$ exist.

Proof. Let us assume that $C \in \mathbb{F}^{p\times q}$ satisfies $A = UCV^\mathsf{T}$. This equality implies $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$ with $Y = VC^\mathsf{T} \in \mathbb{F}^{m\times q}$ and $X = UC \in \mathbb{F}^{n\times p}$, which yields the "only if" implication claimed.

To prove the "if" implication claimed, we consider left inverses $\mathit{\Psi}$ and $\mathit{\Omega}$ of U and V (such exist by part (a) of lemma II.5.9.17).

Let us first treat the particular case of $\operatorname{rank} U = p$ and $\operatorname{rank} V = q$. Consider the representations $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$ with $Y \in \mathbb{F}^{n\times p}$ and $X \in \mathbb{F}^{m\times q}$. Since $\mathit{\Psi}$ is a left inverse of U, the first representation yields (by definition II.5.9.13 and proposition II.5.9.14) $Y^\mathsf{T} = (\mathit{\Psi} U)Y^\mathsf{T} = \mathit{\Psi}(UY^\mathsf{T}) = \mathit{\Psi} A$, and then the second representation leads to $Y^\mathsf{T} = \mathit{\Psi}(XV^\mathsf{T})$. Substituting this expression into the first representation, we obtain $A = UCV^\mathsf{T}$ with $C = \mathit{\Psi} X$. This proves the "if" implication claimed under the additional assumption that $\operatorname{rank} U = p$ and $\operatorname{rank} V = q$.

Let us now turn to the case when $\operatorname{rank} U < p$ or $\operatorname{rank} V < q$ and reduce it to the full-rank case, a proof for which we have given above. Consider $Y \in \mathbb{F}^{n\times p}$ and $X \in \mathbb{F}^{m\times q}$ satisfying $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$. Let $\hat{p} = \operatorname{rank} U$ and $\hat{q} = \operatorname{rank} V$. When U or V is zero, the respective one of the equalities $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$ implies that A is zero, so that any matrix $C \in \mathbb{F}^{p\times q}$ satisfies the conclusion of the "if" implication claimed. For the remainder of the proof, we therefore assume that both U and V are nonzero.

Note that we have (see corollary II.4.4.8) the following. First, $1 \leq \hat{p} \leq \min\{m, p\}$ and $1 \leq \hat{p} \leq \min\{n, q\}$. Second, the factorizations $U = \widehat{U}\widetilde{U}$ and $V = \widehat{V}\widetilde{V}$ hold with factors

$\widehat{U} \in \mathbb{F}^{m\times\hat{p}}$ and $\widetilde{U} \in \mathbb{F}^{\hat{p}\times p}$, both of rank $\hat{p}$, and $\widehat{V} \in \mathbb{F}^{n\times\hat{q}}$ and $\widetilde{V} \in \mathbb{F}^{\hat{q}\times q}$, both of rank $\hat{q}$. Then the equalities $A = UY^\mathsf{T}$ and $A = XV^\mathsf{T}$ imply $A = \widehat{U}\widehat{Y}^\mathsf{T}$ and $A = \widehat{X}\widehat{V}^\mathsf{T}$ with $\widehat{Y} = Y\widetilde{U}^\mathsf{T}$ and $\widehat{X} = X\widetilde{V}^\mathsf{T}$. Since the matrices $\widehat{U}$ and $\widehat{V}$ are full rank, we can apply the "if" implication of this . That yields the existence of $\widehat{C} \in \mathbb{F}^{\hat{p}\times\hat{q}}$ such that $A = \widehat{U}\widehat{C}\widehat{V}^\mathsf{T}$. By part (c) of lemma II.5.9.17, the full-rank non-tall matrices $\widetilde{U}$ and $\widetilde{V}$ have right inverses $\widehat{\Psi}$ and $\widehat{\Omega}$. Then the factorizations $U = \widehat{U}\widetilde{U}$ and $V = \widehat{V}\widetilde{V}$ lead to $\widehat{U} = U\widehat{\Psi}$ and $\widehat{V} = V\widehat{\Omega}$. As a result, $A = \widehat{U}\widehat{C}\widehat{V}^\mathsf{T}$ gives $A = UCV^\mathsf{T}$ with $C = \widehat{\Psi}\widehat{C}\widehat{\Omega}^\mathsf{T}$, which completes the proof of the "if" implication claimed.

We can augment lemma II.5.9.20 with the following result.

Lemma II.5.9.20. Let $m, n, p, q \in \mathbb{N}$, $U \in \mathbb{F}^{m\times p}$, $V \in \mathbb{F}^{n\times q}$ and $A \in \mathbb{F}^{m\times n}$. When a matrix C such that $A = UCV^\mathsf{T}$ exists, it is unique *if and only if* $\operatorname{rank} U = p$ and $\operatorname{rank} V = q$, in which case it satisfies $C = \Psi A \Omega^\mathsf{T}$ with any left inverses Ψ and Ω of U and V.

Before proving lemma II.5.9.20, we note the following. First, the lemma, complementing lemma II.5.9.19, also follows up on lemma II.5.9.18. Second, in the context of lemma II.5.9.20, each of the matrices U and V does have a left inverse by part (a) of lemma II.5.9.17.

Proof of lemma II.5.9.20. To prove the "if" implication claimed, we consider a matrix C such that $A = UCV^\mathsf{T}$. Since Ψ and Ω are left inverses of U and V, we immediately obtain $\Psi A \Omega^\mathsf{T} = \Psi(UCV^\mathsf{T})\Omega^\mathsf{T} = C$ (by definition II.5.9.13 and proposition II.5.9.14), so there is no other such a matrix.

Let us now prove the "only if" implication claimed. To this end, we assume that $\operatorname{rank} U < p$ or $\operatorname{rank} < q$. Let us consider the case of $\operatorname{rank} U < p$. Then there exists a nonzero column vector $x \in \mathbb{F}^q$ such that Ux is zero (see the auxiliary argument in the proof of lemma II.5.9.2). For any $y \in \mathbb{F}^q$ nonzero, we then have $U(xy^\mathsf{T})V^\mathsf{T} = (Ux)y^\mathsf{T}V^\mathsf{T} = 0$ and hence $A = UCV^\mathsf{T}$ with $C \in \mathbb{F}^{p\times q}$ implies $U\widetilde{C}V^\mathsf{T}$ with $\widetilde{C} = C + xy^\mathsf{T} \neq C$. The case of $\operatorname{rank} V < q$ is treated analogously (or by transposition).

To illustrate the power of lemmata II.5.9.19 and II.5.9.20, we consider the following problem.

Example II.5.9.21. Let I denote the identity matrix of order $n \in \mathbb{N}$ and $\Delta \in \mathbb{F}^{n\times n}$. We consider Δ as a perturbation of I and are interested in whether the perturbed matrix $I + \Delta$ is invertible and, when it is so, how the inverse can be expressed in terms of U and V.

For any $E \in \mathbb{F}^{n\times n}$, we have

$$I - (I - E)(I + \Delta) = E(I + \Delta) - \Delta \quad \text{and} \quad I - (I + \Delta)(I - E) = (I + \Delta)E - \Delta\,. \tag{II.5.9.8}$$

In particular, if the matrix $I + \Delta$ is invertible and $E = I - (I + \Delta)^{-1}$, so that $(I + \Delta)^{-1} = I - E$, then equalities (II.5.9.8) yield

$$E = \Delta\,(I + \Delta)^{-1} \quad \text{and} \quad E = (I + \Delta)^{-1}\Delta\,.$$

Relating the perturbation of the inverse to the perturbed inverse, these equalities do not provide a closed-form expression for $(I + \Delta)^{-1}$ but are nevertheless very informative. Indeed, by lemma II.5.9.19, they imply the existence of $\Lambda \in \mathbb{F}^{n\times n}$ such that $E = \Delta\Lambda\Delta$.

In this example, we are interested in the case when the perturbation Δ is, loosely speaking, localized "along relatively few directions". To be precise, we consider a low-rank perturbation: $\mathit{\Delta} = UV^\mathsf{T}$, where $U, V \in \mathbb{F}^{n\times r}$ with $r \in \mathbb{N}$ that is less (and possibly much less) than n. Let us assume additionally that both U and V are full rank.

First, we continue assuming that the matrix $I + \mathit{\Delta}$ is invertible and $E = I - (I + \mathit{\Delta})^{-1}$. The equality $E = \mathit{\Delta}\mathit{\Lambda}\mathit{\Delta}$ obtained above, regardless of $\mathit{\Lambda} \in \mathbb{F}^{n\times n}$, implies $E = UCV^\mathsf{T}$ with $C = V^\mathsf{T}\mathit{\Lambda}U \in \mathbb{F}^{r\times r}$, The first of equalities (II.5.9.8) then leads to

$$I - (I - E)(I + \mathit{\Delta}) = UCV^\mathsf{T}(I + UV^\mathsf{T}) - UV^\mathsf{T} = UGV^\mathsf{T} \tag{II.5.9.9}$$

with $G = C(I' + V^\mathsf{T}U) - I'$, where I' is the identity matrix of order r. Applying lemma II.5.9.20, we conclude that $E = I - (I + \mathit{\Delta})^{-1}$ requires that G be zero, i.e., that $C(I' + V^\mathsf{T}U) = I'$. By lemma II.5.9.4, that can be the case only when the matrix $I' + V^\mathsf{T}U$ is invertible and $C = (I' + V^\mathsf{T}U)^{-1}$.

Let us now assume that the matrix $I' + V^\mathsf{T}U$ is invertible and define $C = (I' + V^\mathsf{T}U)^{-1} \in \mathbb{F}^{r\times r}$ and $E = UCV^\mathsf{T} \in \mathbb{F}^{n\times n}$. Then the matrix $G = C(I' + V^\mathsf{T}U) - I'$ is zero. Exactly as in (II.5.9.9), we obtain from the first of equalities (II.5.9.8) that $I - (I - E)(I + \mathit{\Delta}) = UGV^\mathsf{T}$, which results in $(I - E)(I + \mathit{\Delta}) = I$. Applying lemma II.5.9.4 again, we conclude that the matrix $I + \mathit{\Delta}$ is invertible and $I - E = (I + \mathit{\Delta})^{-1}$.

Summarizing our findings: the matrix $I + UV^\mathsf{T}$ with full-rank matrices $U, V \in \mathbb{F}^{n\times r}$, is invertible if and only if the matrix $I' + V^\mathsf{T}U$ is invertible, in which case $(I + UV^\mathsf{T})^{-1} = I - U\,(I' + V^\mathsf{T}U)^{-1}V^\mathsf{T}$. The right-hand side expression, in spite of involving matrix inversion, simplifies the left-hand side expression insofar as r is less than n. Indeed, it is an explicit expression of the perturbed inverse, the order of which is n, in terms of the inverse of another matrix, the order of which is r.

Note that the assumption that U and V are full rank is not restrictive of $\mathit{\Delta}$: it merely requires the factorization $\mathit{\Delta} = UV^\mathsf{T}$ to be of the least rank possible (which is $\operatorname{rank}\Delta$).

For the the obtained criterion for the invertibility of $I + UV^\mathsf{T}$, we gave above distinct proofs for necessity and sufficiency. Is the assumption that U and V are full rank necessary for the above necessity proof? Is it necessary for the above sufficiency proof? These questions are left to the reader as an exercise.

Finally, the following is an immediate corollary of the above result due to proposition II.4.1.22. For any $n, r \in \mathbb{N}$, any invertible matrix $A \in \mathbb{F}^{n\times n}$ and any full-rank matrices $U, V \in \mathbb{F}^{n\times r}$, the matrix $A + UV^\mathsf{T}$ is invertible if and only if the matrix $I' + V^\mathsf{T}A^{-1}U$ is invertible, in which case $(A + UV^\mathsf{T})^{-1} = A^{-1} - A^{-1}U(I' + V^\mathsf{T}A^{-1}U)^{-1}V^\mathsf{T}A^{-1}$.

Lastly, let us state a corollary of theorem II.5.9.11, which will serve as a useful auxiliary result at later stages of the course.

Corollary II.5.9.22 (row or column pivoting is unnecessary when the rows or columns involved form a full-rank submatrix)**.** Let $m, n, r \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$ be such that $r \le \operatorname{rank} A$.

(a) Assume that the leading submatrix of A of size $r \times n$ is of rank r. Then A has an r-step pivoted LU decomposition corresponding to row- and column-exchange indices $\pi_1, \dots, \pi_r$ and $\sigma_1, \dots, \sigma_r$ such that $\pi_k = k$ for each $k \in \{1, \dots, r\}$.

(b) Assume that the leading submatrix of A of size $m \times r$ is of rank r. Then A has an r-step pivoted LU decomposition corresponding to row- and column-exchange indices $\pi_1, \dots, \pi_r$ and $\sigma_1, \dots, \sigma_r$ such that $\sigma_k = k$ for each $k \in \{1, \dots, r\}$.

Proof. The proof is left to the reader as an exercise.

§ II.5.10. The fundamental subspaces of a matrix

In this section, we consider what is known as the fundamental subspaces of a matrix. We will soon formally introduce the notion of *subspace* and relate it to the notion of *vector space*. In this § II.5.10, the words *subspace* and *space* are used merely as parts of specific terms, such as *fundamental subspace*.

Definition II.5.10.1 (the fundamental subspaces of a matrix)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. By $\operatorname{Ker} A$ and $\operatorname{Im} A$ we denote the following sets:

$$\operatorname{Ker} A = \{x \in \mathbb{F}^n \colon Ax = 0\} \subseteq \mathbb{F}^n \quad \text{and} \quad \operatorname{Im} A = \{Ax \colon x \in \mathbb{F}^n\} \subseteq \mathbb{F}^m \,.$$

The set $\operatorname{Ker} A$ is usually referred to as the *kernel* or as the *null space* of A. The set $\operatorname{Im} A$ is usually referred to as the *image* or as the *range* of A. The sets $\operatorname{Ker} A$, $\operatorname{Im} A$, $\operatorname{Ker} A^{\mathsf{T}}$ and $\operatorname{Im} A^{\mathsf{T}}$ are called the *four fundamental subspaces* of A.

Remark II.5.10.2 (fundamental subspaces and linear combinations)**.** In the context of definition II.5.10.1, recalling definition II.2.2.7 and applying rule (a) of remark II.4.1.5, we notice that $\operatorname{Im} A$ is the set of all possible linear combinations of the columns of A and $\operatorname{Ker} A$ is the set of all possible coefficients of all trivial linear combinations of the columns of A.

The image and kernel subspaces $\operatorname{Im} A$ and $\operatorname{Ker} A$ of a matrix A are closely related to the mapping induced by the matrix, which we will now define.

Definition II.5.10.3 (the mapping induced by a matrix)**.** For any $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$, the mapping $\phi_A \colon \mathbb{F}^n \to \mathbb{F}^m$ induced by the matrix A is given by $\phi_A(x) = Ax$ for each $x \in \mathbb{F}^n$.

We remark that the relation between a matrix $A \in \mathbb{F}^{m\times n}$ and $\phi_A \colon \mathbb{F}^n \to \mathbb{F}^m$ in the context of definition II.5.10.3 is one to one: indeed, $Ax = Bx$ for all $x \in \mathbb{F}^n$ with $A, B \in \mathbb{F}^{m\times n}$ trivially implies $A = B$.

Note that, in the context of the notations of definitions II.5.10.1 and II.5.10.3, the sets $\operatorname{Ker} A$ and $\operatorname{Im} A$ are exactly what is generally referred to as the *null set* and the *image* of a function (see § I.1.7 for the latter notion). The mapping ϕ_A is therefore *surjective* (see § I.1.18) if and only if $\operatorname{Im} A = \mathbb{F}^m$. For *injectivity* (see § I.1.17), we have the following result.

Proposition II.5.10.4. Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Then the mapping ϕ_A induced by A is injective if and only if $\operatorname{Ker} A = \{0\}$.

Proof. Let us prove the necessity first. Assume that ϕ_A is injective and consider $v \in \operatorname{Ker} A$. Then $\phi_A(0) = 0$ and $\phi_A(v) = 0$, so $v = 0$ due to the injectivity assumption. This shows that $\operatorname{Ker} A = \{0\}$.

Let us now prove the sufficiency. Suppose $\operatorname{Ker} A = \{0\}$ and consider $v, v' \in \mathbb{F}^n$ such that $v' \neq v$. Then $v' - v$ is a nonzero element of $\mathbb{F}^n$, and hence $\phi_A(v' - v) \neq 0$. On the other hand, we have $\phi_A(v' - v) = \phi_A(v') - \phi_A(v)$, so we obtain $\phi_A(v') \neq \phi_A(v)$. This proves the injectivity of ϕ_A.

Example II.5.10.5. For $n \in \mathbb{N}$, let I denote the identity matrix of order n. Then definition II.5.10.1 yields $\operatorname{Im} I = \mathbb{F}^n$ and $\operatorname{Ker} I = \{0\}$.

Proposition II.5.10.6. Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Then the following statements are equivalent.

- **(i)** A is the zero matrix from $\mathbb{F}^{m\times n}$;
- **(ii)** $\operatorname{Im} A = \{0\}$;
- **(iii)** $\operatorname{Ker} A = \mathbb{F}^n$.

Proof. The proof is left to the reader as an exercise.

Example II.5.10.7. Consider the following matrices:

$$U = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix} \in \mathbb{F}^{2\times 2} \quad \text{and} \quad W = \begin{bmatrix} 0 & \\ & 1 \end{bmatrix} \in \mathbb{F}^{2\times 2}.$$

Then $\operatorname{Im} U = \{(\alpha, 0) \colon \alpha \in \mathbb{R}\} = \operatorname{Ker} W$ and $\operatorname{Ker} U = \{(0, \alpha) \colon \alpha \in \mathbb{R}\} = \operatorname{Im} W$.

Example II.5.10.8. Consider the following matrix:

$$A = \begin{bmatrix} 1 & 2 \\ & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

Then $\operatorname{Im} A = \{(\alpha, 0) \colon \alpha \in \mathbb{R}\}$ and $\operatorname{Ker} A = \{(2\alpha, -\alpha) \colon \alpha \in \mathbb{R}\}$.

Proposition II.5.10.9 (the fundamental subspaces of a matrix are closed under addition and multiplication by a scalar)**.** Consider $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Let $0_n \in \mathbb{F}^n$ and $0_m \in \mathbb{F}^m$ be zero column vectors. Then $0_n \in \operatorname{Ker} A$ and $0_m \in \operatorname{Im} A$. Furthermore, the following statements hold.

- **(a)** For any $x, x' \in \operatorname{Ker} A$ and $\alpha \in \mathbb{F}$, we have $x + x' \in \operatorname{Ker} A$ and $\alpha x \in \operatorname{Ker} A$.
- **(b)** For any $y, y' \in \operatorname{Im} A$ and $\alpha \in \mathbb{F}$, we have $y + y' \in \operatorname{Im} A$ and $\alpha y \in \operatorname{Im} A$.

Proof. First, we have $A \cdot 0_n = 0_m$, so that $0_n \in \operatorname{Ker} A$ and $0_m \in \operatorname{Im} A$ by definition II.5.10.1.

To prove part (a), we consider $x, x' \in \operatorname{Ker} A$ and $\alpha \in \mathbb{F}$. Then we have $x + x' \in \mathbb{F}^n$ by definition II.2.1.4, $\alpha x \in \mathbb{F}^n$ by definition II.2.1.5 and $Ax = 0_m = Ax'$ by definition II.5.10.1. We therefore obtain $A(x + x') = Ax + Ax' = 0_m + 0_m = 0_m$ (by part (c) of proposition II.4.1.10 and part (c) of proposition II.2.1.8) and $A(\alpha x) = \alpha(Ax) = \alpha \cdot 0_m = 0_m$ (by part (b) of proposition II.4.1.19 and definitions II.2.1.5 and II.2.1.6), so that $x + x' \in \operatorname{Ker} A$ and $\alpha x \in \operatorname{Ker} A$ by definition II.5.10.1.

To prove part (b), we consider $y, y' \in \operatorname{Im} A$ and $\alpha \in \mathbb{F}$. By definition II.5.10.1, there exist $x, x' \in \mathbb{F}^n$ such that $y = Ax$ and $y' = Ax'$. Then $x + x' \in \mathbb{F}^n$ by definition II.2.1.4 and $\alpha x \in \mathbb{F}^n$ by definition II.2.1.5. Further, we have $A(x + x') = Ax + Ax'$ (by part (c) of proposition II.4.1.10) and $A(\alpha x) = \alpha(Ax)$ (by part (b) of proposition II.4.1.19), which yield $A(x + x') = y + y'$ and $A(\alpha x) = \alpha y$. We therefore conclude that $y + y' \in \operatorname{Im} A$ and $\alpha y \in \operatorname{Im} A$ by definition II.5.10.1.

Proposition II.5.10.9 shows that, for any matrix A, the sets $\operatorname{Ker} A$ and $\operatorname{Im} A$ are nonempty and closed under the linear operations of the respective column-vector spaces, which are vector spaces by corollary II.2.2.4. In the following chapter III, we will treat such subsets of vector spaces in full generality, keeping in mind the fundamental subspaces of a matrix as important examples.

Proposition II.5.10.10 (fundamental subspaces of matrices under matrix multiplication)**.** For $m, n, r \in \mathbb{N}$, consider $U \in \mathbb{F}^{m \times r}$ and $W \in \mathbb{F}^{r \times n}$. Then the following statements hold.

(a) $\operatorname{Im} UW \subseteq \operatorname{Im} U$. Furthermore, $\operatorname{Im} UW = \operatorname{Im} U$ if $\operatorname{Im} W = \mathbb{F}^r$.

(b) $\operatorname{Ker} W \subseteq \operatorname{Ker} UW$. Furthermore, $\operatorname{Ker} W = \operatorname{Ker} UW$ if $\operatorname{Ker} U = \{0\}$.

Proof. Applying definition II.5.10.1, we obtain

$$\operatorname{Im} A = \{UWx \colon x \in \mathbb{F}^n\} = \{Uz \colon z \in \operatorname{Im} W\} \subseteq \{Uz \colon z \in \mathbb{F}^r\} = \operatorname{Im} U\,,$$

where the inclusion becomes an equality when $\operatorname{Im} W = \mathbb{F}^r$. This proves part (a).

Similarly, we have

$$\operatorname{Ker} W = \{x \in \mathbb{F}^n \colon Wx = 0\} \subseteq \{x \in \mathbb{F}^n \colon UWx = 0\} = \operatorname{Ker} UW\,,$$

where the inclusion becomes an equality when $\operatorname{Ker} U = \{0\}$. This shows part (b).

Example II.5.10.11. The inclusions asserted in parts (a) and (b) of proposition II.5.10.10 may be proper (strict). Indeed, for the matrices U and W considered in example II.5.10.7, we find that UW is a zero matrix, so that $\operatorname{Im} UW = \{0\}$ and $\operatorname{Ker} UW = \mathbb{F}^2$. These are, respectively, a strict subset and a strict superset of $\operatorname{Im} U = \{(\alpha, 0) \colon \alpha \in \mathbb{F}\} = \operatorname{Ker} W$.

Example II.5.10.12. The equalities given in parts (a) and (b) of proposition II.5.10.10 may hold when the respective assumptions do not hold. Indeed, the matrix $U \in \mathbb{F}^{2\times 2}$ considered in example II.5.10.7 satisfies $U^2 = U$ and therefore $\operatorname{Im} U^2 = \operatorname{Im} U$ and $\operatorname{Ker} U^2 = \operatorname{Ker} U$, even though, as we showed in example II.5.10.7, $\operatorname{Im} U \neq \mathbb{F}^2$ and $\operatorname{Ker} U \neq \{0\}$.

In general, the fundamental subspaces of a product of two matrices are determined by how the images and the kernels of the factors are related to each other. For example, in the product $U^2 = U \cdot U$, the left factor maps into zero only the zero vector produced by the right factor, which is why the kernel of U^2 is equal to that of U.

Let us now see how lemma II.5.3.1 can be used for finding the fundamental subspaces of non-square full-rank matrices.

Lemma II.5.10.13 (fundamental subspaces of full-rank matrices)**.**

(a) Assume that $n \in \mathbb{N}$ and a matrix $A \in \mathbb{F}^{n\times n}$ is invertible. Then $\operatorname{Im} A = \mathbb{F}^n$ and $\operatorname{Ker} A = \{0\}$.

(b) Assume that $m, r \in \mathbb{N}$ are such that $r < m$ and $U \in \mathbb{F}^{m\times r}$. Let $P \in \mathbb{F}^{m\times m}$ be a permutation matrix such that the leading principal submatrix $\widetilde{U}_1$ of order r of

$$\widetilde{U} = PU = \begin{bmatrix} \widetilde{U}_1 \\ \widetilde{U}_2 \end{bmatrix}$$

is invertible. Then $\operatorname{Ker} U = \{0\}$ and $\operatorname{Im} U = \operatorname{Im} R$ for

$$R = P^{\mathsf{T}} \cdot \begin{bmatrix} I_1 \\ \widetilde{U}_2 \, \widetilde{U}_1^{-1} \end{bmatrix} \in \mathbb{F}^{m \times r},$$

where I_1 is the identity matrix of order r.

(c) Assume that $n, r \in \mathbb{N}$ are such that $r < n$ and $W \in \mathbb{F}^{r \times n}$. Let $Q \in \mathbb{F}^{n \times n}$ be a permutation matrix such that the leading principal submatrix $\widetilde{W}_1$ of order r of

$$\widetilde{W} = WQ^{\mathsf{T}} = \begin{bmatrix} \widetilde{W}_1 & \widetilde{W}_2 \end{bmatrix}$$

is invertible. Then $\operatorname{Im} W = \mathbb{F}^r$ and $\operatorname{Ker} W = \operatorname{Im} K$ for

$$K = Q^{\mathsf{T}} \cdot \begin{bmatrix} -\widetilde{W}_2 \widetilde{W}_1^{-1} \\ I_2 \end{bmatrix} \in \mathbb{F}^{n \times (n-r)},$$

where I_2 is the identity matrix of order $n - r$.

Proof. Let us first prove part (a). The invertibility of A implies the following (cf. part (a) of lemma II.5.3.1). First, the equality $Ax = y$ holds for any $y \in \mathbb{F}^n$ with $x = A^{-1}y \in \mathbb{F}^n$, so $\operatorname{Im} A = \mathbb{F}^n$. Second, the equality $Ax = 0$ is equivalent to $x = 0$ for any $x \in \mathbb{F}^n$, so $\operatorname{Ker} A = \{0\}$.

Let us now prove part (b). First, for any $z \in \mathbb{F}^r$, we have $Uz = 0 \Leftrightarrow PUz = 0 \Leftrightarrow \widetilde{U}z = 0 \Leftrightarrow z = 0$, the first equivalence being due to proposition II.4.3.7 and the third, for example, due to part (b) of lemma II.5.3.1. As a result, we have $\operatorname{Ker} U = \{0\}$. Second, we note that $U = R\widetilde{U}_1$ and $R = U\widetilde{U}_1^{-1}$. These equalities imply $\operatorname{Im} U \subseteq \operatorname{Im} R$ and $\operatorname{Im} R \subseteq \operatorname{Im} U$ by part (a) of proposition II.5.10.10, so we have $\operatorname{Im} U = \operatorname{Im} R$.

Finally, we prove part (c). First, for any $z \in \mathbb{F}^r$, part (c) of lemma II.5.3.1 establishes the existence of $\widetilde{x} \in \mathbb{F}^n$ such that $\widetilde{W}\widetilde{x} = z$, which is equivalent to $Wx = z$ for $x = Q^{\mathsf{T}}\widetilde{x} \in \mathbb{F}^n$. This proves $\operatorname{Im} W = \mathbb{F}^r$. Second, we have

$$Wx = 0 \Leftrightarrow \widetilde{W}(Qx) = 0 \Leftrightarrow x = Q^{\mathsf{T}} \cdot \begin{bmatrix} -\widetilde{W}_1^{-1}\widetilde{W}_2 \, \widetilde{x}_2 \\ \widetilde{x}_2 \end{bmatrix} \text{ with some } \widetilde{x}_2 \in \mathbb{F}^{n-r},$$

where the last equivalence is due to part (c) of lemma II.5.3.1. The last condition, in turn, is equivalent to $x = K\widetilde{x}_2$ with some $\widetilde{x}_2 \in \mathbb{F}^{n-r}$, i.e., to $x \in \operatorname{Im} K$. This proves $\operatorname{Ker} W = \operatorname{Im} K$.

Remark II.5.10.14. Note that the assumptions made in part (b) and part (c) of lemma II.5.10.13 are equivalent (by part (a) of theorem II.5.9.7, proposition II.4.3.7 and proposition II.4.1.22) to that U and W are of rank r (i.e., full-rank matrices).

Remark II.5.10.15. An important consequence of proposition II.5.10.10 is that low-rank matrix factorization allows for finding the four fundamental subspaces of any matrix. Consider $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$ nonzero. Let $r = \operatorname{rank} A$. Assume that $U \in \mathbb{F}^{m \times r}$ and $W \in \mathbb{F}^{r \times n}$ are such that $A = UW$. Then $\operatorname{rank} U = r = \operatorname{rank} W$ by proposition II.4.4.11, so we can call the factorization *rank-revealing*. By proposition II.5.10.10, we then have $\operatorname{Im} A = \operatorname{Im} U$ and $\operatorname{Ker} A = \operatorname{Ker} W$, and these two subspaces can be found using lemma II.5.10.13.

A rank-revealing factorization $A = UW$ can be obtained, in particular, by cross approximation or, more specifically, by the pivoted LU decomposition (see theorem II.5.9.11).

Let us now follow up on lemma II.5.10.13 and prove the converses of some of the implications given therein.

Lemma II.5.10.16. Assume that $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Then the following holds:

(a) if $\operatorname{Im} A = \mathbb{F}^m$, then $\operatorname{rank} A = m \leq n$;

(b) if $\operatorname{Ker} A = \{0\}$, then $\operatorname{rank} A = n \leq m$.

Proof. Let us set $r = \operatorname{rank} A$. Then $r \leq \min\{m, n\}$ by corollary II.4.4.8. If $\operatorname{Im} A = \mathbb{F}^m$ or $\operatorname{Ker} A = \{0\}$, it follows from definition II.5.10.1 that the matrix A is nonzero and hence $r \in \mathbb{N}$ (by definition II.4.4.1). By definition II.4.4.1 and proposition II.4.4.4, there exist matrices $U \in \mathbb{F}^{m\times r}$ and $W \in \mathbb{F}^{r\times n}$ such that $A = UW$. Proposition II.4.4.11 then yields $\operatorname{rank} U = r = \operatorname{rank} W$.

First we prove part (a), assuming $\operatorname{Im} A = \mathbb{F}^m$. Suppose $r < m$. Then the matrix U is tall. By part (a) of theorem II.5.9.7, proposition II.4.3.7 and proposition II.4.1.22, the assumption of part (b) of lemma II.5.10.13 is satisfied, and the conclusion then yields $\operatorname{Im} U \neq \mathbb{F}^m$. Applying part (a) of proposition II.5.10.10, we then obtain $\operatorname{Im} A \neq \mathbb{F}^m$, which contradicts the assumption. This proves $m \leq r$ and therefore $\operatorname{rank} A = m \leq n$.

Now we prove part (b), assuming $\operatorname{Ker} A = \{0\}$. Suppose $r < n$. Then the matrix W is wide. By part (a) of theorem II.5.9.7, proposition II.4.3.7 and proposition II.4.1.22, the assumption of part (c) of lemma II.5.10.13 is satisfied, and the conclusion then yields $\operatorname{Ker} W \neq \{0\}$. Applying part (b) of proposition II.5.10.10, we then obtain $\operatorname{Ker} A \neq \{0\}$, which contradicts the assumption. This results in $n \leq r$ and therefore $\operatorname{rank} A = n \leq m$.

The following corollary, following from lemma II.5.10.16 with $m = n$ and part (a) of lemma II.5.10.13, relates the fundamental subspaces and the invertibility of a square matrix.

Corollary II.5.10.17. Let $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Then the following statements are equivalent:

(i) A is invertible;

(ii) $\operatorname{Im} A = \mathbb{F}^n$;

(iii) $\operatorname{Ker} A = \{0\}$.

Proof. The implications (i) ⇒ (ii) and (i) ⇒ (iii) follow immediately from part (a) of lemma II.5.10.13. Conversely, any of statements (ii) and (iii) implies that the matrix A is full rank by lemma II.5.10.16, and lemma II.5.9.3 then yields that the matrix A is invertible.

We close this chapter II by relating the low-rank factorization (§ II.4.4) and the image of a matrix.

First, we note that lemma II.5.9.18 can be equivalently expressed in terms of images as follows. In effect, this equivalent expression strengthens proposition II.5.10.10, which we established above independently.

Proposition II.5.10.18 (minimal factorization and image equality)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$ be of rank $r \in \mathbb{N}$.

(a) Let $U \in \mathbb{F}^{m\times r}$. Then a matrix $Y \in \mathbb{F}^{n\times r}$ such that $A = UY^\mathsf{T}$ exists if and only if $\operatorname{Im} A = \operatorname{Im} U$;

(b) Let $V \in \mathbb{F}^{n\times r}$. Then a matrix $X \in \mathbb{F}^{m\times r}$ such that $A = XV^\mathsf{T}$ exists if and only if $\operatorname{Im} A^\mathsf{T} = \operatorname{Im} V$;

Further, we can interpret any matrix factorization in terms of images. We do this by translating lemma II.5.9.19.

Proposition II.5.10.19 (factorization and image inclusion)**.** Consider $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$.

(a) Let $p \in \mathbb{N}$ and $U \in \mathbb{F}^{m\times p}$. Then a matrix $Y \in \mathbb{F}^{n\times p}$ such that $A = UY^\mathsf{T}$ exists if and only if $\operatorname{Im} A \subseteq \operatorname{Im} U$;

(b) Let $q \in \mathbb{N}$ and $V \in \mathbb{F}^{n\times q}$. Then a matrix $X \in \mathbb{F}^{m\times q}$ such that $A = XV^\mathsf{T}$ exists if and only if $\operatorname{Im} A^\mathsf{T} \subseteq \operatorname{Im} V$;

(c) Let $p, q \in \mathbb{N}$ and $U \in \mathbb{F}^{m\times p}$, $V \in \mathbb{F}^{n\times q}$. Then a matrix $C \in \mathbb{F}^{p\times q}$ such that $A = UCV^\mathsf{T}$ exists if and only if $\operatorname{Im} A \subseteq \operatorname{Im} U$ and $\operatorname{Im} A^\mathsf{T} \subseteq \operatorname{Im} V$.

Proof. Let us first prove part (a). The “only if” implication holds by part (a) of proposition II.5.10.10, so it remains to prove the “if” implication. Let us assume that $\operatorname{Im} A \subseteq \operatorname{Im} U$ and denote the columns of A by $a_1, \ldots, a_n$. These columns belong to $\operatorname{Im} A$ and therefore to $\operatorname{Im} U$, so that there, for every $j \in \{1, \ldots, n\}$, there exists $w_j \in \mathbb{F}^p$ such that $a_j = Uw_j$. Introducing $W = [\, w_1 \cdots w_n \,] \in \mathbb{F}^{p\times n}$, we obtain $A = UW$ (by rule (a) of remark II.4.1.5). Then $A = UX^\mathsf{T}$ holds with $X = W^\mathsf{T} \in \mathbb{F}^{n\times p}$.

Part (b) follows from part (a) by transposition. Developing a detailed proof is left to the reader as an exercise.

Part (c) follows from parts (a) and (b) of this proposition II.5.10.19 and from lemma II.5.9.19.

We remark that part (c) of proposition II.5.10.19 can be proven using lemma II.5.10.13 and proposition II.5.10.10 instead of lemma II.5.9.19, which is left to the reader as an exercise.

Due to definition II.4.4.1 and proposition II.4.4.4, the following is an immediate corollary of propositions II.5.10.18 and II.5.10.19.

Corollary II.5.10.20 (matrix rank and image inclusion)**.** Let $m, n, r \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$.

(a) The following statements are equivalent:

(i) $\operatorname{rank} A \leq r$;

(ii) there exists $U \in \mathbb{F}^{m\times r}$ such that $\operatorname{Im} A \subseteq \operatorname{Im} U$;

(iii) there exists $V \in \mathbb{F}^{n\times r}$ such that $\operatorname{Im} A^\mathsf{T} \subseteq \operatorname{Im} V$.

(b) Assume additionally that $r = \operatorname{rank} A$. Then the following holds:

(iv) $\operatorname{Im} A \subseteq \operatorname{Im} U$ implies $\operatorname{Im} A = \operatorname{Im} U$ for any $U \in \mathbb{F}^{m\times r}$;

(v) $\operatorname{Im} A^{\mathsf{T}} \subseteq \operatorname{Im} V$ implies $\operatorname{Im} A^{\mathsf{T}} = \operatorname{Im} V$ for any $V \in \mathbb{F}^{n \times r}$.

Comparing this § II.5.10 with § II.5.9, one can realize that the notions of the fundamental subspaces of matrices, at this point, give no more than a language for describing relations between matrices alternative to that of matrix factorization. For example, the above proof of part (c) of proposition II.5.10.19 consists in using parts (a) and (b) to equivalently express the statement of lemma II.5.9.19 in terms of image inclusion. It remains, however, to amplify our language of images with an intrinsic way of expressing the redundancy and non-redundancy of the columns of a matrix, parallel to the notions of the rank deficiency and of the full-rank property of the matrix, which would lead to a counterpart of lemma II.5.9.20. Lemmata II.5.9.18 and II.5.9.20 give a glimpse of what is essential in that regard: the full-rank property of U is related to the uniqueness of the representation of the elements of $\operatorname{Im} U$ and the minimality of the number r of columns of U is related to the minimality of $\operatorname{Im} U$ as a superset of $\operatorname{Im} A$. In the following chapter, we will introduce the notions of spanning sets, bases and dimensions. We will do that in a more abstract setting, where the language of matrix factorizations is not applicable directly — but is nevertheless useful once we have reduced the abstract setting to that of column vectors.

Chapter III. Vector spaces. Linear mappings

§ III.1. Subspaces of vector spaces

§ III.1.1. Definitions and properties

Definition III.1.1.1 (a subspace of a vector space)**.** Let V be a vector space over a field $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$ of addition and multiplication by a scalar. Consider a *nonempty subset* U of V. Then U is called a *subspace* of the vector space V if the set U is closed with respect to its linear operations:

$$u \oplus u' \in U \quad \text{and} \quad \alpha \odot u \in U \quad \text{for all} \quad u, u' \in U \quad \text{and} \quad \alpha \in \mathbb{F}\,. \tag{III.1.1.1}$$

A good starting point for checking whether a subset U of a set V that is a vector space is checking whether the additive identity element $\vec{0}$ of V is contained in U. Indeed, a nonempty subset U of V contains at least one element $u \in U$, for which (III.1.1.1) implies $0 \odot u \in U$, where 0 is the additive identity element of $\mathbb{F}$. As we mentioned in remark II.2.2.2, $0 \odot u = \vec{0}$.

The fact that a subset U of a vector space V is a subspace of V is often expressed by the equivalent notations $U \subseteq V$ and $V \supseteq U$. When this can lead to confusion, it is advisable to specify whether the subset or the subspace relation is meant by the use of such notations — or to avoid it altogether.

Example III.1.1.2 (the fundamental subspaces of a matrix are indeed subspaces)**.** Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. Proposition II.5.10.9 implies that the four "fundamental subspaces" (see definition II.5.10.1) of the matrix A are indeed subspaces in the sense of definition III.1.1.1: $\operatorname{Im} A$ and $\operatorname{Ker} A^{\mathsf{T}}$ are subspaces of $\mathbb{F}^m$, whereas $\operatorname{Ker} A$ and $\operatorname{Im} A^{\mathsf{T}}$ are subspaces of $\mathbb{F}^n$.

The term *subspace* is justified by the following proposition.

Proposition III.1.1.3 (subspace of a vector space is a vector space)**.** Let U be a subset of a set V.

(a) Assume that V is a vector space over the field $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$ of addition and multiplication by a scalar. Let U be a subspace of V and $\vec{0}$ denote the additive identity element of V. Consider the functions $\boxplus \colon U \times U \to U$ and $\boxdot \colon \mathbb{F} \times U \to U$ obtained by restricting $\oplus$ and $\odot$:

$$u \boxplus u' = u \oplus u' \quad \text{and} \quad \alpha \boxdot u = \alpha \odot u \quad \text{for all } u, u' \in U \text{ and } \alpha \in \mathbb{F}.$$

Then U is a vector space over the field $\mathbb{F}$ and $\vec{0}$ is the additive identity element of U.

(b) Assume that U is a vector space over the field $\mathbb{F}$ with respect to operations $\boxplus$ and $\boxdot$ of addition and multiplication by a scalar. Let $\vec{0}$ denote the additive identity element of U. Consider functions $\oplus \colon V \times V \to V$ and $\odot \colon \mathbb{F} \times V \to V$ extending $\boxplus$ and $\boxdot$ in the sense that

$$u \oplus u' = u \boxplus u' \quad \text{and} \quad \alpha \odot u = \alpha \boxdot u \quad \text{for all } u, u' \in U \text{ and } \alpha \in \mathbb{F}.$$

Assume that V is a vector space over the field $\mathbb{F}$ with respect to operations $\oplus$ and $\odot$. Then U is a subspace of V and $\vec{0}$ is the additive identity element of V.

Proof. In part (b), we note that the restrictions are correctly defined due to the closedness (III.1.1.1) of U under the operations of V. The vector-space axioms (conditions (a) to (f) in definition II.2.2.1) for U then follow by the restriction to U of the same vector-space axioms for V.

In part (a), we first note that the definition of $\oplus$ and $\odot$ implies that U is closed under these operations in the sense of (III.1.1.1). Since U is nonempty, it is therefore a subspace of V by definition III.1.1.1.

Finally, let us prove simultaneously for parts (a) and (b) that the additive identity elements coincide. One can show that the multiplication of any vector from a vector space by the additive identity element of the field produces the additive identity element of the vector space (see Problem 2(c) in Assignment 1). Then, taking any $u \in U$, we obtain that $0 \odot u$ and $0 \boxdot u$ are the additive identity elements of V and U, respectively. Since the operations agree on U, these identity elements coincide.

Definition III.1.1.1 and proposition III.1.1.3 immediately imply that the subspace relation is transitive, just as the subset relation: $W \subseteq V$ and $V \subseteq U$ imply $W \subseteq U$.

Proposition III.1.1.4. Let U be a vector space, a subset V of the set U be a subspace of the vector space U and a subset W of the set V be a subspace of the vector space V. Then the subset W of U is a subspace of the vector space U.

Proof. The proof is left to the reader as an exercise.

§ III.1.2. Examples of subspaces

Example III.1.2.1. Let 0 and 1 denote the additive and multiplicative identity elements of the field $\mathbb{R}$ of real numbers. Consider $V = \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y)\colon x, y \in \mathbb{R}\}$ and its subset

$$U = \mathbb{R} \times \{0\} = \{(x, 0)\colon x \in \mathbb{R}\}\,.$$

(a) As we noted in example II.1.2.3, set V is a field (the field of complex numbers, denoted by $\mathbb{C}$, see § I.2) with additive and multiplicative identity elements $\vec{0} = (0, 0)$ and $\vec{1} = (1, 0)$ with respect to the operations $+\colon V \times V \to V$ and $\cdot : V \times V \to V$ of addition and multiplication given by (I.2.2.1) and (I.2.3.1). The notation $\mathbb{C}$ is used to refer to V as a field, i.e., in conjunction with the specific two field operations.

As follows from definitions II.1.1.1 and II.2.2.1, any field $\mathbb{F}$ is a vector space over itself with respect to its field operations: in the case of $V = \mathbb{F}$, the closedness conditions of the two definitions are identical and the vector-space axioms trivially follow from the field axioms. In particular, set $V = \mathbb{R}^2$ is a vector space over the field $\mathbb{F} = \mathbb{C}$ with respect to the operations $\oplus = +$ and $\odot = \cdot$ of addition and multiplication by a scalar from $\mathbb{C}$.

Subset U of set V is not a subspace of the vector space V by definition III.1.1.1. Indeed, for the scalar $\mathsf{i} = (0, 1) \in \mathbb{C}$ and for the vector $\vec{1} = (1, 0) \in U$, we have $\mathsf{i} \odot \vec{1} = \mathsf{i} \cdot \vec{1} = \mathsf{i}$ by (I.2.3.1), so that $\mathsf{i} \odot \vec{1} \notin U$. Addition, however, is not the issue here: due to (I.2.2.1), we have $u \oplus u' = u + u' \in U$ for any $u, u' \in U$.

(b) As follows from corollary II.2.2.4, set $V = \mathbb{R}^2$ is a vector space over $\mathbb{F} = \mathbb{R}$ with respect to the componentwise operations $+$ and "$\cdot$" of addition and multiplication by a real scalar, given by definitions II.2.1.4 and II.2.1.5. In this case, subset U of set V is a subspace of the vector space V by definition III.1.1.1: due to definitions II.2.1.4 and II.2.1.5, for any vectors $u, u' \in U \in U$ and for any scalar $\alpha \in \mathbb{R}$, we have $u + u' \in U$ and $\alpha \cdot u \in U$. The difference from part (a) is that, considering V as a real vector space, the multiplication of the elements of U by scalars is confined to real scalars, in which case the subset is closed under the operation.

Example III.1.2.2 (trivial subspaces)**.** Let V be a vector space over the field $\mathbb{F}$ and $\vec{0}$ denote its additive identity element of V. Then the subset $\{\vec{0}\}$ of set V is a subspace of the vector space V. Furthermore, the subspace $\{\vec{0}\}$ of V is called the *trivial subspace* of the vector space V (see definition II.2.2.1). Every subspace U of the vector space V is a vector space with $\vec{0}$ as additive identity element (by part (a) of proposition III.1.1.3), so that $\{\vec{0}\}$ is a subset of set U (and, by the same logic, a subspace of the vector space U). The trivial subspace of V (see definition II.2.2.1) is therefore the smallest of all subspaces of V.

Here is another trivial example (of a subspace that is not necessarily trivial, though).

Example III.1.2.3 (any vector space is its own subspace)**.** Any vector space V is a subspace of itself.

Example III.1.2.4. The set $V = \mathbb{R}^2$ is a vector space over $\mathbb{R}$ with respect to the componentwise operations of addition and multiplication by a scalar, given by definitions II.2.1.4 and II.2.1.5, and the additive identity element of V is $\vec{0} = (0, 0)$ (see corollary II.2.2.4). One can verify that each of the sets

$$U = \{(\alpha, -\alpha) \colon \alpha \in \mathbb{R}\} = \{(x, y) \in V \colon x + y = 0\} \subset V$$

and

$$W = \{(\alpha, 2\alpha) \colon \alpha \in \mathbb{R}\} = \{(x, y) \in V \colon 2x - y = 0\} \subset V$$

of V, shown in figures III.1.2.2(a) and III.1.2.2(c), is closed under the operations of the vector space V and is therefore a subspace of the vector space V by definition III.1.1.1.

In contrast, the set

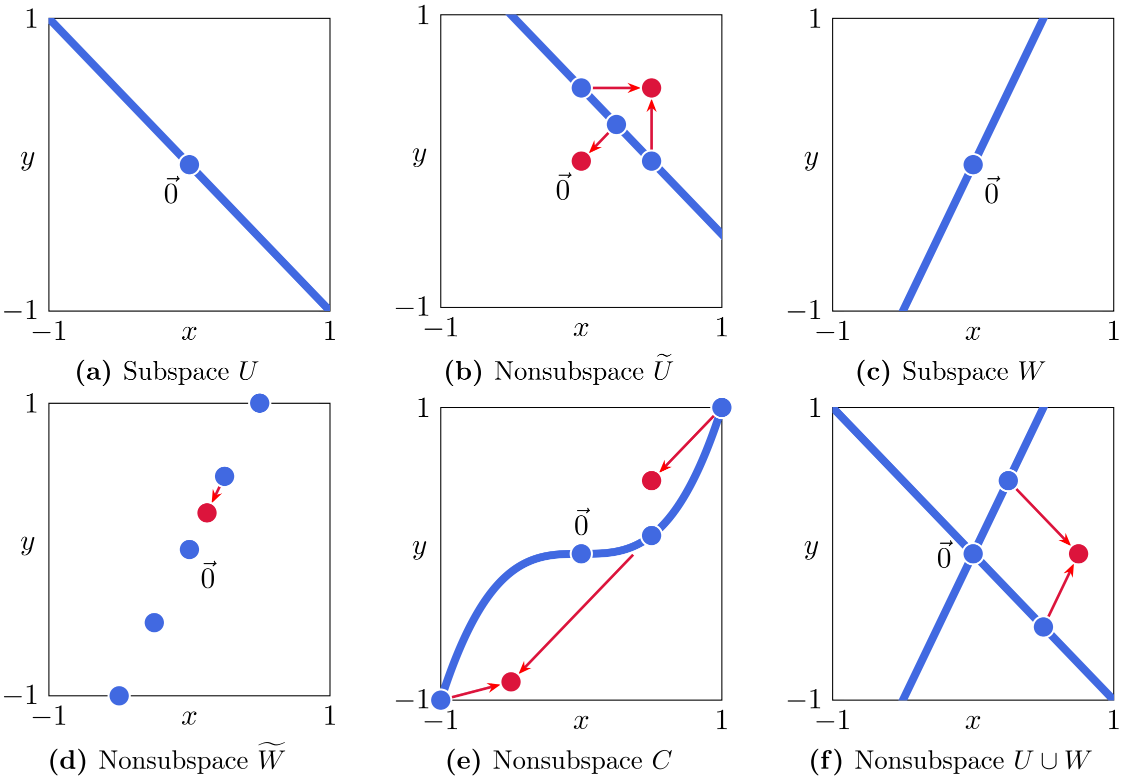
$$\widetilde{U} = \left\{\left(\alpha, \frac{1}{2} - \alpha\right) \colon x \in \mathbb{R}\right\} = \left\{(x, y) \in V \colon x + y = \frac{1}{2}\right\} \subset V$$

is not a subspace of the vector space V since it is closed under neither of its two operations. First, we have $0 \cdot (\frac{1}{4}, \frac{1}{4}) = \vec{0} \notin \widetilde{U}$ while $(\frac{1}{4}, \frac{1}{4}) \in \widetilde{U}$ and $0 \in \mathbb{R}$. Second, we have $(\frac{1}{2}, 0) + (0, \frac{1}{2}) = (\frac{1}{2}, \frac{1}{2}) \notin \widetilde{U}$ while $(\frac{1}{2}, 0) \in \widetilde{U}$ and $(0, \frac{1}{2}) \in \widetilde{U}$. These counterexamples to the closedness of $\widetilde{U}$ under the operations of V are shown in figure III.1.2.2(b).

Similarly, the set

$$C = \{(\alpha, \alpha^3) \colon \alpha \in \mathbb{R}\} \subset V$$

is not a subspace of the vector space V since it is closed under neither of its two operations, even though the inclusion $\vec{0} \in C$ holds. First, we note $\frac{1}{2} \cdot (1, 1) = (\frac{1}{2}, \frac{1}{2}) \notin C$ while $(1, 1) \in C$ and $\frac{1}{2} \in \mathbb{R}$. Second, we have $(\frac{1}{2}, \frac{1}{8}) + (-1, -1) = (\frac{1}{2}, -\frac{7}{8}) \notin C$ while $(\frac{1}{2}, \frac{1}{8}) \in C$ and $(-1, -1) \in C$. These counterexamples to the closedness of C under the operations of V are shown in figure III.1.2.2(e).



(a) Subspace U **(b)** Nonsubspace $\widetilde{U}$ **(c)** Subspace W

(d) Nonsubspace $\widetilde{W}$ **(e)** Nonsubspace C **(f)** Nonsubspace $U \cup W$

Figure III.1.2.2. Illustration for example III.1.2.4. The sets considered are depicted in blue. The axes are restricted to $[-1, 1]$ for both the variables. In each case, individual points contained in the set under consideration are depicted in blue and individual points not contained, in red.

The set

$$\widetilde{W} = \left\{ \left(\frac{k}{4}, \frac{k}{2} \right) : k \in \mathbb{Z} \right\} \subset W \subset V$$

is not a subspace of the vector space V (or W) since it is not closed under the multiplication by a scalar, even though it is closed under the addition. Indeed, we note that $\frac{1}{2} \cdot (\frac{1}{4}, \frac{1}{2}) = (\frac{1}{8}, \frac{1}{4}) \notin \widetilde{W}$ while $(\frac{1}{4}, \frac{1}{2}) \in \widetilde{W}$ and $\frac{1}{2} \in \mathbb{R}$, see figure III.1.2.2(d).

Similarly, the set $U \cup W \subset V$ is not a subspace of the vector space V since it is not closed under the addition, even though it is closed under the multiplication by a scalar. Indeed, we note that $(\frac{1}{4}, \frac{1}{2}) + (\frac{1}{2}, -\frac{1}{2}) = (\frac{3}{4}, 0) \notin W \cup U$ while $(\frac{1}{2}, -\frac{1}{2}) \in U \cup W$ and $(\frac{1}{4}, \frac{1}{2}) \in U \cup W$, see figure III.1.2.2(f).

Example III.1.2.5. As follows from corollary II.2.2.4, the set $V = \mathbb{R}^3$ is a vector space over $\mathbb{R}$ with respect to the componentwise operations of addition and multiplication by a scalar, given by definitions II.2.1.4 and II.2.1.5. One can verify that the set

$$W = \{(\alpha, \beta, \alpha + \beta) : \alpha, \beta \in \mathbb{R}\} = \{(x, y, z) \in V : x + y - z = 0\} \subset V$$

is a subspace of the vector space V since it is closed under its operations. On the other hand, the set

$$\widetilde{W} = \{(\alpha, \beta, \alpha + \beta - 1) : \alpha, \beta \in \mathbb{R}\} = \{(x, y, z) \in V : x + y - z = 1\} \subset V$$

is not a subspace of the vector space V since it closed under neither of its two operations. For example, we have $0 \cdot (1,1,1) = (0,0,0) = \vec{0} \notin \widetilde{W}$ while $(1,1,1) \in \widetilde{W}$ and $0 \in \mathbb{R}$. Another counter-example is as follows: $(1,0,0) + (1,1,1) = (2,1,1) \notin \widetilde{W}$ while $(1,0,0) \in \widetilde{W}$ and $(1,1,1) \in \widetilde{W}$.

Example III.1.2.6 (the set of lower-triangular matrices is a subspace)**.** For any $n \in \mathbb{N}$, the set $\mathcal{M}$ of real lower-triangular matrices of size $n \times n$, which is a subset of $\mathbb{R}^{n\times n}$, is also a subspace of $\mathbb{R}^{n\times n}$.

Indeed, the set $\mathcal{M}$ is clearly nonempty: for example, the identity matrix of order n belongs to it (we could also consider the zero matrix or any other lower-triangular matrix). Further, if $A, B \in \mathbb{R}^{n\times n}$ are lower-triangular matrices and $\alpha \in \mathbb{R}$, then $A_{ij} = B_{ij} = 0$ for any $i, j \in \{i, \ldots, n\}$ such that $i < j$. This implies $(\alpha \cdot A + B)_{ij} = \alpha A_{ij} + B_{ij} = \alpha \cdot 0 + 0 = 0$ for any $i, j \in \{i, \ldots, n\}$ such that $i < j$, i.e., $\alpha \cdot A + B \in \mathcal{M}$. Applying definition III.1.1.1, we therefore conclude that $\mathcal{M}$ is a subspace of $\mathbb{R}^{n\times n}$.

Example III.1.2.7 (the set of permutation matrices is not a subspace)**.** Let $n \in \mathbb{N}$. The set $\mathscr{P}_n$ of permutation matrices of order n over $\mathbb{F}$, which is a subset of $\mathbb{F}^{n\times n}$, is *not* a subspace of $\mathbb{F}^{n\times n}$.

Indeed, if $P \in \mathscr{P}_n$, then $P + P$ is not: $0 + 0 = 0$ holds, but $1 + 1 \neq 1$ does not since $1 \neq 0$ by definition II.1.1.1, so every row of $P + P$ contains $n - 1$ zeros and one more element, which is distinct from 1. This shows that the set $\mathscr{P}_n$ is not closed under the linear operations of $\mathbb{F}^{n\times n}$.

Example III.1.2.8 (polynomial spaces)**.** For a nonempty subset $\mathcal{D}$ of the field $\mathbb{F}$, we consider the vector space $\mathcal{F}(\mathcal{D}, \mathbb{F})$ of $\mathbb{F}$-valued functions defined on $\mathcal{D}$ (see example II.2.2.6). For notational convenience, we denote the operations of $\mathcal{F}(\mathcal{D}, \mathbb{F})$ by $+$ and "$\cdot$".

Consider the set of *algebraic polynomials over the field* $\mathbb{F}$ *defined on* $\mathcal{D}$: for each $n \in \mathbb{N}_0$, we denote by $\mathcal{P}_n(\mathcal{D}, \mathbb{F})$ the set of algebraic polynomials over $\mathbb{F}$ of degree at most n:

$$\mathcal{P}_n(\mathcal{D}, \mathbb{F}) = \Big\{ p \in \mathcal{F}(\mathcal{D}, \mathbb{F}) \colon p(t) = \alpha_0 + \sum_{k=1}^{n} \alpha_k t^k \text{ for all } t \in \mathcal{D} \text{ with } \alpha_0, \ldots, \alpha_n \in \mathbb{F} \Big\} . \tag{III.1.2.1}$$

Let us also define

$$\mathcal{P}(\mathcal{D}, \mathbb{F}) = \bigcup_{n \in \mathbb{N}_0} \mathcal{P}_n(\mathcal{D}, \mathbb{F}) . \tag{III.1.2.2}$$

These sets form an infinite chain of nested *subsets* of $\mathcal{F}(\mathcal{D}, \mathbb{F})$:

$$\mathcal{P}_{n-1}(\mathcal{D}, \mathbb{F}) \subseteq \mathcal{P}_n(\mathcal{D}, \mathbb{F}) \subseteq \mathcal{P}(\mathcal{D}, \mathbb{F}) \subseteq \mathcal{F}(\mathcal{D}, \mathbb{F}) \quad \text{for each} \quad n \in \mathbb{N} . \tag{III.1.2.3}$$

The first inclusion can actually be shown to be proper (strict) when $\#\mathcal{D} \geq n+1$, in which case there are enough points in $\mathcal{D}$ for $\mathcal{P}_n(\mathcal{D}, \mathbb{F})$ to contain functions not contained in $\mathcal{P}_{n-1}(\mathcal{D}, \mathbb{F})$.

In any case, the above polynomial sets are closed under the (pointwise) operations of the vector space $\mathcal{F}(\mathcal{D}, \mathbb{F})$, so (III.1.2.3) is also an infinite chain of nested *subspaces*.

Example III.1.2.9 (function spaces as subspaces)**.** Consider the vector space $V = \mathscr{F}(\mathbb{R}, \mathbb{R})$ of real-valued functions defined on the real line (see example II.2.2.6 in the case of $\mathcal{D} = \mathbb{R}$). One of the elements of V is the function $f\colon \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \begin{cases} 0 & \text{if } t \in \mathbb{Q}, \\ 1 & \text{otherwise} \end{cases} \qquad \text{for each} \quad t \in \mathbb{R}\,.$$

The subsets

$$V_{\text{odd}} = \{f \in V\colon f(-t) = -f(t) \text{ for every } t \in \mathbb{R}\}$$

and

$$V_{\text{even}} = \{f \in V\colon f(-t) = f(t) \text{ for every } t \in \mathbb{R}\}$$

of odd and even real-valued functions defined on the real line are subsets of set V closed under the operations of the vector space V and are therefore subspaces of the vector space V.

For $\tau \in \mathbb{R}$, consider the subsets

$$U_\tau = \{f \in V\colon f(\tau) = 0\} \quad \text{and} \quad \widetilde{U}_\tau = \{f \in V\colon f(\tau) = 1\}$$

of set V. Then U_τ is a subspace of the vector space V, while $\widetilde{U}_\tau$ is not: it does not contain the zero function $\vec{0}$.

Further, for $\tau \in \mathbb{R} \cup \{+\infty, -\infty\}$, consider the subsets

$$W_\tau = \{f \in V\colon \lim_{t\to\tau} f(t) = 0\} \quad \text{and} \quad \widetilde{W}_\tau = \{f \in V\colon \lim_{t\to\tau} f(t) = 1\}$$

of set V. Then W_τ is a subspace of the vector space V, while $\widetilde{W}_\tau$ is not: it does not contain the zero function $\vec{0}$.

The closedness of W_τ under the operations of the vector space V is a consequence of what is often referred to as the *linearity of function limit*:

$$\lim_{t\to\tau}\big(f(t) + g(t)\big) = \lim_{t\to\tau} f(t) + \lim_{t\to\tau} g(t) \quad \text{and} \quad \lim_{t\to\tau}\big(\alpha \cdot f(t)\big) = \alpha \cdot \lim_{t\to\tau} f(t)$$

for any $f, g \in V$ and $\alpha \in \mathbb{R}$ and $\tau \in \mathbb{R} \cup \{+\infty, -\infty\}$ if $\lim_{t\to\tau} f(t)$ and $\lim_{t\to\tau} g(t)$ exist in $\mathbb{R}$. The linearity of function limit also implies that the addition of two continuous functions and the multiplication of such a function by a real scalar produces such a function. As a result, the set of real-valued functions defined and continuous on the real line,

$$\mathscr{C} = \{f \in V\colon f \text{ is continuous}\}\,,$$

is a subset of set V that is closed under the operations of the vector space V and is therefore a subspace of the vector space V. Note that, for every $n \in \mathbb{N}_0$, the vector space $\mathcal{P}_n(\mathbb{R}, \mathbb{R})$ of polynomials over $\mathbb{R}$ of degree at most n (see example III.1.2.8) is a subspace of $\mathscr{C}$.

In § III.2, we introduce the notion of dimension for vector spaces, which allows to quantify their size.

§ III.2. Generating systems, dimension, bases, linear independence

§ III.2.1. Generating systems, spans, finite- and infinite-dimensional spaces

Recall that in definition II.2.2.7 we defined linear combinations in an abstract vector space.

Definition III.2.1.1. Let V be a vector space over $\mathbb{F}$ with respect to operations $+$ and "$\cdot$" of addition and multiplication by a scalar. Consider $n \in \mathbb{N}$ vectors $v_1, \ldots, v_n \in V$. The set of all linear combinations of $v_1, \ldots, v_n$ obtained using the operations of the vector space V

is referred to as the *span*, or the *linear hull*, of $v_1, \dots, v_n$ and is denoted by $\operatorname{span}\{v_1, \dots, v_n\}$:

$$\operatorname{span}\{v_1, \dots, v_n\} = \{\alpha_1 \cdot v_1 + \dots + \alpha_n \cdot v_n \colon \alpha_1, \dots, \alpha_n \in \mathbb{F}\}\,.$$

Let $U = \operatorname{span}\{v_1, \dots, v_n\}$. The vectors $v_1, \dots, v_n$ are said to *span*, or to *generate*, the set U, and the set $\{v_1, \dots, v_n\}$ is called a *spanning set*, or a *generating system*, for the set U.

Note that the associativity and commutativity of addition in vector spaces (conditions (a) and (b) of definition II.2.2.1) render the above notation for the span of vectors correct: the span does not depend on the ordering of the vectors and is defined in definition III.2.1.1 by the *set* formed by $v_1, \dots, v_n$. The same vector-space axioms imply the following result.

Proposition III.2.1.2 (spans are subspaces)**.** Let V be a vector space. Consider $n \in \mathbb{N}$ vectors $v_1, \dots, v_n \in V$. Then $\operatorname{span}\{v_1, \dots, v_n\}$ is a subspace of V.

The following is an immediate consequence of propositions III.2.1.2 to III.1.1.4.

Remark III.2.1.3 (spans are subspaces of superspaces)**.** Let V be a subspace of a vector space U, $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. Then V is a vector space (by proposition III.1.1.3), $\operatorname{span}\{v_j\}_{j=1}^n$ is a subspace of the vector space V (by proposition III.2.1.2) and also a subspace of the vector space U (by proposition III.1.1.4 since V is a subspace of U or, alternatively, by proposition III.2.1.2 since V is a subset of U and hence $v_1, \dots, v_n \in U$).

In the context of remark III.2.1.3, the subspace $W = \operatorname{span}\{v_1, \dots, v_n\}$ is a vector space with respect to the operations of V restricted to W. As a result, the linear combinations of $v_1, \dots, v_n$ assembled in U, V and W (using the operations of the respective vector spaces) from any coefficient tuple are equal. In many of the following results, we may therefore consider a set of vectors from a vector space spanning an entire vector space (such as W in the context of remark III.2.1.3). Such results therefore remain valid in any larger vector space (such as V or U in the context of remark III.2.1.3) of which W is a subspace.

Proposition III.2.1.4. Let V be a vector space over a field $\mathbb{F}$, $n \in \mathbb{N}$ and $v_1, \dots, v_n, v \in V$. Then $\operatorname{span}\{v_1, \dots, v_n\} \subseteq \operatorname{span}\{v_1, \dots, v_n, v\}$.

Proof. The proof immediately follows from definition III.2.1.1 and is left to the reader as an exercise.

Lemma III.2.1.5. Let V be a vector space over a field $\mathbb{F}$, $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$ and $v \in \operatorname{span}\{v_1, \dots, v_n\}$. Then $\operatorname{span}\{v_1, \dots, v_n, v\} = \operatorname{span}\{v_1, \dots, v_n\}$.

Proof. Due to proposition III.2.1.4, we only need to show the inclusion

$$\operatorname{span}\{v_1, \dots, v_n, v\} \subseteq \operatorname{span}\{v_1, \dots, v_n\}\,.$$

Let $u \in \operatorname{span}\{v_1, \dots, v_n, v\}$. By definition III.2.1.1, there exist $\alpha_1, \dots, \alpha_n \in F$ such that $v = \sum_{k=1}^n \alpha_k v_k$ and $\beta_1, \dots, \beta_n, \beta \in \mathbb{F}$ such that $u = \sum_{k=1}^n \beta_k v_k + \beta v$. Substituting the former linear combination into the latter, we obtain

$$u = \sum_{k=1}^{n} \beta_k v_k + \beta \sum_{k=1}^{n} \alpha_k v_k = \sum_{k=1}^{n} (\beta_k + \beta \alpha_k)\, v_k \,,$$

where $\beta_k + \beta\alpha_k \in \mathbb{F}$. By definition III.2.1.1, this implies that $u \in \operatorname{span}\{v_1, \ldots, v_n\}$. Since $u \in \operatorname{span}\{v_1, \ldots, v_n, v\}$ has been arbitrary, we have proved that $\operatorname{span}\{v_1, \ldots, v_n, v\} \subseteq \operatorname{span}\{v_1, \ldots, v_n\}$.

Definition III.2.1.6 (finite- and infinite-dimensional vector spaces)**.** A vector space V is called *finite dimensional* if it has a finite spanning set (in other words, if there exist $n \in \mathbb{N}$ and $v_1, \ldots, v_n \in V$ such that $V = \operatorname{span}\{v_k\}_{k=1}^n$ holds) and *infinite dimensional*, otherwise. When the vector space V is a subspace of another vector space U, it is called a *finite-dimensional subspace* of U and an *infinite-dimensional subspace* of U in the same respective cases.

Example III.2.1.7 (spanning sets of $\mathbb{R}^2$)**.** The set $V = \mathbb{R}^2$ is a vector space over $\mathbb{R}$ with respect to the componentwise operations of addition and multiplication by a scalar, given by definitions II.2.1.4 and II.2.1.5, and the additive identity element of V is $\vec{0} = (0, 0)$ (see corollary II.2.2.4). Clearly, the vector space V is nontrivial. Let us consider the vectors

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in V \,, \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in V \,, \quad u = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \in V \,.$$

From example II.2.2.11 it follows that

$$\operatorname{span}\{e_1, e_2\} = V = \operatorname{span}\{e_1, e_2, u\} \,.$$

At the same time, for any $w \in V$, there exists $v \in V$ such that $v \notin \operatorname{span}\{w\}$. Indeed, we can choose $v = e_2$ if $w \in \operatorname{span}\{e_1\}$ and $v = e_1$ otherwise. So V has a spanning set with two elements, a spanning set with three elements and no spanning set with fewer than two elements.

Example III.2.1.8 (a spanning set of $\mathbb{F}^n$)**.** Let $n \in \mathbb{N}$. Then we have $\mathbb{F}^n = \operatorname{span}\{e_j\}_{j=1}^n$, where $e_1, \ldots, e_n$ are the columns of the identity matrix I of order n over $\mathbb{F}$: $e_j = [\delta_{ij}]_{i=1}^n \in \mathbb{F}^n$ for each $j \in \{1, \ldots, n\}$. Indeed, any column vector $v \in \mathbb{F}^n$ satisfies

$$v = I \cdot v = \sum_{j=1}^{n} v_j e_j \in \operatorname{span}\{e_j\}_{j=1}^n$$

and any vector $v \in \operatorname{span}\{e_j\}_{j=1}^n$ satisfies $v \in \mathbb{F}^n$ since the vector space $\mathbb{F}^n$ is closed under its operations.

Remark III.2.1.9 (the image is the span of the columns)**.** Consider $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$ nonzero. As we pointed out in example III.1.1.2, the set $\operatorname{Im} A$ is a subspace of $\mathbb{F}^m$. With definition III.2.1.1, we can now add that the columns $a_1, \ldots, a_n$ of A span the subspace $\operatorname{Im} A$, i.e., $\operatorname{Im} A = \operatorname{span}\{a_j\}_{j=1}^n$ by definitions II.5.10.1 and III.2.1.1 and rule (a) of remark II.4.1.5. So the subspace $\operatorname{Im} A$ is finite-dimensional. Furthermore, for $r = \operatorname{rank} A$, we have $r \in \mathbb{N}$ and matrices $U \in \mathbb{F}^{m\times r}$ with columns $u_1, \dots, u_r \in \mathbb{F}^m$ and $W \in \mathbb{F}^{r\times n}$ such that $A = UW^\mathsf{T}$ exist (by corollary II.4.4.12) and satisfy (by proposition II.5.10.18) $\operatorname{Im} A = \operatorname{span}\{u_k\}_{k=1}^r$.

Example III.2.1.10 (polynomial spaces in terms of monomials)**.** Let us proceed in the setting of example III.1.2.8. For every $k \in \mathbb{N}_0$, the function $q_k \in V$ given at every $t \in \mathcal{D}$ by

$$q_k(t) = 1 \quad \text{if } k = 0 \quad \text{and} \quad q_k(t) = t^k \quad \text{if } k \in \mathbb{N} \tag{III.2.1.1}$$

in terms of the multiplicative identity element 1 and of the operations of $\mathbb{F}$ is called the *monomial* of degree k. For every $n \in \mathbb{N}_0$, we can equivalently express the definition (III.1.2.1) of $\mathcal{P}_n(\mathcal{D}, \mathbb{F})$ in terms of monomials as follows:

$$\mathcal{P}_n(\mathcal{D}, \mathbb{F}) = \left\{ \sum_{k=0}^n \alpha_k q_k \colon \alpha_0, \dots, \alpha_n \in \mathbb{F} \right\} = \operatorname{span}\{q_k\}_{k=0}^n \,, \tag{III.2.1.2}$$

where the linear combinations are constructed using the (pointwise) operations of the vector space $\mathcal{F}(\mathcal{D}, \mathbb{F})$. In other words, for any $n \in \mathbb{N}_0$, the space $\mathcal{P}_n(\mathcal{D}, \mathbb{F})$ is the span of the first $n + 1$ monomials and is therefore a finite-dimensional vector space.

In the following §§ III.2.2 and III.2.3, we rigorously develop the notion of dimension for finite-dimensional vector spaces, which will allow us, in particular, to classify other subspaces considered in example III.1.2.9 as finite- and infinite-dimensional.

§ III.2.2. Bases and dimension

Example III.2.1.7 showcases that a finite-dimensional vector space can have spanning sets of different cardinality. As example II.2.2.11 demonstrates in the case of the vector space $\mathbb{R}^2$, spanning sets of the minimum possible cardinality may stand out in representing every vector from the vector space as ones not just involving the least possible number of coefficients but also ensuring the uniqueness of the representation. We made a similar observation regarding $\mathbb{F}^n$ at the end of § II.5.10. Now we turn to the notions of a smallest spanning set and of the dimension of a vector space.

Definition III.2.2.1. Let V be a finite-dimensional vector space over the field $\mathbb{F}$ and $n \in \mathbb{N}$. If V is nontrivial, the least cardinality of its spanning set is called the *dimension* of the vector space V and is denoted by $\dim V$, and any spanning set of that cardinality is called a *basis* for V. For a trivial vector space V, the dimension is defined to be zero: $\dim V = 0$.

The wording of definition III.2.2.1 has the following intended effect for two special cases. First, a trivial vector space has no basis. Second, any nonzero vector from a nontrivial vector space spanning the entire vector space forms a basis for that vector space. For vector spaces that are larger (specifically, that are not generated by a single vector), the definition is more substantial: it calls for verifying that a given spanning set is *minimal* in the sense of cardinality. Finally, a vector space V is trivial if and only if $\dim V = 0$.

Example III.2.2.2 (a basis of $\mathbb{R}^2$)**.** In example III.2.1.7, we showed that $\mathbb{R}^2$ is spanned by $e_1, e_2 \in \mathbb{R}^2$ (the columns of the real identity matrix of order two), by e_1, u with any $u \in \mathbb{R}^2$ such that $u_2 \neq 0$, by e_2, u with any $u \in \mathbb{R}^2$ such that $u_1 \neq 0$ and by e_1, e_2, u with any $u \in \mathbb{R}^2$ but has no spanning set with fewer than two elements. We therefore conclude that $\dim \mathbb{R}^2 = 2$. As a result, all the aforementioned two-element spanning sets are bases for $\mathbb{R}^2$, whereas e_1, e_2, u do not form a basis for $\mathbb{R}^2$ for any $u \in \mathbb{R}^2$.

The conclusion of example III.2.2.2 is related to the observation, made in example II.2.2.11, that every vector $v \in \mathbb{R}^2$ has a unique representation (II.2.2.2) in the form of a linear combination of e_1 and e_2. We will soon establish that such representation uniqueness is actually equivalent, for any vectors in any vector space, to that the vectors form a basis for the vector space.

Example III.2.2.3. Following up on example III.2.1.8, for $V = \mathbb{F}^n$ with $n \in \mathbb{N}$, we note that $\dim V = n$ and the columns of the identity matrix of order n over $\mathbb{F}$ form a basis for V. since they constitute a spanning set for V while the vector space V is nontrivial and has no spanning sets with fewer elements.

Indeed, let V be spanned by $v_1, \ldots, v_r \in V$ with $r \in \mathbb{N}$. Consider $U = [u_1 \cdots u_r] \in \mathbb{F}^{n \times r}$. Then, since $e_1, \ldots, e_n \in V$, there exist $W \in \mathbb{F}^{r \times n}$ such that

$$e_j = \sum_{k=1}^{r} W_{kj} \cdot u_k \quad \text{for each} \quad j \in \{1, \ldots, n\} \,,$$

i.e., $I = UW$ and hence $n = \operatorname{rank} I \leq r$ by lemma II.5.9.2, example II.4.1.16 and proposition II.4.4.11.

Let us also note that any vector $v \in \operatorname{span}\{e_j\}_{j=1}^n = \mathbb{F}^n$ has a unique coefficient with respect to $e_1, \ldots, e_n$ in the sense of definition II.2.2.7: for any $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{F}^n$, we have

$$v = \sum_{j=1}^{n} \alpha_j e_j \quad \Leftrightarrow \quad \alpha_j = v_j \quad \text{for each} \quad j \in \{1, \ldots, n\} \,,$$

which in terms of matrix-vector multiplication amounts to solving the linear system $v = I \cdot \alpha$ with respect to $\alpha \in \mathbb{F}^n$.

The following result, relating the notions of space dimension and matrix ranks, follows in large part from the analysis we developed in § II.5.10.

Lemma III.2.2.4. Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$. Then $\dim \operatorname{Im} A = \operatorname{rank} A$.

Proof. Let us first consider the case when the matrix A is zero. Then the claim follows trivially: $r = 0$ by definition II.4.4.1, the subspace $\operatorname{Im} A$ is trivial and $\operatorname{Ker} A = \mathbb{F}^n$ (see proposition II.5.10.6) and hence $\dim \operatorname{Im} A = 0$ (by definition III.2.2.1) and $\dim \operatorname{Ker} A = n$ (see example III.2.2.3).

For the remainder of the proof, we assume that the matrix A is nonzero, so that $\operatorname{rank} A \neq 0$ by definition II.4.4.1 and $\dim \operatorname{Im} A \neq 0$ by definition III.2.2.1 since the subspace $\operatorname{Im} A$ is nontrivial by proposition II.5.10.6.

First, let $r = \dim \operatorname{Im} A$. By definition III.2.2.1, there exist $u_1, \ldots, u_r \in \mathbb{F}^m$ spanning $\operatorname{Im} A$. Applying remark III.2.1.9 for $U = [u_1 \cdots u_r] \in \mathbb{F}^{m \times r}$, we obtain $\operatorname{Im} A \subseteq \operatorname{Im} U$. Then part (a) of corollary II.5.10.20 yields $\operatorname{rank} A \leq \dim \operatorname{Im} A$.

Second, let $r = \operatorname{rank} A$. By corollary II.4.4.12, there exist matrices $U \in \mathbb{F}^{m \times r}$ and $W \in \mathbb{F}^{r \times n}$ such that $A = UW$. Then part (a) of proposition II.5.10.18 yields $\operatorname{Im} A = \operatorname{Im} U$. Applying remark III.2.1.9, we conclude that $\operatorname{Im} A$ is spanned by the r columns of U. Then $\dim \operatorname{Im} A \leq r$ by definition III.2.2.1. So we have $\dim \operatorname{Im} A \leq \operatorname{rank} A$.

Combining the two inequalities proven above, we obtain $\dim \operatorname{Im} A = \operatorname{rank} A$.

Combining lemma III.2.2.4 with part (c) of lemma II.5.10.13 and using part (a) of theorem II.5.9.7 and proposition II.5.10.10, we establish that the rank of a matrix determines also the dimension of its kernel.

Corollary III.2.2.5. Let $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$. Then $\dim \operatorname{Ker} A = n - \operatorname{rank} A$.

Remark III.2.2.6 (image of a matrix and representation uniqueness)**.** Let $m, n, r \in \mathbb{N}$ and a matrix $A \in \mathbb{F}^{m\times n}$ be nonzero. Then $\operatorname{Im} A$ is nontrivial (see proposition II.5.10.6). Consider $r \in \mathbb{N}$ and $u_1, \ldots, u_r \in \mathbb{F}^m$ spanning $\operatorname{Im} A$, so that the matrix $U = [u_1 \cdots u_r] \in \mathbb{F}^{m\times r}$ satisfies $\operatorname{Im} A = \operatorname{Im} U$ by remark III.2.1.9 and $\dim \operatorname{Im} U \leq r$ by definition III.2.2.1. Applying lemma III.2.2.4, we conclude $\dim \operatorname{Im} A = \operatorname{rank} A$ and $\dim \operatorname{Im} U = \operatorname{rank} U$. Then $\operatorname{rank} A = \operatorname{rank} U \leq r$, and the matrix U is therefore square or tall.

Consider $y \in \operatorname{Im} A$. By definition II.5.10.1, there exists a coefficient $w \in \mathbb{F}^r$ satisfying $y = Uw$. Such a coefficient is unique if and only if $\operatorname{Ker} U$ is trivial (by proposition II.5.10.4), which is equivalent to that $\operatorname{rank} U = r$ (by part (b) of lemma II.5.10.16, parts (a) and (b) of lemma II.5.10.13 and part (a) of theorem II.5.9.7). So a coefficient w of y is unique if and only if $\operatorname{rank} A = r$, i.e., by definition III.2.2.1 and lemma III.2.2.4, if and only if $u_1, \ldots, u_r$ form a basis.

So the uniqueness of the representation of column vectors from $\operatorname{Im} A$ with respect to a spanning set of $\operatorname{Im} A$ is equivalent to that the spanning set is a basis.

§ III.2.3. Linear dependence and independence

Following examples III.2.2.2 and III.2.2.3 and remark III.2.2.6, we introduce the notions of linear dependence and independence as the uniqueness and, respectively, nonuniqueness of coefficients of linear combinations (see definition II.2.2.7).

Definition III.2.3.1 (linear dependence and independence)**.** Let V be a vector space over a field $\mathbb{F}$ and $n \in \mathbb{N}$. Vectors $v_1, \ldots, v_n \in V$ are called *linearly independent* if every $v \in \operatorname{span}\{v_j\}_{j=1}^n$ has a unique coefficient $\alpha \in \mathbb{F}^n$ with respect to $v_1, \ldots, v_n$ and *linearly dependent*, otherwise.

Lemma III.2.3.2 (characterization of linear independence)**.** Let V be a vector space over a field $\mathbb{F}$ and $v_1, \ldots, v_n \in V$ for $n \in \mathbb{N}$. Then the following statements are equivalent:

(i) the vectors $v_1, \ldots, v_n$ are linearly independent;

(ii) a linear combination of the vectors is trivial only if its coefficient is trivial:

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \quad \Rightarrow \quad (\alpha_1, \ldots, \alpha_n) = 0$$

for any $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$.

In the context of lemma III.2.3.2, negating the conclusion of the lemma, we obtain that vectors $v_1, \ldots, v_n \in V$ are linearly dependent if and only if there exists a trivial linear combination of the vectors with a nontrivial coefficient, i.e., if and only if there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ such that

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \quad \text{and} \quad (\alpha_1, \ldots, \alpha_n) \neq 0\,.$$

Proof of lemma III.2.3.2. Let us first assume statement (ii) and derive statement (i). Consider $v \in \operatorname{span}\{v_1, \dots, v_n\}$ with coefficients $\alpha, \beta \in \mathbb{F}^n$ with respect to $v_1, \dots, v_n$, so that

$$v = \sum_{j=1}^{n} \alpha_j v_j = \sum_{j=1}^{n} \beta_j v_j \,.$$

Such coefficients exist by definition III.2.1.1. To show that they cannot be distinct, we combine these two representations of v and obtain

$$\sum_{j=1}^{n} (\alpha_j - \beta_j)\, v_j = 0 \,,$$

which implies $\alpha - \beta = 0$ by the assumption of the "if" implication, so that $\alpha = \beta$. The vectors $v_1, \dots, v_n$ are therefore linearly independent by definition III.2.3.1.

Conversely, assuming statement (i) and considering a coefficient $\alpha \in \mathbb{F}^n$ producing a trivial linear combination of $v_1, \dots, v_n$, we obtain

$$\sum_{j=1}^{n} \alpha_j v_j = 0 = \sum_{j=1}^{n} 0 \cdot v_j \,,$$

which yields $\alpha = 0$ by definition III.2.3.1 since $v_1, \dots, v_n$ are linearly independent. This proves statement (ii).

Corollary III.2.3.3. Let V be a vector space over a field $\mathbb{F}$ and $v \in V$. The vector v is linearly dependent if and only if v is zero.

Proof. The proof follows from lemma III.2.3.2 and definition II.2.2.1 and is left to the reader as an exercise.

The following is also an immediate consequence of lemma III.2.3.2, expressing the monotonicity of the linear-independence property of a set of vectors with respect to the set and the fact that it characterizes indeed the set itself, with no regard to the order in which we list the vectors.

Proposition III.2.3.4. Let V be a vector space over a field $\mathbb{F}$ and $n \in \mathbb{N}$ vectors $v_1, \dots, v_n \in V$ be linearly independent. Consider distinct indices $\pi_1, \dots, \pi_m \in \{1, \dots, n\}$, where $m \in \{1, \dots, n\}$. Then the vectors $v_{\pi_1}, \dots, v_{\pi_m}$ are linearly independent.

Proof. The proof is left to the reader as an exercise.

Proposition III.2.3.4 and corollary III.2.3.3 trivially imply the following.

Corollary III.2.3.5. Let V be a vector space over a field $\mathbb{F}$ and $v_1, \dots, v_n \in V$ with $n \in \mathbb{N}$ be linearly independent. Then $v_1, \dots, v_n$ are all nonzero.

An important interpretation of the linear dependence of vectors consists in the possibility of expressing one of the vectors as a linear combination of the others and reducing thereby the spanning set without modifying the span.

Lemma III.2.3.6 (linear-dependence lemma)**.** Consider a vector space V over a field $\mathbb{F}$ and $n \in \mathbb{N}$ vectors $v_1, \ldots, v_n \in V$.

(a) When $n = 1$, the vector v_1 is linearly dependent if and only if it is zero.

(b) When $n > 1$, the vectors $v_1, \ldots, v_n$ are linearly dependent if and only if there exists $k \in \{1, \ldots, n\}$ such that $v_k \in \operatorname{span}\{v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$, in which case

$$\operatorname{span}\{v_j\}_{j=1}^n = \operatorname{span}\{v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}\,.$$

Proof. Part (a) follows trivially from lemma III.2.3.2.

The "if" implication of part (b) follows from definition III.2.1.1 and lemma III.2.3.2: $v_k \in \operatorname{span}\{v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$ implies by definition III.2.1.1 that there exist $\alpha_1, \ldots, \alpha_{k-1}, \alpha_{k+1}, \ldots, \alpha_n \in \mathbb{F}$ such that

$$v_k = \sum_{j=1}^{k-1} \alpha_j v_j + \sum_{j=k+1}^{n} \alpha_j v_j\,, \quad \text{i.e.,} \quad \sum_{j=1}^{n} \alpha_j v_j = 0 \quad \text{with} \quad \alpha_k = 1 \neq 0\,.$$

Applying lemma III.2.3.2, we conclude that the vectors $v_1, \ldots, v_n$ are linearly dependent. Furthermore, for any $\gamma_1, \ldots, \gamma_n \in \mathbb{F}$, we have

$$\begin{aligned}\sum_{j=1}^{n} \gamma_j v_j &= \sum_{j=1}^{k-1} \gamma_j v_j + \gamma_k v_k + \sum_{j=k+1}^{n} \gamma_j v_j \\ &= \sum_{j=1}^{k-1} \gamma_j v_j + \gamma_k \sum_{j=1}^{k-1} \alpha_j v_j + \gamma_k \sum_{j=k+1}^{n} \alpha_j v_j + \sum_{j=k+1}^{n} \gamma_j v_j \\ &= \sum_{j=1}^{k-1} (\gamma_j + \gamma_k \alpha_j) v_j + \sum_{j=k+1}^{n} (\gamma_j + \gamma_k \alpha_j) v_j\,,\end{aligned}$$

which leads to $\operatorname{span}\{v_j\}_{j=1}^n \subseteq \operatorname{span}\{v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$ by definition III.2.1.1. The reverse inclusion follows immediately from definition III.2.1.1, and together the two yield $\operatorname{span}\{v_j\}_{j=1}^n = \operatorname{span}\{v_1, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$.

To prove the "only if" implication of part (b), we reason conversely: the linear dependence of $v_1, \ldots, v_n$ implies by lemma III.2.3.2 that there exist $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ and $k \in \{1, \ldots, n\}$ such that $\alpha_k \neq 0$ and

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \quad \text{and hence} \quad v_k = \sum_{j=1}^{k-1} \beta_j v_j + \sum_{j=k+1}^{n} \beta_j v_j$$

with $\beta_j = -\frac{\alpha_j}{\alpha_k}$ for every $j \in \{1, \ldots, k-1, k+1, \ldots, n\}$, which completes the proof.

It is left to the reader as an exercise to note what specific parts of definitions II.1.1.1 and II.2.2.1 have been used in the above proof.

Example III.2.3.7. Revisiting example II.2.2.11, we conclude the following by definition III.2.3.1:

(i) e_1, e_2 are linearly independent since (II.2.2.3) has a unique solution for any $v \in \mathbb{R}^2$;

(ii) e_1, e_2, u are linearly dependent for any $u \in \mathbb{R}^2$ since (II.2.2.3) has multiple solutions for at least one (in fact, any) $v \in \mathbb{R}^2$;

(iii) e_1, u with $u \in \mathbb{R}^2$ are linearly independent if and only if $u_2 \neq 0$ since $u_2 \neq 0$ is necessary and sufficient for (II.2.2.3) to have a unique solution with $\alpha_2 = 0$ for any $v \in \mathbb{R}^2$;

(iv) e_2, u with $u \in \mathbb{R}^2$ are linearly independent if and only if $u_1 \neq 0$ since $u_1 \neq 0$ is necessary and sufficient for (II.2.2.3) to have a unique solution with $\alpha_1 = 0$ for any $v \in \mathbb{R}^2$.

Example III.2.3.8 (the columns of an identity matrix form a basis)**.** The last conclusion of example III.2.2.3 means that the columns of an identity matrix are linearly independent (by definition III.2.3.1) and that adding any other column from $\mathbb{F}^n$ results in a linearly dependent vectors (by part (b) of lemma III.2.3.6).

In fact, it follows from lemma III.2.3.2 that the columns of any invertible matrix of order n over $\mathbb{F}$ are linearly independent as vectors of $\mathbb{F}^n$.

We have already observed in several settings (examples III.2.2.2 and III.2.2.3 and remark III.2.2.6) that, for a spanning set, the basis property may be equivalent to linear independence. We will now prove that to be generally true, reducing the less trivial one of the two implications to a result we established for matrices in chapter II.

Lemma III.2.3.9 (linear independence and the basis property)**.** Let V be a vector space over a field $\mathbb{F}$ spanned by $n \in \mathbb{N}$ vectors $v_1, \ldots, v_n$. Then $v_1, \ldots, v_n$ form a basis for V if and only if $v_1, \ldots, v_n$ are linearly independent.

Proof. Let us first prove the "only if" implication claimed. Assume that the vectors $v_1, \ldots, v_n$ form a spanning set for V and are linearly dependent. When $n = 1$ part (a) of lemma III.2.3.6 yields that the vector space V is trivial, and it therefore has no basis by definition III.2.2.1. When $n > 1$, part (b) of lemma III.2.3.6 yields that V has a spanning set with $n-1$ elements, so $v_1, \ldots, v_n$ do not form a basis for V by definition III.2.2.1. In any case, if the vectors $v_1, \ldots, v_n$ are linearly dependent, they do not form a basis for V. This proves the "only if" implication of the lemma III.2.3.9.

Let us now prove the "if" implication claimed. Assume that $v_1, \ldots, v_n$ are linearly independent. Then $v_1, \ldots, v_n$ are all nonzero by corollary III.2.3.5, so the vector space V is nontrivial. Suppose that V is spanned by $u_1, \ldots, u_r$ with $r \in \{1, \ldots, n-1\}$. By definition III.2.1.1, there exists a matrix $A = [a_{kj}]_{k=1,\, j=1}^{r \;\;\; n} \in \mathbb{F}^{r \times n}$ such that

$$v_j = \sum_{k=1}^{r} a_{kj} u_k \quad \text{for each} \quad j \in \{1, \ldots, n\}\,. \tag{III.2.3.1}$$

Note that $\operatorname{rank} A \leq r$ by corollary II.4.4.8 and, since the matrix A is wide ($r < n$), we have $\operatorname{rank} A < n$. As we showed in the proof of lemma II.5.9.2, that implies the existence of a *nonzero* $\alpha \in \mathbb{F}^n$ such that $A\alpha = 0$ (in other terms, $\operatorname{Ker} A \neq \{0\}$, see part (b) of lemma II.5.10.16). For the corresponding linear combination of $v_1, \ldots, v_n$, using (III.2.3.1), we obtain

$$\sum_{j=1}^{n} \alpha_j v_j = \sum_{j=1}^{n} \alpha_j \sum_{k=1}^{r} a_{kj} u_k = \sum_{k=1}^{r} (A\alpha)_k u_k = 0\,.$$

Due to lemma III.2.3.2, this contradicts the linear independence of $v_1, \ldots, v_n$. So the assumption that V has a spanning set with fewer than n elements is incorrect. By definition III.2.2.1, since V is spanned by $v_1, \ldots, v_n$, we conclude that that $\dim V = n$ and that $v_1, \ldots, v_n$ form a basis for V.

To consider an example of a different kind, which falls out of the immediate scope of chapter II, let us now revisit the polynomial spaces introduced in example III.1.2.8. Our first goal is to show that, for any $n \in \mathbb{N}_0$, the $n+1$ monomials $q_0, \ldots, q_n$ from $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$, introduced in example III.2.1.10, are linearly independent.

To this end, we consider a *difference operator* defined on a polynomial space. Essentially, we need a "degree-lowering" mapping defined on $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ with $n \in \mathbb{N}$ that maps every element of $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ into an element of $\mathcal{P}_{n-1}(\mathbb{F}, \mathbb{F})$ in such a way that only the elements of the subspace $\mathcal{P}_0(\mathbb{F}, \mathbb{F})$ are mapped into zero. We need to have such a mapping defined pointwise because, at this stage, we are yet to establish the uniqueness of the coefficients of an element of $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ with respect to the monomials $q_0, \ldots, q_n$. For $\mathbb{F} = \mathbb{R}$, the differentiation operator would be a possible choice, but we will instead invoke a construction that is even simpler and is also suitable for any field $\mathbb{F}$.

Example III.2.3.10 (difference operator defined on a polynomial space)**.** Let us revisit example III.1.2.8 in the case of $\mathcal{D} = \mathbb{F}$ and, for convenience, denote $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ by $\mathcal{P}_n$ for each $n \in \mathbb{N}_0$.

Consider $n \in \mathbb{N}_0$ and the corresponding *difference operator* $\Delta \colon \mathcal{P}_n \to \mathcal{P}_n$ defined as follows:

$$(\Delta p)(t) = p(t+1) - p(t) \quad \text{for every} \quad t \in \mathbb{F}\,. \tag{III.2.3.2}$$

The mapping Δ satisfies

$$\begin{aligned}(\Delta(\alpha p + q))(t) &= (\alpha p + q)(t+1) - (\alpha p + q)(t) \\ &= \alpha p(t+1) - \alpha p(t) + q(t+1) - q(t) = \alpha(\Delta p)(t) + (\Delta q)(t)\end{aligned}$$

at every $t \in \mathbb{F}$ for all $p, q \in \mathcal{P}_n$ and $\alpha \in \mathbb{F}$, i.e.,

$$\Delta(\alpha p + q) = \alpha \Delta p + \Delta q \quad \text{for all} \quad p, q \in \mathcal{P}_n \text{ and } \alpha \in \mathbb{F}. \tag{III.2.3.3}$$

Let us consider an arbitrary vector $p \in \mathcal{P}_n$, which can be represented as follows due to (III.2.1.2):

$$p = \sum_{k=0}^{n} \alpha_k q_k \tag{III.2.3.4}$$

with $\alpha_0, \ldots, \alpha_n \in \mathbb{F}$. Using the binomial theorem in the form

$$(t+1)^\ell = \sum_{k=0}^{\ell} \binom{\ell}{k} t^k = t^\ell + \sum_{k=0}^{\ell-1} \binom{\ell}{k} t^k \quad \text{for all} \quad t \in \mathbb{F} \text{ and } \ell \in \mathbb{N},$$

where $\binom{\ell}{k} = \frac{\ell!}{k!(\ell-k)!}$ for all $k, \ell \in \mathbb{N}$ such that $0 \leq k \leq \ell$, and changing the order of summation, we obtain

$$(\mathit{\Delta} p)(t) = \sum_{\ell=1}^{n} \alpha_\ell \big((t+1)^\ell - t^\ell\big) = \sum_{\ell=1}^{n} \alpha_\ell \sum_{k=0}^{\ell-1} \binom{\ell}{k} t^k = \sum_{k=0}^{n-1} \sum_{\ell=k+1}^{n} \alpha_\ell \binom{\ell}{k} t^k = \sum_{k=0}^{n-1} \beta_k t^k \tag{III.2.3.5}$$

for each $t \in \mathbb{F}$ with

$$\beta_k = \sum_{\ell=k+1}^{n} \alpha_\ell \binom{\ell}{k} \in \mathbb{F} \quad \text{for all} \quad k \in \{0, \dots, n-1\}\,. \tag{III.2.3.6}$$

This shows $\mathit{\Delta} p \in \mathcal{P}_{n-1}$ and is equivalent to

$$\mathit{\Delta} p = \sum_{k=0}^{n-1} \beta_k q_k\,. \tag{III.2.3.7}$$

Lemma III.2.3.11 (monomials are linearly independent)**.** For any $n \in \mathbb{N}_0$, the monomials $q_0, \dots, q_n \in \mathcal{P}_n(\mathbb{F}, \mathbb{F})$ introduced as in example III.2.1.10 for $\mathcal{D} = \mathbb{F}$ are linearly independent.

Proof. For convenience, we denote $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ by $\mathcal{P}_n$ for each $n \in \mathbb{N}_0$.

Let us consider $n \in \mathbb{N}$ and the corresponding difference operator $\mathit{\Delta} \colon \mathcal{P}_n \to \mathcal{P}_n$ defined by (III.2.3.2) and analyzed in example III.2.3.10.

As we showed in example III.2.3.10, we have $\mathit{\Delta} p = 0$ for any $p \in \mathcal{P}_0$ and also $\mathit{\Delta} p \in \mathcal{P}_{k-2}$ for any $p \in \mathcal{P}_{k-1}$ with $k \in \{2, \dots, n\}$. Iterating this statement, we find that the nth power $\mathit{\Delta}^n = \mathit{\Delta} \circ \cdots \circ \mathit{\Delta}$ of $\mathit{\Delta}$ (see § I.1.11) satisfies $\mathit{\Delta}^n p = 0$ for any $p \in \mathcal{P}_{n-1}$.

Further, for each $k \in \{1, \dots, n\}$, we find that (III.2.3.4) with $\alpha_{k+1} = \cdots = \alpha_n = 0$ implies (III.2.3.6) with $\beta_{k-1} = k\alpha_k$. Iterating this relation, for any p given by (III.2.3.4) with $\alpha_0, \dots, \alpha_n \in \mathbb{F}$, since it satisfies $p = \alpha_n q_n + q$ with and $q = \alpha_0 q_0 + \cdots + \alpha_{n-1} q_{n-1} \in \mathcal{P}_{n-1}$, we find that $\mathit{\Delta}^n p = n!\alpha_n q_0$ and therefore

$$\alpha_n = \frac{1}{n!} (\mathit{\Delta}^n p)(0)\,, \tag{III.2.3.8}$$

so that the leading coefficient α_n of p_n is uniquely defined. The remaining n coefficients $\alpha_0, \dots, \alpha_{n-1}$ of p_n are coefficients of $p_{n-1} \in \mathcal{P}_{n-1}$ with respect to $q_0, \dots, q_{n-1}$.

Applying the result iteratively (in other words, by induction), we find that

$$\alpha_k = \frac{1}{k!} \mathit{\Delta}^k \Big(p - \sum_{j=k+1}^{n} \alpha_j q_j\Big)(0)$$

for every $k \in \{0, \dots, n\}$, so that the coefficients $\alpha_0, \dots, \alpha_n$ satisfying (III.2.3.4) are all uniquely defined.

By definition III.2.3.1, the monomials $q_0, \dots, q_n$ are linearly independent in the vector space $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ for every $n \in \mathbb{N}_0$.

The following corollary is obtained immediately by combining lemmata III.2.3.9 and III.2.3.11.

Corollary III.2.3.12 (monomials form bases)**.** For any $n \in \mathbb{N}_0$, the polynomial space $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$, defined in example III.1.2.8, has dimension $n+1$, and the monomials $q_0, \dots, q_n \in \mathcal{P}_n(\mathbb{F}, \mathbb{F})$, defined in example III.2.1.10, form a basis for $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$.

§ III.2.4. Synthesis and analysis operators

For convenience, we will now introduce a synthesis operator associated with given vectors, which does nothing else than assembling linear combinations of those vectors with given coefficients.

Definition III.2.4.1. (synthesis operator) Let V be a vector space over the field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. We refer to the mapping $\mathbf{\Phi}\colon \mathbb{F}^n \to \operatorname{span}\{v_j\}_{j=1}^n$ given by

$$\mathbf{\Phi}_{v_1,\dots,v_n}(\alpha) = \sum_{j=1}^{n} \alpha_j \cdot v_j \quad \text{for each} \quad \alpha \in \mathbb{F}^n$$

as the *synthesis operator associated with the vectors* $v_1, \dots, v_n$.

In the context of definition III.2.4.1, the synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n}$ apparently depends on the vectors $v_1, \dots, v_n$ (the ordering matters) and on the operations of the vector space V.

Recalling the general notion of function surjectivity (see § I.1.18), we trivially obtain the following statement from definitions III.2.4.1 and III.2.1.1.

Proposition III.2.4.2. Let V be a vector space over the field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. Then the synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n} : \mathbb{F}^n \to \operatorname{span}\{v_j\}_{j=1}^n$ is surjective

Using the notion of injectivity presented in § I.1.17, we can equivalently characterize the linear dependence and independence of any vectors from a vector space in terms of the injectivity of the corresponding synthesis operator, introduced in definition III.2.4.1.

Proposition III.2.4.3 (linear independence and synthesis-operator injectivity)**.** Let V be a vector space over a field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. Then the synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n}$ is injective if and only if the vectors $v_1, \dots, v_n$ are linearly independent.

Recalling the notions of invertibility and bijectivity of functions (see §§ I.1.15 and I.1.19) and the equivalence thereof (see § I.1.20), we obtain the following statement form propositions III.2.4.2 and III.2.4.3.

Corollary III.2.4.4. Let V be a vector space over a field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. The synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n}$ is bijective (equivalently, invertible) if and only if the vectors $v_1, \dots, v_n$ are linearly independent.

Definition III.2.4.5 (analysis operator)**.** Let V be a vector space over a field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \dots, v_n \in V$. We refer to the inverse (see § I.1.15) of the synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n}$, whenever it exists, as the *analysis operator associated with the vectors* $v_1, \dots, v_n$ and denote it by $\mathbf{\Psi}_{v_1,\dots,v_n}$.

The role of an analysis operator consists in extracting the unique coefficient (the coordinates) of a vector with respect to the given spanning set, the latter being a basis by corollary III.2.4.4 whenever the analysis operator exists. In the context of definition III.2.4.5, due to definition III.2.4.1, the domain of $\mathbf{\Psi}_{v_1,\dots,v_n}$ is $\operatorname{span}\{v_j\}_{j=1}^n$ and the co-domain is $\mathbb{F}^n$. Any analysis operator is obviously invertible and therefore bijective (by § I.1.20) and consequently also injective and surjective (by § I.1.19).

The notion of a *linear combination* is a cornerstone of linear algebra, and introducing an operator (a function) assembling linear combinations is a convenient but not universal practice. The very terms *synthesis operator* and *analysis operator* are not so commonly used in the literature and are introduced here for notational and terminological convenience. In fact, synthesis operators (under various name) are introduced in the literature even less often than their inverses, and the latter, which we call analysis operators here, are often referred to as *coordinate mappings.*

Let us close this § III.2.4 with noting the following property of the synthesis operators, which follows from definitions III.2.4.1 and II.2.2.1.

Proposition III.2.4.6. Let V be a vector space over a field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \ldots, v_n \in V$. The synthesis operator $\mathbf{\Phi}_{v_1,\ldots,v_n}$ satisfies $\mathbf{\Phi}_{v_1,\ldots,v_n}(\eta + \eta') = \mathbf{\Phi}_{v_1,\ldots,v_n}(\eta) + \mathbf{\Phi}_{v_1,\ldots,v_n}(\eta')$ and $\mathbf{\Phi}_{v_1,\ldots,v_n}(\alpha\eta) = \alpha\,\mathbf{\Phi}_{v_1,\ldots,v_n}(\eta)$ for all $\eta, \eta' \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$.

Remark III.2.4.7. The conclusion of proposition III.2.4.6 can be equivalently expressed in the following form: $\mathbf{\Phi}_{v_1,\ldots,v_n}(\alpha\eta + \eta') = \alpha\,\mathbf{\Phi}_{v_1,\ldots,v_n}(\eta) + \mathbf{\Phi}_{v_1,\ldots,v_n}(\eta')$ for all $\eta, \eta' \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$. Verifying the two implications that constitute the equivalence is left to the reader as an exercise.

§ III.3. Linear mappings between vector spaces

§ III.3.1. Basic definitions and examples

The property of synthesis maps asserted in proposition III.2.4.6 and remark III.2.4.7, very particular but also very important, is referred to as *linearity.* In this § III.3, we start studying mappings between vector spaces that exhibit that property and are therefore called *linear.*

Definition III.3.1.1. Let V and U be vector spaces over $\mathbb{F}$, $\boxplus$ and $\boxdot$ denote the operations of V of addition and multiplication by scalars and $\oplus$ and $\odot$ denote the respective operations of U. A mapping $\varphi \colon V \to U$ is called *linear* if

$$\varphi(\alpha \boxdot v \boxplus v') = \alpha \odot \varphi(v) \oplus \varphi(v') \quad \text{for all } v, v' \in V \text{ and } \alpha \in \mathbb{F}. \tag{III.3.1.1}$$

The set of all linear mappings from V to U is denoted by $\mathcal{L}(V, U)$.

Example III.3.1.2 (the zero and identity mappings are linear)**.** Let V and U be vector spaces over $\mathbb{F}$ and $\vec{0}$ denote the additive identity element of U. Applying definition III.3.1.1, we conclude the following.

- **(i)** The zero mapping $\varphi \colon V \to U$, defined pointwise by $\varphi(v) = \vec{0}$ for all $v \in V$, is linear: $\varphi \in \mathcal{L}(V, U)$.
- **(ii)** The identity mapping $\psi \colon V \to V$, given by $\psi(v) = v$ for all $v \in V$, is linear: $\psi \in \mathcal{L}(V, V)$.

Example III.3.1.3 (the synthesis operators are linear)**.** Let V be a vector space over a field $\mathbb{F}$. Consider $n \in \mathbb{N}$ and $v_1, \ldots, v_n \in V$. It follows trivially from proposition III.2.4.6, remark III.2.4.7 and definition III.3.1.1 that the synthesis operator $\mathbf{\Phi}_{v_1,\ldots,v_n} : \mathbb{F}^n \to \operatorname{span}\{v_j\}_{j=1}^n$ is a linear mapping.

Example III.3.1.4 (the mapping induced by a matrix is linear)**.** For any $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$, the mapping $\phi_A \colon \mathbb{F}^n \to \mathbb{F}^m$ induced by the matrix A (see definition II.5.10.3) is linear: $\phi_A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$. Condition (III.3.1.1) of definition III.3.1.1, expressing in this case the linearity of the matrix-vector multiplication, follows from parts (a) and (c) of proposition II.4.1.10, proposition II.4.1.19 and remark II.4.1.4.

Example III.3.1.5 (the difference operator on $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ is linear)**.** Let $n \in \mathbb{N}$ and V denote the vector space $\mathcal{P}_n(\mathbb{F}, \mathbb{F})$ of polynomials introduced in example III.1.2.8. As (III.2.3.3) shows, the difference operator $\varDelta \colon V \to V$ defined by (III.2.3.2) in example III.2.3.10 is linear by definition III.3.1.1: $\varDelta \in \mathcal{L}(V, V)$.

Example III.3.1.6. Let $n \in \mathbb{N}$ and V denote the vector space $\mathcal{P}_n(\mathbb{R}, \mathbb{R})$ of polynomials introduced in example III.1.2.8. The differential operator $\varphi \colon V \to V$ given by

$$\big(\varphi(p)\big)(t) = p''(t) + 2p'(t) - 7p(t) \quad \text{for each} \quad t \in \mathbb{R}$$

at every $p \in V$, where the prime sign denotes the differentiation operation, is linear by definition III.3.1.1. The verification of condition (III.3.1.1) is based on the linearity of differentiation.

Let us now generalize definition II.5.10.1 from the case of a matrix to that of a linear mapping between vector spaces.

Definition III.3.1.7 (the fundamental subspaces of a linear mapping)**.** Let V and U be vector spaces over a field $\mathbb{F}$ and $\varphi \in \mathcal{L}(V, U)$. By $\operatorname{Ker} \varphi$ and $\operatorname{Im} \varphi$ we denote the following sets:

$$\operatorname{Ker} \varphi = \{v \in V \colon \varphi(v) = 0\} \subseteq V \quad \text{and} \quad \operatorname{Im} \varphi = \{\varphi(v) \colon v \in V\} \subseteq U\,.$$

The set $\operatorname{Ker} \varphi$ is usually referred to as the *kernel* or as the *null space* of φ. The set $\operatorname{Im} \varphi$ is usually referred to as the *image* or as the *range* of φ.

The notion of image is standard for functions (see § I.1.7), so definition III.3.1.7 merely introduces a specific notation for the image of a linear function between vector spaces. The notion of the kernel of a linear mapping coincides with the general notion of the null set of a function once a zero element has been defined in the co-domain.

Remark III.3.1.8 (consistency with the matrix setting)**.** For $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m \times n}$, consider the mapping $\phi_A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ induced by the matrix A (see definition II.5.10.3 and example III.3.1.4), the fundamental subspaces of A in the sense of definition II.5.10.1 and the fundamental subspaces of ϕ_A in the sense of definition III.3.1.7. Then $\operatorname{Im} \phi_A = \operatorname{Im} A$ and $\operatorname{Ker} \phi_A = \operatorname{Ker} A$.

Proposition III.3.1.9 (linearity is preserved under inversion)**.** Let V and U be vector spaces over a field $\mathbb{F}$ and a mapping $\varphi \in \mathcal{L}(V, U)$ be invertible. Then $\varphi^{-1} \in \mathcal{L}(U, V)$.

Proof. By the definition of the inverse function (see § I.1.15), the inverse $\varphi^{-1}\colon U \to V$ is such that $\varphi^{-1} \circ \varphi$ and $\varphi \circ \varphi^{-1}$ are the identity transformations of V and U.

Consider $u, u' \in U$ and $\alpha \in \mathbb{F}$. Let $v = \varphi^{-1}(u)$ and $v' = \varphi^{-1}(u')$. The linearity of φ then yields $\varphi(\alpha v + v') = \alpha\varphi(v) + \varphi(v')$, i.e., $\varphi(\alpha v + v') = \alpha u + u'$. Then $\varphi^{-1}(\alpha u + u') = (\varphi^{-1} \circ \varphi)(\alpha v + v') = \alpha v + v' = \alpha\varphi^{-1}(u) + \varphi^{-1}(u')$. Since we have showed this for arbitrary $u, u' \in U$ and $\alpha \in \mathbb{F}$, we conclude that $\phi^{-1} \in \mathcal{L}(U, V)$ by definition III.3.1.1.

Example III.3.1.3 and proposition III.3.1.9 yield the following.

Corollary III.3.1.10 (the analysis operators are linear)**.** Any analysis operator (in the sense of definition III.2.4.5) is linear.

Let us now introduce pointwise linear operations on a set of linear mappings between two vector spaces.

Definition III.3.1.11 (linear operations on linear mappings)**.** Let V and U be vector spaces over a field $\mathbb{F}$. For any $\varphi, \psi \in \mathcal{L}(V, U)$ and $\alpha \in \mathbb{F}$, the pointwise sum $(\varphi + \psi)\colon V \to U$ of φ and ψ and the pointwise multiple $(\alpha \cdot \varphi)\colon V \to U$ of φ with the coefficient α are given by

$$(\varphi + \psi)(v) = \varphi(v) \oplus \psi(v) \quad \text{and} \quad (\alpha \cdot \varphi)(v) = \alpha \odot \varphi(v) \quad \text{for each} \quad v \in V\,.$$

First of all, we note that the addition of linear mappings and the multiplication of linear mappings by scalars produces linear mappings. In other words, the set $\mathcal{L}(V, U)$ is closed under the pointwise operations given by definition III.3.1.11.

Proposition III.3.1.12. Let V and U be vector spaces. Then $\varphi + \psi \in \mathcal{L}(V, U)$ and $\alpha \cdot \varphi \in \mathcal{L}(V, U)$ for all $\varphi, \psi \in \mathcal{L}(V, U)$ and $\alpha \in \mathbb{F}$.

Proof. The proof consists in combining definitions III.3.1.1 and III.3.1.11 and is left to the reader as an exercise.

By verifying all the vector-space axioms from definition II.2.2.1, one obtains the following.

Proposition III.3.1.13. Let V and U be vector spaces over a field $\mathbb{F}$. Then $\mathcal{L}(V, U)$ is a vector space over $\mathbb{F}$ with respect to the pointwise operations $+\colon \mathcal{L}(V, U) \times \mathcal{L}(V, U) \to \mathcal{L}(V, U)$ and $\cdot\colon \mathbb{F} \times \mathcal{L}(V, U) \to \mathcal{L}(V, U)$ of addition and multiplication by scalars given by definition III.3.1.11 and with the zero function (see example III.3.1.2) as an additive identity element.

Further, we note that the linearity of linear mappings is preserved under composition.

Proposition III.3.1.14. Let U, V and W be vector spaces over a field $\mathbb{F}$. Then $\varphi \circ \psi \in \mathcal{L}(W, U)$ for any $\varphi \in \mathcal{L}(V, U)$ and $\psi \in \mathcal{L}(W, V)$.

Proof. Consider $\varphi \in \mathcal{L}(V, U)$ and $\psi \in \mathcal{L}(W, V)$. First, we note that $\varphi \circ \psi \colon W \to U$ is defined (see § I.1.11). To prove the linearity of the composition, we consider $w, w' \in W$ and $\alpha \in \mathbb{F}$. Then

$$\begin{aligned}(\varphi \circ \psi)(\alpha w + w') &= \varphi\big(\psi(\alpha w + w')\big) = \varphi(\alpha\psi(w) + \psi(w')) \\ &= \alpha\varphi(\psi(w)) + \varphi(\psi(w')) = \alpha(\varphi \circ \psi)(w) + (\varphi \circ \psi)(w').\end{aligned}$$

We therefore have $\varphi \circ \psi \in \mathcal{L}(W, U)$ by definition III.3.1.1.

Generalizing proposition II.5.10.9, we can show that the kernel and the image of a linear mapping are indeed subspaces of its domain and co-domain.

Proposition III.3.1.15 (the kernel and image of a linear mapping are indeed subspaces)**.** Let V and U be vector spaces and $\varphi \in \mathcal{L}(V, U)$. Then $\operatorname{Ker}\varphi$ is a subspace of V and $\operatorname{Im}\varphi$ is a subspace of U.

Proof. Note first that $\operatorname{Ker}\varphi$ and $\operatorname{Im}\varphi$ are subsets of V and U and do contain the additive identity elements 0_V and 0_U of the respective spaces by **??**. Indeed, the mapping φ is linear by assumption. That implies $\varphi(0_V) = 0_U$ due to definition III.3.1.1, so that $0_V \in \operatorname{Ker}\varphi$ and $0_U \in \operatorname{Im}\varphi$. It therefore remains to prove the closedness conditions (III.1.1.1) of definition III.1.1.1 for each of the two subsets.

Consider $v, v' \in \operatorname{Ker}\varphi$ and $\alpha \in \mathbb{F}$. Then we have $\varphi(v) = 0_U = \varphi(v')$ by **??**. Then the linearity of φ (definition III.3.1.1) yields $\varphi(v + v') = \varphi(v) + \varphi(v') = 0_U + 0_U = 0_U$ (the last equality is due to condition (c) of definition II.2.2.1) and $\varphi(\alpha v) = \alpha\varphi(x) = \alpha \cdot 0_U = 0_U$ (see remark II.2.2.2 regarding the last equality). Applying **??**, we conclude that $v + v' \in \operatorname{Ker}\varphi$ and $\alpha v \in \operatorname{Ker}\varphi$. By definition III.1.1.1, the set $\operatorname{Ker}\varphi$ is therefore a subspace of V.

Consider $u, u' \in \operatorname{Im}\varphi$ and $\alpha \in \mathbb{F}$. By **??**, there exist $v, v' \in \mathbb{F}^n$ such that $u = \varphi(v)$ and $u' = \varphi(v')$. Then the linearity of φ (definition III.3.1.1) yields $\varphi(v + v') = \varphi(v) + \varphi(v') = u + u'$ and $\varphi(\alpha v) = \alpha\varphi(v) = \alpha u$. As a result, $u + u' \in \operatorname{Im}\varphi$ and $\alpha u \in \operatorname{Im}\varphi$ by **??**. Applying definition III.1.1.1, we conclude that $\operatorname{Im}\varphi$ is a subspace of U.

§ III.3.2. Matrix representation of a linear mapping

In definition II.5.10.3, we introduced the mapping $\phi_A \colon \mathbb{F}^n \to \mathbb{F}^m$ induced by a given matrix $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F}$ is a field and $m, n \in \mathbb{N}$. In example III.3.1.4, we pointed out that the mapping ϕ_A is linear. The relation between A and ϕ_A is very straightforward: the mapping multiplies by the matrix, and the matrix represents the mapping as a matrix by which it multiplies. This relation is clearly one to one.

We will now see that any linear mapping between any two finite-dimensional vector spaces induces matrices that represent it, every such a matrix being unique for any choice of bases for the domain and co-domain.

Definition III.3.2.1. Let V and U be finite-dimensional vector spaces and $\varphi \in \mathcal{L}(V, U)$. Consider bases $v_1, \ldots, v_n$ for V and $u_1, \ldots, u_m$ for U, where $n = \dim V$ and $m = \dim U$. Then the matrix $A = [a_{ij}]_{i=1,\,j=1}^{m\quad n} \in \mathbb{F}^{m\times n}$ given by

$$\varphi(v_j) = \sum_{i=1}^{m} a_{ij} u_i \quad \text{for each} \quad j \in \{1, \ldots, n\} \tag{III.3.2.1}$$

is called *the matrix of the linear mapping φ with respect to the bases $v_1, \dots, v_n$ and $u_1, \dots, u_m$.*

Remark III.3.2.2. Note that, in the context of definition III.3.2.1, the matrix A exists and is unique: for every $j \in \{1, \dots, n\}$, column j of the matrix is the unique coefficient of the image of the jth vector of the basis chosen for the domain with respect to the basis chosen for the co-domain.

Example III.3.2.3. For $m, n \in \mathbb{N}$, consider $A \in \mathbb{F}^{m \times n}$ and the mapping ϕ_A induced by A in the sense of definition II.5.10.3. The columns $e_1, \dots, e_m$ of the identity matrix of order m form a basis for $U = \mathbb{F}^m$ (see example III.2.3.8). Likewise, the columns $f_1, \dots, f_n$ of the identity matrix of order n form a basis for $V = \mathbb{F}^n$. Then A is the matrix of ϕ_A with respect to the bases $f_1, \dots, f_n$ and $e_1, \dots, e_m$.

Let us note that, in the context of definition III.3.2.1, once a pair of bases for V and U has been fixed, not only the matrix is uniquely defined by the linear mapping but also vice versa.

Proposition III.3.2.4. Let V and U be finite-dimensional vector spaces and $\varphi \in \mathcal{L}(V, U)$. Consider bases $v_1, \dots, v_n$ for V and $u_1, \dots, u_m$ for U, where $n = \dim V$ and $m = \dim U$. Then any $\varphi \in \mathcal{L}(V, U)$ and $A \in \mathbb{F}^{m \times n}$ satisfy the condition (III.3.2.1) if and only if

$$\varphi = \mathbf{\Phi}_{u_1,\dots,u_m} \circ \phi_A \circ \mathbf{\Psi}_{v_1,\dots,v_n} . \tag{III.3.2.2}$$

Here, $\mathbf{\Phi}_{u_1,\dots,u_m} \in \mathcal{L}(\mathbb{F}^m, U)$ and $\mathbf{\Psi}_{v_1,\dots,v_n} \in \mathcal{L}(V, \mathbb{F}^n)$ are synthesis and analysis operators, given by definitions III.2.4.1 and III.2.4.5.

In the context of proposition III.3.2.4, the synthesis operator $\mathbf{\Phi}_{v_1,\dots,v_n}$ is bijective (by corollary III.2.4.4) since $v_1, \dots, v_n$ form a basis for V. The analysis operator $\mathbf{\Psi}_{v_1,\dots,v_n} = \mathbf{\Phi}_{v_1,\dots,v_n}^{-1}$ therefore exists.

Remark III.3.2.5. Composing both sides of (III.3.2.2) with $\mathbf{\Psi}_{u_1,\dots,u_m}$ on the left and with $\mathbf{\Phi}_{v_1,\dots,v_n}$ on the right, we obtain the following equivalent condition relating φ and A:

$$\mathbf{\Psi}_{u_1,\dots,u_m} \circ \varphi \circ \mathbf{\Phi}_{v_1,\dots,v_n} = \phi_A . \tag{III.3.2.3}$$

This means that the coefficient $\eta = \mathbf{\Psi}_{v_1,\dots,v_n}(v)$ of any $v \in V$ with respect to $v_1, \dots, v_n$ and the coefficient $\xi = \mathbf{\Psi}_{u_1,\dots,u_m}(u)$ of $u = \varphi(v)$ with respect to $u_1, \dots, u_m$ are related via the multiplication by A:

$$\xi = A\eta . \tag{III.3.2.4}$$

Proof of proposition III.3.2.4. For brevity, let us denote the analysis and synthesis operators $\mathbf{\Psi}_{v_1,\dots,v_n}$ and $\mathbf{\Phi}_{u_1,\dots,u_m}$ by $\mathbf{\Psi}$ and $\mathbf{\Phi}$.

Consider $\varphi \in \mathcal{L}(V, U)$ and $A \in \mathbb{F}^{m \times n}$ satisfying (III.3.2.1). The linearity of φ and (III.3.2.1) yield

$$\varphi(v) = \sum_{j=1}^{n} (\mathbf{\Psi} v)_j \varphi(v_j) = \sum_{j=1}^{n} (\mathbf{\Psi} v)_j \sum_{i=1}^{m} a_{ij} u_i = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} (\mathbf{\Psi} v)_j u_i$$

$$= \sum_{i=1}^{m} \big(A \cdot (\mathbf{\Psi} v)\big)_i u_i = (\mathbf{\Phi} \circ \phi_A \circ \mathbf{\Psi})(v) \qquad \text{(III.3.2.5)}$$

for every $v \in V$, so that φ satisfies (III.3.2.2).

Let us now consider $\varphi \in \mathcal{L}(V, U)$ and $A \in \mathbb{F}^{m\times n}$ satisfying (III.3.2.2). First, we have $\mathbf{\Psi} \in \mathcal{L}(V, \mathbb{F}^n)$ by proposition III.2.4.6, $\phi_A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ due to example III.3.1.4 and $\mathbf{\Phi} \in \mathcal{L}(\mathbb{F}^m, U)$ by propositions III.2.4.6 and III.3.1.9 and definition III.2.4.5. Then proposition III.3.1.14 yields $\varphi \in \mathcal{L}(V, U)$. Denoting the columns of the identity matrix of order n by $e_1, \dots, e_n$, we obtain that $\mathbf{\Psi} v_j = e_j$ and hence

$$\varphi(v_j) = (\mathbf{\Phi} \circ \phi_A)(e_j) = \mathbf{\Phi}(Ae_j) = \sum_{i=1}^{m} a_{ij} u_i$$

for every $j \in \{1, \dots, n\}$, so that (III.3.2.1) holds.

Example III.3.2.6. Let us revisit example III.3.1.6 in the case of $n = 2$. In terms of the monomials (III.2.1.1), we obtain

$$\begin{aligned} \varphi(q_0)(t) &= -7\,, \\ \varphi(q_1)(t) &= 2 - 7t\,, \\ \varphi(q_2)(t) &= 2 + 4t - 7t^2 \end{aligned}$$

for every $t \in \mathbb{R}$. As a result, we have

$$\begin{aligned} \varphi(q_0) &= -7q_0\,, \\ \varphi(q_1) &= 2q_0 - 7q_1\,, \\ \varphi(q_2) &= 2q_0 + 4q_1 - 7q_2\,. \end{aligned}$$

Then definition III.3.2.1 yields that

$$A = \begin{bmatrix} -7 & 2 & 2 \\ & -7 & 4 \\ & & -7 \end{bmatrix} \qquad \text{(III.3.2.6)}$$

is the matrix of φ with respect to the bases q_0, q_1, q_2 and q_0, q_1, q_2. We note immediately that $\operatorname{Ker} A = \operatorname{Ker} \phi_A = \{0\}$ and $\operatorname{Im} A = \operatorname{Im} \phi_A = \mathbb{R}^3$.

Notes missing, see from 0:09:00 of Lecture 2 to 1:14:18 of Lecture 4

§ III.4. Matrix determinant as an indicator of linear dependence

§ III.4.1. Motivating example: $\mathbb{R}^2$

Notes missing, see from 1:15:00 of Lecture 4 to 0:39:10 of Lecture 5

§ III.4.2. Determinant functions

Our plan is to introduce determinant functions axiomatically, as a natural generalization of the notion of area that we considered for $\mathbb{R}^2$. Specifically, we will list three perfectly reasonable properties and use them as axioms to defining *determinant functions*. In the following sections, we will do the following:

- use the axioms to derive certain properties of *any* determinant function, regardless of whether such a function exists at all;
- consider one specific function for each $n \in \mathbb{N}$ and, by induction and using the derived properties, show that it is a determinant function defined on the set of square matrices of order n (which means the existence of such a function) and that any determinant function defined on the set of square matrices of order n coincides with it (which means the uniqueness of such a function).

Definition III.4.2.1 (determinant function)**.** Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Then $\phi\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ is called a *determinant function* if the following conditions are satisfied:

(a) ϕ is multilinear with respect to the columns of its matrix argument, i.e., the restriction of ϕ to a function of any single column of the argument matrix, the others being arbitrary and fixed, is linear:

$$\phi([v_1,\ldots,v_{j-1},\underline{\alpha v_j + v},v_{j+1},\ldots,v_n]) \\ = \alpha\phi([v_1,\ldots,v_{j-1},\underline{v_j},v_{j+1},\ldots,v_n]) + \phi([v_1,\ldots,v_{j-1},\underline{v},v_{j+1},\ldots,v_n])$$

for all $j \in \{1,\ldots,n\}$, $v_1,\ldots,v_n, v \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$;

(b) ϕ vanishes at all matrices with two identical columns:

$$\phi([v_1,\ldots,v_n]) = 0$$

for all $v_1,\ldots,v_n \in \mathbb{F}^n$ and $j,k \in \{1,\ldots,n\}$ such that $j < k$ and $v_j = v_k$,

(c) if I is the identity matrix of order n, then $\phi(I) = 1$.

§ III.4.3. Properties of determinant functions

Lemma III.4.3.1 (determinant function under elementary column transformations)**.** Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Assume that $\phi\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ is a determinant function. Consider $v_1,\ldots,v_n \in \mathbb{F}^n$ and $v = \sum_{j=1}^n \alpha_j v_j$ with $\alpha_1,\ldots,\alpha_n \in \mathbb{F}$. Assume that $k \in \{1,\ldots,n\}$. Then

$$\phi([v_1,\ldots,v_{k-1},v,v_{k+1},\ldots,v_n]) = \alpha_k\phi([v_1,\ldots,v_{k-1},v_k,v_{k+1},\ldots,v_n])\,.$$

Proof. Using the multilinearity property (condition (a) of definition III.4.2.1), we obtain

$$\phi([v_1,\ldots,v_{k-1},v,v_{k+1},\ldots,v_n]) = \sum_{j=1}^n \alpha_j\phi([v_1,\ldots,v_{k-1},v_j,v_{k+1},\ldots,v_n])\,.$$

Each term on the right-hand side with $j \neq k$ is zero by condition (b) of definition III.4.2.1 since column k of $[v_1,\ldots,v_{k-1},v_j,v_{k+1},\ldots,v_n]$ is identical to one of the other columns. This completes the proof.

Corollary III.4.3.2 (determinant of a rank-deficient matrix)**.** Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Assume that $\phi \colon \mathbb{F}^{n\times n} \to \mathbb{F}$ is a determinant function and $V \in \mathbb{F}^{n\times n}$ is such that $\operatorname{rank} V < n$. Then $\phi(V) = 0$.

Proof. Let $v_1, \dots, v_n \in \mathbb{F}^n$ be the columns of V. Since $\operatorname{rank} V < n$, the columns of V are linearly dependent. By lemma III.2.3.6, there exists $k \in \{1, \dots, n\}$ such that $v_k \in \operatorname{span}\{v_1, \dots, v_{k-1}, v_{k+1}, \dots, v_n\}$. By definition III.2.1.1, this means that there exist $\alpha_1, \dots, \alpha_{k-1}, \alpha_{k+1}, \dots, \alpha_n \in \mathbb{F}$ such that

$$v_k = \sum_{j=1}^{k-1} \alpha_j v_j + 0 \cdot v_k + \sum_{j=k+1}^{n} \alpha_j \cdot v_j \,.$$

Applying lemma III.4.3.1, we obtain

$$\phi([v_1, \dots, v_{k-1}, v_k, v_{k+1}, \dots, v_n]) = 0 \cdot \phi([v_1, \dots, v_{k-1}, v_k, v_{k+1}, \dots, v_n]) = 0 \,,$$

which completes the proof.

Lemma III.4.3.3 (alternation of a determinant function)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $\phi \colon \mathbb{F}^{n\times n} \to \mathbb{F}$ be a determinant function. Consider $j, k \in \{1, \dots, n\}$ such that $k > j$ and let Σ denote the (j,k)-exchange matrix of order n. Then $\phi(\Sigma) = -1$ and $\phi(V\Sigma) = \phi(\Sigma)\,\phi(V)$ for every $V \in \mathbb{F}^{n\times n}$.

Proof. Consider an arbitrary matrix $V \in \mathbb{F}^{n\times n}$ and let $v_1, \dots, v_n$ denote its columns. Then the multilinearity property (condition (a) of definition III.4.2.1) gives

$$\begin{aligned}
&\phi([v_1, \dots, v_{j-1}, \underline{v_k + v_j}, v_{j+1}, \dots, v_{k-1}, \underline{v_k + v_j}, v_{k+1}, \dots, v_n]) \\
&\qquad = \phi([v_1, \dots, v_{j-1}, \underline{v_k}, v_{j+1}, \dots, v_{k-1}, \underline{v_k}, v_{k+1}, \dots, v_n]) \\
&\qquad + \phi([v_1, \dots, v_{j-1}, \underline{v_k}, v_{j+1}, \dots, v_{k-1}, \underline{v_j}, v_{k+1}, \dots, v_n]) \\
&\qquad + \phi([v_1, \dots, v_{j-1}, \underline{v_j}, v_{j+1}, \dots, v_{k-1}, \underline{v_k}, v_{k+1}, \dots, v_n]) \\
&\qquad + \phi([v_1, \dots, v_{j-1}, \underline{v_j}, v_{j+1}, \dots, v_{k-1}, \underline{v_j}, v_{k+1}, \dots, v_n]) \,.
\end{aligned}$$

Using condition (b) of definition III.4.2.1, we rewrite the above equality

$$0 = 0 + \phi(V\Sigma) + \phi(V) + 0 \,.$$

Applying this equality in the case when V is the identity matrix of order n and using condition (c) of definition III.4.2.1, we obtain $\phi(\Sigma) = -1$. As a result, we then have $\phi(V\Sigma) = \phi(\Sigma)\phi(V)$.

For the function of n arguments from $\mathbb{F}^n$, the property that swapping two arguments amounts to inverting the sign is called *alternation*, or *anti-symmetry*. So lemma III.4.3.3 shows that any determinant function, considered as a function of the n columns of its original matrix argument, is *alternating*, or *anti-symmetric*.

Multilinear anti-symmetric functions are interesting for us as indicators of linear dependence but have many particular applications. For example, *Slater determinants* are widely used for representing wave functions of systems of fermions in quantum mechanics.

Lemma III.4.3.4 (determinant functions and triangular structure)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $\phi\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ be a determinant function. Then the following properties hold:

(a) if $L \in \mathbb{F}^{n\times n}$ is a lower-triangular matrix, then $\phi(L) = L_{11}\cdots L_{nn}$ and $\phi(VL) = \phi(V)\phi(L)$ for all $V \in \mathbb{F}^{n\times n}$;

(b) if $U \in \mathbb{F}^{n\times n}$ is an upper-triangular matrix, then $\phi(U) = U_{11}\cdots U_{nn}$ and $\phi(VU) = \phi(V)\phi(U)$ for all $V \in \mathbb{F}^{n\times n}$.

Proof. Let us consider $V \in \mathbb{F}^{n\times n}$ and a lower-triangular matrix $L \in \mathbb{F}^{n\times n}$. Denoting the columns of VL by $v'_1, \ldots, v'_n$, we obtain by definition II.4.1.3, that $v'_k = \sum_{j=k}^n L_{jk} v_j$ for each $k \in \{1,\ldots,n\}$ since L is lower triangular. Then lemma III.4.3.1 gives

$$\phi([v'_1,\ldots,v'_{k-1},v'_k,v_{k+1},\ldots,v_n]) = L_{kk}\,\phi([v'_1,\ldots,v'_{k-1},v_k,v_{k+1},\ldots,v_n])$$

for each $k \in \{1,\ldots,n\}$. Applying this equality iteratively with $k \in \{1,\ldots,n\}$, we obtain

$$\phi([v'_1,\ldots,v'_n]) = L_{11}\cdots L_{nn}\,\phi([v_1,\ldots,v_n])\,,$$

which can be equivalently expressed as $\phi(VL) = L_{11}\cdots L_{nn}\,\phi(V)$. In the particular case when V is the identity matrix of order n, this result and condition (c) of definition III.4.2.1 give $\phi(L) = L_{11}\cdots L_{nn}$. Then, for an arbitrary matrix $V \in \mathbb{F}^{n\times n}$, we obtain $\phi(VL) = \phi(V)\phi(L)$.

The proof of part (b) is left to the reader as an exercise.

Lemma III.4.3.5 (determinant functions and the pivoted LU decomposition)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $\phi\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ be a determinant function, and $A \in \mathbb{F}^{n\times n}$ be nonzero. Consider a complete r-step decomposition $A = P^{\mathsf{T}} LUQ$ corresponding to row- and column-exchange indices $\pi_1,\ldots,\pi_r$ and $\sigma_1,\ldots,\sigma_r$ with $r \in \{1,\ldots,n\}$. Then $\phi(A) = \phi(A^{\mathsf{T}}) = \phi(P)\,\phi(U)\,\phi(Q)$, where $\phi(P) = (-1)^\mu = \phi(P^{\mathsf{T}})$ and $\phi(Q) = (-1)^\nu = \phi(Q^{\mathsf{T}})$ with

$$\mu = \#\bigl\{k \in \{1,\ldots,r\}\colon \pi_k > k\bigr\} \quad \text{and} \quad \nu = \#\bigl\{k \in \{1,\ldots,r\}\colon \sigma_k > k\bigr\}\,. \tag{III.4.3.1}$$

Furthermore, every $V \in \mathbb{F}^{n\times n}$ satisfies $\phi(VA) = \phi(V)\,\phi(A)$.

Proof. From definitions II.5.8.1 and II.5.6.1, we obtain the following:

(i) $L \in \mathbb{F}^{n\times n}$ is a unit lower-triangular matrix;

(ii) $U \in \mathbb{F}^{n\times n}$ is an upper-triangular matrix (since the decomposition is complete);

(iii) $\pi_k, \sigma_k \in \{k,\ldots,n\}$ for each $k \in \{1,\ldots,r\}$;

(iv) $P = \mathit{\Pi}_r\cdots\mathit{\Pi}_1$ and $Q = \mathit{\Sigma}_r\cdots\mathit{\Sigma}_1$, where $\mathit{\Pi}_k$ is the (k,π_k)-exchange matrix of order n and $\mathit{\Sigma}_k$ is the (k,σ_k)-exchange matrix of order n for each $k \in \{1,\ldots,r\}$.

Applying lemma III.4.3.3 to each of the equalities

$$P = \mathit{\Pi}_r\cdots\mathit{\Pi}_1\,,\qquad Q = \mathit{\Sigma}_r\cdots\mathit{\Sigma}_1\,,$$
$$P^{\mathsf{T}} = \mathit{\Pi}_1\cdots\mathit{\Pi}_r\,,\quad Q^{\mathsf{T}} = \mathit{\Sigma}_1\cdots\mathit{\Sigma}_r$$

r times, we obtain $\phi(P) = (-1)^\mu$, $\phi(Q) = (-1)^\nu$, $\phi(P^{\mathsf{T}}) = (-1)^\mu$, $\phi(Q^{\mathsf{T}}) = (-1)^\nu$ with μ and ν given by (III.4.3.1).

Applying lemma III.4.3.3 r times, then part (b) of lemma III.4.3.4, then part (a) of lemma III.4.3.4 and then again lemma III.4.3.3 r times to each of the equalities

$$A = \mathit{\Pi}_1 \cdots \mathit{\Pi}_r \, L U \, \mathit{\Sigma}_r \cdots \mathit{\Sigma}_1 \quad \text{and} \quad A^\mathsf{T} = \mathit{\Sigma}_1 \cdots \mathit{\Sigma}_r \, U^\mathsf{T} L^\mathsf{T} \, \mathit{\Pi}_r \cdots \mathit{\Pi}_1 \,,$$

we obtain

$$\phi(A) = \phi(\mathit{\Pi}_1) \cdots \phi(\mathit{\Pi}_r) \cdot \phi(U) \cdot \phi(\mathit{\Sigma}_r) \cdots \phi(\mathit{\Sigma}_1) = \phi(P)\phi(U)\phi(Q)$$

and

$$\phi(A^\mathsf{T}) = \phi(\mathit{\Sigma}_1) \cdots \phi(\mathit{\Sigma}_r) \cdot \phi(U^\mathsf{T}) \cdot \phi(\mathit{\Pi}_r) \cdots \phi(\mathit{\Pi}_1) = \phi(P)\phi(U)\phi(Q)$$

with $\phi(U^\mathsf{T}) = U_{11} \cdots U_{nn} = \phi(U)$. Thence follows the equality $\phi(A) = \phi(A^\mathsf{T})$.

Finally, for every $V \in \mathbb{F}^{n\times n}$, applying the same steps to the equality

$$VA = V \mathit{\Pi}_1 \cdots \mathit{\Pi}_r \, L U \, \mathit{\Sigma}_r \cdots \mathit{\Sigma}_1 \,, \tag{III.4.3.2}$$

we obtain

$$\phi(VA) = \phi(V)\phi(\mathit{\Pi}_1) \cdots \phi(\mathit{\Pi}_r)\phi(U)\phi(\mathit{\Sigma}_r) \cdots \phi(\mathit{\Sigma}_1) = \phi(V)\phi(P)\phi(U)\phi(Q) = \phi(V)\phi(A)\,, \tag{III.4.3.3}$$

which completes the proof.

Lemma III.4.3.6 (determinant functions under matrix transposition, multiplication and inversion)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $\phi\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ be a determinant function. Then the following statements hold.

(a) $\phi(A^\mathsf{T}) = \phi(A)$ for every $A \in \mathbb{F}^{n\times n}$.

(b) $\phi(BA) = \phi(B)\,\phi(A)$ for all $A, B \in \mathbb{F}^{n\times n}$.

(c) Assume that $A \in \mathbb{F}^{n\times n}$ is an invertible matrix. Then $\phi(A) \neq 0$ and $\phi(A^{-1}) = \big(\phi(A)\big)^{-1}$.

Proof. Let $A, B \in \mathbb{F}^{n\times n}$. If A is the zero matrix of size $n \times n$, the statements of parts (a) and (b) hold trivially since $\phi(A) = 0$ by condition (b) and the matrices A^T and BA are both equal to A. When $A \in \mathbb{F}^{n\times n}$ is nonzero, lemma II.5.8.6 guarantees that A has a complete pivoted LU decomposition, and then lemma III.4.3.5 immediately yields the claims of parts (a) and (b).

Consider an invertible matrix $A \in \mathbb{F}^{n\times n}$. Then $\phi(A^{-1}A) = \phi(A^{-1})\,\phi(A)$ by part (b). Noting that $A^{-1}A$ is the identity matrix of order n by definition II.4.1.14 and applying condition (c) of definition III.4.2.1, we obtain the claim.

Lemma III.4.3.7 (determinant functions and block matrices)**.** Let $n, r \in \mathbb{N}$ be such that $r < n$ and $\phi_r\colon \mathbb{F}^{r\times r} \to \mathbb{F}$, $\phi_{n-r}\colon \mathbb{F}^{(n-r)\times(n-r)} \to \mathbb{F}$ and $\phi_n\colon \mathbb{F}^{n\times n} \to \mathbb{F}$ be determinant functions. Consider a matrix $A \in \mathbb{F}^{n\times n}$ partitioned as follows:

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

where $A_{11} \in \mathbb{F}^{r\times r}$. Assume that the matrix A_{11} is invertible and consider the corresponding Schur complement $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$. Then

$$\phi_n(A) = \phi_r(A_{11})\,\phi_{n-r}(S)\,.$$

Proof. Our plan is to use lemma III.4.3.5, which expresses the value of a determinant function at an arbitrary matrix with a pivoted LU decomposition in terms of the values of the same function at the factors of the decomposition. In the present setting, we are dealing with the functions ϕ_n, ϕ_r and ϕ_{n-r} evaluated at the respective matrices A, A_{11} and S. We can consider pivoted LU decompositions of the three matrices to evaluate the respective three values of determinant functions. We should, however, ensure that the decompositions are related in such a way that the values of the determinant functions at the respective factors can be related.

First, note that $\operatorname{rank} A_{11} = r$ by lemma II.5.9.2. Then theorem II.5.9.8 yields that there exists an r-step pivoted LU decomposition

$$A_{11} = \widehat{P}^{\mathsf{T}} \widehat{L} \widehat{U} \widehat{Q} \tag{III.4.3.4}$$

corresponding to row- and column-exchange indices $\pi_1, \ldots, \pi_r$ and $\sigma_1, \ldots, \sigma_r$. By definition II.5.8.1, we have $\widehat{P} = \widehat{\varPi}_r \cdots \widehat{\varPi}_1$ and $\widehat{Q} = \widehat{\varSigma}_r \cdots \widehat{\varSigma}_1$, where $\widehat{\varPi}_k$ and $\widehat{\varSigma}_k$ are the (k, π_k)- and (k, σ_k)-exchange matrices of order r for every $k \in \{1, \ldots, r\}$ and $\widehat{P} A_{11} \widehat{Q}^{\mathsf{T}} = \widehat{L} \widehat{U}$ is an r-step LU decomposition. In particular, the factor $\widehat{L}$ is unit lower triangular and is therefore invertible by part (a) of lemma II.5.5.4 and the factor $\widehat{U}$ is upper triangular and has no zeros on the diagonal and is therefore invertible by part (b) of corollary II.5.5.5. Let $\varPi_k$ and $\varSigma_k$ be the (k, π_k)- and (k, σ_k)-exchange matrices of order n for every $k \in \{1, \ldots, r\}$ and consider $P = \varPi_r \cdots \varPi_1$ and $Q = \varSigma_r \cdots \varSigma_1$. Denoting the identity matrix of order $n-r$ by I_2, we observe that

$$\varPi_k = \begin{bmatrix} \widehat{\varPi}_k & \\ & I_2 \end{bmatrix} \quad \text{and} \quad \varSigma_k = \begin{bmatrix} \widehat{\varSigma}_k & \\ & I_2 \end{bmatrix} \quad \text{for each} \quad k \in \{1, \ldots, r\},$$

and block matrix multiplication then yields

$$P = \begin{bmatrix} \widehat{P} & \\ & I_2 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} \widehat{Q} & \\ & I_2 \end{bmatrix}$$

and $PAQ^{\mathsf{T}} = LU$ with

$$L = \begin{bmatrix} \widehat{L} & \\ L_2 & I_2 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} \widehat{U} & U_2 \\ & S \end{bmatrix},$$

where $L_2 = A_{21} \widehat{Q}^{\mathsf{T}} \widehat{U}^{-1}$, $U_2 = \widehat{L}^{-1} \widehat{P} A_{12}$ and $S = A_{22} - A_{21} A_{11}^{-1} A_{12}$. Indeed, we have $A_{22} - S = A_{21} \widehat{Q}^{\mathsf{T}} (\widehat{P} A_{11} \widehat{Q}^{\mathsf{T}})^{-1} \widehat{P} A_{12} = A_{21} \widehat{Q}^{\mathsf{T}} (\widehat{L} \widehat{U})^{-1} \widehat{P} A_{12} = L_2 U_2$ and hence

$$\begin{aligned} PAQ^{\mathsf{T}} = \begin{bmatrix} \widehat{L}\widehat{U} & \widehat{P} A_{12} \\ A_{21} \widehat{Q}^{\mathsf{T}} & A_{22} \end{bmatrix} = \begin{bmatrix} \widehat{L}\widehat{U} & \widehat{L} U_2 \\ L_2 \widehat{U} & A_{22} \end{bmatrix} &= \begin{bmatrix} \widehat{L} & \\ & I_2 \end{bmatrix} \begin{bmatrix} \widehat{U} & U_2 \\ L_2 \widehat{U} & A_{22} \end{bmatrix} \\ &= \begin{bmatrix} \widehat{L} & \\ & I_2 \end{bmatrix} \begin{bmatrix} I_1 & \\ L_2 & I_2 \end{bmatrix} \begin{bmatrix} \widehat{U} & U_2 \\ & S \end{bmatrix} = LU, \end{aligned}$$

where I_1 is the identity matrix of order r. By definition II.5.6.1, the factorization $PAQ^{\mathsf{T}} = LU$ is an r-step LU decomposition.

By lemma II.5.8.6 and theorem II.5.9.8, the Schur complement S has a complete q-step pivoted LU decomposition

$$S = \widetilde{P}^{\mathsf{T}} \widetilde{L} \widetilde{U} \widetilde{Q} \tag{III.4.3.5}$$

corresponding to row- and column-exchange indices $\widetilde{\pi}_1, \ldots, \widetilde{\pi}_q$ and $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_q$, where $q = \operatorname{rank} S$. Then the factor $\widetilde{U}$ is upper triangular. Applying part (b) of lemma II.5.8.5, we obtain that A has an $(r+q)$-step pivoted LU decomposition

$$A = P_\star^{\mathsf{T}} L_\star U_\star Q_\star \tag{III.4.3.6}$$

corresponding to the row- and column-exchange indices $\pi_1, \ldots, \pi_{r+q}$ and $\sigma_1, \ldots, \sigma_{r+q}$ given by (II.5.8.2) and with the factors $L_\star$ and $U_\star$ given by (II.5.8.3). Since the matrices $\widehat{U}$ and $\widetilde{U}$ are upper triangular, the factor $U_\star$ is also upper triangular due to (II.5.8.3). By part (b) of lemma III.4.3.4, we obtain $\phi_n(U_\star) = \phi_r(\widehat{U}) \phi_{n-r}(\widetilde{U})$.

Applying lemma III.4.3.5 to the pivoted LU decompositions of A, A_{11} and S, we obtain the following:

$$\begin{aligned} \phi_n(A) &= (-1)^{\delta_\star}\, \phi_n(U_\star) = (-1)^{\delta_\star}\, \phi_r(\widehat{U}) \phi_{n-r}(\widetilde{U})\,, \\ \phi_r(A_{11}) &= (-1)^{\widehat{\delta}}\, \phi_r(\widehat{U})\,, \qquad\qquad\qquad\qquad\qquad , \\ \phi_{n-r}(S) &= (-1)^{\widetilde{\delta}}\, \phi_{n-r}(\widetilde{U})\,. \end{aligned}$$

where the respective numbers of “proper” exchanges are

$$\begin{aligned} \delta_\star &= \#\big\{k \in \{1, \ldots, r+q\} \colon \pi_k > k\big\} + \#\big\{k \in \{1, \ldots, r+q\} \colon \sigma_k > k\big\}\,, \\ \widehat{\delta} &= \#\big\{k \in \{1, \ldots, r\} \colon \pi_k > k\big\} + \#\big\{k \in \{1, \ldots, r\} \colon \sigma_k > k\big\}\,, \\ \widetilde{\delta} &= \#\big\{k \in \{1, \ldots, q\} \colon \widetilde{\pi}_k > k\big\} + \#\big\{k \in \{1, \ldots, q\} \colon \widetilde{\sigma}_k > k\big\}\,. \end{aligned}$$

Due to (II.5.8.2), these numbers satisfy $\delta_\star = \widehat{\delta} + \widetilde{\delta}$. As a result, we obtain $\phi_n(A) = \phi_r(A_{11}) \phi_{n-r}(S)$.

Corollary III.4.3.8. Let $m, n \in \mathbb{N}$ and $\phi_m \colon \mathbb{F}^{m \times m} \to \mathbb{F}$, $\phi_n \colon \mathbb{F}^{n \times n} \to \mathbb{F}$ and $\phi_{m+n} \colon \mathbb{F}^{(m+n)\times(m+n)} \to \mathbb{F}$ be determinant functions. Consider matrices $A \in \mathbb{F}^{m \times m}$, $B \in \mathbb{F}^{n \times m}$, $C \in \mathbb{F}^{m \times n}$ and $D \in \mathbb{F}^{n \times n}$. Then

$$L = \begin{bmatrix} A & \\ B & D \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} A & C \\ & D \end{bmatrix}.$$

Then $\phi_{m+n}(L) = \phi_m(A)\, \phi_n(D) = \phi_{m+n}(U)$.

Proof. If $\operatorname{rank} A < m$, then the claim follows trivially. Indeed, $\phi_m(A) = 0$ by corollary III.4.3.2. On the other hand, since the rows of A are linearly dependent, so are the rows of L. One can give an analogous proof to that $\phi_{m+n}(U) = 0$ in this case. Then $\operatorname{rank} L < m + n$, and hence $\phi_{m+n}(L) = 0$ by corollary III.4.3.2. This shows that the claimed equalities hold trivially in this case.

If $\operatorname{rank} A = m$, then A is invertible by part (a) of theorem II.5.9.7, and lemma III.4.3.7 yields the claim.

Definition III.4.3.9 (matrix minor)**.** Let $\mathbb{F}$ be a field and $m, n \in \mathbb{N}$ be such that $m, n \geq 2$, $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$. For every matrix $A = [a_{ij}]_{i=1,\, j=1}^{m \quad n} \in \mathbb{F}^{m \times n}$, the submatrix $\mathscr{M}_{ij}(A) = [a_{\pi_k \sigma_\ell}]_{k=1,\, \ell=1}^{m-1\, n-1} \in \mathbb{F}^{(m-1)\times(n-1)}$ formed by the intersection of rows $\pi = (1, \dots, i-1, i+1, \dots, m)$ and columns $\sigma = (1, \dots, j-1, j+1, \dots, n)$ of A (in other words, obtained from A by removing row i and column j) is called *minor* (i, j) *of* A:

$$\mathscr{M}_{ij}(A) = \left[\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ \hline a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ a_{m1} & \cdots & a_{m,j-1} & a_{m,j+1} & \cdots & a_{mn} \end{array}\right] \in \mathbb{F}^{(m-1)\times(n-1)} .$$

Lemma III.4.3.10 (expansion of a determinant function along a column)**.** Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$ be such that $n \geq 2$. Assume that $\phi_n \colon \mathbb{F}^{n \times n} \to \mathbb{F}$ and $\phi_{n-1} \colon \mathbb{F}^{(n-1)\times(n-1)} \to \mathbb{F}$ are determinant functions. Then

$$\phi_n(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij}\, \phi_{n-1}\big(\mathscr{M}_{ij}(A)\big) \tag{III.4.3.7}$$

for all $A = [a_{ik}]_{i=1,\, k=1}^{n \quad n} \in \mathbb{F}^{n \times n}$ and $j \in \{1, \dots, n\}$.

Proof. Consider $A = [a_{ik}]_{i=1,\, k=1}^{n \quad n} \in \mathbb{F}^{n \times n}$ and $j \in \{1, \dots, n\}$. Let $e_1, \dots, e_n$ denote the columns of the identity matrix of order n. Since column j of A is equal to $\sum_{i=1}^{n} a_{ij}\, e_i$, condition (a) of definition III.4.2.1 yields

$$\phi_n(A) = \sum_{i=1}^{n} a_{ij}\, \phi_n(\widetilde{A}_{ij}) \, ,$$

where $\widetilde{A}_{ij}$ is obtained from A by replacing column j with e_i.

$$\widetilde{A}_{ij} = \left[\begin{array}{ccc|c|cc} a_{11} & \cdots & a_{1,j-1} & & a_{1,j+1} & a_{1n} \\ \vdots & \cdots & \vdots & & \vdots & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & & a_{i-1,j+1} & a_{i-1,n} \\ \hline a_{i1} & \cdots & a_{i,j-1} & 1 & a_{i,j+1} & a_{i,n} \\ \hline a_{i+1,1} & \cdots & a_{i+1,j-1} & & a_{i+1,j+1} & a_{i+1,n} \\ \vdots & \cdots & \vdots & & \vdots & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & & a_{n,j+1} & a_{nn} \end{array}\right] \tag{III.4.3.8}$$

for each $i \in \{1, \dots, n\}$.

Let $\mathit{\Pi}_k$ denote the $(k, k+1)$-exchange matrix of order n for each $k \in \{1, \ldots, n-1\}$ and consider the permutation matrices $Q_j = I \cdot \mathit{\Pi}_1 \cdots \mathit{\Pi}_{j-1}$ and $P_i = I \cdot \mathit{\Pi}_1 \cdots \mathit{\Pi}_{i-1}$ with $i \in \{1, \ldots, n-1\}$, where I is the identity matrix of order n.

For every $i \in \{1, \ldots, n\}$, introducing

$$\widehat{A}_{ij} = \left[\begin{array}{c|c} 1 & b_{ij} \\ \hline & \mathscr{M}_{ij}(A) \end{array}\right] \quad \text{with} \quad b_{ij} = \begin{bmatrix} a_{i1} & \cdots & a_{i,j-1} & a_{i,j+1} & \cdots & a_{in} \end{bmatrix} \in \mathbb{F}^{1\times(n-1)},$$

we obtain $\widetilde{A}_{ij} = \mathit{\Pi}_{i-1} \cdots \mathit{\Pi}_1 \widehat{A}_{ij} \mathit{\Pi}_1 \cdots \mathit{\Pi}_{j-1}$. Applying lemma III.4.3.3 and part (a) of lemma III.4.3.6, we arrive at

$$\begin{aligned} \phi_n(\widetilde{A}_{ij}) &= \phi_n(\mathit{\Pi}_{i-1} \cdots \mathit{\Pi}_1 \widehat{A}_{ij} \mathit{\Pi}_1 \cdots \mathit{\Pi}_{j-1}) = \cdots = (-1)^{j-1}\, \phi_n(\mathit{\Pi}_{i-1} \cdots \mathit{\Pi}_1 \widehat{A}_{ij}) \\ &= (-1)^{j-1}\, \phi_n(\widehat{A}_{ij}^{\mathsf{T}} \mathit{\Pi}_1 \cdots \mathit{\Pi}_{i-1}) = \cdots = (-1)^{i-1+j-1}\, \phi_n(\widehat{A}_{ij}^{\mathsf{T}}) = (-1)^{i+j}\, \phi_n(\widehat{A}_{ij}). \end{aligned}$$

Finally, by corollary III.4.3.8 and condition (c) of definition III.4.2.1, we have $\phi_n(\widehat{A}_{ij}) = \phi_{n-1}\big(\mathscr{M}_{ij}(A)\big)$, so that

$$\phi_n(\widetilde{A}_{ij}) = (-1)^{i+j}\, \phi_{n-1}\big(\mathscr{M}_{ij}(A)\big).$$

Substituting this result into the expression for $\phi(A)$ obtained at the beginning of the proof, we obtain (III.4.3.7).

§ III.4.4. Matrix determinants as unique determinant functions

In § III.4.3, we derived various properties and formulae of abstract determinant functions, blindly assuming that they exist. In particular, (III.4.3.7) features an expansion of a determinant function along a column of its argument. The analysis given in § III.4.3 shows that determinant functions are very convenient *if* they exist, but it does not explicitly demonstrate that they actually do.

For each $n \in \mathbb{N}$, we need to find out whether there is a function satisfying all conditions of definition III.4.2.1. In the case of $n = 1$, the answer is obvious. On the other hand, lemma III.4.3.10 shows what form *any* determinant function on $\mathbb{F}^{n\times n}$ has to take once there exists a determinant function on $\mathbb{F}^{(n-1)\times(n-1)}$. We will now use such a recursive definition to prove that, for each $n \in \mathbb{N}$, there exists a unique determinant function on $\mathbb{F}^{n\times n}$.

Definition III.4.4.1 (matrix determinant)**.** Let $\mathbb{F}$ be a field. For each $n \in \mathbb{N}$, the function $\det_n \colon \mathbb{F}^{n\times n} \to \mathbb{F}$ defined as follows is referred to as the *matrix determinant* of order n.

First, $\det_1$ is defined by setting $\det_1([a]) = a$ for all $a \in \mathbb{F}$.

For each $n \in \mathbb{N}$ such that $n \geq 2$, $\det_n$ is defined through $\det_{n-1}$ by setting

$$\det_n(A) = \sum_{i=1}^{n} (-1)^{i+1} A_{i1} \det_{n-1}\big(\mathscr{M}_{i1}(A)\big) \tag{III.4.4.1}$$

for all $A \in \mathbb{F}^{n\times n}$.

Remark III.4.4.2 (notation)**.** In the context of definition III.4.4.1, the subscript indices specifying matrix orders and the parentheses enclosing arguments are often omitted, so all $\det_n$ with $n \in \mathbb{N}$ are often denoted by “det” and $\det_n(A)$ for $A \in \mathbb{F}^{n\times n}$ with $n \in \mathbb{N}$ is often written as $\det A$.

Theorem III.4.4.3 (matrix determinants are unique determinant functions)**.** Let $\mathbb{F}$ be a field. For each $n \in \mathbb{N}$, the function $\det_n$ is a unique determinant function on $\mathbb{F}^{n\times n}$.

Proof. Consider first the case of $n = 1$. The function $\det_1$ trivially satisfies definition III.4.2.1 and is therefore a determinant function. On the other hand, if $\phi_1 \colon \mathbb{F}^{1\times 1} \to \mathbb{F}$ is a determinant function, then conditions (a) and (c) of definition III.4.2.1 imply that $\phi_1([a]) = a \cdot \phi_1([1]) = a \cdot 1 = a = \det_1([a])$ for every $a \in \mathbb{F}$, which proves that $\phi_1 = \det_1$.

For the remainder of the proof, we consider $n \in \mathbb{N}$ such that $n \geq 2$ and assume that $\det_1, \ldots, \det_{n-1}$ are unique determinant functions on $\mathbb{F}^{1\times 1}, \ldots, \mathbb{F}^{(n-1)\times(n-1)}$ respectively. Our goal is to show that the function $\det_n \colon \mathbb{F}^{n\times n} \to \mathbb{F}$ defined by (III.4.4.1) is a unique determinant function on $\mathbb{F}^{n\times n}$.

(i) Consider $j \in \{1, \ldots, n\}$ and matrices

$$A = [a_1, \ldots, a_n]\,, \quad B = [a_1, \ldots, a_{j-1}, b, a_{j+1}, \ldots, a_n]$$

and

$$C = [a_1, \ldots, a_{j-1}, \alpha a_j + b, a_{j+1}, \ldots, a_n]$$

with $a_1, \ldots, a_n, b \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$.

If $j = 1$, then it follows immediately from the definition of $\det_n$ that

$$\det_n C = \sum_{i=1}^{n} (-1)^{i+1} (\alpha A_{i1} + B_{i1}) \det_{n-1} \mathscr{M}_{i1}(A) = \alpha \det_n A + \det_n B\,.$$

If $j > 1$, then, since $\det_{n-1}$ is a determinant function, condition (a) of definition III.4.2.1 yields

$$\det_{n-1} \mathscr{M}_{i1}(C) = \alpha \det_{n-1} \mathscr{M}_{i1}(A) + \det_{n-1} \mathscr{M}_{i1}(B)$$

for each $i \in \{1, \ldots, n\}$. This implies that

$$\det_n C = \sum_{i=1}^{n} (-1)^{i+1} A_{i1} \det_{n-1} \mathscr{M}_{i1}(C) = \alpha \det_n A + \det_n B\,.$$

In either case, $\det_n$ satisfies condition (a) of definition III.4.2.1.

(ii) Evaluating $\det_n$ at the identity matrix I_n of order n, we obtain $\det_n I_n = 1 \cdot \det_{n-1} I_{n-1}$, where I_{n-1} is the identity matrix of order $n-1$. Since $\det_{n-1} I_{n-1} = 1$ by condition (c) of definition III.4.2.1, we conclude that $\det_n I_n = 1$, i.e., $\det_n$ satisfies condition (c) of definition III.4.2.1.

(iii) Let us consider a matrix $A = [a_{ij}]_{i=1,\,j=1}^{n\quad n} \in \mathbb{F}^{n\times n}$ and assume that its columns with indices $j, k \in \{1, \ldots, n\}$ such that $j < k$ are identical: $a_{ij} = a_{ik}$ for every $i \in \{1, \ldots, n\}$.

First, let us consider the case of $j > 1$. Then, for each $i \in \{1, \ldots, n\}$, columns $j-1$ and $k-1$ of $\mathscr{M}_{i1}(A)$ are identical and, thus, $\det_{n-1} \mathscr{M}_{i1}(A) = 0$ by condition (b) of definition III.4.2.1 since $\det_{n-1}$ is a determinant function. This immediately leads to $\det_n(A) = 0$.

The remainder of the proof will focus on the case of $j = 1$. If $n = 2$, then $k = 2$ and

$$\det_2 A = a_{11} a_{22} - a_{21} a_{12} = a_{11} a_{22} - a_{11} a_{22} = 0\,.$$

If $n > 2$, then both $\det_{n-1}$ and $\det_{n-2}$ are determinant functions by assumption, and lemma III.4.3.10 gives the following expansion along column $k-1$:

$$\det_{n-1} \mathscr{M}_{i1}(A) = \sum_{\ell=1}^{n-1} (-1)^{\ell+k-1} \big(\mathscr{M}_{i1}(A)\big)_{\ell\, k-1} \det_{n-2} \mathscr{M}_{\ell\, k-1}\big(\mathscr{M}_{i1}(A)\big)$$

for each $i \in \{1, \ldots, n\}$.

Let us now look at $\big(\mathscr{M}_{i1}(A)\big)_{\ell\, k-1}$ with $i \in \{1, \ldots, n\}$ and $\ell \in \{1, \ldots, n-1\}$: we have $\big(\mathscr{M}_{i1}(A)\big)_{\ell\, k-1} = a_{\ell k}$ for $\ell < i$ and $\big(\mathscr{M}_{i1}(A)\big)_{\ell\, k-1} = a_{\ell+1\, k}$ for $\ell \geq i$.

Now we will express in terms of A the iterated minors of A appearing in the above exppansion. Let us, for any distinct $i, \ell \in \{1, \ldots, n\}$, denote by $\widetilde{A}_{i\ell}$ the matrix obtained from A by removing its rows i and ℓ and columns 1 and k. Note that we then have $\widetilde{A}_{\ell i} = \widetilde{A}_{i\ell}$ for any distinct $i, \ell \in \{1, \ldots, n\}$. Further, for any $i \in \{1, \ldots, n\}$ and $\ell \in \{1, \ldots, n-1\}$, we have $\mathscr{M}_{\ell\, k-1}\big(\mathscr{M}_{i1}(A)\big) = \widetilde{A}_{i\ell}$ if $\ell < i$ and $\mathscr{M}_{\ell\, k-1}\big(\mathscr{M}_{i1}(A)\big) = \widetilde{A}_{i\, \ell+1}$ if $\ell \geq i$.

These observations show that we can simplify the above sum by splitting it into two:

$$\begin{aligned}
&\det_{n-1} \mathscr{M}_{i1}(A) \\
&= \sum_{\ell=1}^{i-1} (-1)^{\ell+k-1} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell} + \sum_{\ell=i}^{n-1} (-1)^{\ell+k-1} a_{\ell+1\, k} \det_{n-2} \widetilde{A}_{i\, \ell+1} \\
&\qquad = -\sum_{\ell=1}^{i-1} (-1)^{\ell+k} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell} + \sum_{\ell=i+1}^{n} (-1)^{\ell+k} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell}
\end{aligned}$$

for each $i \in \{1, \ldots, n\}$, where we shifted the summation index at the last step in order to bring factors appearing in the second sum into a form similar to the form of the factors present in the first sum.

Substituting this into the definition of $\det_n A$ and swapping the summation indices i and ℓ (which we are free to denote as we prefer) in the second sum, we obtain

$$\begin{aligned}
\det_n A &= \sum_{\substack{i,\ell \in \{1,\ldots,n\}:\\ \ell < i}} (-1)^{i+\ell+k} a_{i1} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell} - \sum_{\substack{i,\ell \in \{1,\ldots,n\}:\\ \ell > i}} (-1)^{i+\ell+k} a_{i1} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell} \\
&= \sum_{\substack{i,\ell \in \{1,\ldots,n\}:\\ \ell < i}} (-1)^{i+\ell+k} a_{i1} a_{\ell k} \det_{n-2} \widetilde{A}_{i\ell} - \sum_{\substack{\ell,i \in \{1,\ldots,n\}:\\ i > \ell}} (-1)^{\ell+i+k} a_{\ell 1} a_{ik} \det_{n-2} \widetilde{A}_{\ell i} \\
&= \sum_{\substack{i,\ell \in \{1,\ldots,n\}:\\ \ell < i}} (-1)^{i+\ell+k} (a_{i1} a_{\ell k} - a_{\ell 1} a_{ik}) \det_{n-2} \widetilde{A}_{i\ell} \,.
\end{aligned}$$

Since columns 1 and k of A are identical, we have $a_{i1} a_{\ell k} - a_{\ell 1} a_{ik} = a_{i1} a_{\ell 1} - a_{\ell 1} a_{i1} = 0$ for all $i, \ell \in \{1, \ldots, n\}$, which leads to $\det_n(A) = 0$. This completes the proof of that $\det_n$ satisfies condition (c) of definition III.4.2.1.

In items (i) to (iii) above, we proved that $\det_n$ satisfies the conditions stated in definition III.4.2.1. It is therefore a determinant function on $\mathbb{F}^{n\times n}$. By assumption, $\det_{n-1}$ is a unique determinant function on $\mathbb{F}^{(n-1)\times(n-1)}$. Then any determinant function ϕ_n on $\mathbb{F}^{n\times n}$ satisfies lemma III.4.3.10 with $\phi_{n-1} = \det_{n-1}$, and we conclude that $\phi_n = \det_n$ due to (III.4.3.7).

The claim follows by induction.

Theorem III.4.4.3 shows that the matrix determinant of order $n \in \mathbb{N}$ is a unique determinant function on $\mathbb{F}^{n\times n}$. As a consequence, all properties established in § III.4.3 for *any* determinant function on $\mathbb{F}^{n\times n}$ hold for $\det_n$ in particular.

The following combinatorial formula for the matrix determinant can be derived immediately from lemma III.4.3.10 and theorem III.4.4.3.

Lemma III.4.4.4. Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $\mathscr{P}_n$ denote the set of permutations of $1, \ldots, n$. Then

$$\det A = \sum_{\pi \in \mathscr{P}_n} \operatorname{sign}(\pi) \cdot A_{\pi_1 1} \cdots A_{\pi_n n} \tag{III.4.4.2}$$

for every $A \in \mathbb{F}^{n\times n}$, where, for every $\pi \in \mathscr{P}_n$, $\operatorname{sign}(\pi) = (-1)^{\operatorname{inv}(\pi)}$ is the sign of the permutation π, encoding the parity of the permutation π, i.e., the parity of the number

$$\operatorname{inv}(\pi) = \#\big\{(i,j) \in \{1,\ldots,n\} \colon i < j \text{ and } \pi_i > \pi_j\big\}$$

of inversions in the permutation π.

Proof. Proof is left to the reader as an exercise.

§ III.4.5. Some applications of the matrix determinant

Combining lemmata II.5.9.2 to III.2.2.4, III.2.3.9, II.5.10.13 and II.5.10.16 and definition III.2.2.1 and part (a) of theorem II.5.9.7 with lemma III.4.3.6 and corollary III.4.3.2, we obtain the following meta-theorem, which connects our results on the full-rank property and invertibility of matrices and the basis property and linear independence of vectors.

Theorem III.4.5.1 (meta-theorem)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. Then the following conditions are equivalent:

(i) A is invertible;
(ii) $\operatorname{rank} A = n$;
(iii) the columns of A form a basis for $\operatorname{Im} A$;
(iv) the columns of A^T form a basis for $\operatorname{Im} A^\mathsf{T}$;
(v) $\operatorname{Ker} A = \{0\}$;
(vi) $\operatorname{Im} A = \mathbb{F}^n$;
(vii) the columns of A are linearly independent;
(viii) the rows of A are linearly independent;
(ix) $\det A \neq 0$;
(x) $\det A^\mathsf{T} \neq 0$;
(xi) $\det AA^\mathsf{T} \neq 0$;
(xii) $\det A^\mathsf{T}A \neq 0$.

We finally turn to a few impressive results on the solution of linear systems and matrix inversion, which are straightforward consequences of our analysis of matrix determinants.

Proposition III.4.5.2 (Cramer's rule for linear systems)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$, a matrix $A \in \mathbb{F}^{n\times n}$ be invertible, $b \in \mathbb{F}^n$ and $x = A^{-1}b$. Then, for every $j \in \{1, \dots, n\}$,

$$x_j = \frac{\det \widetilde{A}_j}{\det A} \,,$$

where $\widetilde{A}_j \in \mathbb{F}^{n\times n}$ denotes the matrix obtained from A by replacing column j with b.

Proof. Note first that $\det A \neq 0$ by part (c) of lemma III.4.3.6, so the expression given for the components of x is defined.

The equality $Ax = b$ means that b is a linear combination of the columns of A with the coefficients $x_1, \dots, x_n$. Then lemma III.4.3.1 yields $\det A_j = x_j \cdot \det A$ for every $j \in \{1, \dots, n\}$.

Definition III.4.5.3 (cofactor and adjugate matrices)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$. The mutually transpose matrices

$$\operatorname{cof} A = \left[(-1)^{i+j} \det \mathscr{M}_{ij}(A)\right]_{i=1,\,j=1}^{n\quad\;\, n} \in \mathbb{F}^{n\times n}$$

and

$$\operatorname{adj} A = \left[(-1)^{i+j} \det \mathscr{M}_{ji}(A)\right]_{i=1,\,j=1}^{n\quad\;\, n} \in \mathbb{F}^{n\times n}$$

are called the *cofactor matrix* of A and the *adjugate matrix* of A.

Corollary III.4.5.4 (Cramer's rule for matrix inversion)**.** Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$, a matrix $A \in \mathbb{F}^{n\times n}$ be invertible and $e_1, \dots, e_n$ be the columns of the identity matrix of order n. Then

$$A^{-1} = \frac{\operatorname{adj} A}{\det A} \,.$$

Proof. Let $x_1, \dots, x_n$ denote the columns of $X = A^{-1}$. Then we have $AX = [e_1 \cdots e_n]$, i.e., $Ax_i = e_i$ and hence $x_i = A^{-1}e_i$ for every $i \in \{1, \dots, n\}$.

Let us consider $i, j \in \{1, \dots, n\}$ and denote by $\widetilde{A}_{ij}$ be the matrix obtained from A by replacing column j with e_i, as in (III.4.3.8). The entry $(A^{-1})_{ji}$ is the jth component of the column vector x_i, so proposition III.4.5.2 yields

$$(A^{-1})_{ji} = (x_i)_j = \frac{\det \widetilde{A}_{ij}}{\det A} \,.$$

In the proof of lemma III.4.3.10 we showed that $\det \widetilde{A}_{ij} = (-1)^{i+j} \det \mathscr{M}_{ij}(A)$. Recalling definition III.4.5.3, we obtain $\det \widetilde{A}_{ij} = (\operatorname{adj} A)_{ji}$.

Notes missing, see from 0:18:30 of Lecture 8 to 0:40:00 of Lecture 8 (the complexity of evaluating the determinant: recurrent exppansions and the combinatorial formula vs. Gaussian elimination)

Notes missing, see from 0:40:00 of Lecture 8 to 0:50:00 of Lecture 8 (determinants of linear transformations)

Notes missing, see from 0:50:00 of Lecture 8 to 1:03:00 of Lecture 10 (similarity transformation for matrices)

Notes missing, see from 1:03:00 of Lecture 8 to 0:22:00 of Lecture 9 (sums of subspaces)

Notes missing, see from 0:22:00 of Lecture 9 to 1:09:50 of Lecture 9 (Grassmann's dimension formula)

Notes missing, see from 1:09:50 of Lecture 9 to 0:22:10 of Lecture 10 (linear independence — or direct sums — of subspaces)

Chapter IV. Eigenvalues and eigenvectors

In chapter IV, we will focus on transformations, which are maps the domains and codomains of which are identical.

§ IV.1. Eigenpairs and eigenspaces of linear transformations and matrices

§ IV.1.1. Eigenpairs and eigenspaces of a linear transformation

Definition IV.1.1.1. Let V be a vector space over a field $\mathbb{F}$ and a transformation $\varphi\colon V \to V$ be linear. If a scalar $\lambda \in \mathbb{F}$ and a *nonzero* vector $v \in V$ are such that $\varphi(v) = \lambda v$, then we refer to

(a) (λ, v) as an *eigenpair* of φ,

(b) λ as an *eigenvalue* of φ,

(c) v as an *eigenvector* of φ.

Proposition IV.1.1.2. Let V be a vector space over a field $\mathbb{F}$ and a transformation $\varphi\colon V \to V$ be linear. If scalars $\lambda, \mu \in \mathbb{F}$ and a *nonzero* vector $v \in V$ are such that (λ, v) and (μ, v) are both eigenpairs of φ, then $\lambda = \mu$.

Proof. By definition IV.1.1.1, we have $\lambda v = \varphi(v) = \mu v$. Then $(\lambda - \mu)v = 0$, which implies $\lambda - \mu = 0$ since v is nonzero.

Proposition IV.1.1.2 shows that every eigenvector corresponds to a unique eigenvalue. For this reason, in the context of definition IV.1.1.1, the vector v is *an eigenvector of* φ *corresponding to the eigenvalue* λ and the scalar λ is *the eigenvalue of* φ *corresponding to the eigenvector* v.

Proposition IV.1.1.3. Let V be a vector space over a field $\mathbb{F}$, $\mathsf{id}\colon V \to V$ be the identity transformation of V and a transformation $\varphi\colon V \to V$ be linear.

(a) A scalar $\lambda \in \mathbb{F}$ is an eigenvalue of φ if and only if $\operatorname{Ker}(\varphi - \lambda\mathsf{id}) \neq \{0\}$.

(b) If a scalar $\lambda \in \mathbb{F}$ is an eigenvalue of φ, then $\operatorname{Ker}(\varphi - \lambda\mathsf{id}) \setminus \{0\}$ is the set of eigenvectors of φ corresponding to the eigenvalue λ.

Proof. Assume that $\lambda \in \mathbb{F}$ is an eigenvalue of φ. By definition IV.1.1.1, this means that there exists a nonzero vector $v \in V$ such that $\varphi(v) = \lambda v$, i.e., $(\varphi - \lambda\mathsf{id})(v) = 0$. This means that $v \in \operatorname{Ker}(\varphi - \lambda\mathsf{id})$, so that $\operatorname{Ker}(\varphi - \lambda\mathsf{id}) \neq \{0\}$.

On the other hand, if $\operatorname{Ker}(\varphi - \lambda\mathsf{id}) \neq \{0\}$ for some $\lambda \in \mathbb{F}$, then there exists a nonzero vector $v \in V$ such that $v \in \operatorname{Ker}(\varphi - \lambda\mathsf{id})$, i.e., $(\varphi - \lambda\mathsf{id})(v) = 0$. This is equivalent to that $\varphi(v) = \lambda v$.

So we have proved part (a). Let us now show part (b).

Assume that $\lambda \in \mathbb{F}$ is an eigenvalue of φ. If $v \in \operatorname{Ker}(\varphi - \lambda\mathsf{id}) \setminus \{0\}$, then we know that v is nonzero and that $v \in \operatorname{Ker}(\varphi - \lambda\mathsf{id})$. The latter means that $v \in V$ and $(\varphi - \lambda\mathsf{id})(v) = 0$, i.e., $\varphi(v) = \lambda v$. By definition IV.1.1.1, v is then an eigenvector corresponding to the eigenvalue λ.

Conversely, if $v \in V$ is then an eigenvector corresponding to the eigenvalue $\lambda \in \mathbb{F}$ in the sense of definition IV.1.1.1, then v is nonzero and $\varphi(v) = \lambda v$. The latter means that $(\varphi - \lambda \mathsf{id})(v) = 0$, i.e., that $v \in \operatorname{Ker}(\varphi - \lambda \mathsf{id})$. This gives $v \in \operatorname{Ker}(\varphi - \lambda \mathsf{id}) \setminus \{0\}$.

Definition IV.1.1.4. In the context of definition IV.1.1.1, the subspace $\operatorname{Ker}(\varphi - \lambda \mathsf{id})$ of V is called the *eigenspace of φ corresponding to the eigenvalue λ.*

Lemma IV.1.1.5. Let V be a vector space over a field $\mathbb{F}$ and a transformation $\varphi\colon V \to V$ be linear. Assume that $r \in \mathbb{N}$ and $\lambda_1, \ldots, \lambda_r \in \mathbb{F}$ are *distinct* eigenvalues of φ with associated the respective eigenvectors $v_1, \ldots, v_r$. Then the vectors $v_1, \ldots, v_r$ are linearly independent.

Proof. Assume that, for some $k \in \{2, \ldots, r\}$, the first $k-1$ vectors $v_1, \ldots, v_{k-1}$ are linearly independent. Since $v_1, \ldots, v_r$ are all nonzero by definition IV.1.1.1, this certainly holds for $k = 2$.

Let us assume additionally that the first k vectors $v_1, \ldots, v_{k-1}, v_k$ are linearly dependent. Then there exist $\alpha_1, \ldots, \alpha_k \in \mathbb{F}$ not all zeros and such that

$$\sum_{i=1}^{k} \alpha_i v_i = 0\,.$$

Assume that $\alpha_k = 0$. Then $\alpha_1, \ldots, \alpha_{k-1}$ are not all zero since $\alpha_1, \ldots, \alpha_k$ are not all zero, and the above trivial linear combination is a linear combination of $v_1, \ldots, v_{k-1}$. This contradicts that $v_1, \ldots, v_{k-1}$ are linearly independent. So we have $\alpha_k \neq 0$ and can express v_k as follows:

$$v_k = \sum_{i=1}^{k-1} \beta_i v_i \quad \text{with} \quad \beta_i = -\frac{\alpha_i}{\alpha_k} \in \mathbb{F} \quad \text{for each} \quad i \in \{1, \ldots, k-1\}\,.$$

Think how we could have arrived at the same conclusion using lemma III.2.3.6!

Since φ is a linear mapping and $v_1, \ldots, v_k$ are eigenvectors of φ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$, we have

$$\begin{aligned} 0 = \varphi(v_k) - \lambda_k v_k = \varphi\left(\sum_{i=1}^{k-1} \beta_i v_i\right) - \lambda_k \sum_{i=1}^{k-1} \beta_i v_i = \sum_{i=1}^{k-1} \beta_i \varphi(v_i) - \sum_{i=1}^{k-1} \lambda_k \beta_i v_i \\ = \sum_{i=1}^{k-1} \lambda_i \beta_i v_i - \sum_{i=1}^{k-1} \lambda_k \beta_i v_i = \sum_{i=1}^{k-1} (\lambda_i - \lambda_k) \beta_i v_i\,. \end{aligned}$$

Since, by assumption, $v_1, \ldots, v_{k-1}$ are linearly independent, we conclude that $(\lambda_i - \lambda_k)\beta_i = 0$ for each $i \in \{1, \ldots, k-1\}$. Since the scalars $\lambda_1, \ldots, \lambda_k$ are distinct, that implies $\beta_i = 0$ for each $i \in \{1, \ldots, k-1\}$. Then $v_k = 0$, which contradicts that v_k is nonzero by definition IV.1.1.1 as an eigenvector.

Our additional assumption is therefore false, and $v_1, \ldots, v_k$ are linearly independent. This conclusion implies that $v_1, \ldots, v_r$ are linearly independent since we considered an arbitrary index $k \in \{2, \ldots, r\}$.

§ IV.1.2. Invariant subspaes of linear transformations

Notes missing, see from 1:10:55 of Lecture 10 to 0:32:30 of Lecture 11 (invariant subspaces, diagonal matrix and the decomposition of the space into one-dimensional invariant subspaces)

§ IV.1.3. Eigenpairs and eigenspaces of matrices

Definition IV.1.3.1. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Consider $A \in \mathbb{F}^{n\times n}$ and the corresponding transformation $\varphi_A \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^n)$ given by definition II.5.10.3. Then the eigenpairs, eigenvalues, eigenvectors and eigenspaces of φ_A in the sense of definitions IV.1.1.1 and IV.1.1.4, are called eigenpairs, eigenvalues, eigenvectors and eigenspaces of the matrix A.

Definition IV.1.3.2. Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$ and I be the identity matrix of order n. Then the function $\chi\colon \mathbb{F} \to \mathbb{F}$ given by

$$\chi(\lambda) = \det(A - \lambda I) \quad \text{for all} \quad \lambda \in \mathbb{F}$$

is called the *characteristic polynomial of* A.

Proposition IV.1.3.3. In the context of definition IV.1.3.2, χ is an algebraic polynomial of degree n with the leading coefficient $(-1)^n$.

Proof. Let I and $e_1, \ldots, e_n$ denote the identity matrix of order n and its columns. By definition IV.1.3.2, the characteristic polynomial $\chi[A]$ of a matrix $A \in \mathbb{F}^{n\times n}$ is given by

$$\chi[A](\lambda) = \det(A - \lambda I)\,,$$

where the matrix determinant det of order n is a determinant function by theorem III.4.4.3.

Let us prove the statement of the proposition by induction with respect to a parameter $k \in \{0, 1, \ldots, n\}$, for a *fixed* order n. To this end, for each $k \in \{0, 1, \ldots, n\}$, let us introduce

$$I_k = \sum_{i=1}^{k} e_i e_i^{\mathsf{T}} \in \mathbb{F}^{n\times n}$$

and define $\chi_k[A]\colon \mathbb{F} \to \mathbb{F}$ by setting

$$\chi_k[A](\lambda) = \det(A - \lambda I_k) \quad \text{for each} \quad \lambda \in \mathbb{F}$$

for any $A \in \mathbb{F}^{n\times n}$. Clearly, we have that $I_n = I$ and that I_0 is the zero matrix of size $n \times n$. Then we have $\chi_n[A] = \chi[A]$ for every $A \in \mathbb{F}^{n\times n}$. We will prove by induction with respect to k that $\chi_k[A]$ with $k \in \{1, \ldots, n\}$ is an algebraic polynomial of degree k. In the particular case of $k = n$, this result is equivalent to the claim of the proposition regarding $\chi_k[A]$ with $A \in \mathbb{F}^{n\times n}$.

Consider $A \in \mathbb{F}^{n\times n}$. For each $k \in \{1, \ldots, n\}$, using the multilinearity of matrix determinant with respect to the columns of its argument (condition (a) of definition III.4.2.1), we obtain the following:

$$\begin{aligned}\chi_k[A](\lambda) = \det(A - \lambda I_k) &= \det(A - \lambda I_{k-1} - \lambda e_k e_k^{\mathsf{T}}) \\ &= \det(A - \lambda I_{k-1}) - \lambda \det(A - A e_k e_k^{\mathsf{T}} + e_k e_k^{\mathsf{T}} - \lambda I_{k-1})\end{aligned}$$

$$= \chi_{k-1}[A](\lambda) - \lambda \chi_{k-1}\Big[A - Ae_k e_k^{\mathsf{T}} + e_k e_k^{\mathsf{T}}\Big](\lambda) \tag{IV.1.3.1}$$

for every $\lambda \in \mathbb{F}$. The middle step consists in applying the linearity property of the matrix determinant of order n with respect to column k of its argument and relies upon the fact that, for every $\lambda \in \mathbb{F}$,

$$A - Ae_k e_k^{\mathsf{T}} + e_k e_k^{\mathsf{T}} - \lambda I_{k-1}$$

is obtained from $A - \lambda I_k$ by replacing its column k, which is equal to $(A - \lambda I_k)e_k = Ae_k$, with e_k.

Clearly, $\chi_0[A]$ is an algebraic polynomial of degree zero for any $A \in \mathbb{F}^{n\times n}$. For every $k \in \{1, \dots, n\}$, assuming that $\chi_{k-1}[A]$ is an algebraic polynomial of degree $k-1$ for any $A \in \mathbb{F}^{n\times n}$, we obtain by (IV.1.3.1) that $\chi_k[A]$ is an algebraic polynomial of degree k for any $A \in \mathbb{F}^{n\times n}$. We therefore conclude that $\chi_n[A] = \chi[A]$ is an algebraic polynomial of degree n for any $A \in \mathbb{F}^{n\times n}$.

Let us work out the leading coefficient of $\chi[A]$. Denoting the leading coefficient of $\chi_k[A]$ by $c_k[A]$ for all $k \in \{0, 1, \dots, n\}$ and $A \in \mathbb{F}^{n\times n}$, we obtain from (IV.1.3.1) that

$$c_k\Big[A - \sum_{i=k+1}^{n} Ae_i e_i^{\mathsf{T}} + \sum_{i=k+1}^{n} e_i e_i^{\mathsf{T}}\Big] = (-1)\, c_{k-1}\Big[A - \sum_{i=k}^{n} Ae_i e_i^{\mathsf{T}} + \sum_{i=k}^{n} e_i e_i^{\mathsf{T}}\Big] \tag{IV.1.3.2}$$

for all $k \in \{1, \dots, n\}$ and $A \in \mathbb{F}^{n\times n}$. For each $k \in \{0, \dots, n\}$ and every $A \in \mathbb{F}^{n\times n}$, the matrix

$$A - \sum_{i=k+1}^{n} Ae_i e_i^{\mathsf{T}} + \sum_{i=k+1}^{n} e_i e_i^{\mathsf{T}} = A - \sum_{i=k+1}^{n} (A - I)e_i e_i^{\mathsf{T}} = AI_k + (I - I_k)$$

is obtained from A by replacing its columns $k+1, \dots, n$ with $e_{k+1}, \dots, e_n$. So iterating the above relation (IV.1.3.2) for $k = n, n-1, \dots, 1$ yields

$$c_n[A] = (-1)^n c_0[I] = (-1)^n$$

for every $A \in \mathbb{F}^{n\times n}$ since $\chi_0[I](\lambda) = \det I = 1$ for any $\lambda \in \mathbb{F}$.

Proposition IV.1.3.4. Consider a vector space V of dimension $n \in \mathbb{N}$ over a field $\mathbb{F}$ with a basis $v_1, \dots, v_n$ Let $I \in \mathbb{F}^{n\times n}$ be the identity matrix of order n over $\mathbb{F}$. Consider $\varphi \in \mathcal{L}(V, V)$ and the matrix $A \in \mathbb{F}^{n\times n}$ of φ with respect to the bases $v_1, \dots, v_n$ and $v_1, \dots, v_n$. Let χ denote the characteristic polynomial of the mapping φ and of the matrix A. Then the following conditions are equivalent for any $\lambda \in \mathbb{F}$:

(i) λ is an eigenvalue of the mapping φ;

(ii) λ is an eigenvalue of the matrix A;

(iii) $\varphi - \lambda\mathsf{id}$ is not injective;

(iv) $\varphi - \lambda\mathsf{id}$ is not surjective;

(v) $\operatorname{Ker}(\varphi - \lambda\mathsf{id}) \neq \{0\}$;

(vi) $\operatorname{Im}(\varphi - \lambda\mathsf{id}) \neq V$;

(vii) $\operatorname{Ker}(A - \lambda I) \neq \{0\}$;

(viii) $\operatorname{Im}(A - \lambda I) \neq \mathbb{F}^n$;

(ix) $\operatorname{rank}(A - \lambda I) < n$;

(x) $\det(A - \lambda I) = 0$;

(xi) $\chi(\lambda) = 0$.

Proposition IV.1.3.4 allows, in particular, finding the set of all eigenvalues of a matrix as the set of roots of its characteristic polynomial.

> Notes missing, see from 0:13:40 of Lecture 12 to 0:57:00 of Lecture 12 (an example of diagonalization, over $\mathbb{C}$, in three dimensions)

§ IV.1.4. Algebraic properties of characteristic polynomials

Definition IV.1.4.1. Let $n \in \mathbb{N}$ and $a_0, \dots, a_n \in \mathbb{C}$. Assume that $a_n \neq 0$. Consider the algebraic polynomial $p\colon \mathbb{F} \to \mathbb{F}$ given by

$$p(t) = a_0 + \sum_{k=1}^{n} a_k t^k \quad \text{for all} \quad t \in \mathbb{F}\,.$$

Let $\lambda \in \mathbb{F}$ and $m \in \mathbb{N}$. Assume that $b_0, \dots, b_{n-m} \in \mathbb{F}$ are such that the algebraic polynomial $q\colon \mathbb{F} \to \mathbb{F}$ given by

$$q(t) = b_0 + \sum_{k=1}^{n-m} b_k t^k \quad \text{for all} \quad t \in \mathbb{F}$$

satisfies $q(\lambda) \neq 0$ and

$$p(t) = (t - \lambda)^m \, q(t) \quad \text{for all} \quad t \in \mathbb{F}\,.$$

Then λ is called a *root* of p of *multiplicity* m.

Definition IV.1.4.2. Let $\mathbb{F}$ be a field, $n \in \mathbb{N}$, $A \in \mathbb{F}^{n \times n}$, χ be the characteristic polynomial of A and $\lambda \in \mathbb{F}$ be an eigenvalue of A. Then the multiplicity of the root λ of χ is called the *algebraic multiplicity of the eigenvalue* λ *of* A and $\dim \operatorname{Ker}(A - \lambda I)$ is called the *geometric multiplicity of the eigenvalue* λ *of* A.

> Notes missing, see from 0:57:00 of Lecture 12 to 1:08:15 of Lecture 12 (roots of polynomials and their multiplicity, algebraically complete (closed) fields

In the case of $\mathbb{F} = \mathbb{C}$, the following important result guarantees that the characteristic polynomial of a matrix $A \in \mathbb{C}^{n \times n}$ with $n \in \mathbb{N}$ has exactly n roots counted with multiplicity. The proof of this theorem is beyond the scope of the present course.

Theorem IV.1.4.3 (fundamental theorem of algebra)**.** Let $n \in \mathbb{N}$ and $a_0, \dots, a_n \in \mathbb{C}$. Assume that $a_n \neq 0$. Then the algebraic polynomial $p\colon \mathbb{C} \to \mathbb{C}$ given by

$$p(t) = a_0 + \sum_{k=1}^{n} a_k t^k \quad \text{for all} \quad t \in \mathbb{C}$$

has a root in $\mathbb{C}$.

Corollary IV.1.4.4. Let $n \in \mathbb{N}$ and $a_0, \dots, a_n \in \mathbb{C}$. Assume that $a_n \neq 0$. Consider the algebraic polynomial $p\colon \mathbb{C} \to \mathbb{C}$ given by

$$p(t) = a_0 + \sum_{k=1}^{n} a_k t^k \quad \text{for all} \quad t \in \mathbb{C}\,.$$

Then there exist $r \in \{0, \dots, n\}$, distinct $\lambda_1, \dots \lambda_r \in \mathbb{C}$ and $m_1, \dots, m_r \in \{1, \dots, n\}$ such that $m_1 + \dots + m_r = n$ and

$$p(t) = a_n\, (t - \lambda_1)^{m_1} \cdots (t - \lambda_r)^{m_r} \quad \text{for all} \quad t \in \mathbb{C}\,.$$

§ IV.1.5. Algebra meets geometry: the algebraic and geometric multiplicity of an eigenvalue, diagonalizable and defective transformations

Notes missing, see from 1:15:15 of Lecture 12 to 0:08:30 of Lecture 13 (the algebraic and geometric multiplicity of an eigenvalue, algebra meets geometry, remarks on the lecture proof of proposition IV.1.3.3)

Notes missing, see from 0:08:30 of Lecture 13 to 0:28:15 of Lecture 13 (defective and diagonalizable transformation, diagonalization, the diagonalization of a matrix by a similarity transformation)

Definition IV.1.5.1. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Two matrices $A, B \in \mathbb{F}^{n \times n}$ are called *similar* if there exists an invertible matrix $V \in \mathbb{F}^{n \times n}$ such that $A = VBV^{-1}$.

Definition IV.1.5.2. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. A matrix $A \in \mathbb{F}^{n \times n}$ is called *diagonalizable* if it is similar to a diagonal matrix.

Proposition IV.1.5.3 (diagonalization and eigenvalue decomposition)**.** Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Consider a matrix $A \in \mathbb{F}^{n \times n}$, a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1, \dots, \lambda_n \in \mathbb{F}$ and and $V = [v_1 \ \cdots v_n] \in \mathbb{F}^{n \times n}$ with $v_1, \dots, v_n \in \mathbb{F}^n$. Assume that V is invertible. Then $A = V\Lambda V^{-1}$ if and only if $\lambda_1, \dots, \lambda_n$ are eigenvalues of A and $v_1, \dots, v_n$ are corresponding eigenvectors of A.

Proof. The equality $A = V\Lambda V^{-1}$ implies $AV = V\Lambda$. Considering the latter equality columnwise, we obtain $Av_k = \lambda_k v_k$ for each $k \in \{1, \dots, n\}$. Clearly, $v_1, \dots, v_n$ are all nonzero since V is invertible. Applying definitions IV.1.3.1 and IV.1.1.1, we conclude that $\lambda_1, \dots, \lambda_n$ are eigenvalues of A and $v_1, \dots, v_n$ are eigenvectors of A corresponding to these eigenvalues.

Conversely, if $\lambda_1, \dots, \lambda_n$ are eigenvalues of A and $v_1, \dots, v_n$ are eigenvectors of A corresponding to these eigenvalues, definitions IV.1.3.1 and IV.1.1.1 yield $Av_k = \lambda_k v_k$ for each $k \in \{1, \dots, n\}$. Rewriting this in a matrix form, we obtain $AV = V\Lambda$. Since the matrix V is invertible, this implies $A = V\Lambda V^{-1}$.

Notes missing, see from 0:34:00 of Lecture 13 to 0:56:05 of Lecture 13 (example: the infinite-dimensional setting, the existence of eigenvalues under restriction to a finite-dimensional invariant subspace)

Notes missing, see from 0:56:05 of Lecture 13 to 1:09:15 of Lecture 13 (summary: the three ways in which a linear transformation can fail to be diagonalizable)

§ IV.2. Generalized eigenpairs and eigenspaces of linear transformations and matrices

§ IV.2.1. Generalized eigenpairs and eigenspaces of a transformations map

Notes missing, see from 1:09:15 of Lecture 13 to the end of Lecture 13 (generalized eigenspaces and their properties)

Notes missing, see Lecture 15 (nilpotent transformations: basic results)

Chapter V. Norms and orthogonality

§ V.1. Norms and normed spaces

§ V.1.1. Norms on vector spaces

Notes missing, see from 0:00:00 of Lecture 16 to the end of Lecture 17

§ V.1.2. Operator norms. Matrix norms

Notes missing, see Lecture 18

§ V.2. Orthogonality, inner products and inner-product spaces

Throughout § V.2, we consider the case of $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. In the former case, complex conjugation is to be ignored since it does not affect real scalars.

§ V.2.1. Inner product

Definition V.2.1.1 (inner products and inner-product spaces)**.** Let V be a vector space over the field $\mathbb{F}$ and a function $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{F}$ satisfy the following conditions.

(a) Nonnegativity: $\langle v, v \rangle \in \mathbb{R}$ and $\langle v, v \rangle \geq 0$ for every $v \in V$.

(b) Nondegeneracy: for every $v \in V$, the condition $\langle v, v \rangle = 0$ implies $v = 0$.

(c) Conjugate symmetry (when $\mathbb{F} = \mathbb{C}$) or symmetry (when $\mathbb{F} = \mathbb{R}$): $\langle v, u \rangle = \overline{\langle u, v \rangle}$ for all $u, v \in V$.

(d) Linearity with respect to the second argument: $\langle u, \alpha v + w \rangle = \alpha \langle u, v \rangle + \langle u, w \rangle$ for all $u, v, w \in V$ and $\alpha \in \mathbb{F}$.

Then the function $\langle \cdot, \cdot \rangle$ is called an *inner product* and the vector space V is called an *inner-product space.* In particular, the vector space V is called a *Euclidean space* in the case of $\mathbb{F} = \mathbb{R}$ and a *unitary space* in the case of $\mathbb{F} = \mathbb{C}$.

Proposition V.2.1.2 (conjugate linearity of inner products with respect to the first argument)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle \cdot, \cdot \rangle$. Then $\langle \cdot, \cdot \rangle$ is linear (when $\mathbb{F} = \mathbb{R}$) or conjugate linear (when $\mathbb{F} = \mathbb{C}$) with respect to the first argument: $\langle \alpha v + w, u \rangle = \overline{\alpha} \langle v, u \rangle + \langle w, u \rangle$ for all $u, v, w \in V$ and $\alpha \in \mathbb{F}$.

Proof. Using the conjugate symmetry of inner products (part (c) of definition V.2.1.1) and their linearity with respect to the second argument (part (d) of definition V.2.1.1), we obtain

$$\langle \alpha v + w, u \rangle = \overline{\langle u, \alpha v + w \rangle} = \overline{\alpha \langle u, v \rangle + \langle u, w \rangle} = \overline{\alpha} \langle v, u \rangle + \langle w, u \rangle$$

for all $u, v, w \in V$ and $\alpha \in \mathbb{F}$.

Example V.2.1.3 ($\mathbb{R}^n$ as a Euclidean space)**.** Consider the vector space $V = \mathbb{R}^n$ with $n \in \mathbb{N}$ over the field $\mathbb{R}$. The function $\langle \cdot, \cdot \rangle \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ given by

$$\langle u, v \rangle = \sum_{i=1}^{n} u_i v_i \quad \text{for all} \quad u, v \in \mathbb{R}^n \tag{V.2.1.1}$$

is an inner product on $\mathbb{R}^n$, and $\mathbb{R}^n$ is therefore a Euclidean space. This inner product is ofter referred to as the *standard inner product of* $\mathbb{R}^n$.

Proof. The proof is left to the reader as an exercise.

Example V.2.1.4 ($\mathbb{C}^n$ as a unitary space)**.** Consider the vector space $V = \mathbb{C}^n$ with $n \in \mathbb{N}$ over the field $\mathbb{C}$. The function $\langle \cdot, \cdot \rangle \colon \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ given by

$$\langle u, v \rangle = \sum_{i=1}^{n} \overline{u_i} v_i \quad \text{for all} \quad u, v \in \mathbb{C}^n \tag{V.2.1.2}$$

is an inner product on $\mathbb{C}^n$, and $\mathbb{C}^n$ is therefore a unitary space. This inner product is ofter referred to as the *standard inner product of* $\mathbb{C}^n$.

Proof. The proof is left to the reader as an exercise.

Example V.2.1.5 ($\mathbb{C}^n$ as a Euclidean space)**.** Consider the vector space $V = \mathbb{C}^n$ with $n \in \mathbb{N}$ over the field $\mathbb{R}$, see example II.2.2.5. The function $\langle\cdot,\cdot\rangle\colon \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{R}$ given by

$$\langle u, v\rangle = \sum_{i=1}^{n} (\operatorname{Re} u_i \operatorname{Re} v_i + \operatorname{Im} u_i \operatorname{Im} v_i) \quad \text{for all} \quad u, v \in \mathbb{C}^n$$

is an inner product on $\mathbb{C}^n$, and $\mathbb{C}^n$ as a real vector space is therefore a Euclidean space.

Proof. The proof is left to the reader as an exercise.

Definition V.2.1.6 (orthogonality of vectors)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. and $u, v \in V$. If $\langle u, v\rangle = 0$, then we say that *the vector u is orthogonal to the vector v with respect to the inner product $\langle\cdot,\cdot\rangle$*. The indication of a particular inner product is often omitted when it is clear from the context what inner product is meant.

Remark V.2.1.7 (the orthogonality relation is reflexive)**.** In the context of definition V.2.1.6, due to the conjugate symmetry of inner products, u is orthogonal to v if and only if v is orthogonal to u.

Definition V.2.1.8 (induced norm)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. The function $\|\cdot\|\colon V \to \mathbb{R}$ given by

$$\|v\| = \sqrt{\langle v, v\rangle} \quad \text{for each} \quad v \in V$$

is called *the norm induced by the inner product* $\langle\cdot,\cdot\rangle$, or simply *the induced norm* when it is clear from the context what inner product is meant.

Remark V.2.1.9. In the context of definition V.2.1.8, the induced norm is correctly defined as stated due to the nonnegativity property of inner products (part (a) of definition V.2.1.1). We, however, do not claim at this moment that an induced norm is a norm. We will prove this fact later, in lemma V.2.1.13.

Theorem V.2.1.10 (Pythagorean theorem)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and the corresponding induced norm $\|\cdot\|$. Let $u, v \in V$ be such that $\langle u, v\rangle = 0$. Then

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

Proof. Using definition V.2.1.8, part (d) of definition V.2.1.1 and proposition V.2.1.2, we obtain

$$\|u + v\|^2 = \langle u + v, u + v\rangle = \langle u, u\rangle + \langle v, u\rangle + \langle u, v\rangle + \langle v, v\rangle = \|u\|^2 + 0 + 0 + \|v\|^2 = \|u\|^2 + \|v\|^2 .$$

Theorem V.2.1.11 (Cauchy–Bunyakovsky–Schwarz inequality)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and the corresponding induced norm $\|\cdot\|$ and $u, v \in V$. Then

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

and equality is attained if and only u and v are linearly dependent.

Proof. In the case when any of u and v is zero, both the sides of the inequality are zero, and it trivially holds. Also, u and v are linearly dependent in this case.

Let us assume for the remainder of the proof that both u and v are nonzero. Then, by the nondegeneracy property of inner products, we have $\langle v, v\rangle \neq 0$. So we can consider

$$\alpha = \frac{\langle v, u\rangle}{\langle v, v\rangle} \in \mathbb{F} , \quad \xi = \alpha v \in V \quad \text{and} \quad \eta = u - \xi \in V .$$

Let us establish that η is orthogonal to ξ. Indeed,

$$\langle \xi, \eta\rangle = \langle \alpha v, u - \alpha v\rangle = \overline{\alpha} \langle v, u\rangle - \alpha\overline{\alpha} \langle v, v\rangle = \alpha\overline{\alpha} \langle v, v\rangle - \alpha\overline{\alpha} \langle v, v\rangle = 0 .$$

Applying theorem V.2.1.10, we obtain $\|u\|^2 = \|\xi\|^2 + \|\eta\|^2$, which in the present setting, due to part (d) of definition V.2.1.1 and proposition V.2.1.2, means

$$\|u\|^2 = \|\xi\|^2 + \|\eta\|^2 = \langle \alpha v, \alpha v\rangle + \langle \eta, \eta\rangle = \alpha\overline{\alpha} \langle v, v\rangle + \langle \eta, \eta\rangle = \frac{|\langle v, u\rangle|^2}{\|v\|^4} \|v\|^2 + \langle \eta, \eta\rangle .$$

By part (a) of definition V.2.1.1, this implies

$$\|u\|^2 \geq \frac{|\langle v, u\rangle|^2}{\|v\|^2} ,$$

where equality is attained if and only if $\langle \eta, \eta\rangle = 0$, which, by part (b) of definition V.2.1.1, holds if and only if $\eta = 0$. The condition $\eta = 0$ is equivalent to that $u = \xi$, i.e., that u is a multiple of v. Since u is nonzero, the latter condition is equivalent to that u and v are linearly dependent.

Definition V.2.1.12. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$, $n \in \mathbb{N}$ and $v_1, \ldots, v_n \in V$.

(a) The vectors $v_1, \ldots, v_n \in V$ are called *orthogonal with respect to the inner product* $\langle\cdot,\cdot\rangle$ if they are all nonzero and $\langle v_i, v_j\rangle = 0$ for any $i, j \in \{1, \ldots, n\}$ such that $i \neq j$.

(b) The vectors $v_1, \ldots, v_n \in V$ are called *orthonormal with respect to the inner product* $\langle\cdot,\cdot\rangle$ if $\langle v_i, v_j\rangle = \delta_{ij}$ for any $i, j \in \{1, \ldots, n\}$.

The indication of a particular inner product is often omitted when it is clear from the context what inner product is meant.

The only defining property of norms that does not follow for induced norms from the defining properties of inner products is the triangle inequality. We can obtain it using theorem V.2.1.11.

Lemma V.2.1.13. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and the corresponding induced norm $\|\cdot\|$. Then $\|\cdot\|$ is a norm on V.

Definition V.2.1.14 (orthogonal complement)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and U be a subspace of V. The set of all vectors from V that are orthogonal to all vectors from U with respect to the inner product $\langle\cdot,\cdot\rangle$ is called the *orthogonal complement of U in V with respect to the inner product $\langle\cdot,\cdot\rangle$* and is denoted by $U^{\perp}$:

$$U^{\perp} = \{v \in V \colon \langle v, u\rangle = 0 \quad \text{for each} \quad u \in U\}\,.$$

Proposition V.2.1.15 (an orthogonal complement is a subspace)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and U be a subspace of V. Then the orthogonal complement $U^{\perp}$ of U in V with respect to the inner product $\langle\cdot,\cdot\rangle$ is a subspace of V.

Proof. The proof is left to the reader as an exercise.

Proposition V.2.1.16 (orthogonal complements and reflexivity)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and U be a subspace of V. Let W be the orthogonal complement $U^{\perp}$ of U in V with respect to the inner product $\langle\cdot,\cdot\rangle$. Then the orthogonal compleent $W^{\perp}$ of W in V with respect to the inner product $\langle\cdot,\cdot\rangle$ is equal to U: $(U^{\perp})^{\perp} = U$

Proof. The proof is left to the reader as an exercise.

§ V.2.2. Orthogonal bases

Lemma V.2.2.1. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the norm induced by the inner product $\langle\cdot,\cdot\rangle$. For $n \in \mathbb{N}$, consider vectors $v_1, \dots, v_n \in V$ orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$. Let $\alpha_1, \dots, \alpha_n \in \mathbb{F}$ and

$$v = \sum_{k=1}^{n} \alpha_k v_k\,.$$

Then

$$\alpha_k = \frac{\langle v_k, v\rangle}{\langle v_k, v_k\rangle}$$

for each $k \in \{1, \dots, n\}$ and

$$\|v\|^2 = \sum_{k=1}^{n} |\alpha_k|^2 \|v_k\|^2\,.$$

Proof. Consider $k \in \{1, \dots, n\}$. Using part (d) of definition V.2.1.1, we obtain

$$\langle v_k, v\rangle = \left\langle v_k, \sum_{j=1}^{n} \alpha_j v_j \right\rangle = \sum_{j=1}^{n} \alpha_j \langle v_k, v_j\rangle\,.$$

Since $v_1, \dots, v_n$ are orthogonal in the sense of definition V.2.1.12, we have $v_k \neq 0$ and $\langle v_k, v_j \rangle = \delta_{jk} \langle v_k, v_k \rangle$ for each $j \in \{1, \dots, n\}$, which gives

$$\langle v_k, v \rangle = \sum_{j=1}^{n} \alpha_j \delta_{jk} \langle v_k, v_k \rangle = \alpha_k \langle v_k, v_k \rangle \quad \text{and hence} \quad \alpha_k = \frac{\langle v_k, v \rangle}{\langle v_k, v_k \rangle}$$

by part (b) of definition V.2.1.1. This holds for every $k \in \{1, \dots, n\}$ and therefore proves the stated expressions for the coefficients of the arbitrary vector $v \in \operatorname{span}\{v_1, \dots, v_n\}$.

Similarly, for every $k \in \{2, \dots, n\}$, we can use part (d) of definition V.2.1.1, to arrive at

$$\left\langle \alpha_k v_k, \sum_{j=1}^{k-1} \alpha_j v_j \right\rangle = \overline{\alpha_k} \sum_{j=1}^{k-1} \alpha_j \langle v_k, v_j \rangle = 0$$

due to that $v_1, \dots, v_n$ are orthogonal in the sense of definition V.2.1.12. Then theorem V.2.1.10 and proposition V.2.1.2 and part (d) of definition V.2.1.1 yield

$$\begin{aligned} \left\| \sum_{j=1}^{k} \alpha_j v_j \right\|^2 = \left\| \sum_{j=1}^{k-1} \alpha_j v_j \right\|^2 + \langle \alpha_k v_k, \alpha_k v_k \rangle = \left\| \sum_{j=1}^{k-1} \alpha_j v_j \right\|^2 + \alpha_k \overline{\alpha_k} \, \|v_k\|^2 \\ = \left\| \sum_{j=1}^{k-1} \alpha_j v_j \right\|^2 + |\alpha_k|^2 \, \|v_k\|^2 . \end{aligned}$$

Applying this equality iteratively for $k = n, n-1, \dots, 2$, we obtain

$$\|v\|^2 = \sum_{k=1}^{n} |\alpha_k|^2 \|v_k\|^2 .$$

Corollary V.2.2.2. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the norm induced by the inner product $\langle \cdot, \cdot \rangle$. For $n \in \mathbb{N}$, consider vectors $v_1, \dots, v_n \in V$ orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle$. Then

$$v = \sum_{k=1}^{n} \frac{\langle v_k, v \rangle}{\langle v_k, v_k \rangle} v_k \quad \text{and} \quad \|v\|^2 = \sum_{k=1}^{n} \frac{|\langle v_k, v \rangle|^2}{\|v_k\|^2}$$

for every $v \in \operatorname{span}\{v_1, \dots, v_n\}$.

The following is an immediate consequence of definition III.2.3.1 and corollary V.2.2.2.

Corollary V.2.2.3. In the context of corollary V.2.2.2, the vectors $v_1, \dots, v_n$ are linearly independent.

Remark V.2.2.4. In the context of corollary V.2.2.2, let the vector space V be finite dimensional. Then definition III.2.2.1 implies $n \leq \dim V$.

Example V.2.2.5. Let $V = \mathbb{R}^m$ for $m \in \mathbb{N}$ and $\langle \cdot, \cdot \rangle$ be the standard inner product (V.2.1.1) on $\mathbb{R}^m$. Let $n \in \mathbb{N}$ and $v_1, \dots, v_n \in \mathbb{R}^m$ be column vectors orthonormal with respect to the inner product $\langle\cdot,\cdot\rangle$. Consider the matrix $Q = [v_1, \dots, v_n] \in \mathbb{R}^{m\times n}$ and a column vector

$$v = \sum_{k=1}^{n} \alpha_k v_k = Q\alpha$$

with $\alpha \in \mathbb{R}^n$. Then corollary V.2.2.2 (or immediately lemma V.2.2.1) gives $\alpha = Q^{\mathsf{T}} v$ and $v = Q\alpha = QQ^{\mathsf{T}} v$. Note: this equality implies $v \in \operatorname{Im} Q$, as we assume in this example!

§ V.2.3. Matrices and orthogonality

For the fields of real and complex numbers, we introduce Hermitian conjugation. This operation is different from transposition (definition II.3.3.1) only in that complex conjugation is additionally applied to the matrix entries.

Definition V.2.3.1 (conjugate transpose)**.** Let $m, n \in \mathbb{N}$, $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ and $A \in \mathbb{F}^{m\times n}$. The matrix $[\,\overline{A_{ji}}\,]_{j=1,\, i=1}^{n\quad m} \in \mathbb{F}^{n\times m}$ is called the *conjugate transpose* (*Hermitian adjoint*, *conjugate*) of A and is denoted by A^{H}. In other terms, we set

$$A^{\mathsf{H}} = [\,\overline{A_{ji}}\,]_{j=1,\, i=1}^{n\quad m} \in \mathbb{F}^{n\times m}.$$

The following results are direct analogues of propositions II.3.3.2, II.3.3.3 and II.4.1.11.

Proposition V.2.3.2 (reflexivity of conjugate transposition)**.** Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. The operation of conjugate transposition is reflexive: for any $m, n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$, we have $(A^{\mathsf{H}})^{\mathsf{H}} = A$.

Proof. The proof is left to the reader as an exercise.

Proposition V.2.3.3 (conjugate transposition under linear operations)**.** Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, $m, n \in \mathbb{N}$, $\alpha \in \mathbb{F}$ and $A, B \in \mathbb{F}^{m\times n}$. Then $(A + B)^{\mathsf{H}} = A^{\mathsf{H}} + B^{\mathsf{H}}$ and $(\alpha \cdot A)^{\mathsf{H}} = \overline{\alpha} \cdot A^{\mathsf{H}}$.

Proof. The proof is left to the reader as an exercise.

Proposition V.2.3.4 (conjugate transposition under matrix multiplication)**.** Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, $m, n, r \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times r}, B \in \mathbb{F}^{r\times n}$. Then $(AB)^{\mathsf{H}} = B^{\mathsf{H}} A^{\mathsf{H}}$.

Proof. The proof is left to the reader as an exercise.

Proposition V.2.3.5 (matrix inversion under conjugate transposition)**.** Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, $n \in \mathbb{N}$ and $A \in \mathbb{F}^{n\times n}$ be an invertible matrix. Then A^{H} is invertible and $(A^{\mathsf{H}})^{-1} = (A^{-1})^{\mathsf{H}}$.

Proof. The proof is left to the reader as an exercise.

Notes missing, see from 0:25:10 to 1:15:44 of Lecture 20

§ V.2.4. Orthogonal projections

Definition V.2.4.1 (orthogonal projection)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. For $n\in\mathbb{N}$, consider a subspace W of V of dimension $n\in\mathbb{N}$ and assume that vectors $v_1,\ldots,v_n\in V$ form a basis for W that is orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$. The mapping $\mathbf{\Pi}_{v_1,\ldots,v_n}\colon V\to V$ given by

$$\mathbf{\Pi}_{v_1,\ldots,v_n}(v)=\sum_{k=1}^{n}\frac{\langle v_k,v\rangle}{\langle v_k,v_k\rangle}\,v_k$$

for every $v\in V$ is called *the operator of orthogonal projection with respect to the basis $v_1,\ldots,v_n$ and the inner product $\langle\cdot,\cdot\rangle$*, or simply *the orthogonal projection with respect to the basis $v_1,\ldots,v_n$* when it is clear from the context what inner product is meant.

For every $v\in V$, the vector $\mathbf{\Pi}_{v_1,\ldots,v_n}(v)\in V$ is called *the orthogonal projection of v with respect to the basis $v_1,\ldots,v_n$ and the inner product $\langle\cdot,\cdot\rangle$*, or simply *the orthogonal projection of v with respect to the basis $v_1,\ldots,v_n$* when it is clear from the context what inner product is meant.

Lemma V.2.4.2 (properties of orthogonal projections)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. For $n\in\mathbb{N}$, consider a subspace W of V of dimension $n\in\mathbb{N}$ and assume that vectors $v_1,\ldots,v_n\in V$ form a basis for W that is orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$. The operator $\mathbf{\Pi}_{v_1,\ldots,v_n}$ of orthogonal projection with respect to the basis $v_1,\ldots,v_n$ and the inner product $\langle\cdot,\cdot\rangle$, given by definition V.2.4.1, is linear and satisfies the following:

(a) $\mathbf{\Pi}^2_{v_1,\ldots,v_n}=\mathbf{\Pi}_{v_1,\ldots,v_n}$;

(b) $\langle w,v-\mathbf{\Pi}_{v_1,\ldots,v_n}v\rangle=0$ for all $v\in V$ and $w\in W$;

(c) $\operatorname{Im}\mathbf{\Pi}_{v_1,\ldots,v_n}=\operatorname{span}\{v_1,\ldots,v_n\}$;

(d) $\operatorname{Ker}\mathbf{\Pi}_{v_1,\ldots,v_n}=\operatorname{span}\{v_1,\ldots,v_n\}^{\perp}$.

Proof. The proof is left to the reader as an exercise.

Lemma V.2.4.3 (orthogonal projections are invariant under basis transformations)**.** Consider a vector space V over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. For $n\in\mathbb{N}$, consider a subspace W of V of dimension $n\in\mathbb{N}$ and assume that vectors $v_1,\ldots,v_n\in V$ form a basis for W that is orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$. Assume that vectors $u_1,\ldots,u_n\in V$ also form a basis for W that is orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$. Then $\mathbf{\Pi}_{u_1,\ldots,u_n}=\mathbf{\Pi}_{v_1,\ldots,v_n}$ for every $v\in V$.

Proof. Using corollary V.2.2.2, we obtain

$$u_k = \sum_{j=1}^{n} \frac{\langle v_j, u_k\rangle}{\langle v_j, v_j\rangle} v_j \quad \text{and} \quad v_k = \sum_{j=1}^{n} \frac{\langle u_j, v_k\rangle}{\langle u_j, u_j\rangle} u_j$$

for each $k \in \{1, \dots, n\}$.

Consider $i, j \in \{1, \dots, n\}$. Together with the linearity and conjugate symmetry of the inner product, the orthogonality of $u_1, \dots, u_n$ gives

$$\langle v_i, v_j\rangle = \sum_{k=1}^{n} \frac{\overline{\langle u_k, v_i\rangle}}{\langle u_k, u_k\rangle} \sum_{\ell=1}^{n} \frac{\langle u_\ell, v_j\rangle}{\langle u_\ell, u_\ell\rangle} \langle u_k, u_\ell\rangle$$

$$= \sum_{k=1}^{n} \frac{\langle v_i, u_k\rangle}{\langle u_k, u_k\rangle} \sum_{\ell=1}^{n} \frac{\langle u_\ell, v_j\rangle}{\langle u_\ell, u_\ell\rangle} \delta_{k\ell} \langle u_k, u_k\rangle = \sum_{k=1}^{n} \frac{\langle v_i, u_k\rangle \langle u_k, v_j\rangle}{\langle u_k, u_k\rangle} .$$

On the other hand, the orthogonality of $v_1, \dots, v_n$ implies

$$\langle v_i, v_j\rangle = \delta_{ij} \langle v_i, v_i\rangle .$$

So we arrive at

$$\sum_{k=1}^{n} \frac{\langle v_i, u_k\rangle \langle u_k, v_j\rangle}{\langle u_k, u_k\rangle} = \delta_{ij} \langle v_i, v_i\rangle ,$$

which holds for any $i, j \in \{1, \dots, n\}$.

For every $v \in V$, using the linearity and conjugate symmetry of the inner product definition V.2.4.1 in the expressions given for $\mathbf{\Pi}_{v_1,\dots,v_n}$ and $\mathbf{\Pi}_{u_1,\dots,u_n}$ in definition V.2.4.1, we obtain

$$\mathbf{\Pi}_{u_1,\dots,u_n}(v) = \sum_{k=1}^{n} \frac{\langle u_k, v\rangle}{\langle u_k, u_k\rangle} u_k = \sum_{k=1}^{n} \left\{ \sum_{j=1}^{n} \frac{\overline{\langle v_j, u_k\rangle}}{\langle v_j, v_j\rangle} \frac{\langle v_j, v\rangle}{\langle u_k, u_k\rangle} \right\} \sum_{i=1}^{n} \frac{\langle v_i, u_k\rangle}{\langle v_i, v_i\rangle} v_i$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\langle u_k, v_j\rangle}{\langle v_j, v_j\rangle} \frac{\langle v_j, v\rangle}{\langle u_k, u_k\rangle} \frac{\langle v_i, u_k\rangle}{\langle v_i, v_i\rangle} v_i = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\langle v_j, v\rangle}{\langle v_i, v_i\rangle \langle v_j, v_j\rangle} v_i \sum_{k=1}^{n} \frac{\langle u_k, v_j\rangle \langle v_i, u_k\rangle}{\langle u_k, u_k\rangle}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\langle v_j, v\rangle}{\langle v_i, v_i\rangle \langle v_j, v_j\rangle} \delta_{ij} \langle v_i, v_i\rangle v_i = \sum_{j=1}^{n} \frac{\langle v_j, v\rangle}{\langle v_j, v_j\rangle} v_j = \mathbf{\Pi}_{v_1,\dots,v_n}(v) ,$$

which proves the claim.

Remark V.2.4.4 (orthogonal projections are associated with subspaces)**.** In the context of definition V.2.4.1, we use $v_1, \dots, v_n$ as an orthogonal basis for W. Clearly, due to the way the mapping $\mathbf{\Pi}_{v_1,\dots,v_n}$ is defined, it does not depend on the order of the basis vectors. Furthermore, as lemma V.2.4.3 shows, $\mathbf{\Pi}_{v_1,\dots,v_n}$ does not depend on the particular orthogonal basis of W used to define the projection.

§ V.2.5. Gram–Schmidt orthogonalization

Theorem V.2.5.1 (Gram–Schmidt orthogonalization process)**.** Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle \cdot, \cdot \rangle$. For $n \in \mathbb{N}$, let $n \in \mathbb{N}$ vectors $v_1, \dots, v_n \in V$ be linearly independent. Let vectors $u_1, \dots, u_n \in V$ be defined recursively as follows:

$$u_1 = \lambda_1 v_1 \quad \text{and} \quad u_k = \lambda_k (v_k - \mathbf{\Pi}_{u_1,\dots,u_{k-1}} v_k) \quad \text{for each} \quad k \in \{2, \dots, n\}$$

with nonzero $\lambda_1, \dots, \lambda_n \in \mathbb{F}$. Then, for each $k \in \{1, \dots, n\}$, the vectors $u_1, \dots, u_k$ form a basis for $\operatorname{span}\{v_1, \dots, v_k\}$ that is orthogonal with respect to the inner product $\langle \cdot, \cdot \rangle$.

Proof. From the assumption that the vectors $v_1, \dots, v_n$ are linearly independent, we conclude that they are all nonzero and that the vectors $v_1, \dots, v_k$ are linearly independent for every $k \in \{1, \dots, n\}$. In particular, we have $v_1 \neq 0$. For every $k \in \{1, \dots, n\}$, let us define $W_k = \operatorname{span}\{v_1, \dots, v_k\}$ and note that the vectors $v_1, \dots, v_k$ form a basis for W_k.

Since $\lambda_1 \neq 0$ by assumption, for $u_1 = \lambda_1 v_1$ we immediately obtain $u_1 \neq 0$ and $v_1 = \lambda^{-1} u_1$. Then the inclusion $W_1 \subseteq \operatorname{span}\{u_1\}$ holds as well as the inclusion $\operatorname{span}\{u_1\} \subseteq W_1$, and the two together yield $\operatorname{span}\{u_1\} = W_1$. Since $u_1 \neq 0$, this vector forms an orthogonal basis for W_1.

Let us consider $k \in \{2, \dots, n\}$, assume that the vectors $u_1, \dots, u_{k-1}$ form an orthogonal basis for W_{k-1} and define $u_k = \lambda_k (v_k - \mathbf{\Pi}_{u_1,\dots,u_{k-1}} v_k)$. Then part (c) of lemma V.2.4.2 implies that $\mathbf{\Pi}_{u_1,\dots,u_{k-1}} v_k \in \operatorname{span}\{u_1, \dots, u_{k-1}\} = W_{k-1}$, so that $u_k \in \operatorname{span}\{v_1, \dots, v_{k-1}, v_k\} = W_k$. Then $\operatorname{span}\{u_1, \dots, u_k\} \subseteq W_k$ by lemma III.2.1.5. On the other hand, since $\lambda_k \neq 0$, we can express v_k as follows: $v_k = \lambda_k^{-1} u_k + \mathbf{\Pi}_{u_1,\dots,u_{k-1}} v_k$. This shows that $v_k \in \operatorname{span}\{u_1, \dots, u_k\}$. Then $W_k = \operatorname{span}\{v_1, \dots, v_k\} \subseteq \operatorname{span}\{u_1, \dots, u_k\}$. Combining the two inclusions, we obtain $\operatorname{span}\{u_1, \dots, u_k\} = W_k$.

If $u_k = 0$ were the case, $v_k \in W_{k-1}$ would hold, contradicting with the linear independence of $v_1, \dots, v_n$. We therefore conclude that $u_k \neq 0$. For each $j \in \{1, \dots, k-1\}$, part (b) of lemma V.2.4.2 yields

$$\langle u_j, v_k - \mathbf{\Pi}_{u_1,\dots,u_{k-1}} v_k \rangle = 0$$

since $u_j \in W_{k-1}$ and $u_1, \dots, u_{k-1}$ form an orthogonal basis for W_{k-1}. Then, since $u_1, \dots, u_{k-1}$ are orthogonal and u_k is nonzero and orthogonal to each of them, the vectors $u_1, \dots, u_k$ are orthogonal. By lemma V.2.4.2, $u_1, \dots, u_k$ are linearly independent; they therefore form an orthogonal basis for $\operatorname{span}\{u_1, \dots, u_k\} = W_k$.

Applying the above argument iteratively, we obtain the claim.

Corollary V.2.5.2. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle \cdot, \cdot \rangle$. Let W be a subspace of V of dimension $n \in \mathbb{N}$. Then W has an orthogonal basis.

Proof. Since $\dim W = n$, the definition of the dimension of a vector space yields that there exists a basis $v_1, \dots, v_n$ of W. Applying theorem V.2.5.1, we obtain that there exists an orthogonal basis $u_1, \dots, u_n$.

Definition V.2.5.3. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle \cdot, \cdot \rangle$. For $n \in \mathbb{N}$, consider a subspace W of V of dimension $n \in \mathbb{N}$. Let $\mathbf{\Pi}_W \colon V \to V$ given by $\mathbf{\Pi}_W = \mathbf{\Pi}_{v_1,\dots,v_n}$, where $v_1, \dots, v_n$ is an orthogonal basis for W and $\mathbf{\Pi}_{v_1,\dots,v_n} \colon V \to V$ is given by definition V.2.4.1. Then $\mathbf{\Pi}_W$ is called *the operator of orthogonal projection onto W with respect to the inner product $\langle \cdot, \cdot \rangle$*, or simply *the orthogonal projection onto W* when it is clear from the context what inner product is meant.

For every $v \in V$, the vector $\mathbf{\Pi}_W(v) \in V$ is called *the orthogonal projection of v onto W with respect to the inner product $\langle\cdot,\cdot\rangle$*, or simply *the orthogonal projection of v onto W* when it is clear from the context what inner product is meant.

Proposition V.2.5.4. Let V be a vector space over the field $\mathbb{F}$ with an inner product $\langle\cdot,\cdot\rangle$. For $n \in \mathbb{N}$, consider a subspace W of V of dimension $n \in \mathbb{N}$. Then the operator of orthogonal projection onto W with respect to the inner product $\langle\cdot,\cdot\rangle$ is well defined.

Proof. As corollary V.2.5.2 shows, there exist an orthogonal basis $v_1,\ldots,v_n$ for W, so we can consider the corresponding mapping $\mathbf{\Pi}_{v_1,\ldots,v_n} \colon V \to V$ given by definition V.2.4.1. On the other hand, lemma V.2.4.3 shows that $\mathbf{\Pi}_{v_1,\ldots,v_n}$ does not depend on what orthogonal basis is used to introduce it in definition V.2.4.1. So definition V.2.5.3 defines a unique mapping $\mathbf{\Pi}_W \colon V \to V$ that depends only on V, $\langle\cdot,\cdot\rangle$ and W.

Remark V.2.5.5. The nonzero scaling factors $\lambda_1,\ldots,\lambda_n$ appearing in theorem V.2.5.1 allow, for example, to normalize the basis in the course of orthogonalization by setting

$$\tilde{u}_k = v_k - \mathbf{\Pi}_{u_1,\ldots,u_{k-1}} v_k \quad \text{and} \quad \lambda_k = \frac{1}{\sqrt{\langle \tilde{u}_k, \tilde{u}_k\rangle}} \quad \text{for each} \quad k \in \{1,\ldots,n\}\,.$$

This choice ensures that the resulting vectors $u_k = \lambda_k(v_k - \mathbf{\Pi}_{u_1,\ldots,u_{k-1}} v_k) = \lambda_k \tilde{u}_k$ with $k \in \{1,\ldots,n\}$ for an *orthonormal* basis for $\operatorname{span}\{v_1,\ldots,v_k\}$.

§ V.2.6. QR decomposition of matrices

Definition V.2.6.1. Let $m,n,r \in \mathbb{N}$ be such that $r \leq \min\{m,n\}$. Assume that $Q \in \mathbb{F}^{m\times r}$ is a matrix with columns orthonormal with respect to the standard inner product of $\mathbb{F}^m$: $Q^{\mathsf{H}}Q = I$, where I is the identity matrix of order r. Further, assume that $R \in \mathbb{F}^{r\times n}$ is an upper-triangular matrix. Let $A = QR$. Then this equality, as a representation of A in terms of Q and R, is called a *QR decomposition of A*.

Theorem V.2.6.2. Let $m,n \in \mathbb{N}$ and $A \in \mathbb{F}^{m\times n}$ be such that $\operatorname{rank} A = n$. Then A has a QR decomposition.

Proof. Consider the vector space $V = \mathbb{F}^m$ over the field $\mathbb{F}$ with the standard inner product $\langle\cdot,\cdot\rangle$: $\langle u,v\rangle = u^{\mathsf{H}}v$ for all $u,v \in \mathbb{F}^m$. Let us denote the columns of A by $a_1,\ldots,a_n$. Since $\operatorname{rank} A = n$, the column vectors $a_1,\ldots,a_n \in V$ are linearly independent.

Let us apply the Gram–Schmidt orthogonalization process (theorem V.2.5.1) to V with the inner product $\langle\cdot,\cdot\rangle$ and to the linearly independent vectors $a_1,\ldots,a_n \in V$ with the scaling factors given by

$$\lambda_k = \frac{1}{\sqrt{\langle \tilde{q}_k, \tilde{q}_k\rangle}} = \frac{1}{\|\tilde{q}_k\|_2} \quad \text{with} \quad \tilde{q}_k = a_k - \mathbf{\Pi}_{q_1,\ldots,q_{k-1}} a_k \quad \text{for each} \quad k \in \{1,\ldots,n\}\,.$$

This yields vectors $q_k = \lambda_k (a_k - \mathbf{\Pi}_{q_1,\ldots,q_{k-1}} a_k) = \lambda_k \tilde{q}_k$ with $k \in \{1, \ldots, n\}$ that satisfy the following properties. First, they are orthonormal with respect to the inner product $\langle \cdot, \cdot \rangle$, which means that $q_i^{\mathsf{H}} q_j = \delta_{ij}$ for all $i, j \in \{1, \ldots, n\}$. Second, for each $k \in \{1, \ldots, n\}$, the vectors $q_1, \ldots, q_k$ form a basis for $\operatorname{span}\{a_1, \ldots, a_k\}$, so that $a_k \in \operatorname{span}\{q_1, \ldots, q_k\}$.

Let us introduce $Q = [q_1, \ldots, q_n] \in \mathbb{F}^{m \times n}$ and $R = Q^{\mathsf{H}} A = [Q^{\mathsf{H}} a_1, \ldots, Q^{\mathsf{H}} a_n] \in \mathbb{F}^{n \times n}$. The orthonormality of $q_1, \ldots, q_n$ with respect to the particular inner product we use here means that $Q^{\mathsf{H}} Q$ is the identity matrix of order n. Further, for each $k \in \{1, \ldots, n\}$, since $a_k \in \operatorname{span}\{q_1, \ldots, q_n\}$, we have $Q Q^{\mathsf{H}} a_k = \mathbf{\Pi}_{q_1,\ldots,q_n} a_k = a_k$ (cf. example V.2.2.5). This immediately gives $QR = QQ^{\mathsf{H}} A = A$. On the other hand, by lemma V.2.2.1, the fact that $a_k \in \operatorname{span}\{q_1, \ldots, q_k\}$ for each $k \in \{1, \ldots, n\}$ implies that the matrix R is upper triangular. So $A = QR$ is a QR decomposition of A.

Remark V.2.6.3. In the context of theorem V.2.6.2, for each $k \in \{1, \ldots, n\}$, we have $R_{kk} = \lambda_k^{-1} = \|\tilde{q}_k\|_2^{-1}$ since

$$a_k = \lambda_k^{-1} q_k + \mathbf{\Pi}_{q_1,\ldots,q_{k-1}} a_k \,,$$

where $\mathbf{\Pi}_{q_1,\ldots,q_{k-1}} a_k \in \operatorname{span}\{q_1, \ldots, q_{k-1}\}$ by part (c) of lemma V.2.4.2.