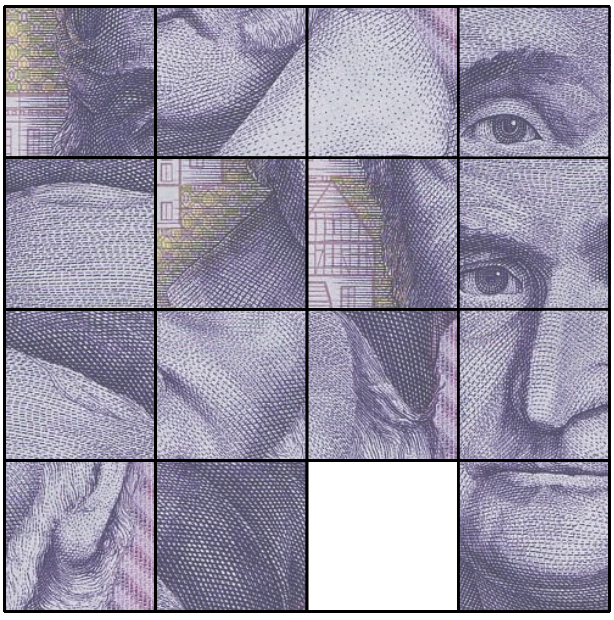
Linear Algebra

Benjamin Sambale
Leibniz Universität Hannover

Version: March 31, 2026



Contents

Introduction **5**
Preface 5
Motivation 5
Notation 6
Conventions 8

Linear Algebra I **11**

1 Propositional Logic and Set Theory **11**
1.1 Propositions 11
1.2 Sets 13
1.3 Mathematical Induction 16

2 Cartesian Products and Functions **17**
2.1 Pairs and Tuples 17
2.2 Injective and Surjective Functions 19

3 Fields and Vector Spaces **23**
3.1 Groups and Fields 23
3.2 Vector Spaces and Subspaces 25

4 Bases and Dimension **29**
4.1 Linear Independence and Generating Sets 29
4.2 Characterization and Existence of Bases 31
4.3 Dimension 33

5 Matrices **35**
5.1 The Matrix Vector Space 35
5.2 Matrix Multiplication 37
5.3 The Rank of a Matrix 39

6 The Gaussian Algorithm **41**
6.1 Systems of Equations 41
6.2 Elementary Row Operations 43
6.3 Applications 45

7 Linear Maps **50**
7.1 Definitions and Examples 50
7.2 Representation Matrices 54
7.3 Dual Spaces 59

8 Eigenvalues and Eigenvectors **63**
8.1 Definitions and Examples 63
8.2 Diagonalizability 64

9 Determinants **67**
9.1 Recursive Definition 67
9.2 Properties 70
9.3 Laplace Expansion 72
9.4 The Leibniz Formula 75

Exercises **80**

Linear Algebra II **86**

10 Polynomials **86**
10.1 The Vector Space of Polynomials 86
10.2 Roots 89
10.3 Characteristic Polynomials 91
10.4 Minimal Polynomials 95

11 Euclidean Geometry **99**
11.1 Scalar Products 99
11.2 Orthonormal Bases 102
11.3 Symmetric and orthogonal maps 104
11.4 Complex Numbers 107
11.5 The Principal Axis Theorem 110

12 Bilinear Forms **113**
12.1 Gram Matrices 113
12.2 Sylvester's Law of Inertia 116
12.3 Positive definite matrices 120

13 Unitary Spaces **125**
13.1 Sesquilinear Forms 125
13.2 Adjoint Maps 126
13.3 The Spectral Theorem 129

14 The Jordan Normal Form **134**
14.1 Generalized Eigenspaces 134
14.2 Jordan Blocks 138
14.3 Applications 142

15 The Frobenius Normal Form **146**
15.1 Irreducible Polynomials 146
15.2 Companion Matrices 149
15.3 Centralizers 154
15.4 Splitting Fields 156

16 The Jordan-Chevalley Decomposition **158**
16.1 The Chinese Remainder Theorem 158

16.2 Separable and semisimple maps . . . 161
16.3 Generalized Jordan Blocks . . . 162

Exercises **166**

Linear Algebra III 172

17 Numerical Methods **172**
17.1 Efficient Arithmetic . . . 172
17.2 The Condition Number . . . 176
17.3 Stable Variants of the Gaussian Elimination . . . 180
17.4 Iterative Methods . . . 185
17.5 Matrix Norms . . . 189
17.6 Eigenvalue Computation . . . 192
17.7 Orthonormalization . . . 197

18 Analytical Aspects **200**
18.1 Eigenvalue Estimates . . . 200
18.2 The Spectral Radius . . . 201
18.3 The Exponential Function of a Matrix . . . 203
18.4 Non-negative Matrices . . . 206
18.5 The PageRank . . . 211

19 Linear Optimization **213**
19.1 Linear Programs . . . 213
19.2 Convex Sets . . . 215
19.3 The Simplex Algorithm . . . 218

20 Lattices and Quadratic Forms **223**
20.1 Lattices . . . 223
20.2 The Minimal Norm . . . 225
20.3 Integer Matrices . . . 228
20.4 Theorems of Hermite and Minkowski . . . 232
20.5 The LLL Algorithm . . . 234
20.6 Quadratic Forms . . . 238
20.7 Successive Minima . . . 243

Exercises **245**

Appendix 251

Index **257**

Introduction

Warning: This is an AI-translated version of my German lectures notes, performed by *Gemini 3 Flash Preview*. I have not checked whether Gemini introduced errors. Use with care!

Preface

These notes are an extension of my lecture Linear Algebra A&B in the winter semester 2020/21 and summer semester 2021 at Leibniz University Hannover. While the lecture was primarily aimed at computer science students, the present notes are directed at mathematicians. The third part was added at the end of 2025 and deals mainly with practical applications. Some of the presented theorems (including those by Fillmore, Mirsky, Mazur-Ulam, Schur-Horn, and Frobenius) are difficult to find in the standard literature. Furthermore, the Frobenius normal form is treated in full generality. The exercises are limited to theoretical aspects and should be supplemented by practical examples. I thank Stefanos Aivazidis, Annika Bartelt, Gereon Koßmann and Claude Sonnet (4.6) for pointing out errors (further hints are welcome). The following book covers approximately the topics of Part I and II:

Hoffman, Kunze, *Linear algebra*, 2nd edition, Prentice-Hall, New Jersey, 1971

Motivation

You have obtained measurement data $d_1 = -2, d_2 = 3, \dots$ through physical experiments at various times $t_1 = 1, t_2 = 2, \dots$. From theoretical considerations, it is known that these data follow a law, that is, there is a function f with $f(t_i) = d_i$ for $i = 1, 2, \dots$. Here, f depends (linearly) on unknown parameters $x_1, x_2, \dots$, for example $f(t) = t^2x_1 - tx_2 + x_3$. Determining these parameters based on the measurement data leads to a system of linear equations:

$$\begin{aligned} x_1 - x_2 + x_3 &= -2 \\ 4x_1 - 2x_2 + x_3 &= 3 \\ &\vdots \end{aligned} \tag{S}$$

We answer, among others, the following questions:

- When is the system (S) solvable? (Theorem 6.4)
- How many solutions are there? (Remark 6.7(b))
- What structure does the solution set have? (Theorem 6.6)
- How does one calculate all solutions in practice? (Theorem 6.15)

The developed methods (vector spaces, matrices, and linear maps) have numerous applications in other fields:

- Image processing: How does one recognize faces in photos?
- Search engines: According to which criteria does Google rank internet pages? (section 18.5)
- Coding theory: How does one detect and correct errors in the transmission of digital data?
- Cryptography: How does one encrypt data resistant to attacks by quantum computers? (Remark 20.11)
- Electrical engineering: How does one calculate resistances in circuits?
- Meteorology: How does one predict tomorrow's weather?
- Stochastics: With what probability does one reach the goal after a random walk?
- Artificial intelligence: How are Large Language Models trained?

Notation

i. e.	id est (that is)
cf.	confer (compare)
wlog.	without loss of generality
wrt.	with respect to
$\mathbf{t}, \mathbf{f}$	true, false
$\wedge, \vee, \neg, \Rightarrow, \Leftrightarrow, \exists, \forall$	logical expressions
$:=$, $:\Leftrightarrow$	left side is defined by right side
$\square$	end of proof
$\varnothing$	empty set
$\mathcal{P}(M)$	power set of M
$\cup$, $\dot\cup$, $\cap$, $\setminus$	union (disjoint), intersection, difference of sets
$\lvert A\rvert$	cardinality of A (number of elements)
(a, b), $(x_1, \ldots, x_n)$	pair, n-tuple
$A_1 \times \ldots \times A_n$	Cartesian product of sets $A_1, \ldots, A_n$
$[a]$	equivalence class of $a \in A$
$A \to B$, $a \mapsto b$	map from A to B
$\mathrm{Fun}(A, B)$	set of all maps $A \to B$
$\mathbb{N}, \mathbb{N}_0, \mathbb{Z}$	natural numbers (without and with 0), integers
$\mathbb{Q}, \mathbb{R}, \mathbb{R}_{>0}, \mathbb{R}_{\geq 0}, \mathbb{C}$	rational, real (positive, non-negative), and complex numbers
$\mathbb{F}_2$	field with two elements
$K^\times$	$= K \setminus \{0\}$ (multiplicative group)
$f_{\lvert A}$, f^{-1}, $f \circ g$	restriction, inverse function, and composition of functions
f^*	dual or adjoint map to f
$f(A), f^{-1}(B), \mathrm{Ker}(f)$	image, preimage, kernel of f
id, id_V	identity (on V)
0_K, 0_V, 1_G	zero element, zero vector, identity element in G
δ_{ij}	Kronecker delta
$\sum_{i=1}^n \lambda_i v_i$	linear combination
$U \leq V$, $U < V$, V/U	subspace (proper), factor space
$H \leq G$, $H < G$	subgroup (proper) of G

$U \cong V$	isomorphic vector spaces
V^*, V^{**}	dual space, bidual space
U^0, U_0	dual complements
$U \oplus W$, $U \times W$	direct sum/product of U and W
$e_1, \ldots, e_n$	standard basis of K^n
$b_1^*, \ldots, b_n^*$	dual basis
$\langle S \rangle$, $\langle s_1, \ldots, s_n \rangle$	span of S
$\dim_K V = \dim V$	dimension of V over K
${}_B[v]$	coordinate representation of v wrt. the basis B
$\mathrm{Hom}(V, W)$	vector space of all linear maps $V \to W$
$\mathrm{End}(V)$	$= \mathrm{Hom}(V, V)$
$K^{n \times m}$	vector space of $n \times m$ matrices over K
$\mathrm{GL}(V)$, $\mathrm{GL}(n, K)$	general linear group
$\mathrm{SL}(n, K)$	special linear group
$\mathrm{O}(V)$, $\mathrm{O}(n, K)$	orthogonal group
$\mathrm{SO}(n, K)$	special orthogonal group
$\mathrm{U}(V)$, $\mathrm{U}(n, \mathbb{C})$	unitary group
$\mathrm{SU}(n, \mathbb{C})$	special unitary group
$\mathrm{Aff}(V)$	affine group
$0_{n \times m}$, 0_n, 1_n	zero matrix, identity matrix
E_{st}	standard matrix with 1 at position (s, t)
$A \sim B$	A row-equivalent to B
$A \approx B$	A similar to B
$(A \mid b)$	augmented coefficient matrix
A_{st}	$= (a_{ij} : i \neq s, j \neq t)$ (deletion of row/column)
A^{-1}, A^+	inverse and pseudoinverse of A
A^{t}, $A^{-\mathrm{t}}$	transpose and transpose-inverse of A
$\widehat{A}$, $\widetilde{A}$	row echelon form, complementary matrix of A
$\overline{A}$, A^*	complex-conjugate, adjoint matrix of A
$\mathrm{rk}(A)$, $\mathrm{tr}(A)$, $\det(A)$	rank, trace, determinant of A
$\mathrm{vol}(D)$	volume (Jordan measure) of $A \subseteq \mathbb{R}^n$
${}_C[f]_B$, $[f]$, ${}_C\Delta_B$	representation matrix, basis change matrix
$E_\lambda(f)$, $H_\lambda(f)$	eigenspace, generalized eigenspace for the eigenvalue λ of f
S_n	symmetric group of degree n
A_n	alternating group of degree n
$\mathrm{sgn}(\sigma)$, P_σ	sign, permutation matrix of $\sigma \in S_n$
$K[X]$	vector space of polynomials with coefficients in K
$K(X)$	field of rational functions
α'	derivative of $\alpha \in K[X]$
$\deg(\alpha)$	degree of $\alpha \in K[X]$
$\alpha \mid \beta$	α divides β
$\alpha \equiv \beta \pmod{\delta}$	$\delta \mid \alpha - \beta$
χ_A, χ_f, μ_A, μ_f	characteristic polynomial, minimal polynomial of A, f
μ_v	minimal polynomial of the cyclic subspace
$[v, w]$, $\lvert v \rvert$	inner product, Euclidean norm of v
$v \perp w$	v and w are orthogonal, i. e. $[v, w] = 0$
π	length of the semicircle arc with radius 1
$\cos\varphi$, $\sin\varphi$	cosine, sine of φ
$S^\perp$	orthogonal complement of $S \subseteq V$

$v \times w$	cross product of v and w
$D(\varphi)$, $S(\varphi)$	rotation, reflection in $\mathbb{R}^2$
$D_{st}(\phi)$	Givens rotation
S_v	reflection across the hyperplane $v^\perp$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$, $\overline{z}$	real part, imaginary part, complex conjugation of z
i	imaginary unit
$\mathrm{Bil}(V)$	vector space of bilinear forms on V
${}_B[\beta]_B$	Gram matrix of a bilinear form β
$\mathrm{ind}(\beta)$, $\mathrm{ind}(A)$	index of $\beta \in \mathrm{Bil}(V)$, $A \in K^{n\times n}$
$J_n(\lambda)$	Jordan block for $\lambda \in K$ of size $n \times n$
$J_k(\gamma)$	generalized Jordan block for $\gamma \in K[X]$
$B(\alpha)$	companion matrix of $\alpha \in K[X]$
$\mathrm{C}(f)$, $\mathrm{C}(A)$	centralizer of $f \in \mathrm{End}(V)$, $A \in K^{n\times n}$
W_n	n-th Fourier matrix
H_n	n-th Hilbert matrix
$\mathcal{F}_n$	discrete Fourier transform
$\kappa(A)$	condition number of A
$\lvert A\rvert$	Frobenius norm of A
$\lVert v\rVert_p$, $\lVert A\rVert_p$	p-(matrix-)norm
$\lVert v\rVert_{\max}$, $\lVert A\rVert_{\max}$	maximum norm, row sum norm
$\lambda_k(A)$	the k-th largest eigenvalue of $A = A^*$
$\rho(A)$	spectral radius of A
$\exp(A)$	exponential function of A
A_+	non-negative matrix $(\lvert a_{ij}\rvert)_{ij}$
L^*	dual linear program to L
$\mathrm{con}(\Delta)$	convex hull of $\Delta \subseteq \mathbb{R}^n$
$\mathrm{supp}(x)$	support of $x \in \mathbb{R}^n$
$\lfloor x\rfloor$, $\lfloor x\rceil$	floor and rounded value of $x \in \mathbb{R}$
$\mathrm{disc}(\Delta)$	discriminant of the lattice Δ
Δ^*	dual lattice to Δ
E_8	even lattice in $\mathbb{R}^8$ with full rank
$\Delta \perp \Lambda$	orthogonal decomposition of lattices
γ_n	n-th Hermite constant
cd, gcd	(greatest) common divisor
$\det(q)$	determinant of the quadratic form q
$\mu_i(q)$	successive minimum of q

Conventions

- Proper names are written in Small Caps upon first use.
- K is always a field, V a finite-dimensional K-vector space, subspaces are usually called U, W, V_1 etc.
- Sets and matrices are denoted by uppercase Latin letters $(A, B, \ldots, M, \ldots)$.
- Elements of sets are denoted by lowercase letters, vectors by u, v, w, natural numbers by n, m, k, l, maps by f, g, h etc.
- For sets of sets, "calligraphic" letters are often used ($\mathcal{M}$, $\mathcal{P}$).

- For polynomials, scalars (field elements in the context of vector spaces), and bilinear forms, we use Greek letters. The most common are:

α	β	γ, Γ	δ, Δ	ϵ, ε	ζ	η	$\theta, \vartheta, \Theta$	λ, Λ	μ
alpha	beta	gamma	delta	epsilon	zeta	eta	theta	lambda	mu
ν	ξ	π, Π	ρ, ϱ	σ, Σ	τ	φ, ϕ, Φ	χ	ψ, Ψ	ω, Ω
nu	xi	pi	rho	sigma	tau	phi	chi	psi	omega

Linear Algebra I

1 Propositional Logic and Set Theory

1.1 Propositions

Remark 1.1. The language of mathematics is based on logical principles that must ultimately be accepted as given. All "higher" mathematical objects can be traced back to set-theoretic constructs. We treat these topics here only as far as they are needed for the understanding of linear algebra. More information can be found in my notes on logic and set theory.

Definition 1.2.

- A *proposition* (or simply a *statement*) A is a sentence that takes either the *truth value true* ($\mathbf{t}$) or *false* ($\mathbf{f}$). One then says A *holds* or A *does not hold*.
- For propositions A and B, $\neg A$ (*not* A), $A \wedge B$ (A *and* B), $A \vee B$ (A *or* B), $A \Rightarrow B$ (A *implies* B) and $A \Leftrightarrow B$ (A *if and only if* B) are also propositions with the following truth values:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftrightarrow B$
$\mathbf{t}$	$\mathbf{t}$	$\mathbf{f}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{t}$
$\mathbf{t}$	$\mathbf{f}$	$\mathbf{f}$	$\mathbf{f}$	$\mathbf{t}$	$\mathbf{f}$	$\mathbf{f}$
$\mathbf{f}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{f}$	$\mathbf{t}$	$\mathbf{t}$	$\mathbf{f}$
$\mathbf{f}$	$\mathbf{f}$	$\mathbf{t}$	$\mathbf{f}$	$\mathbf{f}$	$\mathbf{t}$	$\mathbf{t}$

- Two propositions A and B are called *equivalent* if $A \Leftrightarrow B$ is true, i.e., if A and B have the same truth value.
- A *predicate* is a property $A = A(x)$ that only becomes a proposition by substituting a variable x. If applicable, $\forall x : A(x)$ (*for all* x, $A(x)$ holds) and $\exists x : A(x)$ (*there exists* an x such that $A(x)$ holds) are propositions.

Example 1.3. The following sentences are propositions (even if we do not know the truth value):

- All blue cats can fly ($\mathbf{t}$).
- $1 + 1 = 3$ ($\mathbf{f}$).
- Every even number greater than 2 is the sum of two prime numbers (?).[1]

In contrast, the following are not propositions:

- Let $\epsilon > 0$ (*assumption*).
- $a^2 + b^2 = c^2$ (*equation*).
- This sentence is false (*paradox*).

[1] Goldbach's conjecture

From the predicate $x > 0$, one can form the true proposition $\forall x > 4 : x > 0$.

Remark 1.4.

(a) In contrast to everyday language, the mathematical *or* differs from *either or*. That is, the proposition $\mathbf{t} \vee \mathbf{t}$ is true. We do not introduce a separate symbol for *either or*.[2] Furthermore, distinguish between the formulations "There exists a…" and "There exists exactly one…".

(b) The truth of the proposition $\mathbf{f} \Rightarrow \mathbf{f}$ irritates many beginners (see Example 1.3). Interpretation: If the premise is not satisfied, there is nothing to show. Furthermore, one must distinguish the proposition $A \Rightarrow B$ from its *converse* $B \Rightarrow A$.[3]

(c) For propositions $A_1, \ldots, A_n$, one defines $A_1 \wedge \ldots \wedge A_n$ by $\forall i : A_i$ and $A_1 \vee \ldots \vee A_n$ by $\exists i : A_i$.

(d) To determine the truth value of a proposition A, one performs *equivalence transformations*, i. e., one replaces A with an equivalent proposition. For this purpose, the following rules of inference are useful.[4]

Lemma 1.5. *Let A, B, and C be propositions. Then:*

(a) The following propositions are equivalent to A:

$$\neg\neg A, \qquad A \wedge \mathbf{t}, \qquad A \vee \mathbf{f}, \qquad A \wedge A, \qquad A \vee A, \qquad \mathbf{t} \Rightarrow A$$

(b) $A \wedge B$ and $B \wedge A$ are equivalent, as are $A \vee B$ and $B \vee A$ (commutative law).

(c) $A \vee \neg A$ holds (law of excluded middle) *and $\neg(A \wedge \neg A)$* (law of non-contradiction).

(d) $A \wedge (B \vee C)$ and $(A \wedge B) \vee (A \wedge C)$ are equivalent, as are $A \vee (B \wedge C)$ and $(A \vee B) \wedge (A \vee C)$ (distributive law).

(e) $\neg(A \wedge B)$ and $\neg A \vee \neg B$ are equivalent, as are $\neg(A \vee B)$ and $\neg A \wedge \neg B$ (De Morgan*'s* laws).

(f) $A \Rightarrow B$, $\neg A \vee B$, and $(\neg B) \Rightarrow (\neg A)$ are equivalent (contraposition).

(g) $(A \Rightarrow B) \wedge (B \Rightarrow C)$ implies $A \Rightarrow C$ (transitivity).

(h) From $A \wedge (A \Rightarrow B)$ follows B (modus ponens).

(i) $A \Leftrightarrow B$ is equivalent to $(A \Rightarrow B) \wedge (B \Rightarrow A)$.

Proof. All assertions can be easily verified using truth tables. For three variables, one must distinguish $2^3 = 8$ cases (anyone who finds a faster way can earn a million dollars[5]). Alternatively, some of the assertions can be derived from those already proven. For instance, the second De Morgan's law follows from the first:

$$\neg(A \vee B) \overset{\text{(a)}}{\Longleftrightarrow} \neg((\neg\neg A) \vee (\neg\neg B)) \Longleftrightarrow \neg(\neg(\neg A \wedge \neg B)) \overset{\text{(a)}}{\Longleftrightarrow} (\neg A \wedge \neg B). \qquad \square^6$$

[2] In computer science, one speaks of `XOR`.
[3] One could also write $A \Leftarrow B$.
[4] A *lemma* is an auxiliary theorem with little significance of its own.
[5] The SAT *problem* of theoretical computer science is NP-complete. One of the seven *Millennium Problems* asks whether P = NP.
[6] This box marks the end of a proof.

Remark 1.6.

(a) De Morgan's laws can be formulated more generally for predicates in the form $(\neg\forall x : A(x)) \Leftrightarrow (\exists x : (\neg A(x)))$ and $(\neg\exists x : A(x)) \Leftrightarrow (\forall x : (\neg A(x)))$.

(b) Lemma 1.5 shows that one can express all further terms using only the symbols $\neg$ and $\wedge$. For the sake of readability, however, one should use all symbols sparingly.

1.2 Sets

Definition 1.7 (Cantor)**.** A *set* M is a collection of definite, distinct objects x of our perception or of our thought into a whole.[7] One then says: x is an *element* of M and writes $x \in M$ as well as $M = \{x : x \in M\}$ (resp. $x \notin M$ for $\neg(x \in M)$). The number $|M|$ of elements of M is called the *cardinality* or *size* of M. In the case $|M| < \infty$, M is called *finite* and otherwise *infinite*.

Remark 1.8.

(a) Definition 1.7 is imprecise, because it allows sets that lead to logical contradictions. For example, let

$$M := \{x : x \notin x\} \qquad (\text{Russell}'s\,Antinomy)^{8}$$

The proposition $M \in M$ can then be neither true nor false. In modern mathematics, such contradictions are prevented by introducing an *axiom system*, i.e., one prescribes the truth value of as few "elementary" propositions (axioms) as possible. The Zermelo-Fraenkel system is widely used. One of its axioms states:

Sets are equal if and only if they contain the same elements.

This implies that the elements of a set have no fixed order. Thus, $\{2, 1, 1, 2, 2\} = \{1, 2\}$ holds.

(b) In some situations, the so-called *axiom of choice* (see Example 2.3) is additionally required. It is accepted by most mathematicians, although it allows the construction of counterintuitive sets: The Banach-Tarski *paradox* states, for example, that one can decompose a sphere of volume 1 into five parts which, when reassembled differently, yield two spheres of volume 1.

(c) According to Gödel's second *incompleteness theorem*, it is impossible to prove that the Zermelo-Fraenkel axioms do not yield contradictions.[9] If this is indeed the case (which most mathematicians assume), then Gödel's first incompleteness theorem states that there are propositions whose truth value cannot be determined.[10] The best-known example of this is the *continuum hypothesis* (see Remark 2.11(f)).

Definition 1.9.

(a) For sets A and B let

$$\begin{aligned} \varnothing &:= \{\} \qquad (\textit{empty set}), \\ A \cup B &:= \{x : x \in A \vee x \in B\} \qquad (\textit{union}), \end{aligned}$$

[7] Cantor's wording
[8] The symbol := states that the left side is defined by the right side.
[9] See notes on logic and set theory
[10] The idea of the proof consists of formalizing the proposition "This sentence is not provable."

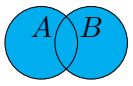
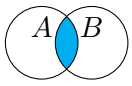
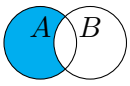
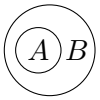
$$A \cap B := \{x : x \in A \wedge x \in B\} \qquad (\textit{intersection}),^{11}$$
$$A \setminus B := \{x : x \in A \wedge x \notin B\} \qquad (\textit{difference}).^{12}$$

(b) In the case $A \cup B = B$, A is a *subset* of B. One then writes $A \subseteq B$ or $A \subsetneq B$, if additionally $A \neq B$ (one then speaks of a *proper* subset[13]). If A is not a subset of B, one writes $A \not\subseteq B$.

(c) One calls A and B *disjoint*, if $A \cap B = \varnothing$. If applicable, one calls $A \dot\cup B := A \cup B$ a *disjoint union*.

Remark 1.10.

(a) Relationships between sets can be illustrated by Venn diagrams:

$A \cup B$:  $A \cap B$:  $A \setminus B$:  $A \subseteq B$: 

Attention: If more than three sets are involved, the general situation can no longer be represented by circles.[14]

(b) Union and intersection of arbitrarily many sets A_i (where i comes from an index set I) can be defined as follows:

$$\bigcup_{i \in I} A_i := \{x : \exists i \in I : x \in A_i\}, \qquad \bigcap_{i \in I} A_i := \{x : \forall i \in I : x \in A_i\}.$$

If A is the disjoint union of sets A_i, one speaks of a *partition* of A.

(c) To prove the equality of sets $A = B$, it is often easier to show the equivalent proposition $(A \subseteq B) \wedge (B \subseteq A)$.

Example 1.11.

(a) The set of *natural numbers* $\mathbb{N} := \{1, 2, 3, \ldots\}$. We set $\mathbb{N}_0 := \{0, 1, 2 \ldots\} = \mathbb{N} \cup \{0\}$.[15] Note: For some authors, $0 \in \mathbb{N}$.

(b) The set of *integers* $\mathbb{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. It holds that $\mathbb{N} = \{n \in \mathbb{Z} : n > 0\}$. The integers of the form $2n$ (resp. $2n + 1$) with $n \in \mathbb{Z}$ are called *even* (resp. *odd*).

(c) The set of *rational numbers* $\mathbb{Q} := \{\frac{a}{b} : a, b \in \mathbb{Z},\ b \neq 0\}$.

(d) The set of *real numbers* $\mathbb{R}$ consists of all *decimal fractions* such as $2 = 2.0$, $\frac{1}{3} = 0.33\ldots$, $\sqrt{2} = 1.4142\ldots$ or $\pi = 3.1415\ldots$ (the decimal expansion can be terminating, periodic or non-periodic). In analysis, real numbers are defined as limits of rational Cauchy sequences. In the following, we assume the usual rules for basic arithmetic operations. These can also be introduced strictly axiomatically.

[11] Note the similarity of the symbols $\cup$ and $\vee$ as well as $\cap$ and $\wedge$.
[12] In some books one writes $A - B$ instead of $A \setminus B$.
[13] The symbol $\subset$ is unfortunately not used consistently in the literature.
[14] see `https://en.wikipedia.org/wiki/Venn_diagram`
[15] Strictly axiomatically, one defines $0 := \varnothing$, $1 := \{\varnothing\}$ and generally $n + 1 := n \cup \{n\}$.

(e) It holds that $\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}$. We show the claim $\mathbb{Q} \neq \mathbb{R}$ indirectly. Assumption: $\mathbb{Q} = \mathbb{R}$. Then $\sqrt{2} \in \mathbb{Q}$ and there exist $a, b \in \mathbb{Z}$ with $\sqrt{2} = \frac{a}{b}$ and $b \neq 0$. wlog. , we can assume that a and b are coprime (otherwise one can simplify $\frac{a}{b}$). Rearranging yields $2b^2 = a^2$. In particular, a^2 is even. Since the square of an odd number is odd ($(2n+1)^2 = 2(2n^2+2n)+1$), a is even, say $a = 2c$. It follows that $b^2 = 2c^2$. By the same argument, b is now also even. Thus 2 is a common divisor of a and b. This contradiction shows that the assumption was false. Therefore $\mathbb{Q} \neq \mathbb{R}$.

(f) The elements of a set can certainly be sets themselves. In such cases, one often uses script letters. For example, $\mathcal{M} := \{\{1,2\},\{1,3\},\{2,3\}\}$ consists of all 2-element subsets of $\{1,2,3\}$.

Lemma 1.12. *For sets A, B and C, the following hold:*

(a) $A \cap B \subseteq A \subseteq A \cup B$.

(b) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ *and* $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (distributive law).

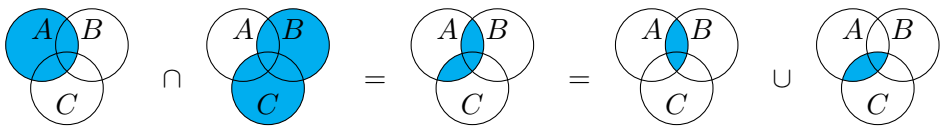
(c) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$ *and* $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$ (De Morgan's laws).

(d) $\boxed{|A \cup B| + |A \cap B| = |A| + |B|}$ *and* $|A \dot\cup B| = |A| + |B|$.

Proof.

(a) Follows directly from the definition.

(b) We only prove the first equality (prove the second one yourself):



(c) This time we use Lemma 1.5 (for the first equation):

$$\begin{aligned} x \in A \setminus (B \cup C) &\iff (x \in A \wedge (x \notin B \cup C)) \iff (x \in A \wedge (x \notin B \wedge x \notin C)) \\ &\iff ((x \in A \wedge x \notin B) \wedge (x \in A \wedge x \notin C)) \iff x \in (A \setminus B) \cap (A \setminus C). \end{aligned}$$

(d) If A or B is infinite, then so is $A \cup B$ and the claim holds if one interprets $\infty + n = \infty$ for $n \in \mathbb{N}_0 \cup \{\infty\}$. Now let A and B be finite, say $A \cap B = \{x_1, \ldots, x_s\}$, $A = \{x_1, \ldots, x_s, a_1, \ldots, a_t\}$ and $B = \{x_1, \ldots, x_s, b_1, \ldots, b_u\}$. Then

$$|A \cup B| + |A \cap B| = s + t + u + s = |A| + |B|.$$

If A and B are disjoint, then $|A \cap B| = |\varnothing| = 0$ and the second claim follows. □

Definition 1.13. The *power set* $\mathcal{P}(M)$ of a set M is the set of all subsets of M, i.e.

$$\mathcal{P}(M) := \{N : N \subseteq M\}.$$

1.3 Mathematical Induction

Theorem 1.14 (Principle of mathematical induction)**.** *Let $A(n)$ be a predicate for $n \in \mathbb{N}$ with the properties:*

- *Base case: $A(1)$ holds.*
- *Inductive step: $\forall n \in \mathbb{N} : (A(n) \Longrightarrow A(n+1))$.*

Then $A(n)$ holds for all $n \in \mathbb{N}$.

Proof. Proof by contradiction: If $A(n)$ does not hold for all $n \in \mathbb{N}$, then there exists a smallest n with $\neg A(n)$. By the base case, $n \neq 1$. By the choice of n, $A(n-1)$ holds. By the inductive step, $A(n-1) \Rightarrow A(n)$ holds. Thus $A(n)$ holds by modus ponens. Contradiction. □

Remark 1.15. One often uses variants of mathematical induction. For example:

- Base case: $A(1) \wedge A(2)$ holds.
- Inductive step: $\forall n \in \mathbb{N} : \big((A(n) \wedge A(n+1)) \Rightarrow A(n+2)\big)$.

Example 1.16. We prove $(1 + 2 + \ldots + n)^2 = 1^3 + 2^3 + \ldots + n^3$ for all $n \in \mathbb{N}$.

Base case: For $n = 1$, it holds that $1^2 = 1 = 1^3$.

Induction hypothesis: Assume that $(1 + 2 + \ldots + n)^2 = 1^3 + 2^3 + \ldots + n^3$ $(*)$ already holds.

Inductive step: We must prove the claim for $n + 1$. First, an auxiliary calculation:

$$\begin{aligned} 2(1 + 2 + \ldots + n) &= (1 + 2 + \ldots + n) + (n + (n-1) + \ldots + 1) \\ &= (1 + n) + (2 + n - 1) + \ldots + (n + 1) = n(n + 1) \end{aligned}$$

(this was recognized by Gauss as a 9-year-old[16]). According to the binomial formula, it now holds that

$$\begin{aligned} ((1 + 2 + \ldots + n) + (n + 1))^2 &= (1 + 2 + \ldots + n)^2 + 2(1 + 2 + \ldots + n)(n + 1) + (n + 1)^2 \\ &\overset{(*)}{=} 1^3 + 2^3 + \ldots + n^3 + n(n + 1)(n + 1) + (n + 1)^2 \\ &= 1^3 + 2^3 + \ldots + n^3 + (n + 1)^3. \end{aligned}$$

□

[16] see American Scientist

2 Cartesian Products and Functions

2.1 Pairs and Tuples

Remark 2.1. According to Remark 1.8, the elements of a set are unordered. We introduce an ordered variant.

Definition 2.2.

- Let A and B be sets. The *Cartesian product* of A and B is the set $A \times B$ consisting of all (*ordered*) *pairs*[1] (a, b) with $a \in A$, $b \in B$, such that

$$(a, b) = (a', b') \iff (a = a' \wedge b = b').$$

  It holds that $\boxed{|A \times B| = |A||B|,}$ provided one uses the rules $\infty \cdot 0 = 0$ and $\infty \cdot n = \infty$ for $n \in \mathbb{N} \cup \{\infty\}$.

- Analogously, one defines *triples* (a, b, c) and *n-tuples* $(a_1, \ldots, a_n)$ for $n \geq 2$. For sets $A_1, \ldots, A_n$, one sets

$$A_1 \times \ldots \times A_n := \big\{(a_1, \ldots, a_n) : a_1 \in A_1, \ldots, a_n \in A_n\big\}.$$

  If $A := A_1 = \ldots = A_n$, then one uses the abbreviation $A^n := A_1 \times \ldots \times A_n$.

- Cartesian products can also be defined for arbitrary families of sets. Let I be an index set and $(A_i : i \in I)$ a family of sets. One defines $\bigtimes_{i \in I} A_i := \{(a_i)_{i \in I} : \forall i \in I : a_i \in A_i\}$.

Example 2.3.

(a) The Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ consists of all coordinates in the 2-dimensional plane.

(b) It holds that

$$\{1, 2\} \times \{2, 3, 4\} = \big\{(1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4)\big\}.$$

(c) The Cartesian product $\bigtimes_{i \in \mathbb{N}} \mathbb{R}$ is the set of all real sequences from analysis.

(d) Let $(A_i : i \in I)$ be an arbitrary family of non-empty sets. The already mentioned *axiom of choice* states that $\bigtimes_{i \in I} A_i \neq \varnothing$, i. e. one can choose an element from each set A_i *simultaneously*.

Definition 2.4. A *relation* on a non-empty set A is a subset $R \subseteq A \times A$. Usually, one chooses a symbol, for example $\sim$, and writes $a \sim b$ if $(a, b) \in R$. One calls R

- *reflexive*, if $\forall a \in A : a \sim a$.
- *symmetric*, if $\forall a, b \in A : (a \sim b \Rightarrow b \sim a)$.

[1] The formal definition of pairs can be reduced to sets: $(a, b) := \{\{a\}, \{a, b\}\}$.

- *antisymmetric*, if $\forall a, b \in A : ((a \sim b \wedge b \sim a) \Rightarrow a = b)$.[2]
- *transitive*, if $\forall a, b, c \in A : ((a \sim b \wedge b \sim c) \Rightarrow a \sim c)$.
- *equivalence relation*, if R is reflexive, symmetric, and transitive.
- (*partial*) *order relation*, if R is reflexive, antisymmetric, and transitive.

If R is an equivalence relation on the set A and $a \in A$, then $[a] := \{b \in A : a \sim b\} \subseteq A$ is called the *equivalence class* of a.

Example 2.5.

(a) The *trivial* relation $R = A \times A$ is an (uninteresting) equivalence relation.

(b) The equality relation $\{(a, a) : a \in A\}$ with the symbol $=$ is the "smallest" reflexive relation on A. Trivially, it is an equivalence relation.[3] Many other equivalence relations can be reduced to equality. For example, let A be the set of all humans and $a \sim b$ if $a, b \in A$ live in the same country. The equivalence classes then correspond to the countries.

(c) One can show through simple examples that the properties reflexive, symmetric, and transitive are independent of each other. For example, the relation

$$\{(1,1), (2,2), (3,3), (1,2), (2,1), (2,3), (3,2)\}$$

on $A = \{1, 2, 3\}$ is reflexive and symmetric, but not transitive.

(d) On $\mathbb{R}$, the less-than-or-equal relation $\leq$ is an order relation. It additionally has the property that any two numbers a and b are related, i. e., $a \leq b$ or $b \leq a$ holds (one speaks of a *total* order relation).

(e) On the power set of any set A, the inclusion relation $\subseteq$ is an order relation. In the case $A = \mathbb{N}$, $\{1\}$ and $\{2\}$ are not related ($\{1\} \not\subseteq \{2\} \not\subseteq \{1\}$). In contrast to $\leq$, $\subseteq$ is therefore not total.

Lemma 2.6. *Let R be an equivalence relation on a set A. Then there exists a subset $T \subseteq A$ such that the equivalence classes $[t]$ with $t \in T$ form a partition of A, i. e., $A = \dot{\bigcup}_{t \in T} [t]$.*

Proof. Let $\sim$ be the symbol of R. Let $a, b \in A$ and $c \in [a] \cap [b]$. Then $a \sim c$ and $b \sim c$ hold. Since $\sim$ is symmetric, $c \sim b$ holds. Since $\sim$ is transitive, $a \sim b$ holds. For every $d \in [b]$, it thus holds that $a \sim b \sim d$ and $a \sim d$. This shows $[b] \subseteq [a]$ and analogously one obtains $[a] \subseteq [b]$. It follows that $[a] = [b]$. Thus, any two equivalence classes are either equal or disjoint. The existence of T now follows from the axiom of choice. □

Remark 2.7.

(a) In the situation of Lemma 2.6, T is called a *system of representatives* for the equivalence classes.

(b) If $A = \dot{\bigcup}_{i \in I} A_i$ is a partition of A, then

$$a \sim b \ :\Longleftrightarrow\ \exists i \in I : a, b \in A_i$$

defines an equivalence relation on A (verify). Therefore, partitions and equivalence relations correspond to each other.

[2] Attention: There is also the stronger property *asymmetric*: $\forall a, b \in A : (a \sim b \Rightarrow b \not\sim a)$.

[3] For symmetric relations, one should choose "symmetric" symbols. Unfortunately, the converse is not always given, e. g., the symbol | for the antisymmetric divisibility relation on $\mathbb{N}$.

Example 2.8. For the equivalence relation on the set of all humans from Example 2.5, the presidents of each country form a system of representatives.

2.2 Injective and Surjective Functions

Definition 2.9.

- Let A and B be sets. A *function* or *map* f from A to B is a rule that assigns to each $a \in A$ exactly one $f(a) \in B$.[4] One then writes[5]

$$f\colon A \to B, \qquad a \mapsto f(a).$$

  We denote the set of all maps $A \to B$ by $\mathrm{Fun}(A, B)$.

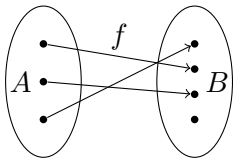
- A is called the *domain* and B the *codomain* of f. Furthermore, $f(a)$ is the *image* of a under f and $f(A) := \{f(a) : a \in A\} \subseteq B$ is the *image* of f. For $B' \subseteq B$,

$$f^{-1}(B') := \{a \in A : f(a) \in B'\} \subseteq A$$

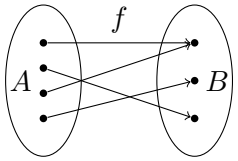
  is the *preimage* of B' under f.

- $f\colon A \to B$ is called

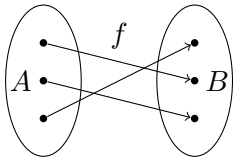
  – *injective*, if $\forall a, a' \in A : (f(a) = f(a') \Longrightarrow a = a')$.



  – *surjective*, if $\forall b \in B : \exists a \in A : f(a) = b$, i. e. $f(A) = B$.



  – *bijective* (or *bijection*), if f is injective and surjective. If applicable, A and B are called *equinumerous*.



- The *restriction* of $f\colon A \to B$ to a subset $A' \subseteq A$ is the function

$$f_{|A'}\colon A' \to B, \qquad a \mapsto f(a).$$

  For another function $g\colon B \to C$, the map

$$g \circ f\colon A \to C, \qquad a \mapsto g(f(a))$$

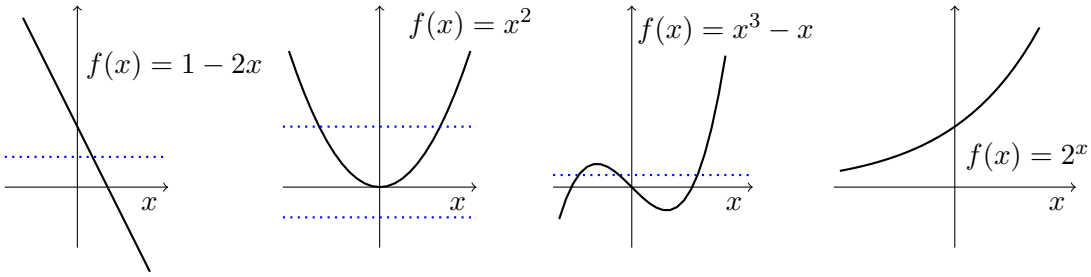
  is called the *composition* (or *composite*, *concatenation*) of f and g.

Example 2.10.

(a) For every set A and $B \subseteq A$, $f\colon B \to A$, $b \mapsto b$ is an injective function, which is called the *inclusion map*. In the case $B = A$, f is even bijective and one calls $f = \mathrm{id}_A$ the *identity* on A.

[4] Formally: A function is a subset $f \subseteq A \times B$ such that for each $a \in A$ there exists exactly one $b \in B$ with $(a, b) \in f$.
[5] Note the different arrows $\to$ and $\mapsto$.

(b) Maps $f\colon \mathbb{R} \to \mathbb{R}$ can be represented graphically:



Injective (resp. surjective) means that the graph of f intersects every horizontal line at most (resp. at least) once. We read off:

Function	injective	surjective	bijective
$f(x) = 1 - 2x$	✓	✓	✓
$f(x) = x^2$	✗	✗	✗
$f(x) = x^3 - x$	✗	✓	✗
$f(x) = 2^x$	✓	✗	✗

Remark 2.11.

(a) If A and B are finite sets, then $|\mathrm{Fun}(A, B)| = |B|^{|A|}$ holds, because for $f\colon A \to B$ and each $a \in A$ one has $|B|$ possibilities to choose $f(a) \in B$.

(b) Attention: Injective is not the opposite of surjective (a common beginner's mistake)!

(c) One can make every function $f\colon A \to B$ surjective by restricting the codomain to the image: $f\colon A \to f(A)$.

(d) For $f\colon A \to B$ it holds that

$$\begin{aligned} f \text{ injective} &\implies |A| = |f(A)| \leq |B| \\ f \text{ surjective} &\implies |B| = |f(A)| \leq |A| \\ f \text{ bijective} &\implies |A| = |B| \end{aligned}$$

(where $\infty \leq \infty$).

(e) Two finite sets A and B are of the same cardinality if and only if $|A| = |B|$. In this case, the properties injective, surjective, and bijective are equivalent according to (d). For infinite sets, this is in general false (Example 2.10).

(f) Although $\mathbb{N}$ contains only "half as many" numbers as $\mathbb{Z}$, $\mathbb{N}$ and $\mathbb{Z}$ are of the same cardinality through the bijection

$$\mathbb{N} \to \mathbb{Z}, \qquad n \mapsto \begin{cases} \frac{n-1}{2} & \text{if } n \text{ is odd,} \\ -\frac{n}{2} & \text{if } n \text{ is even} \end{cases}$$

A set that is of the same cardinality as $\mathbb{N}$ is called *countable*[6]. According to Cantor's *diagonalization arguments*, $\mathbb{Q}$ is countable, but $\mathbb{R}$ is not, i.e., $\mathbb{R}$ is *uncountable*. The (unprovable) *continuum hypothesis* states that every infinite subset of $\mathbb{R}$ is of the same cardinality as either $\mathbb{N}$ or $\mathbb{R}$. The following theorem shows that there are arbitrarily "large" sets (*cardinal numbers*), which, however, are rarely encountered in practice.

[6] In some books, finite sets are also counted among the countable sets.

Theorem 2.12 (Cantor)**.** *Every set M is "smaller" than its power set, i.e., there exists an injective mapping $M \to \mathcal{P}(M)$, but no bijection. If M is finite, then $|\mathcal{P}(M)| = 2^{|M|}$ holds.*

Proof. The mapping $M \to \mathcal{P}(M)$, $a \mapsto \{a\}$ is certainly injective. Suppose there exists a bijection $f\colon M \to \mathcal{P}(M)$. Let

$$A := \{x \in M : x \notin f(x)\} \in \mathcal{P}(M).$$

Then there exists an $a \in M$ with $f(a) = A$. The contradiction $a \in A = f(a) \iff a \notin f(a)$ follows. For the second assertion, let $M = \{x_1, \ldots, x_n\}$. Then the mapping

$$\mathcal{P}(M) \to \{0,1\}^n, \qquad A \mapsto (a_1, \ldots, a_n)$$

with $a_i = 1 \iff x_i \in A$ is a bijection. Thus $|\mathcal{P}(M)| = |\{0,1\}^n| = 2^n = 2^{|M|}$. □

Lemma 2.13. *Let $f\colon A \to B$, $g\colon B \to C$, $h\colon C \to D$ be functions. Then:*

(a) $(h \circ g) \circ f = h \circ (g \circ f)$ (Associative law).

(b) If f and g are injective, then so is $g \circ f$.

(c) If f and g are surjective, then so is $g \circ f$.

(d) If $g \circ f$ is injective, then so is f.

(e) If $g \circ f$ is surjective, then so is g.

(f) f is bijective if and only if there exists a function $g\colon B \to A$ with $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$. If applicable, g is uniquely determined and one calls $f^{-1} := g$ the inverse function *of f.*

Proof.

(a) For $a \in A$, we have $((h \circ g) \circ f)(a) = (h \circ g)(f(a)) = h(g(f(a))) = h((g \circ f)(a)) = (h \circ (g \circ f))(a)$.

(b) For $a, a' \in A$ with $(g \circ f)(a) = (g \circ f)(a')$, it holds that $g(f(a)) = g(f(a'))$, thus $f(a) = f(a')$ and $a = a'$.

(c) We have $(g \circ f)(A) = g(f(A)) = g(B) = C$.

(d) Let $f(a) = f(a')$ for $a, a' \in A$. Then $(g \circ f)(a) = g(f(a)) = g(f(a')) = (g \circ f)(a')$. Since $g \circ f$ is injective, it follows that $a = a'$.

(e) We have $C = (g \circ f)(A) = g(f(A)) \subseteq g(B) \subseteq C$, thus $g(B) = C$.

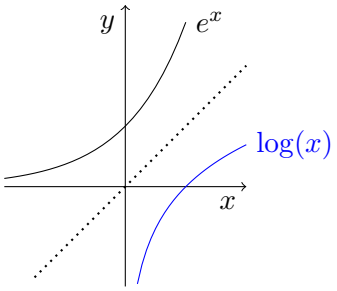
(f) If $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$, then f is injective by (d) and surjective by (e), thus also bijective. Conversely, let f be bijective. For each $b \in B$, there then exists exactly one $g(b) \in A$ with $f(g(b)) = b$. Therefore, $g\colon B \to A$ is the only mapping with $f \circ g = \mathrm{id}_B$. From $f(a) = f(g(f(a)))$ it follows that $g(f(a)) = a$ for all $a \in A$, since f is injective. This shows $g \circ f = \mathrm{id}_A$. □

Remark 2.14. Do not confuse the inverse function with the preimage. The connection between both concepts is $f^{-1}(\{b\}) = \{f^{-1}(b)\}$ for every bijection $f\colon A \to B$ and $b \in B$.

Example 2.15.

(a) The map $f\colon \mathbb{Q} \to \mathbb{Q}$, $x \mapsto 2x + 1$ is a bijection with inverse map $f^{-1}\colon \mathbb{Q} \to \mathbb{Q}$, $x \mapsto \frac{x-1}{2}$ (verify).

(b) The inverse map of the *exponential function* $\exp\colon \mathbb{R} \to \mathbb{R}_{>0}$, $x \mapsto e^x$ is the *natural logarithm* $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$. One obtains the graph of log by reflection across the line $y = x$:



Note that the mere existence of the inverse function does not yet provide a concrete formula for $f^{-1}(x)$. This circumstance is exploited in cryptography (*one-way function*).

3 Fields and Vector Spaces

3.1 Groups and Fields

Remark 3.1. In almost all applications of linear algebra, only the four basic arithmetic operations (addition, subtraction, multiplication, and division) are used. So that one does not have to prove every statement anew for every number system ($\mathbb{Q}, \mathbb{R}, \ldots$), number systems are replaced by abstract *groups* (with one operation) and *fields* (with two operations). To describe solution sets of systems of equations, one introduces *vector spaces*. Note that these are merely models for investigating linear problems that have proven themselves over time (just like metric spaces in analysis or the Bohr model of the atom in chemistry).

Definition 3.2. An *operation* $\cdot$ on a set G is a map $G \times G \to G$, $(x, y) \mapsto x \cdot y$. The pair $(G, \cdot)$ (or just G) is called a *group*, if

- $\forall x, y, z \in G : (x \cdot y) \cdot z = x \cdot (y \cdot z)$ (*associative law*),
- $\exists e \in G : (\forall x \in G : e \cdot x = x = x \cdot e)$ (*identity element*),
- $\forall x \in G : (\exists y \in G : y \cdot x = e = x \cdot y)$ (*inverse element*).

If additionally

- $\forall x, y \in G : x \cdot y = y \cdot x$ (*commutative law*),

then G is called *abelian*.[1]

Remark 3.3. Let G be a group with identity element e.

(a) For convenience, we often write xy instead of $x \cdot y$.

(b) If $e' \in G$ is also an identity element, then $e' = e' \cdot e = e$. Thus e is uniquely determined and we often write $e = 1_G = 1$ or $e = 0_G = 0$ if the operation is $+$.

(c) Let $y, y' \in G$ be inverses of $x \in G$. Then

$$y' = y'e = y'(xy) = (y'x)y = ey = y.$$

Thus x has exactly one inverse and we write $y = x^{-1}$ or $y = -x$ if the operation is $+$. In the latter case, we write $x - y := x + (-y)$ for arbitrary $x, y \in G$.[2]

(d) For $x, y \in G$, we have $\boxed{(x^{-1})^{-1} = x}$ and $\boxed{(xy)^{-1} = y^{-1}x^{-1}}$ (note the order!).

Example 3.4.

(a) Because $e \in G$, a group is never empty. On the other hand, there is the *trivial* group $G = \{e\}$.

[1] Named after N. Abel.

[2] In non-abelian groups, the notation $\frac{x}{y}$ is problematic, because it could mean both xy^{-1} and $y^{-1}x$.

(b) According to the usual calculation rules, $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$ are abelian groups with identity element 0. On the other hand, $(\mathbb{Z}, -)$ is *not* a group, because the associative law is violated:

$$(1 - 2) - 3 = -4 \neq 2 = 1 - (2 - 3).$$

Likewise, $(\mathbb{N}, +)$ has no identity element and in $(\mathbb{N}_0, +)$ not every element has an inverse (e. g., $-1 \notin \mathbb{N}_0$).

(c) Obviously, $(\mathbb{Q}\setminus\{0\}, \cdot)$ and $(\mathbb{R}\setminus\{0\}, \cdot)$ are abelian groups with identity element 1, but not $(\mathbb{Z}\setminus\{0\}, \cdot)$, because $2^{-1} = \frac{1}{2} \notin \mathbb{Z}$.

(d) For groups $G_1, \ldots, G_n$, the set $G_1 \times \ldots \times G_n$ is also a group with

$$(x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) := (x_1 y_1, \ldots, x_n y_n)$$

for $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in G_1 \times \ldots \times G_n$ (Exercise I.8). The identity element is $(1_{G_1}, \ldots, 1_{G_n})$. One then speaks of the *direct product* of $G_1, \ldots, G_n$ (instead of the Cartesian product).

Definition 3.5. A *field* is a set K with operations $+$ and $\cdot$ such that the following properties hold:

- $(K, +)$ is an abelian group with identity element 0.
- $(K \setminus \{0\}, \cdot)$ is an abelian group with identity element 1. One sets $K^\times := K \setminus \{0\}$.
- $\forall x, y, z \in K : x \cdot (y + z) = (x \cdot y) + (x \cdot z)$ (*distributive law*).

Remark 3.6. In the following, let K always be a field.

(a) By the convention "multiplication before addition," we save parentheses. For example, let $xy + z := (x \cdot y) + z$ for $x, y, z \in K$.

(b) For all $x \in K$, it holds that $\boxed{x \cdot 0 = 0 = 0 \cdot x,}$ because $x0 = x(0 + 0) = x0 + x0$. It follows that $(-x)y = -(xy)$ for $x, y \in K$.

(c) For $x, y, z \in K$ and $z \neq 0$, the *cancellation rule* $xz = yz \Longrightarrow x = y$ holds, because

$$x = x \cdot 1 = x(zz^{-1}) = (xz)z^{-1} = (yz)z^{-1} = \ldots = y.$$

Example 3.7.

(a) According to the usual calculation rules, $\mathbb{Q}$ and $\mathbb{R}$ are fields. There are also infinitely many fields "between" $\mathbb{Q}$ and $\mathbb{R}$ (cf. Exercise I.14). On the other hand, $(\mathbb{Z}, +, \cdot)$ is not a field, since $(\mathbb{Z} \setminus \{0\}, \cdot)$ is not a group.

(b) Every field possesses at least the two elements 0 and 1. In fact, $\mathbb{F}_2 = \{0, 1\}$ is already a field if one defines $1 + 1 := 0$. The operation tables are thereby completely determined:

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

On computers, all calculations are performed in $\mathbb{F}_2$ by interpreting 0 and 1 as *bits*. In algebra[3], one constructs for every prime power q a field with exactly q elements (cf. Exercise I.9).

[3] see Algebra notes

3.2 Vector Spaces and Subspaces

Definition 3.8. A *vector space* V over a field K (short: K-vector space) is an abelian group wrt. $+$ together with a *scalar multiplication* $K \times V \to V$, $(\lambda, v) \mapsto \lambda \cdot v$ with the following properties:

- $\forall v \in V : 1 \cdot v = v$,
- $\forall v, w \in V,\ \lambda \in K : \lambda \cdot (v + w) = \lambda \cdot v + \lambda \cdot w$,
- $\forall v \in V,\ \lambda, \mu \in K : (\lambda + \mu) \cdot v = \lambda \cdot v + \mu \cdot v$,
- $\forall v \in V,\ \lambda, \mu \in K : (\lambda \cdot \mu) \cdot v = \lambda \cdot (\mu \cdot v)$.

The elements of V are called *vectors* and the elements in K are called *scalars* (in this context). The neutral element 0_V in V is called the *zero vector*.

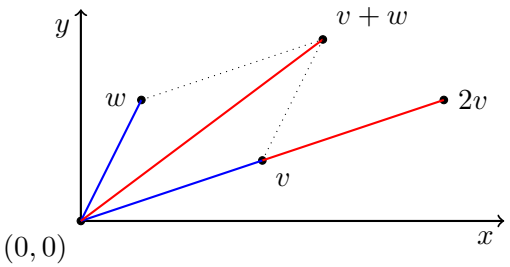
Remark 3.9. Note that $+$ denotes both the addition in K and in V. Likewise, $\cdot$ stands for the multiplication in K and for the scalar multiplication (this is imprecise, but quite common). In both cases, we will often omit the symbol $\cdot$. If misunderstandings are excluded, we also write 0 instead of 0_V. In case of doubt, you must be able to decide whether the zero element in K or V is meant.

Example 3.10.

(a) The *zero space* $V = \{0_V\}$ with the scalar multiplication $\lambda \cdot 0_V := 0_V$ for all $\lambda \in K$.

(b) For K-vector spaces $V_1, \ldots, V_n$, the direct product (wrt. $+$) $V_1 \times \ldots \times V_n$ is also a vector space with component-wise scalar multiplication: $\lambda(v_1, \ldots, v_n) := (\lambda v_1, \ldots, \lambda v_n)$ for $v_i \in V_i$ and $\lambda \in K$ (verify).

(c) Obviously, K itself is a vector space in which the scalar multiplication coincides with the ordinary multiplication. According to (b), K^n is also a vector space for $n \geq 1$. In $\mathbb{R}^2$, vector addition and scalar multiplication can be interpreted geometrically:



(d) If $v_1, \ldots, v_n$ are vectors from V and $\lambda_1, \ldots, \lambda_n \in K$, then the *linear combination* $\lambda_1 v_1 + \ldots + \lambda_n v_n$ also lies in V (proof by induction on n). One uses the summation symbol for this:

$$\boxed{\sum_{i=1}^{n} \lambda_i v_i := \lambda_1 v_1 + \ldots + \lambda_n v_n.}$$

If $v_1, \ldots, v_n$ are pairwise distinct (i.e., $v_i \neq v_j$ for $i \neq j$)[4] and at least one $\lambda_i \neq 0$, then the linear combination is called *non-trivial*. Sometimes the *empty sum* without summands occurs. This is

[4] "pairwise distinct" is stronger than the formulation "not all are equal"

always interpreted as 0_V. For example $\sum_{i=1}^{0} v_i = 0$. Let also $v_i = \mu_{i1}w_{i1} + \mu_{i2}w_{i2} + \ldots + \mu_{im}w_{im}$ be a linear combination for $i = 1, \ldots, n$.[5] Then one obtains a *double sum*:

$$\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij} w_{ij}.$$

Since $(V, +)$ is abelian, one may rearrange the summands arbitrarily and thus swap the summation symbols:

$$\boxed{\sum_{i=1}^{n} \sum_{j=1}^{m} \mu_{ij} w_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{n} \mu_{ij} w_{ij}.}$$

An advantage of algebra over analysis is that all sums are finite and one does not have to consider convergence.

Theorem 3.11. *Let $A \neq \varnothing$ be a set and V a K-vector space. Then* $\mathrm{Fun}(A, V)$ *with the following operations is a K-vector space:*

$$\begin{aligned} (f+g)(a) &:= f(a) + g(a) && (f, g \in \mathrm{Fun}(M, V),\ a \in M) \\ (\lambda f)(a) &:= \lambda f(a) && (\lambda \in K) \end{aligned}$$

Proof. Obviously $f + g$ and λf lie in $\mathrm{Fun}(A, V)$. The trivial map $f(a) = 0$ for all $a \in A$ is the neutral element wrt. $+$. For $f \colon A \to V$, $-f \colon A \to V$, $a \mapsto -f(a)$ is inverse to f wrt. $+$. For $f, g, h \colon A \to V$ and $a \in A$ we have

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

Therefore $+$ is associative. In the same way, the remaining vector space axioms transfer from V to $\mathrm{Fun}(A, V)$. □

Definition 3.12.

- A subset H of a group G is called a *subgroup*, if H with the restricted operation is itself a group, i. e.
  - $1_G \in H$,
  - $\forall g, h \in H : gh \in H$,
  - $\forall h \in H : h^{-1} \in H$.

  We write $H \leq G$ if applicable. In the case $H \neq G$, H is called a *proper* subgroup and we write $H < G$.
- A subset U of a vector space V is called a *subspace*, if U with the restricted operations is itself a vector space, i. e.
  - $(U, +)$ is a subgroup of $(V, +)$,
  - $\forall v \in U,\ \lambda \in K : \lambda v \in U$.

[5] In case of doubt, one should separate double indices by a comma $\mu_{i,1}$.

We then write $U \leq V$ as with subgroups. In the case $U \neq V$, U is called a *proper* subspace and we write $U < V$.

Remark 3.13.

(a) The conditions guarantee that H is *closed* under multiplication and U is *closed* under addition and scalar multiplication, respectively. Thus the operations on H and U are *well-defined*. The remaining group axioms or vector space axioms do not need to be checked, as they already hold in the larger set G or V.

(b) In the following, we restrict ourselves to the study of subspaces. Most statements also apply analogously to (abelian) groups.

(c) For vector spaces, the conditions can be summarized as follows: A *non-empty* subset $U \subseteq V$ is a subspace if and only if for all $u, v \in U$ and $\lambda \in K$ it holds: $\lambda u + v \in U$ (Exercise I.10).

Example 3.14.

(a) Every vector space V possesses the subspaces $\{0_V\}$ and V.

(b) From $U \leq W \leq V$ it follows that $U \leq V$. From $U, W \leq V$ and $U \subseteq W$ it certainly also follows that $U \leq W$.

(c) The intersection of an arbitrary number of subspaces is again a subspace (verify).

(d) We prove $U := \big\{(x, 0) : x \in \mathbb{R}\big\} \leq \mathbb{R}^2$ with the help of Remark 3.13: Because of $(0, 0) \in U$, $U \neq \varnothing$. For $(x_1, 0), (x_2, 0) \in U$ and $\lambda \in \mathbb{R}$ we have

$$\lambda(x_1, 0) + (x_2, 0) = (\lambda x_1, 0) + (x_2, 0) = (\lambda x_1 + x_2, 0) \in U.$$

Geometrically, U corresponds to the x-axis in the plane. Analogously, the xy-plane

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

is a subspace of $\mathbb{R}^3$.

(e) The subset $U := \big\{(x, x^2) : x \in \mathbb{Q}\big\}$ of $\mathbb{Q}^2$ is *not* a subspace, because $(1, 1) \in U$, but $2 \cdot (1, 1) = (2, 2) \notin U$. We show in Remark 7.19 that every subspace can be described by *linear* equations.

(f) Obviously $U := \big\{(0, 0), (1, 0)\big\}$ and $W := \big\{(0, 0), (0, 1)\big\}$ are subspaces of $\mathbb{F}_2^2$, but $U \cup W$ is not (Why?)

Lemma 3.15. *Let V be a K-vector space and $U \leq V$. For $v \in V$ let $v + U := \{v + u : u \in U\} \subseteq V$. Then $V/U := \{v + U : v \in V\}$ becomes a K-vector space with*

$$(v + U) + (w + U) := (v + w) + U \qquad (v, w \in V),$$
$$\lambda(v + U) := \lambda v + U \qquad (\lambda \in K).$$

Proof. Let $\overline{v} := v + U$ for $v \in V$. For $\overline{v} = \overline{v'}$ and $\overline{w} = \overline{w'}$ we have

$$\overline{v + w} = v + w + U = v + w' + U = w' + v + U = w' + v' + U = \overline{v' + w'}.$$

This shows that the addition on V/U is well-defined. The neutral element is $0 + U = U$. For $\lambda \in K$ it holds analogously

$$\overline{\lambda v} = \lambda v + U = \lambda v + \lambda U = \lambda(v + U) = \lambda(v' + U) = \lambda v' + U = \overline{\lambda v'}.$$

Thus the scalar multiplication is also well-defined. The vector space axioms for V/U follow directly from the axioms for V. □

Remark 3.16. One calls V/U the *quotient space* of V by U. The sets $v + U$ are sometimes referred to as *affine spaces* (Exercise III.25). They are the equivalence classes of the relation $v \sim w :\Longleftrightarrow v - w \in U$.

4 Bases and Dimension

4.1 Linear Independence and Generating Sets

Remark 4.1. In order to be able to compare infinitely large vector spaces, we introduce the dimension as a finer characteristic. It will be shown that vector spaces are largely determined by their dimension alone (Theorem 7.10).

Definition 4.2. Let V be a vector space.

(a) For $S \subseteq V$, let $\langle S \rangle \subseteq V$ be the set of all linear combinations of elements from S. One calls $\langle S \rangle$ the *span* of S.[1] In the case $S = \{s_1, \ldots, s_n\}$, we also write $\langle s_1, \ldots, s_n \rangle$ instead of $\langle S \rangle$ (i. e. we omit the set braces).

(b) For subspaces $U, W \leq V$, let

$$U + W := \{u + w : u \in U, w \in W\} \subseteq V$$

be the (Minkowski) *sum* of U and W. In the case $U \cap W = \{0\}$, the sum is called *direct* and one writes $U \oplus W$ instead of $U + W$.[2]

Lemma 4.3. *Let V be a vector space, $S \subseteq V$ and $U, W \leq V$. Then $\langle S \rangle$ and $U + W$ are subspaces of V.*

Proof. Clearly 0 is a linear combination of elements from S, i. e. $0 \in \langle S \rangle$ (in the case $S = \varnothing$, choose the empty sum). Addition and scalar multiplication of linear combinations are again linear combinations. This shows $\langle S \rangle \leq V$. Because of $0 \in U \cap W$, we have $0 = 0 + 0 \in U + W$. Let $u_1 + w_1, u_2 + w_2 \in U + W$ and $\lambda \in K$. Then

$$\lambda(u_1 + w_1) + (u_2 + w_2) = \underbrace{(\lambda u_1 + u_2)}_{\in U} + \underbrace{(\lambda w_1 + w_2)}_{\in W} \in U + W.$$

Thus $U + W \leq V$ as well. □

Example 4.4.

(a) It holds that $\langle \varnothing \rangle = \{0\}$, because the empty sum is the only linear combination from $\varnothing$.

(b) For $U \leq W \leq V$, it holds that $U + W = W$ and $\langle U \rangle = U = U \oplus \{0\}$.

(c) For $s_1, \ldots, s_n \in V$, it holds that $\langle s_i \rangle = \{\lambda s_i : \lambda \in K\} =: Ks_i$ and $\langle s_1, \ldots, s_n \rangle = Ks_1 + \ldots + Ks_n$. In particular, $\mathbb{R}^2 = \mathbb{R}(1, 0) \oplus \mathbb{R}(0, 1)$.

[1] In some books, one writes $\mathrm{Span}(S)$ instead of $\langle S \rangle$.
[2] This replaces the (disjoint) union, see Exercise I.12.

Definition 4.5. A subset S of a vector space V is called

- *generating set*, if $\langle S \rangle = V$. In the case $|S| < \infty$, V is called *finitely generated*.
- *linearly dependent*, if 0_V is a non-trivial linear combination of elements from S.
- *linearly independent*, if not linearly dependent, i. e. for pairwise distinct elements $s_1, \ldots, s_n \in S$ and $\lambda_1, \ldots, \lambda_n \in K$, it holds:

$$\sum_{i=1}^{n} \lambda_i s_i = 0 \qquad \Longrightarrow \qquad \lambda_1 = \ldots = \lambda_n = 0.$$

- *basis*, if S is a linearly independent generating set.

Remark 4.6. Since bases are sets, their elements do not have a fixed ordering. In fact, however, many theorems depend on the order of the basis elements. We therefore introduce the following terminology: vectors $s_1, \ldots, s_n$ are called linearly independent (or form a basis) if they are pairwise distinct and $\{s_1, \ldots, s_n\}$ is linearly independent (or a basis).

Example 4.7.

(a) The empty set is always linearly independent and forms a basis of the zero space.

(b) Because of $1_K \cdot 0_V = 0_V$, the zero vector is never part of a linearly independent set. A single vector $v \neq 0$, on the other hand, is always linearly independent, because from $\lambda v = 0$ with $\lambda \in K^\times$ follows the contradiction

$$v = 1v = (\lambda^{-1}\lambda)v = \lambda^{-1}(\lambda v) = \lambda^{-1}0 = 0.$$

(c) Vectors $v, w \in V \setminus \{0\}$ are linearly dependent if and only if $Kv = Kw$, i. e. v is a scalar multiple of w and vice versa.

(d) Every subset of a linearly independent set is linearly independent.

(e) For $n \geq 1$ let

$$\begin{aligned} e_1 &:= (1, 0, \ldots, 0), \\ e_2 &:= (0, 1, 0, \ldots, 0), \\ &\vdots \\ e_n &:= (0, \ldots, 0, 1) \end{aligned}$$

be vectors from K^n. Since every vector $v = (v_1, \ldots, v_n) \in K^n$ can be written in the form $v = \sum_{i=1}^{n} v_i e_i$, $\{e_1, \ldots, e_n\}$ is a generating system of K^n. From $v = 0 \Longleftrightarrow v_1 = \ldots = v_n = 0$ follows the linear independence of $\{e_1, \ldots, e_n\}$. One calls $e_1, \ldots, e_n$ the *standard basis* of K^n.

4.2 Characterization and Existence of Bases

Theorem 4.8. *Let $b_1, \dots, b_n$ be a basis of a K-vector space V. Then every $v \in V$ can be uniquely written in the form $v = \sum_{i=1}^{n} \lambda_i b_i$ with $\lambda_1, \dots, \lambda_n \in K$. In particular, the map*

$$ {}_B[.] \colon V \to K^n, \qquad v \mapsto {}_B[v] := (\lambda_1, \dots, \lambda_n) $$

is a bijection.

Proof. Because of $V = \langle b_1, \dots, b_n \rangle$, every $v \in V$ is a linear combination of the given form. Let $\lambda_1, \dots, \lambda_n, \mu_1, \dots, \mu_n \in K$ with

$$ v = \sum_{i=1}^{n} \lambda_i b_i = \sum_{i=1}^{n} \mu_i b_i. $$

Then $0 = v - v = \sum_{i=1}^{n} (\lambda_i - \mu_i) b_i$. Since $\{b_1, \dots, b_n\}$ is linearly independent, it follows that $\lambda_i = \mu_i$ for $i = 1, \dots, n$. □

Definition 4.9. In the situation of Theorem 4.8, ${}_B[v]$ is called the *coordinate representation* of v wrt. B.

Lemma 4.10. *For a vector space V and $B \subseteq V$, the following are equivalent:*

(1) B is a basis of V.

(2) B is a minimal generating system, i. e. for all $b \in B$, $B \setminus \{b\}$ is not a generating system.

(3) B is maximally linearly independent, i. e. for all $v \in V \setminus B$, $B \cup \{v\}$ is linearly dependent.

Proof. We perform a *circular proof*.[3]

(1) ⇒ (2): Let B be a basis, so in particular a generating set of V. Suppose that $B \setminus \{b\}$ is also a generating set for some $b \in B$. Then there exist $\lambda_1, \dots, \lambda_n \in K$ and $b_1, \dots, b_n \in B \setminus \{b\}$ with $b = \sum_{i=1}^{n} \lambda_i b_i$. Because of $-b + \sum_{i=1}^{n} \lambda_i b_i = 0$, B would then be linearly dependent. Contradiction.

(2) ⇒ (3): Let B be a minimal generating set. Let $\sum_{i=1}^{n} \lambda_i b_i = 0$ for $\lambda_1, \dots, \lambda_n \in K$ and pairwise distinct $b_1, \dots, b_n \in B$. If $\lambda_i \neq 0$ for some i, then

$$ b_i = -\lambda_i^{-1} \sum_{j \neq i} \lambda_j b_j = \sum_{j \neq i} (-\lambda_i^{-1} \lambda_j) b_j \in \langle B \setminus \{b_i\} \rangle. $$

But then $B \setminus \{b_i\}$ would also be a generating set. Thus $\lambda_1 = \ldots = \lambda_n = 0$ and B is linearly independent. Now let $v \in V \setminus B$. Because of $\langle B \rangle = V$, there exist $\lambda_1, \dots, \lambda_n \in K$ and $b_1, \dots, b_n \in B$ with $v = \sum_{i=1}^{n} \lambda_i b_i$ and $-v + \sum_{i=1}^{n} \lambda_i b_i = 0$. In particular, $B \cup \{v\}$ is linearly dependent.

(3) ⇒ (1): Let B be maximally linearly independent. We must show $\langle B \rangle = V$. Let $v \in V$. In the case $v \in B$, $v \in \langle B \rangle$. So let $v \notin B$. Then $B \cup \{v\}$ is linearly dependent. Thus there exist $\lambda_1, \dots, \lambda_n \in K^\times$, $\mu \in K$, $b_1, \dots, b_n \in B$ with $\mu v + \sum_{i=1}^{n} \lambda_i b_i = 0$. Since B is linearly independent, $\mu \neq 0$ must hold. This yields

$$ v = -\mu^{-1} \sum_{i=1}^{n} \lambda_i b_i = \sum_{i=1}^{n} (-\mu^{-1} \lambda_i) b_i \in \langle B \rangle. $$

Overall, $V = \langle B \rangle$. □

[3] A *circular reasoning* on the other hand is a flawed argument in which the claim is already assumed.

Theorem 4.11 (Basis Extension Theorem)**.** *Let V be a vector space with a finite generating set $E \subseteq V$. Then every linearly independent set $U \subseteq V$ can be extended to a basis of V by adding elements from E.*

Proof. Let $E = \{s_1, \ldots, s_n\}$. In the case $E \subseteq \langle U \rangle$, $V = \langle E \rangle \subseteq \langle U \rangle$, i.e., U is already a basis. So let $E \not\subseteq \langle U \rangle$ and wlog. $s_1 \notin \langle U \rangle$. As usual, $U_1 := U \cup \{s_1\}$ is then linearly independent. We can now repeat the argument with U_1 instead of U. In the case $E \subseteq \langle U_1 \rangle$, U_1 is a basis and otherwise we can assume $s_2 \notin \langle U_1 \rangle$. Then $U_2 := U_1 \cup \{s_2\}$ is linearly independent etc. Since E is finite, one obtains a basis of V after finitely many steps. □

Example 4.12. The linearly independent set $U := \big\{(1,2,0),(2,1,0)\big\} \subseteq \mathbb{R}^3$ can be extended to a basis with the standard basis vector e_3 (but not with e_1 or e_2).

Theorem 4.13 (Steinitz Exchange Lemma)**.** *Let V be a vector space with generating set E. For every linearly independent subset $U \subseteq V$, it holds that $|U| \leq |E|$.*

Proof. Wlog. let E be finite, say $E = \{s_1, \ldots, s_n\}$. Let $u_1, \ldots, u_m \in U$ be pairwise distinct. We must show $m \leq n$. Since U is linearly independent, $0 \neq u_1 = \sum_{i=1}^{n} \lambda_i s_i$, where not all $\lambda_1, \ldots, \lambda_n \in K$ vanish. So let wlog. $\lambda_1 \neq 0$ and therefore

$$s_1 = \lambda_1^{-1} u_1 + \sum_{i=2}^{n} (-\lambda_1^{-1} \lambda_i) s_i \in \langle u_1, s_2, \ldots, s_n \rangle.$$

Consequently, $\{u_1, s_2, \ldots, s_n\}$ is also a generating set with n elements (we have *exchanged* s_1 for u_1). Now write $u_2 = \mu_1 u_1 + \sum_{i=2}^{n} \mu_i s_i$ with $\mu_1, \ldots, \mu_n \in K$. Because $u_2 \notin \langle u_1 \rangle$, at least one μ_i with $i \geq 2$ must be non-zero. Say $\mu_2 \neq 0$. Because of

$$s_2 = -\mu_2^{-1} \mu_1 u_1 + \mu_2^{-1} u_2 - \sum_{i=3}^{n} \mu_2^{-1} \mu_i s_i \in \langle u_1, u_2, s_3, \ldots, s_n \rangle$$

one can exchange s_2 for u_2 in the same way. Repeating this process, one finally obtains the generating set $\{u_1, \ldots, u_m, s_{m+1}, \ldots, s_n\}$ of V. In particular, $m \leq n$. □

Example 4.14. The set $\big\{(1,2,3,4),(-1,4,0,2),(0,5,2,1),(0,0,-7,1),(-3,4,1,0)\big\} \subseteq \mathbb{R}^4$ must be linearly dependent, since $\{e_1, e_2, e_3, e_4\}$ is a generating set of $\mathbb{R}^4$ (note that no calculation is necessary).

Theorem 4.15. *Every finitely generated vector space has a finite basis and any two bases have the same cardinality.*

Proof. Let V be a vector space with a finite generating set E. According to the Basis Extension Theorem, the linearly independent set $\varnothing$ can be extended to a basis B of V using elements from E. In particular, $|B| \leq |E| < \infty$. Let C also be a basis of V. By the Exchange Lemma, $|C| \leq |B| \leq |C|$, hence $|C| = |B|$. Since B and C are finite, they must have the same cardinality according to Remark 2.11(e). □

Corollary 4.16. *Every subspace U of a finitely generated vector space V is finitely generated and possesses a* complement *$W \leq V$, i.e., $V = U \oplus W$ holds.*

Proof. Let B be a basis of V and $S \subseteq U$ be linearly independent. According to the exchange theorem, $|S| \leq |B| < \infty$ holds. In particular, U possesses a maximal linearly independent subset C. According to Lemma 4.10, C is a (finite) basis of U. Thus U is finitely generated. According to the basis extension theorem, C can be extended to a basis D of V. The second assertion then follows with $W := \langle D \setminus C \rangle$. □

Remark 4.17.

(a) In the situation of Corollary 4.16, W is in general not uniquely determined. For example,

$$\mathbb{R}^2 = \mathbb{R}(1,0) \oplus \mathbb{R}(0,1) = \mathbb{R}(1,0) \oplus \mathbb{R}(1,1).$$

(b) Using the axiom of choice (more precisely with Zorn's *Lemma*), one can show that every vector space has a (possibly infinite) basis and any two bases have the same cardinality. For example, $\mathbb{R}$ as a $\mathbb{Q}$-vector space has infinite bases, none of which can be explicitly given. We will consider in Theorem 6.12 how to efficiently calculate bases of finitely generated vector spaces.

4.3 Dimension

Definition 4.18. Let B be a basis of a finitely generated K-vector space V. Then

$$d = \dim_K V = \dim V := |B| \in \mathbb{N}_0$$

is called the *dimension* of V. According to Theorem 4.15, d does not depend on the choice of B. Instead of "finitely generated," one can now say *finite-dimensional* or more precisely *d-dimensional*.

Example 4.19.

(a) For every field K and $n \geq 1$, K^n has dimension n (choose the standard basis). The subspace $U := \{(x,x) \in K^2 : x \in K\} \leq K^2$ is 1-dimensional with basis $\{(1,1)\}$.

(b) In $\mathbb{R}^3$, $\{(x,y,0) : x,y \in \mathbb{R}\}$ describes a 2-dimensional plane. More generally, a $(d-1)$-dimensional subspace of a d-dimensional space is called a *hyperplane*.

(c) Let V be a d-dimensional $\mathbb{F}_2$-vector space. The coordinate representation wrt. a basis shows

$$|V| = |\mathbb{F}_2^d| = |\mathbb{F}_2 \times \ldots \times \mathbb{F}_2| = 2^d.$$

Remark 4.20.

(a) From the above theorems, several useful facts follow:

- For $U \leq V$, it holds that $\dim U \leq \dim V$ with equality if and only if $U = V$ (extend a basis of U to a basis of V).
- Every generating set E of V contains a basis of V (reduce to a minimal generating set). In particular, $|E| \geq \dim V$.
- $d+1$ vectors of a d-dimensional vector space are linearly dependent.

(b) In linear algebra, finite-dimensional vector spaces are the primary focus, while infinite-dimensional vector spaces are the subject of *functional analysis*.

(c) The following formula corresponds to the equation $|A \cup B| = |A| + |B| - |A \cap B|$ for finite sets A and B (Lemma 1.12).

Theorem 4.21 (Dimension Formula)**.** *For subspaces U and W of a finite-dimensional vector space V, it holds that*

$$\boxed{\dim(U+W) = \dim U + \dim W - \dim(U \cap W).}$$

If the sum is direct, then $\dim(U \oplus W) = \dim U + \dim W$.

Proof. Let $\{b_1, \ldots, b_n\}$ be a basis of $U \cap W$. We extend this to a basis $\{b_1, \ldots, b_n, c_1, \ldots, c_s\}$ of U and a basis $\{b_1, \ldots, b_n, d_1, \ldots, d_t\}$ of W. Since $U + W$ consists of elements of the form $u + w$ with $u \in U$ and $w \in W$, $U + W$ is generated by $b_1, \ldots, b_n, c_1, \ldots, c_s, d_1, \ldots, d_t$.

For linear independence, let $\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_s, \rho_1, \ldots, \rho_t \in K$ with

$$\underbrace{\sum_{i=1}^{n} \lambda_i b_i}_{=:v} + \underbrace{\sum_{j=1}^{s} \mu_j c_j}_{=:u} + \underbrace{\sum_{k=1}^{t} \rho_k d_k}_{=:w} = 0.$$

Then $v + u = -w \in U \cap W$. Thus $v + u$ can be expressed as a linear combination of $b_1, \ldots, b_n$. On the other hand, the representation of $v + u$ wrt. the basis $\{b_1, \ldots, b_n, c_1, \ldots, c_s\}$ is unique according to Theorem 4.8. This shows $\mu_1 = \ldots = \mu_s = 0$. Now $v + w = 0$ is a linear combination of the basis $\{b_1, \ldots, b_n, d_1, \ldots, d_t\}$. This is only possible if $\lambda_1 = \ldots = \lambda_n = \rho_1 = \ldots = \rho_t = 0$. Therefore $\{b_1, \ldots, b_n, c_1, \ldots, c_s, d_1, \ldots, d_t\}$ is linearly independent and consequently a basis of $U + W$. We obtain

$$\dim(U+W) = n + s + t = (n+s) + (n+t) - n = \dim U + \dim W - \dim(U \cap W).$$

If the sum is direct, then $U \cap W = \{0\}$ and the second assertion follows. □

Example 4.22. Let

$$U := \langle (1,1,0), (0,2,1) \rangle \leq \mathbb{R}^3,$$
$$W := \langle (1,1,1) \rangle \leq \mathbb{R}^3.$$

Apparently $\dim U = 2$ and $\dim W = 1$ holds. For $v \in U \cap W$ there exist $\lambda, \mu, \rho \in \mathbb{R}$ with

$$v = \lambda(1,1,0) + \mu(0,2,1) = \rho(1,1,1).$$

It follows that $(\lambda, \lambda + 2\mu, \mu) = (\rho, \rho, \rho)$. A comparison of coefficients yields $\lambda = \rho = \mu$ and $3\rho = \lambda + 2\mu = \rho$. This can only hold for $\rho = 0$. Thus $v = 0(1,1,1) = 0$ and $U \cap W = \{0\}$. One obtains $\dim(U+W) = \dim U + \dim W = 2 + 1 = 3$. Because of $U + W \leq \mathbb{R}^3$, it follows that $\mathbb{R}^3 = U \oplus W$.

Theorem 4.23. *For vector spaces $U \leq V$, it holds that* $\boxed{\dim V = \dim U + \dim(V/U).}$

Proof. Let W be a complement of U in V. Let B be a basis of W. It suffices to show that $\overline{B} := \{b + U : b \in B\}$ is a basis of V/U. Every element in V has the form $v = w + u$ with $u \in U$ and $w \in W$. Because of $v + U = w + U$, $\overline{B}$ is a generating set of V/U. Now let $\lambda_b \in K$ with $\sum_{b \in B} \lambda_b (b + U) = 0_{V/U}$. Then $\sum_{b \in B} \lambda_b b \in U$. From $U \cap W = \{0\}$ it follows that $\lambda_b = 0$ for all $b \in B$. Thus $\overline{B}$ is linearly independent. □

5 Matrices

5.1 The Matrix Vector Space

Remark 5.1. Unless stated otherwise, we henceforth tacitly assume that all vector spaces are finite-dimensional. A matrix is a scheme for the explicit calculation of bases of vector spaces and solutions of systems of linear equations. Matrices also appear as independent objects in numerous other fields.

Definition 5.2. Let K be a field and $n, m \in \mathbb{N}$. An $(n \times m\text{-})$*matrix* over K is a rectangularly arranged nm-tuple

$$A = (a_{ij})_{i,j=1}^{n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}$$

with $a_{ij} \in K$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The set of $n \times m$-matrices over K is denoted by $K^{n \times m}$. In the case $n = m$, A is called *square*.

Example 5.3.

(a) The 1×1-matrices correspond exactly to the elements of K. The vectors in K^n can be viewed as $1 \times n$-matrices. One then speaks of *row vectors*. The $n \times 1$-matrices are accordingly called *column vectors*. If misunderstandings are excluded, we use the standard basis $e_1, \ldots, e_n$ as both row and column vectors.

(b) The $n \times m$*-zero matrix* $0_{n \times m} := (0)_{i,j} \in K^{n \times m}$, $0_n := 0_{n \times n}$ (as usual, we omit the indices if misunderstandings are excluded).

(c) The (square) $n \times n$*-identity matrix*

$$1_n = (\delta_{ij}) := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The symbol δ_{ij} is called the Kronecker *delta*. It holds that

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

The rows of 1_n form the standard basis $\{e_1, \ldots, e_n\}$ of K^n.

(d) For $\lambda_1, \ldots, \lambda_n \in K$, one calls

$$\operatorname{diag}(\lambda_1, \ldots, \lambda_n) := (\delta_{ij}\lambda_i) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

a *diagonal matrix*. The entries $\lambda_1, \ldots, \lambda_n$ form the *main diagonal*. In the special case $\lambda_1 = \ldots = \lambda_n$, one speaks of *scalar matrices*.

(e) The $n \times m$-matrix E_{st} with a 1 at position (s, t) and otherwise only zeros. These are called *standard matrices*. With the Kronecker delta, $E_{st} = (\delta_{is}\delta_{jt})_{i,j}$ holds.

(f) For $A \in K^{n\times m}$, $A^{\mathrm{t}} := (a_{ji})_{i,j} \in K^{m\times n}$ is the *transpose matrix* of A. It is obtained from A by reflection across the main diagonal:

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \quad \longrightarrow \quad A^{\mathrm{t}} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$

Obviously $(A^{\mathrm{t}})^{\mathrm{t}} = A$. By transposing, row vectors become column vectors and vice versa.

(g) Square matrices A are called *symmetric* if $A^{\mathrm{t}} = A$.

Lemma 5.4. *With component-wise operations, $K^{n\times m}$ becomes an nm-dimensional K-vector space:*

$$A + B := (a_{ij} + b_{ij})_{i,j},$$
$$\lambda \cdot A := (\lambda a_{ij})_{i,j}$$

for $A = (a_{ij}), B = (b_{ij}) \in K^{n\times m}$ and $\lambda \in K$. The standard matrices E_{st} form a basis of $K^{n\times m}$. In particular, $\dim(K^{n\times m}) = nm$.

Proof. The defined operations on $K^{n\times m}$ correspond exactly to the operations in K^{nm} by arranging the vectors from K^{nm} as an $n \times m$ matrix. Since K^{nm} is a vector space, $K^{n\times m}$ must also be a vector space. The standard basis vectors $e_1, \ldots, e_{nm}$ of K^{nm} correspond (up to order) exactly to the standard matrices. □

Example 5.5.

(a) Attention: Only matrices of the same format can be added. For example:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} + \begin{pmatrix} 0 & -1 & 1 \\ 1 & -2 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 4 \\ 5 & 3 & 8 \end{pmatrix} \qquad 2 \cdot \begin{pmatrix} 1 & 2 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 0 & -6 \end{pmatrix}$$

The scalar matrices are exactly the scalar multiples of the identity matrix.

(b) The symmetric matrices form a subspace S of $K^{n\times n}$. A basis is obtained by the matrices $E_{11}, \ldots, E_{nn}$ and $E_{ij} + E_{ji}$ for $i < j$. In particular,

$$\dim S = n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}.$$

5.2 Matrix Multiplication

Definition 5.6. For $A = (a_{ij}) \in K^{n\times m}$ and $B = (b_{ij}) \in K^{m\times k}$, let $A \cdot B := (c_{ij})_{i,j} \in K^{n\times k}$ with

$$\boxed{c_{ij} := \sum_{l=1}^{m} a_{il}b_{lj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{im}b_{mj}.}$$

Remark 5.7.

(a) Mnemonic: c_{ij} is formed by "calculating" the i-th row of A with the j-th column of B:

$$\begin{pmatrix} * & * & * \\ * & * & * \\ a & b & c \\ * & * & * \end{pmatrix} \cdot \begin{pmatrix} * & a' \\ * & b' \\ * & c' \end{pmatrix} = \begin{pmatrix} * & * \\ * & * \\ * & aa' + bb' + cc' \\ * & * \end{pmatrix}$$

(the stars denote arbitrary entries). It is often helpful to imagine the following scheme:

$$\boxed{4 \times 3} \cdot \boxed{3 \times 2} = \boxed{4 \times 2}$$

(b) The multiplication of diagonal matrices is simple:

$$\operatorname{diag}(\lambda_1, \ldots, \lambda_n) \cdot \operatorname{diag}(\mu_1, \ldots, \mu_n) = \operatorname{diag}(\lambda_1\mu_1, \ldots, \lambda_n\mu_n).$$

(c) As a vector space, $(K^{n\times n}, +)$ is an abelian group. The following lemma shows that $(K^{n\times n}, +, \cdot)$ satisfies some, but not all field axioms.[1]

Lemma 5.8. *For all matrices A, B, C with "matching" format and $\lambda \in K$, the following hold:*

$$\boxed{\begin{array}{lll} A \cdot 1_m = A = 1_n \cdot A, & (AB)^{\mathrm{t}} = B^{\mathrm{t}}A^{\mathrm{t}}, & \lambda(AB) = (\lambda A)B = A(\lambda B), \\ A(BC) = (AB)C, & A(B + C) = AB + AC, & (A + B)C = AC + BC. \end{array}}$$

Proof. As usual, let $A = (a_{ij})$, $B = (b_{ij})$, and $C = (c_{ij})$. For an arbitrary matrix M, let M_{ij} be the entry at position (i, j). Then

$$\begin{aligned} (A1_m)_{ij} &= \sum_{k=1}^{m} a_{ik}\delta_{kj} = a_{ij} = \sum_{k=1}^{n} \delta_{ik}a_{kj} = (1_nA)_{ij}, \\ \big((AB)^{\mathrm{t}}\big)_{ij} &= \sum_{k=1}^{m} a_{jk}b_{ki} = \sum_{k=1}^{m} (B^{\mathrm{t}})_{ik}(A^{\mathrm{t}})_{kj} = (B^{\mathrm{t}}A^{\mathrm{t}})_{ij}, \\ \big(\lambda(AB)\big)_{ij} &= \lambda \sum_{k=1}^{m} a_{ik}b_{kj} = \sum_{k=1}^{m} (\lambda a_{ik})b_{kj} = \big((\lambda A)B\big)_{ij} = \sum_{k=1}^{m} a_{ik}(\lambda b_{kj}) = \big(A(\lambda B)\big)_{ij}, \\ \big(A(BC)\big)_{ij} &= \sum_{k=1}^{m} a_{ik}(BC)_{kj} = \sum_{k=1}^{m} a_{ik} \sum_{l=1}^{n} b_{kl}c_{lj} = \sum_{k=1}^{m}\sum_{l=1}^{n} a_{ik}b_{kl}c_{lj} \end{aligned}$$

[1] This weaker structure is called a *ring*.

$$= \sum_{l=1}^{n} \Big(\sum_{k=1}^{m} a_{ik} b_{kl} \Big) c_{lj} = \sum_{l=1}^{n} (AB)_{il} c_{lj} = ((AB)C)_{ij},$$

$$\big(A(B+C)\big)_{ij} = \sum_{k=1}^{m} a_{ik}(B+C)_{kj} = \sum_{k=1}^{m} a_{ik}(b_{kj} + c_{kj}) = \sum_{k=1}^{m} (a_{ik} b_{kj} + a_{ik} c_{kj})$$

$$= \sum_{k=1}^{m} a_{ik} b_{kj} + \sum_{k=1}^{m} a_{ik} c_{kj} = (AB)_{ij} + (AC)_{ij},$$

$$\big((A+B)C\big)_{ij} = \sum_{k=1}^{m} (a_{ik} + b_{ik}) c_{kj} = \sum_{k=1}^{m} a_{ik} c_{kj} + \sum_{k=1}^{m} b_{ik} c_{kj} = (AC)_{ij} + (BC)_{ij}. \qquad \square$$

Remark 5.9.

(a) For $A \in K^{n\times m}$ it holds, of course, that $0_{k\times n} \cdot A = 0_{k\times m}$ and $A \cdot 0_{m\times k} = 0_{n\times k}$.

(b) We call matrices $A, B \in K^{n\times n}$ *commuting*, if $AB = BA$. This is generally not satisfied for $n \geq 2$:

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = 0_2 \neq \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Therefore $(K^{n\times n}, +, \cdot)$ is *not* a field. We also see that not every matrix (different from 0_n) possesses an inverse wrt. $\cdot$ (even the cancellation rule does not hold for matrices).

(c) A square matrix $A \in K^{n\times n}$ is called *invertible*, if a matrix $B \in K^{n\times n}$ with

$$AB = 1_n = BA$$

exists.[2] If $AC = 1_n = CA$ also holds, then $C = C1_n = C(AB) = (CA)B = 1_n B = B$. Therefore B is uniquely determined and one writes, as with groups, $A^{-1} := B$. One calls A^{-1} the *inverse* matrix of A. We show in Lemma 5.15 that $AB = 1_n$ already implies invertibility, i. e. $BA = 1_n$ does not need to be checked.

(d) If A is invertible, then so is A^{t}, because

$$A^{\mathrm{t}}(A^{-1})^{\mathrm{t}} = (A^{-1}A)^{\mathrm{t}} = 1_n^{\mathrm{t}} = 1_n = (AA^{-1})^{\mathrm{t}} = (A^{-1})^{\mathrm{t}} A^{\mathrm{t}}.$$

One therefore sets $\boxed{A^{-\mathrm{t}} := (A^{-1})^{\mathrm{t}} = (A^{\mathrm{t}})^{-1}.}$

(e) Sometimes it is useful to divide matrices into rectangular *blocks*:

$$A = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} \qquad \text{with } A_1 \in K^{n_1\times m_1},\ A_2 \in K^{n_1\times m_2},\ A_3 \in K^{n_2\times m_1},\ A_4 \in K^{n_2\times m_2}$$

For another matrix $B = \left(\begin{smallmatrix} B_1 & B_2 \\ B_3 & B_4 \end{smallmatrix}\right)$ with blocks $B_1 \in K^{m_1\times k_1}$, $B_2 \in K^{m_1\times k_2}$, $B_3 \in K^{m_2\times k_1}$, $B_4 \in K^{m_2\times k_2}$ it then holds that

$$AB = \begin{pmatrix} A_1B_1 + A_2B_3 & A_1B_2 + A_2B_4 \\ A_3B_1 + A_4B_3 & A_3B_2 + A_4B_4 \end{pmatrix}.$$

If A_1 and A_4 are square and $A_2 = A_3 = 0$, then $A = \left(\begin{smallmatrix} A_1 & 0 \\ 0 & A_4 \end{smallmatrix}\right) = \operatorname{diag}(A_1, A_4)$ is called a *block diagonal matrix*.

[2] Invertible (or non-invertible) matrices are also called *regular* (or *singular*).

Lemma 5.10. *The invertible matrices in $K^{n\times n}$ form a group* $\mathrm{GL}(n,K)$ *wrt. multiplication. It is called the* general linear group of degree n *over* K.

Proof. The identity element 1_n is obviously invertible. With A, A^{-1} is also invertible. For invertible matrices A and B, it holds that

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = A1_nA^{-1} = AA^{-1} = 1_n$$

and analogously $(B^{-1}A^{-1})(AB) = 1_n$. This shows that AB is also invertible with $(AB)^{-1} = B^{-1}A^{-1}$. The associative law of multiplication follows from Lemma 5.8. □

Example 5.11. In $\mathrm{GL}(2,\mathbb{F}_2)$ we have

$$\begin{pmatrix}1&1\\1&0\end{pmatrix}\begin{pmatrix}0&1\\1&1\end{pmatrix} = 1_2 = \begin{pmatrix}0&1\\1&1\end{pmatrix}\begin{pmatrix}1&1\\1&0\end{pmatrix}.$$

5.3 The Rank of a Matrix

Theorem 5.12. *Let $z_1,\dots,z_n$ be the rows and $s_1,\dots,s_m$ be the columns of a matrix $A\in K^{n\times m}$. Then* $\dim\langle z_1,\dots,z_n\rangle = \dim\langle s_1,\dots,s_m\rangle$.

Proof. According to Remark 4.20 there exists $I\subseteq\{1,\dots,n\}$ such that $\{z_i : i\in I\}$ is a basis of $\langle z_1,\dots,z_n\rangle$. Let analogously $\{s_j : j\in J\}$ be a basis of $\langle s_1,\dots,s_m\rangle$. Let $A=(a_{ij})$. We show that the rows of the matrix

$$B := (a_{ij})_{i\in I, j\in J} \in K^{|I|\times|J|}$$

are linearly independent. For this, let $\lambda_i\in K$ with $\sum_{i\in I}\lambda_i a_{ij}=0$ for all $j\in J$. For $k\in\{1,\dots,m\}\setminus J$ there exist $\mu_j\in K$ with $s_k=\sum_{j\in J}\mu_j s_j$, i.e., $a_{ik}=\sum_{j\in J}\mu_j a_{ij}$ for all $i\in\{1,\dots,n\}$. It follows that

$$\sum_{i\in I}\lambda_i a_{ik} = \sum_{i\in I}\lambda_i\sum_{j\in J}\mu_j a_{ij} = \sum_{j\in J}\mu_j\sum_{i\in I}\lambda_i a_{ij} = 0.$$

Thus $\sum_{i\in I}\lambda_i a_{ij}=0$ holds for all $j\in\{1,\dots,m\}$, i.e., $\sum_{i\in I}\lambda_i z_i=0$. From the linear independence of $\{z_i : i\in I\}$ it follows that $\lambda_i=0$ for $i\in I$. Therefore the rows of B are linearly independent. Since they lie in the $|J|$-dimensional vector space $K^{|J|}$, it holds that $|I|\le|J|$. The rows (resp. columns) of A are the columns (resp. rows) of A^{t}. Using the above argument with A^{t}, it follows that $|J|\le|I|$. Overall, $\dim\langle z_1,\dots,z_n\rangle = |I| = |J| = \dim\langle s_1,\dots,s_m\rangle$. □

Definition 5.13. In the situation of Theorem 5.12, one calls

$$\mathrm{rk}(A) := \dim\langle z_1,\dots,z_n\rangle = \dim\langle s_1,\dots,s_m\rangle$$

the *rank* of A. In the case $\mathrm{rk}(A)=\min\{n,m\}$, one says: A has *full rank*.

Example 5.14.

(a) The identity matrix 1_n has (full) rank n, because its rows form the standard basis. It holds that $\mathrm{rk}(\begin{smallmatrix}1&2\\2&4\end{smallmatrix})=1$, because the second row is twice the first. We will consider in Remark 6.13 how to calculate the rank of any matrix efficiently.

(b) For every matrix A, it holds that $\operatorname{rk}(A) = \operatorname{rk}(A^{\mathrm{t}})$ and $\operatorname{rk}(A) = 0 \iff A = 0$.

Lemma 5.15.

(a) For matrices A and B with "matching" format, $\operatorname{rk}(AB) \leq \min\{\operatorname{rk}(A), \operatorname{rk}(B)\}$ holds.

(b) A square matrix is invertible if and only if it has full rank.

Proof.

(a) Let $s_1, \ldots, s_m$ be the columns of A and let $(\lambda_1, \ldots, \lambda_m)^{\mathrm{t}}$ be the i-th column of B. Then $\lambda_1 s_1 + \ldots + \lambda_m s_m$ is the i-th column of AB. Thus, the columns of AB are linear combinations of the columns of A. This shows $\operatorname{rk}(AB) \leq \operatorname{rk}(A)$. From Example 5.14 it follows that

$$\operatorname{rk}(AB) = \operatorname{rk}((AB)^{\mathrm{t}}) = \operatorname{rk}(B^{\mathrm{t}}A^{\mathrm{t}}) \leq \operatorname{rk}(B^{\mathrm{t}}) = \operatorname{rk}(B).$$

(b) If $A \in K^{n\times n}$ is invertible, then

$$n = \operatorname{rk}(1_n) = \operatorname{rk}(AA^{-1}) \overset{\text{(a)}}{\leq} \operatorname{rk}(A) \leq n,$$

i. e. A has full rank. Conversely, let $\operatorname{rk}(A) = n$. Then the standard basis vectors $e_1, \ldots, e_n$ can be expressed as linear combinations of the columns of A. Thus, there exists $B \in K^{n\times n}$ with $AB = 1_n$. Because $\operatorname{rk}(A^{\mathrm{t}}) = \operatorname{rk}(A)$, there also exists a $C \in K^{n\times n}$ with $A^{\mathrm{t}}C = 1_n$, i. e. $C^{\mathrm{t}}A = (A^{\mathrm{t}}C)^{\mathrm{t}} = 1_n^{\mathrm{t}} = 1_n$. Because $C^{\mathrm{t}} = C^{\mathrm{t}}(AB) = (C^{\mathrm{t}}A)B = B$, A is invertible. □

6 The Gaussian Algorithm

6.1 Systems of Equations

Definition 6.1. A (*linear*) *system of equations* is a matrix equation of the form $Ax = b$, where the *coefficient matrix* $A \in K^{n\times m}$ and the vector $b \in K^{n\times 1}$ are given. The goal is to find the *solution set*

$$L := \{x \in K^{m\times 1} : Ax = b\} \subseteq K^{m\times 1}.$$

- In the case $L \neq \varnothing$, the system of equations is called *solvable*.
- In the case $b = 0$, the system of equations is called *homogeneous* and otherwise *inhomogeneous*.
- By appending the column b to A, one obtains the *augmented* coefficient matrix $(A|b) \in K^{n\times(m+1)}$.

Example 6.2. The system of equations

$$\begin{aligned} 2x_1 + 3x_2 &= 5, \\ -x_1 \qquad &= 2 \end{aligned}$$

corresponds to the matrix equation

$$\begin{pmatrix} 2 & 3 \\ -1 & 0 \end{pmatrix} x = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

with exactly one solution $x = (-2, 3)^{\mathrm{t}}$.

Remark 6.3. Every homogeneous system of equations is solvable, because the zero vector is a solution.

Theorem 6.4 (Kronecker-Capelli[1])**.** *The system of equations $Ax = b$ is solvable if and only if* $\mathrm{rk}(A) = \mathrm{rk}(A|b)$.

Proof. Let $s_1, \ldots, s_m$ be the columns of A. Then

$$\begin{aligned} \mathrm{rk}(A) = \mathrm{rk}(A|b) &\iff \dim\langle s_1, \ldots, s_m\rangle = \dim\langle s_1, \ldots, s_m, b\rangle \\ &\iff \langle s_1, \ldots, s_m\rangle = \langle s_1, \ldots, s_m, b\rangle \iff b \in \langle s_1, \ldots, s_m\rangle \\ &\iff \exists x_1, \ldots, x_m \in K : b = \sum_{i=1}^{m} x_i s_i \iff \exists x \in K^{m\times 1} : Ax = b. \end{aligned}$$

□

[1] sometimes also named after Rouché and Frobenius

Remark 6.5. Let $A \in K^{n\times m}$ with full rank $n \leq m$. Then

$$\operatorname{rk}(A) \leq \operatorname{rk}(A|b) \leq \min\{n, m+1\} = n = \operatorname{rk}(A)$$

and $Ax = b$ is always solvable. In the case $n = m$, A is invertible according to Lemma 5.15 and $x = A^{-1}b$ is the unique solution. In the case $n < m$, the system of equations $Ax = b$ is called *underdetermined*, i.e., there are fewer equations than unknowns. We show that there are then multiple solutions.

Theorem 6.6. *Let $A \in K^{n\times m}$ and $b \in K^{n\times 1}$.*

(a) The solution set of the homogeneous system of equations $Ax = 0$ is a subspace L_0 of $K^{m\times 1}$ of dimension $m - \operatorname{rk}(A)$.

(b) If the system of equations $Ax = b$ has a solution $\widetilde{x}$, then

$$\widetilde{x} + L_0 := \{\widetilde{x} + y : y \in L_0\}$$

is the solution set.

Proof.

(a) Because of $0 \in L_0$, $L_0 \neq \varnothing$. For $x, y \in L_0$ and $\lambda \in K$, we have

$$A(\lambda x + y) = \lambda Ax + Ay = 0,$$

i.e., $\lambda x + y \in L_0$. This shows $L_0 \leq K^{m\times 1}$ (Remark 3.13). Let $b_1, \ldots, b_k$ be a basis of L_0. We extend it to a basis $b_1, \ldots, b_m$ of $K^{m\times 1}$. The i-th column of A is Ae_i with the standard basis vector e_i. Since e_i is a linear combination of $b_1, \ldots, b_m$, every column of A lies in $\langle Ab_1, \ldots, Ab_m\rangle$. In particular, $\operatorname{rk}(A) = \dim\langle Ab_1, \ldots, Ab_m\rangle$. From $b_1, \ldots, b_k \in L_0$ it follows that $Ab_1 = \ldots = Ab_k = 0$ and

$$\operatorname{rk}(A) = \dim\langle Ab_{k+1}, \ldots, Ab_m\rangle.$$

It suffices to show that the $m - k$ vectors $Ab_{k+1}, \ldots, Ab_m$ are linearly independent, because then $\operatorname{rk}(A) = m - k$. Let $\lambda_i \in K$ with $\sum_{i=k+1}^{m} \lambda_i Ab_i = 0$. Then also $A\sum_{i=k+1}^{m} \lambda_i b_i = 0$, i.e., $\sum_{i=k+1}^{m} \lambda_i b_i \in L_0 = \langle b_1, \ldots, b_k\rangle$. Since $b_1, \ldots, b_m$ is a basis of $K^{m\times 1}$, one obtains $\lambda_{k+1} = \ldots = \lambda_m = 0$ as desired.

(b) We have

$$Ax = b \iff Ax = A\widetilde{x} \iff A(x - \widetilde{x}) = 0 \iff x - \widetilde{x} \in L_0 \iff x \in \widetilde{x} + L_0. \qquad \square$$

Remark 6.7.

(a) If $A \in K^{n\times m}$ has full rank $m \leq n$, then the system of equations $Ax = b$ has at most one solution. In the case $m < n$, one speaks of *overdetermined* systems of equations. In the following, we deal with the explicit construction of the solution set of an arbitrary system of equations.

(b) Since the map $L_0 \to \widetilde{x} + L_0$, $v \mapsto \widetilde{x} + v$ is a bijection, a solvable system of equations has exactly as many solutions as the corresponding homogeneous system of equations. If K is finite (e.g., $K = \mathbb{F}_2$), then the number of these solutions is a power of $|K|$ (Theorem 4.8). For infinite fields such as $\mathbb{Q}$ or $\mathbb{R}$, every system of equations has no, exactly one, or infinitely many solutions.

6.2 Elementary Row Operations

Definition 6.8. The following transformations of a matrix $A \in K^{n\times m}$ are called (*elementary*) *row operations*:

- Multiplication of a row of A by a scalar $\lambda \in K^\times$. This corresponds to the multiplication by an *elementary matrix* of the form

$$\begin{pmatrix} 1_{s-1} & & 0 \\ & \lambda & \\ 0 & & 1_{n-s} \end{pmatrix} = 1_n + (\lambda - 1)E_{ss}$$

  from the left to A.

- Swapping two rows of A. This corresponds to the multiplication by an elementary matrix of the form

$$\begin{pmatrix} 1_{s-1} & & & & \\ & 0 & & 1 & \\ & & 1_{t-s-1} & & \\ & 1 & & 0 & \\ & & & & 1_{n-t} \end{pmatrix} = 1_n - E_{ss} - E_{tt} + E_{st} + E_{ts}$$

  from the left to A.

- Adding a multiple of one row of A to another row. This corresponds to the multiplication by an elementary matrix of the form

$$\begin{pmatrix} 1_{s-1} & & & & \\ & 1 & & \lambda & \\ & & 1_{t-s-1} & & \\ & & & 1 & \\ & & & & 1_{n-t} \end{pmatrix} = 1_n + \lambda E_{st} \qquad (\lambda \in K,\ s \neq t)$$

  from the left to A.

Matrices A and B are called *row-equivalent* if A can be transformed into B by finitely many elementary row operations. If applicable, we write $A \sim B$.

Remark 6.9.

(a) All elementary row operations are reversible. From $A \sim B$ it thus follows that $B \sim A$. Furthermore, the elementary matrices are invertible. According to Lemma 5.10, the product of elementary matrices is also invertible. From $A \sim B$ there follows the existence of a matrix $S \in \mathrm{GL}(n, K)$ with $SA = B$.

(b) According to (a), row equivalence is an equivalence relation on $K^{n\times m}$. In the next theorem, we determine a particularly simple system of representatives for the equivalence classes.

(c) Analogously, one defines (*elementary*) *column operations*. These correspond to the multiplication of elementary matrices from the *right* to A (try it out). Column operations can also be realized by row operations with A^{t}. We use the notation $A \sim B$ also when B results from A through column operations.

Theorem 6.10 (GAUSS algorithm[2])**.** *Every matrix $A \in K^{n\times m}$ is row-equivalent to exactly one matrix $\widehat{A}$ in* reduced row echelon form[3]*, i. e.*

$$\widehat{A} = \begin{pmatrix} 0 & \cdots & 0 & \boxed{1} & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \boxed{1} & * & \cdots & * & 0 & * & \cdots & * \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & \boxed{1} & * & \cdots & * \\ \vdots & & & & & & & & & & & & & & \vdots \end{pmatrix}.$$

Proof. The following algorithm transforms A into $\widehat{A}$:

(1) Set $z := 1$ (row index).

(2) For $s = 1, \ldots, m$ (column index):

- If $\exists i \geq z : a_{is} \neq 0$, then:
  - Swap the i-th with the z-th row. Subsequently, $a_{zs} \neq 0$ holds.
  - Divide the z-th row by a_{zs}. Subsequently, $a_{zs} = 1$ holds.
  - For $j = 1, \ldots, z-1, z+1, \ldots, n$ subtract a_{js} times the z-th row from the j-th row. Subsequently, $a_{js} = 0$ holds.
  - Increase z by 1.

For the uniqueness of $\widehat{A}$, let B and C be matrices in reduced row echelon form with $A \sim B$ and $A \sim C$. Then $B \sim C$ also holds and there exists $S \in \mathrm{GL}(n, K)$ with $SB = C$. Let b_i (resp. s_i) be the i-th column of B (resp. S). Then $c_i = Sb_i$ is the i-th column of C. Let $e_1, \ldots, e_n$ be the standard basis of K^n. Let $b_i \in \langle e_1, \ldots, e_k \rangle$. We show $b_i = c_i$ and $s_k = e_k^{\mathrm{t}}$ by induction on k. In the case $k = 0$, $b_i \in \langle \varnothing \rangle = \{0\}$, so $b_i = 0$. Then certainly $c_i = Sb_i = 0$ as well. Now let the claim be already proven up to $k-1$. The first column (from the left) of B that does not lie in $\langle e_1, \ldots, e_{k-1} \rangle$ is $b_i = e_k$ due to the reduced row echelon form. The columns of S are linearly independent since S is invertible. This shows

$$c_i = Sb_i = Se_k^{\mathrm{t}} = s_k \notin \langle s_1, \ldots, s_{k-1} \rangle = \langle e_1, \ldots, e_{k-1} \rangle.$$

Thus c_i is the first column of C that does not lie in $\langle e_1, \ldots, e_{k-1} \rangle$, i.e., $c_i = e_k = s_k = b_i$ (reduced row echelon form). For every further column $b_j \in \langle e_1, \ldots, e_k \rangle$, it now holds that $c_j = Sb_j = (e_1, \ldots, e_k, s_{k+1}, \ldots, s_n)b_j = b_j$. Thus $b_i = c_i$ for $i = 1, \ldots, m$, i.e., $B = C$. □

Example 6.11.

$$\begin{pmatrix} 0 & 1 & 1 & 3 \\ 2 & 1 & 3 & 0 \\ 3 & -1 & 2 & 1 \end{pmatrix} \overset{\leftarrow\!\rceil}{\leftarrow\!\rceil} \sim \begin{pmatrix} 2 & 1 & 3 & 0 \\ 0 & 1 & 1 & 3 \\ 3 & -1 & 2 & 1 \end{pmatrix} \mid :2 \sim \begin{pmatrix} 1 & 1/2 & 3/2 & 0 \\ 0 & 1 & 1 & 3 \\ 3 & -1 & 2 & 1 \end{pmatrix} \overset{\rceil\!\rceil -3}{\leftarrow\!\rceil +}$$

$$\sim \begin{pmatrix} 1 & 1/2 & 3/2 & 0 \\ 0 & 1 & 1 & 3 \\ 0 & -5/2 & -5/2 & 1 \end{pmatrix} \overset{\leftarrow\!\rceil +}{\rceil\!\rceil -1/2} \overset{5/2}{\leftarrow\!\rceil +} \sim \begin{pmatrix} 1 & 0 & 1 & -3/2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 17/2 \end{pmatrix} \mid :17/2$$

[2] also called *Gauss elimination* or *Gauss-Jordan algorithm*

[3] Every non-zero row contains a *leading one*. All entries to the left, above, and below a leading one are 0. The leading ones move further to the right with each row. Zero rows (if present) are at the bottom.

$$\sim \begin{pmatrix} 1 & 0 & 1 & -3/2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{matrix} \\ \leftarrow + \\ \leftarrow -3 \end{matrix} \sim \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

6.3 Applications

Theorem 6.12. *Let $U := \langle u_1, \ldots, u_n \rangle \leq K^m$. Let $A \in K^{n \times m}$ be the matrix with rows $u_1, \ldots, u_n$. Then the non-zero rows of $\widehat{A}$ form a basis of U. In particular,* $\dim U = \operatorname{rk}(\widehat{A}) = \operatorname{rk}(A)$.

Proof. Through elementary row operations, rows of A are replaced by linear combinations of rows. The rows of $\widehat{A}$ therefore generate a subspace $W \leq U$. Since all row operations are reversible, it even holds that $W = U$ and $\dim U = \operatorname{rk}(A) = \operatorname{rk}(\widehat{A})$. Let k be the number of non-zero rows in $\widehat{A}$. Then $\operatorname{rk}(\widehat{A}) \leq k$. On the other hand, $\widehat{A}$ possesses the linearly independent columns $e_1, \ldots, e_k$. This shows $\operatorname{rk}(\widehat{A}) = k$ and the claim follows. □

Remark 6.13. To determine a basis of U, one does not need to create zeros above the leading ones during the Gaussian algorithm (the non-zero rows are still linearly independent, as their number remains the same). Furthermore, it is advisable to avoid divisions by swapping with rows that already have a leading one in the current column. If one is only interested in $\dim U$ (or more generally in $\operatorname{rk}(A)$), then because of $\operatorname{rk}(A) = \operatorname{rk}(A^{\mathrm{t}})$, one may also use elementary column operations. This can be useful if A has fewer columns than rows (many possibilities lead to the goal).

Example 6.14. A kind of chess puzzle: Rank in two moves!

$$\overset{\overset{+}{\frown}}{\begin{pmatrix} 1 & -1 & 0 \\ 2 & 0 & 2 \\ 3 & -1 & 2 \\ -1 & 2 & 1 \end{pmatrix}} \sim \overset{\overset{-1\ +}{\frown}}{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 2 \\ 3 & 2 & 2 \\ -1 & 1 & 1 \end{pmatrix}} \sim \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 2 & 0 \\ -1 & 1 & 0 \end{pmatrix} \implies \operatorname{rk}(A) = 2$$

Theorem 6.15. *Let $Ax = b$ be a system of equations with $A \in K^{n \times m}$. Let $(\widehat{A}|c)$ be the row echelon form of $(A|b)$. Then:*

(a) $Ax = b$ is solvable if and only if e_{m+1} is not a row of $(\widehat{A}|c)$.

If applicable, the solution set is obtained as follows: Let $(1, n_1), \ldots, (k, n_k)$ be the positions of the leading ones in $(\widehat{A}|c)$. The $n - k$ zero rows are deleted. For all $i \in \{1, \ldots, m\} \setminus \{n_1, \ldots, n_k\}$, we insert the row $-e_i$ at position i, so that the resulting matrix $M \in K^{m \times (m+1)}$ has only entries ± 1 on the main diagonal.

(b) The last column $\widetilde{x}$ of M satisfies $A\widetilde{x} = b$.

(c) The $m - k$ columns of M that do not belong to the indices $n_1, \ldots, n_k$ form a basis of $L_0 := \{x \in K^{m \times 1} : Ax = 0\}$.

(d) The solution set of $Ax = b$ is $\widetilde{x} + L_0$.

Proof. Let $S \in \mathrm{GL}(n, K)$ with $(SA|Sb) = S(A|b) = (\widehat{A}|c)$. For $x \in K^{m\times 1}$ it holds that

$$Ax = b \iff SAx = Sb \iff \widehat{A}x = c.$$

If e_{m+1} is a row of $(\widehat{A}|c)$, then in the equation $\widehat{A}x = c$ one obtains the contradiction $0 = 1$, i. e. there is no solution.

Now let e_{m+1} not be a row of $(\widehat{A}|c)$. We verify the claims using the following example

$$(\widehat{A}|c) = \left(\begin{array}{ccccc|c} . & 1 & a_1 & . & a_2 & c_1 \\ . & . & . & 1 & a_3 & c_2 \\ . & . & . & . & . & . \end{array}\right) \qquad M = \begin{pmatrix} -1 & . & . & . & . & . \\ . & 1 & a_1 & . & a_2 & c_1 \\ . & . & -1 & . & . & . \\ . & . & . & 1 & a_3 & c_2 \\ . & . & . & . & -1 & . \end{pmatrix}$$

(here . stands for 0 for better clarity). One easily sees that $\widehat{A}\widetilde{x} = c$ holds. Thus $A\widetilde{x} = b$ also holds. This proves (a) and (b). Similarly, one sees that the (marked in red) columns s_1, s_3 and s_5 of M lie in L_0. The different positions of the entries -1 in s_i imply the linear independence of $\{s_1, s_3, s_5\}$. On the other hand, according to Theorem 6.6 and Remark 6.13

$$\dim L_0 = m - \mathrm{rk}(A) = m - \mathrm{rk}(\widehat{A}) = m - k = 3.$$

This shows (c). From Theorem 6.6 follows (d). □

Example 6.16.

$$\begin{pmatrix} -1 & 3 & 1 & 1 & 0 \\ -2 & 1 & -3 & 2 & -4 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} x = \begin{pmatrix} 11 \\ -6 \\ 4 \end{pmatrix}$$

$$(A|b) = \left(\begin{array}{ccccc|c} -1 & 3 & 1 & 1 & 0 & 11 \\ -2 & 1 & -3 & 2 & -4 & -6 \\ 0 & 1 & 1 & 0 & 0 & 4 \end{array}\right) \sim \cdots \sim \begin{pmatrix} 1 & 0 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$

$$M = \begin{pmatrix} 1 & 0 & 2 & -1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 4 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$

$$L = \widetilde{x} + L_0 = \begin{pmatrix} 1 \\ 4 \\ 0 \\ 0 \\ 2 \end{pmatrix} + \left\langle \begin{pmatrix} 2 \\ 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 0 \\ -1 \\ 0 \end{pmatrix} \right\rangle$$

Check:

$$A\widetilde{x} = \begin{pmatrix} -1 & 3 & 1 & 1 & 0 \\ -2 & 1 & -3 & 2 & -4 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \\ 0 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 11 \\ -6 \\ 4 \end{pmatrix} = b.$$

Theorem 6.17 (Matrix Inversion)**.** *For $A \in K^{n\times n}$ let $(\widehat{A}|B)$ be the row echelon form of $(A|1_n) \in K^{n\times 2n}$. A is invertible if and only if $\widehat{A} = 1_n$. If so, $A^{-1} = B$.*

Proof. It holds that

$$A \text{ invertible} \overset{5.15}{\iff} \operatorname{rk}(A) = n \overset{6.12}{\iff} \operatorname{rk}(\widehat{A}) = n \iff \widehat{A} = 1_n.$$

Now let $S \in \mathrm{GL}(n, K)$ with

$$(SA|S) = S(A|1_n) = (\widehat{A}|B) = (1_n|B).$$

Then $B = S = S(AA^{-1}) = (SA)A^{-1} = 1_nA^{-1} = A^{-1}$. □

Corollary 6.18. *Every invertible matrix is a product of elementary matrices.*

Proof. For $A \in \mathrm{GL}(n, K)$ it holds that $A \sim 1_n$ according to Theorem 6.17. □

Example 6.19.

$$(A|1_3) = \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 & 1 & 0 \\ -1 & 1 & -2 & 0 & 0 & 1 \end{array}\right) \overset{\text{II}+\text{I}}{\underset{\text{III}+\text{I}}{\sim}} \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \end{array}\right) \overset{\text{III}}{\underset{\text{II}}{\sim}}$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 1 & 0 \end{array}\right) \overset{\text{III}+\text{II}}{\sim} \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & 0 & -3 & 2 & 1 & 1 \end{array}\right) \begin{array}{l} \\ \\ \mid : (-3) \end{array}$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & 0 & 1 & -2/3 & -1/3 & -1/3 \end{array}\right) \overset{\text{I}+\text{III}}{\underset{\text{II}+3\,\text{III}}{\sim}} \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/3 & -1/3 & -1/3 \\ 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & -2/3 & -1/3 & -1/3 \end{array}\right)$$

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

Remark 6.20.

(a) If during the Gaussian algorithm an "offset" of the rows occurs

$$\begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & 1 & \\ & & & & & * \end{pmatrix},$$

then the row echelon form must have a zero row. The matrix then cannot be invertible and one can terminate the algorithm prematurely.

(b) Matrices $A, B \in K^{n\times m}$ are called *equivalent*, if A can be transformed into B by row and column operations (and vice versa). This means there exist $S \in \mathrm{GL}(n,K)$ and $T \in \mathrm{GL}(m,K)$ with $SAT = B$. Obviously, this defines an equivalence relation. It is easy to see that every matrix A is equivalent to a matrix of the form

$$\begin{pmatrix} 1_r & 0 \\ 0 & 0 \end{pmatrix}$$

where $r = \mathrm{rk}(A)$ is uniquely determined. In particular, A and B are equivalent if and only if $\mathrm{rk}(A) = \mathrm{rk}(B)$ holds. Thus, no new symbol needs to be introduced for equivalence. The number of equivalence classes of $n \times n$-matrices is $n + 1$.

(c) The following theorem provides an efficient algorithm to simultaneously determine the sum and intersection of subspaces.

Theorem 6.21 (Zassenhaus algorithm)**.** *Let $U := \langle u_1, \ldots, u_s\rangle \leq K^n$ and $W := \langle w_1, \ldots, w_t\rangle \leq K^n$. Let*

$$A := \begin{pmatrix} u_1 & u_1 \\ \vdots & \vdots \\ u_s & u_s \\ w_1 & 0 \\ \vdots & \vdots \\ w_t & 0 \end{pmatrix} \in K^{(s+t)\times 2n}, \qquad \widehat{A} = \begin{pmatrix} b_1 & * \\ \vdots & \vdots \\ b_k & * \\ 0 & c_1 \\ \vdots & \vdots \\ \vdots & c_l \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix},$$

where $b_1, \ldots, b_k, c_1, \ldots, c_l \in K^n$ with $b_k \neq 0 \neq c_l$. Then $\{b_1, \ldots, b_k\}$ is a basis of $U + W$ and $\{c_1, \ldots, c_l\}$ is a basis of $U \cap W$.

Proof. Because of $U + W = \langle u_1, \ldots, u_s, w_1, \ldots, w_t\rangle$, $\{b_1, \ldots, b_k\}$ is a basis of $U + W$ according to Theorem 6.12. Furthermore, every row of the form $(0, c_m)$ of $\widehat{A}$ is a linear combination of the rows of A, say

$$(0, c_m) = \sum_{i=1}^{s} \lambda_i (u_i, u_i) + \sum_{j=1}^{t} \mu_j (w_j, 0)$$

with $\lambda_1, \ldots, \lambda_s, \mu_1, \ldots, \mu_t \in K$. This shows

$$c_m = \sum_{i=1}^{s} \lambda_i u_i = -\sum_{j=1}^{t} \mu_j w_j \in U \cap W$$

for $m = 1, \ldots, l$. Due to the row echelon form, $\{c_1, \ldots, c_l\}$ is linearly independent. By elementary *column* operations, one transforms A to

$$\begin{pmatrix} 0 & u_1 \\ \vdots & \vdots \\ 0 & u_s \\ w_1 & 0 \\ \vdots & \vdots \\ w_t & 0 \end{pmatrix}.$$

From Remark 6.13 it now follows easily that $\mathrm{rk}(A) = \dim U + \dim W$. The dimension formula therefore yields

$$\dim(U \cap W) = \dim U + \dim W - \dim(U + W) = \mathrm{rk}(A) - k = \mathrm{rk}(\widehat{A}) - k = l.$$

Thus $\{c_1, \ldots, c_l\}$ is a basis of $U \cap W$. □

Example 6.22. Let $U := \langle (1,1,1,0), (0,-4,1,5) \rangle$ and $W := \langle (0,-2,1,2), (1,-1,1,3) \rangle$. As usual, one does not have to perform all steps of the Gaussian algorithm:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & -4 & 1 & 5 & 0 & -4 & 1 & 5 \\ 0 & -2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 3 & 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & -2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & -4 & 1 & 5 & 0 & -4 & 1 & 5 \\ 0 & -2 & 0 & 3 & -1 & -1 & -1 & 0 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & -2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -4 & 1 & 5 \\ 0 & 0 & -1 & 1 & -1 & -1 & -1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & -2 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -4 & 1 & 5 \\ 0 & 0 & 0 & 0 & -1 & 3 & -2 & -5 \end{pmatrix}$$

It follows that $U + W = \langle (1,1,1,0), (0,-2,1,2), (0,0,-1,1) \rangle$ and $U \cap W = \langle (-1,3,-2,-5) \rangle$.

7 Linear Maps

7.1 Definitions and Examples

Remark 7.1. In order to relate different vector spaces V and W, we study maps $V \to W$ that "respect" addition and scalar multiplication. It will be shown that such maps can be described by matrices.

Definition 7.2. A map $f\colon V \to W$ between K-vector spaces V and W is called *linear* or a *homomorphism*, if for all $u, v \in V$ and $\lambda \in K$ the following holds:

$$\boxed{f(\lambda u + v) = \lambda f(u) + f(v).}$$

The set of linear maps $V \to W$ is denoted by $\mathrm{Hom}(V, W)$. If f is linear and bijective, then f is called an *isomorphism*. If applicable, V and W are called *isomorphic* and we write $V \cong W$.

Remark 7.3.

(a) A map $f\colon V \to W$ is linear if and only if

$$\begin{aligned} f(u + v) &= f(u) + f(v), \\ f(\lambda u) &= \lambda f(u) \end{aligned}$$

holds for all $u, v \in V$ and $\lambda \in K$ (set $\lambda = 1$ or $v = 0$ in Definition 7.2; cf. Remark 3.13). Isomorphic vector spaces therefore differ only by the naming of their elements.

(b) For $f \in \mathrm{Hom}(V, W)$, it holds that

$$f(0_V) = f(0_K \cdot 0_V) = 0_K f(0_V) = 0_W.$$

(c) Let $f \in \mathrm{Hom}(U, V)$ and $g \in \mathrm{Hom}(V, W)$. For $u, u' \in U$ and $\lambda \in K$, it holds that

$$(g \circ f)(\lambda u + u') = g(\lambda f(u) + f(u')) = \lambda g(f(u)) + g(f(u')) = \lambda (g \circ f)(u) + (g \circ f)(u').$$

This shows $g \circ f \in \mathrm{Hom}(U, W)$.

(d) If $f\colon V \to W$ is an isomorphism, then $f^{-1}\colon W \to V$ is also an isomorphism[1], because for $w, w' \in W$ and $\lambda \in K$ it holds that

$$\begin{aligned} f^{-1}(\lambda w + w') &= f^{-1}\big(\lambda f(f^{-1}(w)) + f(f^{-1}(w'))\big) \\ &= f^{-1}\big(f(\lambda f^{-1}(w) + f^{-1}(w'))\big) = \lambda f^{-1}(w) + f^{-1}(w'). \end{aligned}$$

(e) The isomorphism of vector spaces is an equivalence relation.[2] Reflexivity follows from the isomorphism id_V (Example 7.4), symmetry follows from (c), and transitivity follows from (d).

[1] In analysis, the inverse map of a bijective continuous function is in general not continuous. Therefore, there is the strange term *homeomorphism* (not a typo).

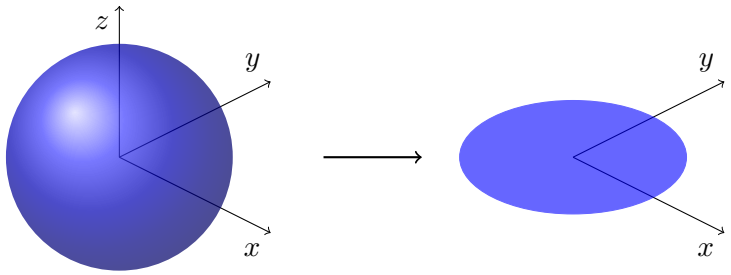
[2] However, the totality of all vector spaces is not a set, but a *class*.

Example 7.4.

(a) The *zero map* $0\colon V \to W$, $v \mapsto 0_W$ is always linear. The identity $\mathrm{id}_V\colon V \to V$ is an isomorphism.

(b) For $f \in \mathrm{Hom}(V,W)$ and $U \leq V$, the restriction $f_{|U}$ is linear. In particular, the inclusion map $U \to V$ is linear as a restriction of id_V.

(c) For $n \leq m$, the *projection* $K^m \to K^n$, $(x_1,\dots,x_m) \mapsto (x_1,\dots,x_n)$ is a surjective homomorphism. The projection $\mathbb{R}^3 \to \mathbb{R}^2$ reduces a 3-dimensional object to its "shadow":



(d) Let $f\colon \mathbb{R} \to \mathbb{R}$ be linear and $a := f(1)$. For $x \in \mathbb{R}$, it holds that $f(x) = f(x1) = xf(1) = ax$. The graph of f therefore describes a line through the origin. Note: In school mathematics, functions of the form $f(x) = ax + b$ are sometimes also referred to as "linear" (such maps are called *affine*[3], see Exercise I.20).

(e) For $A \in K^{n\times m}$, the map $K^{m\times 1} \to K^{n\times 1}$, $x \mapsto Ax$ is linear according to Lemma 5.8. We show in Theorem 7.18 that every linear map (after choosing a basis) has this form.

(f) The transposition $K^{n\times m} \to K^{m\times n}$, $A \mapsto A^{\mathrm{t}}$ is an isomorphism.

(g) For every n-element set $M = \{x_1,\dots,x_n\}$, the map $\mathrm{Fun}(M,V) \to V^n$, $f \mapsto (f(x_1),\dots,f(x_n))$ is an isomorphism.

Lemma 7.5. *For $f \in \mathrm{Hom}(V,W)$, $V_1 \leq V$ and $W_1 \leq W$, it holds that $f(V_1) \leq W$ and $f^{-1}(W_1) \leq V$. In particular, $f(V) \leq W$ and $\mathrm{Ker}(f) := f^{-1}(\{0\}) \leq V$.*

Proof. Because of $0 = f(0) \in f(V_1)$, we have $f(V_1) \neq \varnothing$. For $u, v \in V_1$ and $\lambda \in K$, it holds that

$$\lambda f(u) + f(v) = f(\lambda u + v) \in f(V_1).$$

This shows $f(V_1) \leq W$ (Remark 3.13). Because of $0 \in f^{-1}(\{0\}) \subseteq f^{-1}(W_1)$, we also have $f^{-1}(W_1) \neq \varnothing$. For $u, w \in f^{-1}(W_1)$ and $\lambda \in K$, it holds that $f(\lambda u + w) = \lambda f(u) + f(w) \in W_1$, i. e. $\lambda u + w \in f^{-1}(W_1)$. This shows $f^{-1}(W_1) \leq V$. □

Definition 7.6. In the situation of Lemma 7.5, $\mathrm{Ker}(f)$ is called the *kernel* of f and

$$\mathrm{rk}(f) := \dim f(V)$$

the *rank* of f. For $A \in K^{n\times m}$, let $\mathrm{Ker}(A) := \{x \in K^{m\times 1} : Ax = 0\}$ be the *kernel* of A.

Lemma 7.7. *$f \in \mathrm{Hom}(V,W)$ is injective if and only if $\mathrm{Ker}(f) = \{0\}$. In this case, $V \to f(V)$, $v \mapsto f(v)$ is an isomorphism.*

[3] An example is the conversion from degrees Celsius to Fahrenheit: $f(x) = \frac{9}{5}x + 32$.

Proof. Let f be injective and $v \in \mathrm{Ker}(f)$. From $f(v) = 0 = f(0)$ it follows that $v = 0$, i.e., $\mathrm{Ker}(f) = \{0\}$. Conversely, let $\mathrm{Ker}(f) = \{0\}$ and $u, v \in V$ with $f(u) = f(v)$. Then $f(u - v) = f(u) - f(v) = 0$, so $u - v \in \mathrm{Ker}(f) = \{0\}$. Therefore $u = v$ and f is injective. The second statement is trivial. □

Theorem 7.8. *Let V and W be vector spaces. Let $b_1, \ldots, b_n$ be a basis of V and let $c_1, \ldots, c_n \in W$ be arbitrary. Then there exists exactly one linear map $f\colon V \to W$ with $f(b_i) = c_i$ for $i = 1, \ldots, n$. The following hold:*

(a) f is injective $\iff$ $c_1, \ldots, c_n$ are linearly independent.

(b) f is surjective $\iff$ $W = \langle c_1, \ldots, c_n \rangle$.

(c) f is an isomorphism $\iff$ $c_1, \ldots, c_n$ is a basis of W.

Proof. Every $u \in V$ can be uniquely written in the form $u = \sum_{i=1}^{n} \lambda_i b_i$. We define

$$f(u) := \sum_{i=1}^{n} \lambda_i c_i \in W.$$

For $v = \sum_{i=1}^{n} \mu_i b_i$ and $\rho \in K$ it holds that

$$f(\rho u + v) = f\Big(\sum_{i=1}^{n} (\rho\lambda_i + \mu_i) b_i\Big) = \sum_{i=1}^{n} (\rho\lambda_i + \mu_i) c_i = \rho \sum_{i=1}^{n} \lambda_i c_i + \sum_{i=1}^{n} \mu_i c_i = \rho f(u) + f(v).$$

Thus f is linear with $f(b_i) = c_i$ for $i = 1, \ldots, n$. If $g \in \mathrm{Hom}(V, W)$ is also such that $g(b_i) = c_i$ for $i = 1, \ldots, n$, then

$$g(u) = \sum_{i=1}^{n} \lambda_i g(b_i) = \sum_{i=1}^{n} \lambda_i c_i = \sum_{i=1}^{n} \lambda_i f(b_i) = f(u)$$

for all $u \in V$. Thus $g = f$ and f is uniquely determined.

(a) Let f be injective and $\sum_{i=1}^{n} \lambda_i c_i = 0$ for $\lambda_i \in K$. Then

$$f\Big(\sum_{i=1}^{n} \lambda_i b_i\Big) = \sum_{i=1}^{n} \lambda_i c_i = 0.$$

From Lemma 7.7 it follows that $\sum_{i=1}^{n} \lambda_i b_i \in \mathrm{Ker}(f) = \{0\}$. Since $b_1, \ldots, b_n$ are linearly independent, one obtains $\lambda_1 = \ldots = \lambda_n = 0$. Thus $c_1, \ldots, c_n$ are linearly independent. Conversely, let $c_1, \ldots, c_n$ be linearly independent and $u := \sum_{i=1}^{n} \lambda_i b_i \in \mathrm{Ker}(f)$. Then $\sum_{i=1}^{n} \lambda_i c_i = f(u) = 0$ and one obtains $\lambda_1 = \ldots = \lambda_n = 0$. Therefore $u = 0$ and $\mathrm{Ker}(f) = \{0\}$. By Lemma 7.7, f is injective.

(b) Let f be surjective and $w \in W$. Then there exists a $v = \sum_{i=1}^{n} \lambda_i b_i \in V$ with $f(v) = w$. It follows that

$$w = f(v) = \sum_{i=1}^{n} \lambda_i c_i \in \langle c_1, \ldots, c_n \rangle.$$

Conversely, let $W = \langle c_1, \ldots, c_n \rangle$ and $w \in W$. Then there exist $\lambda_i \in K$ with $w = \sum_{i=1}^{n} \lambda_i c_i$. For $v := \sum_{i=1}^{n} \lambda_i b_i \in V$ it then holds that $f(v) = w$, i. e. f is surjective.

(c) Follows from (a) and (b). □

Remark 7.9. Let $f \in \mathrm{Hom}(V, W)$. In the case $\dim V < \dim W$, f is not surjective, because the image of a basis of V cannot be a generating set of W. In the case $\dim V > \dim W$, f is not injective, because the image of a basis cannot be linearly independent. For $\dim V = \dim W$ one obtains

$$f \text{ injective} \iff f \text{ surjective} \iff f \text{ bijective}$$

(cf. Remark 2.11(e)).

Theorem 7.10. *Two K-vector spaces are isomorphic if and only if they have the same dimension. In particular, every n-dimensional K-vector space is isomorphic to K^n.*

Proof. Let $f\colon V \to W$ be an isomorphism of vector spaces and B a basis of V. According to Theorem 7.8, $f(B)$ is a basis of W. Thus $\dim V = |B| = |f(B)| = \dim W$ holds. Conversely, if V and W have the same dimension, then there exist bases $\{b_1, \ldots, b_n\}$ and $\{c_1, \ldots, c_n\}$ of V and W respectively. According to Theorem 7.8, there exists an isomorphism $f\colon V \to W$ with $f(b_i) = c_i$ for $i = 1, \ldots, n$. The second assertion follows from $\dim K^n = n$. An explicit isomorphism is obtained by the coordinate representation $V \to K^n$, $v \mapsto {}_B[v]$ (it maps B to the standard basis of K^n). □

Remark 7.11.

(a) The vector spaces $\{0\}, K, K^2, \ldots$ form a system of representatives for the isomorphism classes of finite-dimensional K-vector spaces.

(b) For K-vector spaces $V_1, \ldots, V_n$ with $d_i := \dim V_i$, it holds that $V_1 \times \ldots \times V_n \cong K^{d_1} \times \ldots \times K^{d_n} \cong K^{d_1 + \ldots + d_n}$.

(c) Although $\mathbb{Q}$ and $\mathbb{Q}^2$ have the same cardinality, $\mathbb{Q} \ncong \mathbb{Q}^2$ holds according to Theorem 7.10.

(d) Let $U, V \leq W$ be vector spaces. Obviously $V \leq U + V$ and $U \cap V \leq U$ hold. From Theorem 4.23 and the dimension formula, it follows that

$$\dim((U+V)/V) = \dim(U+V) - \dim(V) = \dim(U) - \dim(U \cap V) = \dim(U/(U \cap V)).$$

With Theorem 7.10 one obtains the so-called *first isomorphism theorem*[4][5]

$$\boxed{(U+V)/V \cong U/(U \cap V).}$$

(e) For vector spaces $U \leq V \leq W$, obviously $V/U \leq W/U$ holds. From

$$\begin{aligned}\dim\big((W/U)/(V/U)\big) &= \dim(W/U) - \dim(V/U) \\ &= \dim(W) - \dim(U) - \dim(V) + \dim(U) = \dim(W/V)\end{aligned}$$

one obtains the *second isomorphism theorem*[6]

$$\boxed{(W/U)/(V/U) \cong W/V.}$$

[4] Mnemonic: On one side there are two U, on the other side two V.
[5] Some authors use different numberings for the isomorphism theorems, see Wikipedia.
[6] Mnemonic: Canceling a double fraction.

Theorem 7.12 (Homomorphism Theorem)**.** *For $f \in \mathrm{Hom}(V,W)$, the map*

$$\begin{aligned} \overline{f}\colon V/\operatorname{Ker}(f) &\to f(V), \\ v + \operatorname{Ker}(f) &\mapsto f(v) \end{aligned}$$

is an isomorphism. Thus $\boxed{V/\operatorname{Ker}(f) \cong f(V)}$ *and* $\boxed{\dim V = \operatorname{rk}(f) + \dim \operatorname{Ker}(f)}$ *hold.*

Proof. For $v, w \in V$ it holds that

$$\begin{aligned} v + \operatorname{Ker}(f) = w + \operatorname{Ker}(f) &\iff v - w \in \operatorname{Ker}(f) \iff f(v-w) = 0 \\ &\iff f(v) = f(w) \iff \overline{f}(v + \operatorname{Ker}(f)) = \overline{f}(w + \operatorname{Ker}(f)). \end{aligned}$$

The implication $\Rightarrow$ shows that $\overline{f}$ is well-defined, while the implication $\Leftarrow$ shows that $\overline{f}$ is injective. Obviously $\overline{f}$ is also linear and surjective. The equation for $\dim V$ follows from Theorem 4.23. □

7.2 Representation Matrices

Theorem 7.13. *For K-vector spaces V and W, $\mathrm{Hom}(V,W)$ is a subspace of $\mathrm{Fun}(V,W)$.*

Proof. Certainly the neutral element $f = 0$ of $\mathrm{Fun}(V,W)$ lies in $\mathrm{Hom}(V,W)$. Let $f, g \in \mathrm{Hom}(V,W)$, $u, v \in V$ and $\lambda, \mu \in K$. Due to

$$\begin{aligned} (f+g)(\mu u + v) &= f(\mu u + v) + g(\mu u + v) = \mu f(u) + f(v) + \mu g(u) + g(v) = \mu(f+g)(u) + (f+g)(v), \\ (\lambda f)(\mu u + v) &= \lambda f(\mu u + v) = \lambda(\mu f(u) + f(v)) = \mu\lambda f(u) + \lambda f(v) = \mu(\lambda f)(u) + (\lambda f)(v) \end{aligned}$$

$f + g$ and λf are linear. □

Remark 7.14. If B is a basis of V, then the restriction map $\mathrm{Hom}(V,W) \to \mathrm{Fun}(B,W)$, $f \mapsto f_{|B}$ is an isomorphism according to Theorem 7.8.

Definition 7.15. Let V and W be vector spaces with bases $B = \{b_1, \ldots, b_m\}$ and $C = \{c_1, \ldots, c_n\}$ respectively. Let $f \in \mathrm{Hom}(V,W)$ and $f(b_i) = \sum_{j=1}^{n} a_{ji} c_j$ with $a_{ji} \in K$ for $i = 1, \ldots, m$.

(a) One calls ${}_C[f]_B := (a_{ij}) \in K^{n \times m}$ the *representation matrix* of f wrt. B and C.

(b) In the case $V = W$ and $f = \mathrm{id}_V$, one calls ${}_C\Delta_B := {}_C[\mathrm{id}_V]_B$ the *basis change matrix* wrt. B and C.

(c) If B and C are the standard bases of $V = K^m$ and $W = K^n$, then one sets $[f] := {}_C[f]_B$. Mnemonic: The columns of $[f]$ are the images of the standard basis.

Remark 7.16. In the situation of Definition 7.15, ${}_B\Delta_B = 1_m$ holds.

Example 7.17. The map

$$f\colon \mathbb{R}^2 \to \mathbb{R}^3, \qquad (x,y) \mapsto (2x-y,\ y,\ -3x)$$

is linear with matrix

$$[f] = \begin{pmatrix} 2 & -1 \\ 0 & 1 \\ -3 & 0 \end{pmatrix}.$$

Obviously $B := \big\{(1,-1),(0,2)\big\}$ and $C := \big\{(1,1,1),(0,-1,1),(1,0,1)\big\}$ are bases of $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively. Due to

$$\begin{aligned} f(1,-1) &= (3,-1,-3) = -7(1,1,1) - 6(0,-1,1) + 10(1,0,1), \\ f(0,2) &= (-2,2,0) = 4(1,1,1) + 2(0,-1,1) - 6(1,0,1) \end{aligned}$$

it follows that

$${}_C[f]_B = \begin{pmatrix} -7 & 4 \\ -6 & 2 \\ 10 & -6 \end{pmatrix}$$

(in case of doubt, you must determine the entries by a system of equations).

Theorem 7.18. *Let V be an m-dimensional vector space with basis B and let W be an n-dimensional vector space with basis C. Then the map*

$${}_C[.]_B\colon \operatorname{Hom}(V,W) \to K^{n\times m}, \qquad f \mapsto {}_C[f]_B$$

is an isomorphism with

$$\boxed{{}_C[f(v)]^{\mathrm{t}} = {}_C[f]_B\,{}_B[v]^{\mathrm{t}}}$$

$$\begin{array}{ccc} V & \xrightarrow{\ f\ } & W \\ {\scriptstyle {}_B[.]}\downarrow & & \downarrow{\scriptstyle {}_C[.]} \\ K^m & \xrightarrow[{}_C[f]_B]{} & K^n \end{array}$$

for all $v \in V$. In particular, $\dim \operatorname{Hom}(V,W) = nm$ *and* $\operatorname{rk}(f) = \operatorname{rk}({}_C[f]_B)$.

Proof. Let $B = \{b_1,\dots,b_m\}$ and $C = \{c_1,\dots,c_n\}$. According to Theorem 7.8, ${}_C[.]_B$ is a bijection. Let $f, g \in \operatorname{Hom}(V,W)$ with ${}_C[f]_B = (a_{ij})$ and ${}_C[g]_B = (b_{ij})$. For $i = 1,\dots,m$ and $\lambda \in K$ it holds that

$$(\lambda f + g)(b_i) = \lambda f(b_i) + g(b_i) = \lambda \sum_{j=1}^{n} a_{ji}c_j + \sum_{j=1}^{n} b_{ji}c_j = \sum_{j=1}^{n} (\lambda a_{ji} + b_{ji})c_j.$$

This shows ${}_C[\lambda f + g]_B = \lambda\,{}_C[f]_B + {}_C[g]_B$. Thus ${}_C[.]_B$ is an isomorphism. Let $v = \sum_{i=1}^{m} v_i b_i$. Then

$$f(v) = \sum_{i=1}^{m} v_i f(b_i) = \sum_{i=1}^{m} v_i \sum_{j=1}^{n} a_{ji}c_j = \sum_{j=1}^{n} \Big(\sum_{i=1}^{m} a_{ji}v_i\Big)c_j$$

and it follows that

$${}_C[f]_B\,{}_B[v]^{\mathrm{t}} = \Big(\sum_{i=1}^{m} a_{ji}v_i\Big)_j^{\mathrm{t}} = {}_C[f(v)]^{\mathrm{t}}.$$

Clearly ${}_B[b_i] = e_i$ is the i-th standard basis vector. Thus ${}_C[f(b_i)] = {}_C[f]_B\,{}_B[b_i]^{\mathrm{t}}$ is the i-th column of ${}_C[f]_B$. Since ${}_C[.]$ is an isomorphism, it holds that

$$\operatorname{rk}(f) = \dim\langle f(b_1),\dots,f(b_m)\rangle = \dim\langle {}_C[f(b_1)],\dots,{}_C[f(b_m)]\rangle = \operatorname{rk}({}_C[f]_B). \qquad \square$$

Remark 7.19.

(a) For $V = W$ and $f = \mathrm{id}_V$ one obtains ${}_C[v]^{\mathrm{t}} = {}_C\Delta_{BB}[v]^{\mathrm{t}}$. For $f\colon K^n \to K^m$ it holds that $f(v)^{\mathrm{t}} = [f]v^{\mathrm{t}}$ wrt. the standard bases.

(b) Let $U \leq V$. According to Corollary 4.16, U has a complement W, i.e., $V = U \oplus W$. The projection $f\colon V \to W$, $u + w \mapsto w$ for $u \in U$ and $w \in W$ is a homomorphism with $\mathrm{Ker}(f) = U$. Therefore, every subspace is the kernel of a homomorphism. Let B and C be bases of V and W respectively. For $A := {}_C[f]_B$ it holds that $\mathrm{Ker}(f) = \{v \in V : A_B[v]^{\mathrm{t}} = 0\}$ according to Theorem 7.18. Thus U can be described as the solution set of a linear system of equations.

Example 7.20. Let f, B and C be as in Example 7.17. For $v := (2,4) = 2(1,-1) + 3(0,2) \in \mathbb{R}^2$ it holds that

$$f(v)^{\mathrm{t}} = [f]v^{\mathrm{t}} = \begin{pmatrix} 2 & -1 \\ 0 & 1 \\ -3 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \\ -6 \end{pmatrix},$$

$${}_C[f(v)]^{\mathrm{t}} = {}_C[f]_B{}_B[v]^{\mathrm{t}} = \begin{pmatrix} -7 & 4 \\ -6 & 2 \\ 10 & -6 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} -2 \\ -6 \\ 2 \end{pmatrix}.$$

Check: $-2(1,1,1) - 6(0,-1,1) + 2(1,0,1) = (0,4,-6) = f(v)$.

Corollary 7.21. *Let $f \in \mathrm{Hom}(V,W)$ and $A := {}_C[f]_B \in K^{n\times m}$ wrt. arbitrary bases. Then*

(a) f injective $\iff \mathrm{rk}(A) = m$.

(b) f surjective $\iff \mathrm{rk}(A) = n$.

Proof.

(a) f injective $\overset{7.7}{\iff} \mathrm{Ker}(A) = \mathrm{Ker}(f) = \{0\} \overset{6.6}{\iff} \mathrm{rk}(A) = m.$

(b) f surjective $\iff f(V) = W \iff \mathrm{rk}(A) = \mathrm{rk}(f) = \dim W = n.$ □

Theorem 7.22. *Let U, V and W be vector spaces with bases B, C and D respectively. Let $f \in \mathrm{Hom}(U,V)$ and $g \in \mathrm{Hom}(V,W)$. Then*

$$\boxed{{}_D[g \circ f]_B = {}_D[g]_C{}_C[f]_B.}$$

Proof. According to Remark 7.3, $g \circ f \in \mathrm{Hom}(U,W)$. Let $B = \{b_1, \ldots, b_m\}$, $C = \{c_1, \ldots, c_n\}$ and $D = \{d_1, \ldots, d_k\}$. Let ${}_C[f]_B = (a_{ij})$ and ${}_D[g]_C = (b_{ij})$. Then

$$(g \circ f)(b_i) = g\Big(\sum_{j=1}^{n} a_{ji}c_j\Big) = \sum_{j=1}^{n} a_{ji}g(c_j) = \sum_{j=1}^{n} a_{ji} \sum_{l=1}^{k} b_{lj}d_l = \sum_{l=1}^{k} \Big(\sum_{j=1}^{n} b_{lj}a_{ji}\Big)d_l.$$

Therein, $\sum_{j=1}^{n} b_{lj}a_{ji}$ is the entry of ${}_D[g]_C{}_C[f]_B$ at position (l,i) as desired. □

Remark 7.23. Mnemonic: The composition of linear maps corresponds to the multiplication of matrices. For linear maps f, g, h between "matching" spaces, the distributive laws of matrices carry over:

$$(f+g) \circ h = (f \circ h) + (g \circ h) \qquad f \circ (g+h) = (f \circ g) + (f \circ h).$$

This can, of course, also be verified by direct calculation.

Example 7.24. Let f, B and C be as in Example 7.17. Let $g \in \operatorname{Hom}(\mathbb{R}^3, \mathbb{R}^2)$ with matrix ${}_B[g]_C = -\begin{pmatrix} 3 & 0 & 2 \\ 1 & 1/2 & 1 \end{pmatrix}$. Then

$$ {}_B[g \circ f]_B = {}_B[g]_C {}_C[f]_B = -\begin{pmatrix} 3 & 0 & 2 \\ 1 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} -7 & 4 \\ -6 & 2 \\ 10 & -6 \end{pmatrix} = 1_2 = {}_B[\mathrm{id}_{\mathbb{R}^2}]_B, $$

i. e. $g \circ f = \mathrm{id}_{\mathbb{R}^2}$. Conversely, $f \circ g \neq \mathrm{id}_{\mathbb{R}^3}$, because f cannot be surjective.

Corollary 7.25. *Let $f\colon V \to W$ be an isomorphism between vector spaces V and W with bases B and C respectively. Then* $\boxed{{}_B[f^{-1}]_C = {}_C[f]_B^{-1}.}$ *In the case $V = W$, ${}_C\Delta_B^{-1} = {}_B\Delta_C$.*

Proof. According to Remark 7.3, $f^{-1} \in \operatorname{Hom}(W, V)$. From Theorem 7.22 it follows that

$$ {}_C[f]_B {}_B[f^{-1}]_C = {}_C[\mathrm{id}_W]_C = {}_C\Delta_C = 1_n $$

and ${}_B[f^{-1}]_C = {}_C[f]_B^{-1}$. The second statement follows with $f = \mathrm{id}_V$. □

Remark 7.26. The isomorphisms $f\colon V \to V$ form the *general linear group* $\mathrm{GL}(V)$. They correspond exactly to the invertible matrices, i. e. ${}_B[.]_B\colon \mathrm{GL}(V) \to \mathrm{GL}(n, K)$ is an isomorphism of groups (instead of vector spaces).

Corollary 7.27 (Change of basis)**.** *Let B, B' be bases of V and C, C' be bases of W. For $f \in \operatorname{Hom}(V, W)$ it holds that*

$$ \boxed{{}_{C'}[f]_{B'} = {}_{C'}\Delta_C {}_C[f]_B {}_B\Delta_{B'}.} $$

In the case $V = W$:

$$ \boxed{{}_{B'}[f]_{B'} = {}_{B'}\Delta_B {}_B[f]_{BB'} \Delta_B^{-1}.} $$

Proof. From Theorem 7.22 it follows that

$$ {}_{C'}\Delta_C {}_C[f]_B {}_B\Delta_{B'} = {}_{C'}[\mathrm{id}_W]_C {}_C[f]_B {}_B[\mathrm{id}_V]_{B'} = {}_{C'}[\mathrm{id}_W]_C {}_C[f]_{B'} = {}_{C'}[f]_{B'}. $$

For $V = W$, $C = B$ and $C' = B'$ one obtains ${}_{B'}[f]_{B'} = {}_{B'}\Delta_B {}_B[f]_B {}_B\Delta_{B'} = {}_{B'}\Delta_B {}_B[f]_{BB'} \Delta_B^{-1}$ using Corollary 7.25. □

Example 7.28. Let again $f\colon \mathbb{R}^2 \to \mathbb{R}^3$ be as in Example 7.17. We replace the basis $B = \{(1, -1), (0, 2)\}$ with $B' := \{(0, 1), (1, 1)\}$. Because of $(0, 1) = 0(1, -1) + \frac{1}{2}(0, 2)$ and $(1, 1) = (1, -1) + (0, 2)$, it holds that

$$ {}_B\Delta_{B'} = \begin{pmatrix} 0 & 1 \\ 1/2 & 1 \end{pmatrix}, \quad {}_C[f]_{B'} = {}_C[f]_B {}_B\Delta_{B'} = \begin{pmatrix} -7 & 4 \\ -6 & 2 \\ 10 & -6 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1/2 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ 1 & -4 \\ -3 & 4 \end{pmatrix}. $$

Definition 7.29. Matrices $A, B \in K^{n \times n}$ are called *similar*, if there exists an $S \in \mathrm{GL}(n, K)$ with $B = SAS^{-1}$. We write $A \approx B$ if necessary.

Remark 7.30.

(a) Obviously $A \approx A$ (choose $S = 1_n$). From $B = SAS^{-1}$ it follows that $A = S^{-1}BS$. From $B = SAS^{-1}$ and $C = TBT^{-1}$ with $T \in \mathrm{GL}(n, K)$ it follows that $C = TSAS^{-1}T^{-1} = (TS)A(TS)^{-1}$. Therefore, the similarity of matrices is an equivalence relation.

(b) According to Theorem 7.18 and Corollary 7.27, every endomorphism of V determines, by the choice of a basis, a similarity class of representation matrices in $K^{n\times n}$. This allows concrete calculations with abstract maps. In Linear Algebra II, we construct special bases such that the representation matrices have the "simplest" possible form (for example, diagonal matrices). This speeds up calculations.

Definition 7.31. For $A = (a_{ij})_{i,j} \in K^{n\times n}$, we call $\mathrm{tr}(A) := \sum_{i=1}^{n} a_{ii}$ the *trace* of A. This is the sum of the main diagonal entries.

Lemma 7.32. *The map* $\mathrm{tr}\colon K^{n\times n} \to K$ *is linear with* $\mathrm{tr}(A^{\mathrm{t}}) = \mathrm{tr}(A)$ *and* $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ *for* $A, B \in K^{n\times n}$.

Proof. For $A = (a_{ij})$, $B = (b_{ij})$ and $\lambda \in K$, it holds that

$$\mathrm{tr}(\lambda A + B) = \mathrm{tr}\big((\lambda a_{ij} + b_{ij})_{i,j}\big) = \sum_{i=1}^{n}(\lambda a_{ii} + b_{ii}) = \lambda \sum_{i=1}^{n} a_{ii} + \sum_{i=1}^{n} b_{ii} = \lambda\,\mathrm{tr}(A) + \mathrm{tr}(B).$$

Thus tr is linear. Since a reflection across the main diagonal does not change the diagonal itself, $\mathrm{tr}(A^{\mathrm{t}}) = \mathrm{tr}(A)$ holds. Furthermore,

$$\mathrm{tr}(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} = \mathrm{tr}(BA). \qquad \square$$

Corollary 7.33. *Similar matrices have the same trace.*

Proof. For $A \in K^{n\times n}$ and $S \in \mathrm{GL}(n, K)$, it holds that $\mathrm{tr}(S(AS^{-1})) = \mathrm{tr}((AS^{-1})S) = \mathrm{tr}(A)$. $\square$

Definition 7.34. Let V be a vector space with basis B and $f \in \mathrm{Hom}(V, V)$. We call $\mathrm{tr}(f) := \mathrm{tr}({}_B[f]_B)$ the *trace* of f. According to Corollary 7.27 and Corollary 7.33, $\mathrm{tr}(f)$ does not depend on the choice of the basis.

Theorem 7.35 (Fillmore)**.** *Let* $A \in K^{n\times n} \setminus K1_n$ *and* $d_1, \ldots, d_n \in K$ *with* $\mathrm{tr}(A) = d_1 + \ldots + d_n$. *Then* A *is similar to a matrix with main diagonal* $d_1, \ldots, d_n$.

Proof. Induction on n: Since $A \notin K1_n$, we have $n \geq 2$. Let $A = (a_{ij})$. If $a_{ij} \neq 0$ for some $i \neq j$, then $Ae_j \notin \langle e_j\rangle$. If A is a diagonal matrix, then there exist $i \neq j$ with $a_{ii} \neq a_{jj}$. In this case,

$$A(e_i + e_j) = a_{ii}e_i + a_{jj}e_j \notin \langle e_i + e_j\rangle.$$

In any case, there exists a $b_1 \in K^n$ such that b_1 and Ab_1 are linearly independent. Then b_1 and $b_2 := Ab_1 - d_1b_1$ are also linearly independent. We extend b_1, b_2 to a basis $b_1, \ldots, b_n$ of K^n. wrt. this basis, A has the form $\left(\begin{smallmatrix} d_1 & * \\ * & A_1 \end{smallmatrix}\right)$ with $A_1 = (a'_{ij}) \in K^{(n-1)\times(n-1)}$. In the case $n = 2$, we are finished, because $\mathrm{tr}(A_1) = \mathrm{tr}(A) - d_1 = d_2$.

Now let $n \geq 3$. If $A_1 \in K1_{n-1}$, then

$$A(b_1 + b_3) = d_1 b_1 + b_2 + Ab_3 \in b_2 + \langle b_1, b_3 \rangle.$$

By replacing b_3 with $b_1 + b_3$, we achieve $a'_{23} = 1$ and $A_1 \notin K1_{n-1}$. By induction, there exists $S \in \mathrm{GL}(n-1, K)$ such that SA_1S^{-1} has main diagonal $d_2, \ldots, d_n$. Now

$$\begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} d_1 & * \\ * & A_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & S^{-1} \end{pmatrix} = \begin{pmatrix} d_1 & * \\ * & SA_1S^{-1} \end{pmatrix}$$

has main diagonal $d_1, \ldots, d_n$. □

Example 7.36. Let $A := \left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right) \in \mathbb{Q}^{2\times 2}$ and $(d_1, d_2) = (0, 3)$. We set $b_1 := (1, 1)$ and $b_2 := Ab_1 = (1, 2)$. Let E be the standard basis of $\mathbb{Q}^2$ and $B := \{b_1, b_2\}$. For $S := {}_E\Delta_B = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, we have

$$S^{-1}AS = \begin{pmatrix} 0 & -2 \\ 1 & 3 \end{pmatrix}$$

according to Corollary 7.27.

7.3 Dual Spaces

Definition 7.37. For a K-vector space V, $V^* := \mathrm{Hom}(V, K)$ is called the *dual space* of V. Its elements are called (*linear*) *functionals*.

Lemma 7.38. *Let $b_1, \ldots, b_n$ be a basis of V. For $i = 1, \ldots, n$, let $b_i^* \in V^*$ with $b_i^*(b_j) = \delta_{ij}$ for $j = 1, \ldots, n$. Then $b_1^*, \ldots, b_n^*$ is a basis of V^*.*

Proof. Let $\lambda_1, \ldots, \lambda_n \in K$ with $f := \lambda_1 b_1^* + \ldots + \lambda_n b_n^* = 0$. For $i = 1, \ldots, n$, we have $0 = f(b_i) = \lambda_i$. Therefore, $b_1^*, \ldots, b_n^*$ are linearly independent. The claim follows from $\dim V^* = \dim V$ (Theorem 7.18). □

Example 7.39. If $e_1, \ldots, e_n$ is the standard basis of $V = K^n$, then $e_1^*, \ldots, e_n^*$ are the projections, i. e. $e_i^*(v_1, \ldots, v_n) = v_i$ for $i = 1, \ldots, n$.

Remark 7.40.

(a) In the situation of Lemma 7.38, $b_1^*, \ldots, b_n^*$ is called the *dual basis* to $b_1, \ldots, b_n$.

(b) For infinite-dimensional spaces V, $V^* \ncong V$, because V^* is "larger" than V (without proof).

(c) According to (the proof of) Theorem 7.13, there exists an isomorphism $V \to V^*$ that maps a basis to the corresponding dual basis. However, there is no *canonical* isomorphism that does not depend on a choice of basis (cf. Theorem 12.7). The next theorem shows that the situation between V and the *bidual space* $V^{**} := (V^*)^*$ is better.

Theorem 7.41. *For $v \in V$, let $F_v \colon V^* \to K$, $f \mapsto f(v)$. Then $F \colon V \to V^{**}$, $v \mapsto F_v$ is a canonical isomorphism.*

Proof. For $f_1, f_2 \in V^*$ and $\lambda \in K$, we have

$$F_v(\lambda f_1 + f_2) = (\lambda f_1 + f_2)(v) = \lambda f_1(v) + f_2(v) = \lambda F_v(f_1) + F_v(f_2),$$

i. e. $F_v \in V^{**}$. For $v, w \in V$, it holds that

$$F_{\lambda v+w}(f) = f(\lambda v + w) = \lambda f(v) + f(w) = \lambda F_v(f) + F_w(f) = (\lambda F_v + F_w)(f),$$

i. e. F is linear. Now let $F_v = 0$. In the case $v \neq 0$, one can extend v to a basis of V. For the dual basis, we would then have $0 = F_v(v^*) = v^*(v) = 1$. This contradiction shows $\operatorname{Ker}(F) = \{0\}$, i. e. F is injective. Due to $\dim V^{**} = \dim V^* = \dim V < \infty$, F must also be surjective. □

Definition 7.42. For $U \leq V$ and $W \leq V^*$, let

$$U^0 := \{f \in V^* : f(U) = \{0\}\} \subseteq V^*,$$
$$W_0 := \{v \in V : \forall f \in W : f(v) = 0\} \subseteq V.$$

We call U^0 (resp. W_0) the *dual complement* of U (resp. W).

Lemma 7.43.

(a) For $U \leq V$, $U^0 \leq V^*$ with $\dim V = \dim U + \dim U^0$.

(b) For $U \leq V^*$, $U_0 \leq V$ with $\dim V = \dim U + \dim U_0$.

(c) The maps $U \mapsto U^0$ and $U \mapsto U_0$ are inverse to each other.

(d) For $U, W \leq V$, it holds that $\boxed{(U+W)^0 = U^0 \cap W^0}$ and $\boxed{(U \cap W)^0 = U^0 + W^0.}$

(e) For $U, W \leq V^*$, it holds that $\boxed{(U+W)_0 = U_0 \cap W_0}$ and $\boxed{(U \cap W)_0 = U_0 + W_0.}$

(f) It holds that $V = U \oplus W \iff V^* = U^0 \oplus W^0$.

Proof.

(a) The restriction $F\colon V^* \to U^*$, $f \mapsto f_{|U}$ is a homomorphism with kernel $U^0 \leq V^*$. Since every functional in U^* can be extended to V^* (basis extension), F is surjective. From the homomorphism theorem, it follows that $\dim V = \dim V^* = \dim U^0 + \dim U^* = \dim U^0 + \dim U$.

(b) The construction from (a) first yields $U^0 = \{f \in V^{**} : f(U) = \{0\}\} \leq V^{**}$ with $\dim V = \dim U + \dim U^0$. For the isomorphism $F\colon V \to V^{**}$, $v \mapsto F_v$ from Theorem 7.41, we have

$$v \in U_0 \iff F_v(U) = \{0\} \iff F_v \in U^0.$$

Thus $U_0 = F^{-1}(U^0) \leq V$ with $\dim U_0 = \dim U^0$.

(c) By definition, $U \leq (U^0)_0$ and $U \leq (U_0)^0$. For dimension reasons, equality holds.

(d) We have $f \in (U+W)^0 \iff f(U) = \{0\} = f(W) \iff f \in U^0 \cap W^0$. This shows the first equation. From $U \cap W \subseteq U, W$ it follows that $U^0 + W^0 = \langle U^0 \cup W^0 \rangle \subseteq (U \cap W)^0$. According to (a) and the dimension formula, we have

$$\begin{aligned}\dim(U^0 + W^0) &= \dim U^0 + \dim W^0 - \dim(U^0 \cap W^0) \\ &= 2\dim V - \dim U - \dim W - \dim((U+W)^0)\end{aligned}$$

$$= \dim V - \dim U - \dim W + \dim(U+W)$$
$$= \dim V - \dim(U \cap W) = \dim((U \cap W)^0).$$

Thus $(U \cap W)^0 = U^0 + W^0$.

(e) According to (c) and (d), we have

$$(U+W)_0 = \big((U_0)^0 + (W_0)^0\big)_0 = \big((U_0 \cap W_0)^0\big)_0 = U_0 \cap W_0,$$
$$(U \cap W)_0 = \big((U_0)^0 \cap (W_0)^0\big)_0 = \big((U_0 + W_0)^0\big)_0 = U_0 + W_0.$$

(f) Let $V = U \oplus W$. According to (d), $U^0 + W^0 = (U \cap W)^0 = \{0\}^0 = V^*$ and $U^0 \cap W^0 = (U+W)^0 = V^0 = \{0\}$. Thus $V^* = U^* \oplus W^*$. Conversely, let $V^* = U^0 \oplus W^0$. From (e) it follows that

$$U + W = (U^0)_0 + (W^0)_0 = (U^0 \cap W^0)_0 = \{0\}_0 = V$$

and $U \cap W = (U^0)_0 \cap (W^0)_0 = (U^0 + W^0)_0 = (V^*)_0 = \{0\}$. This shows $V = U \oplus W$. □

Corollary 7.44. *Let V be an n-dimensional vector space and $0 \leq k \leq n$. Then there is a bijection between the set of k-dimensional subspaces and the set of $(n-k)$-dimensional subspaces of V.*

Proof. Let $F\colon V \to V^*$ be an arbitrary isomorphism. Let $U \leq V$ with dimension k. According to Lemma 7.43, $F(U)_0 \leq V$ and $F^{-1}(U^0) \leq V$ have dimension $n-k$. Furthermore, we have

$$F^{-1}\big((F(U)_0)^0\big) = F^{-1}(F(U)) = U,$$
$$F\big(F^{-1}(U^0)\big)_0 = (U^0)_0 = U.$$

Therefore, the maps $U \mapsto F(U)_0$ and $U \mapsto F^{-1}(U^0)$ are inverse bijections to each other between the set of k-dimensional subspaces and the set of $(n-k)$-dimensional subspaces of V. □

Theorem 7.45. *Let V be an n-dimensional vector space over a field K with $q < \infty$ elements. For $1 \leq k \leq n$, V has exactly*

$$\frac{(q^n-1)(q^{n-1}-1)\dots(q^{n-k+1}-1)}{(q^k-1)(q^{k-1}-1)\dots(q-1)}$$

subspaces of dimension k.

Proof. Every k-dimensional subspace $U \leq V$ is spanned by a k-tuple of linearly independent vectors $(v_1, \dots, v_k)$. For $v_1 \in V \setminus \{0\}$, there are $|V| - 1 = q^n - 1$ possibilities. Due to linear independence, v_2 must not lie in $\langle v_1 \rangle \cong K$. Therefore, there are $q^n - q$ possibilities for $v_2 \in V \setminus \langle v_1 \rangle$. In general, there are $q^n - q^{i-1}$ possibilities for the choice of $v_i \in V \setminus \langle v_1, \dots, v_{i-1} \rangle$. In total, there are $D(n,k) := (q^n-1)(q^n-q)\dots(q^n-q^{k-1})$ linearly independent k-tuples $(v_1, \dots, v_k)$ in V. However, many of these span the same space.

The same argument with U instead of V yields exactly $D(k,k) = (q^k-1)(q^k-q)\dots(q^k-q^{k-1})$ linearly independent k-tuples in U. Thus, exactly $D(k,k)$ of the k-tuples in V span the same subspace. The number of k-dimensional subspaces is therefore $\frac{D(n,k)}{D(k,k)}$. The claim follows by canceling all factors of q. □

Remark 7.46. At first glance, it is not clear why the formula given in Theorem 7.45 is an integer at all. We have thus proven an arithmetic statement using linear algebra. Because of

$$\frac{(q^n-1)(q^{n-1}-1)\dots(q^{n-k+1}-1)}{(q^k-1)(q^{k-1}-1)\dots(q-1)} = \frac{(q^n-1)(q^{n-1}-1)\dots(q^{k+1}-1)}{(q^{n-k}-1)(q^{n-k-1}-1)\dots(q-1)}$$

one obtains Corollary 7.44 for finite fields.

Example 7.47. The number of 2-dimensional subspaces of $\mathbb{F}_2^5$ is

$$\frac{(2^5-1)(2^4-1)}{(2^2-1)(2-1)} = \frac{31\cdot 15}{3} = 155.$$

Definition 7.48. Let V, W be vector spaces and $f \in \operatorname{Hom}(V,W)$. Then $f^*\colon W^* \to V^*$, $g \mapsto g \circ f$ is called the *dual map* to f.

Theorem 7.49. *For vector spaces V, W, the map $\operatorname{Hom}(V,W) \to \operatorname{Hom}(W^*,V^*)$, $f \mapsto f^*$ is an isomorphism. Let B and C be bases of V and W, respectively. Let B^* and C^* be the corresponding dual bases. Then*

$$\boxed{{}_{B^*}[f^*]_{C^*} = {}_C[f]_B^{\mathrm{t}}.}$$

In particular,

(a) f is injective $\iff$ f^* is surjective.

(b) f is surjective $\iff$ f^* is injective.

Proof. For $g_1, g_2 \in W^*$ and $\lambda \in K$, it holds that

$$f^*(\lambda g_1 + g_2) = (\lambda g_1 + g_2) \circ f = \lambda(g_1 \circ f) + (g_2 \circ f) = \lambda f^*(g_1) + f^*(g_2)$$

according to Remark 7.23. This shows $f^* \in \operatorname{Hom}(W^*, V^*)$. For $f_1, f_2 \in \operatorname{Hom}(V,W)$, it holds analogously that

$$(\lambda f_1 + f_2)^*(g) = g \circ (\lambda f_1 + f_2) = \lambda(g \circ f_1) + (g \circ f_2) = (\lambda f_1^* + f_2^*)(g).$$

Therefore, $f \mapsto f^*$ is linear. Let $B = \{b_1, \dots, b_m\}$ and $C = \{c_1, \dots, c_n\}$. Let $f(b_i) = \sum_{j=1}^n a_{ji} c_j$ with $a_{ji} \in K$. Then it holds that

$$f^*(c_i^*)(b_j) = (c_i^* \circ f)(b_j) = c_i^*\Big(\sum_{k=1}^n a_{kj} c_k\Big) = a_{ij} = \Big(\sum_{k=1}^n a_{ik} b_k^*\Big)(b_j)$$

for $j = 1, \dots, m$. It follows that $f^*(c_i^*) = \sum_{k=1}^n a_{ik} b_k^*$ for $i = 1, \dots, n$. This shows ${}_C[f]_B^{\mathrm{t}} = (a_{ji})^{\mathrm{t}} = (a_{ij}) = {}_{B^*}[f^*]_{C^*}$. According to Theorem 7.18, $f \mapsto f^*$ is an isomorphism.

With Corollary 7.21, we obtain

$$\begin{aligned} f \text{ injective} &\iff \operatorname{rk}({}_C[f]_B) = m \iff \operatorname{rk}({}_{B^*}[f^*]_{C^*}) = m \iff f^* \text{ surjective}, \\ f \text{ surjective} &\iff \operatorname{rk}({}_C[f]_B) = n \iff \operatorname{rk}({}_{B^*}[f^*]_{C^*}) = n \iff f^* \text{ injective}. \end{aligned}$$

□

Remark 7.50. For $f \in \operatorname{Hom}(V,W)$ and $g \in \operatorname{Hom}(W,U)$, it holds that $(g \circ f)^* = f^* \circ g^*$, because

$$(g \circ f)^*(\varphi) = \varphi \circ (g \circ f) = (\varphi \circ g) \circ f = g^*(\varphi) \circ f = f^*(g^*(\varphi)) = (f^* \circ g^*)(\varphi)$$

for $\varphi \in U^*$. Alternatively, one can use the matrix identity $(AB)^{\mathrm{t}} = B^{\mathrm{t}} A^{\mathrm{t}}$ from Lemma 5.8.

8 Eigenvalues and Eigenvectors

8.1 Definitions and Examples

Remark 8.1. In this chapter, we investigate homomorphisms $f\colon V \to V$ between the same spaces. These are called *endomorphisms* and we write $\mathrm{End}(V) := \mathrm{Hom}(V, V)$. By choosing a suitable basis B of V, we will achieve that ${}_B[f]_B$ has the simplest possible form.[1]

Definition 8.2. Let V be a K-vector space and $f \in \mathrm{End}(V)$. We call $\lambda \in K$ an *eigenvalue* of f if the *eigenspace*

$$E_\lambda(f) := \{v \in V : f(v) = \lambda v\}$$

is not the zero space. If applicable, $\dim E_\lambda(f)$ is called the *geometric multiplicity* of λ. The vectors $v \in E_\lambda(f) \setminus \{0\}$ are called *eigenvectors* for the eigenvalue λ.

Remark 8.3.

(a) For $v \in V$ and $\lambda \in K$ it holds that

$$f(v) = \lambda v \iff f(v) - \lambda v = 0 \iff (f - \lambda\,\mathrm{id})(v) = 0 \iff v \in \mathrm{Ker}(f - \lambda\,\mathrm{id}).$$

Therefore, the eigenspace $E_\lambda(f) = \mathrm{Ker}(f - \lambda\,\mathrm{id})$ is indeed a subspace of V. According to the homomorphism theorem, $\dim(V) - \mathrm{rk}(f - \lambda\,\mathrm{id})$ is the geometric multiplicity of λ.

(b) Let B be a basis of V, $x := {}_B[v]^{\mathrm{t}} \in K^{n\times 1}$ and $A := {}_B[f]_B$. Then it holds that

$$f(v) = \lambda v \overset{7.18}{\iff} Ax = \lambda x \iff (A - \lambda 1_n)x = 0.$$

Therefore, $E_\lambda(f)$ can be calculated by solving the homogeneous system of equations $(A-\lambda 1_n)x = 0$. We then also speak of eigenvalues, eigenspaces, and eigenvectors of A. Since similar matrices describe the same endomorphism (wrt. different bases), they have the same eigenvalues (but not the same eigenvectors).

(c) From Lemma 7.7 and Remark 7.9 it follows: $f \in \mathrm{End}(V)$ is an isomorphism if and only if 0 is *not* an eigenvalue of f.

Example 8.4. Let $f \in \mathrm{End}(\mathbb{R}^3)$ with

$$A := [f] = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Subtracting $\lambda = 1$ on the main diagonal, one obtains a matrix of rank 1 with three identical rows. Obviously, $b_1 := (1, 0, -1)$ and $b_2 := (0, 1, -1)$ form a basis of $E_1(f)$. In particular, $\lambda = 1$ has geometric

[1] According to Remark 6.20, one can always find bases B, C of V such that ${}_C[f]_B = \mathrm{diag}(1_r, 0_{n-r})$ holds. However, the product of such matrices cannot be interpreted meaningfully.

multiplicity 2. Since the row sums of A are constant, $b_3 := (1,1,1)$ is an eigenvector for the eigenvalue $\lambda = 4$. Now $B := \{b_1, b_2, b_3\}$ is a basis of $\mathbb{R}^3$ with ${}_B[f]_B = \operatorname{diag}(1,1,4)$. Thus one easily calculates ${}_B[f \circ f \circ f]_B = {}_B[f]_B^3 = \operatorname{diag}(1,1,4)^3 = \operatorname{diag}(1,1,64)$.

8.2 Diagonalizability

Definition 8.5. One calls $f \in \operatorname{End}(V)$ *diagonalizable*, if a basis B of V exists such that ${}_B[f]_B$ is a diagonal matrix. A matrix $A \in K^{n\times n}$ is called *diagonalizable*, if A is similar to a diagonal matrix.

Remark 8.6. Obviously, $f \in \operatorname{End}(V)$ is diagonalizable if and only if V has a basis consisting of eigenvectors of f. A matrix A is diagonalizable if and only if the corresponding linear map $f\colon K^{n\times 1} \to K^{n\times 1}$, $x \mapsto Ax$ is diagonalizable (Corollary 7.27).

Example 8.7.

(a) The map from Example 8.4 is diagonalizable.

(b) Diagonal matrices are obviously diagonalizable.

(c) Let $A := \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right) \in K^{2\times 2}$. Since $A - \lambda 1_2$ has full rank for $\lambda \neq 0$, $\lambda = 0$ is the only eigenvalue of A. Because of $E_0(A) = \langle (1,0) \rangle$, there exists no basis of eigenvectors and A is *not* diagonalizable.

(d) Let $A := \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right) \in \mathbb{Q}^{2\times 2}$ and $\lambda \in \mathbb{Q}$. Then

$$A - \lambda 1_2 = \begin{pmatrix} -\lambda & 2 \\ 1 & -\lambda \end{pmatrix} \begin{array}{l} \leftarrow \\ \\ \end{array}{}^{+}_{\lambda} \sim \begin{pmatrix} 0 & 2-\lambda^2 \\ 1 & -\lambda \end{pmatrix}.$$

Because of $\sqrt{2} \notin \mathbb{Q}$ (Example 1.11), $2 - \lambda^2 \neq 0$. Thus A has no eigenvalue over $\mathbb{Q}$ and cannot be diagonalizable. On the other hand, A is diagonalizable as an $\mathbb{R}^{2\times 2}$-matrix (Example 8.13).

Definition 8.8. In Definition 4.2, we introduced the (direct) sum $U + W$ (resp. $U \oplus W$) of two subspaces $U, W \leq V$. For $U_1, \ldots, U_n \leq V$, one defines inductively

$$U_1 + \ldots + U_n := (U_1 + \ldots + U_{n-1}) + U_n \leq V.$$

Obviously, $U_1 + \ldots + U_n$ consists of the elements of the form $u_1 + \ldots + u_n$ with $u_i \in U_i$ for $i = 1, \ldots, n$. We call the sum *direct* and write $U_1 \oplus \ldots \oplus U_n$ if $U_1 + \ldots + U_{n-1} = U_1 \oplus \ldots \oplus U_{n-1}$ and $(U_1 + \ldots + U_{n-1}) \cap U_n = \{0\}$. The following characterization is more convenient.

Lemma 8.9. *For subspaces $U_1, \ldots, U_n$ of a vector space V, the following statements are equivalent:*

(1) $U_1 + \ldots + U_n = U_1 \oplus \ldots \oplus U_n$.

(2) $\dim(U_1 + \ldots + U_n) = \dim(U_1) + \ldots + \dim(U_n)$.

(3) The map $U_1 \times \ldots \times U_n \to U_1 + \ldots + U_n$, $(u_1, \ldots, u_n) \mapsto u_1 + \ldots + u_n$ is an isomorphism.

(4) Every element $u \in U_1 + \ldots + U_n$ can be uniquely written in the form $u = u_1 + \ldots + u_n$ with $u_i \in U_i$ for $i = 1, \ldots, n$.

(5) If $u_1 + \ldots + u_n = 0$ with $u_i \in U_i$ for $i = 1, \ldots, n$, then $u_1 = \ldots = u_n = 0$ follows.

Proof.

(1) ⇒ (2): For $n = 1$, (2) is trivial. By induction, we may assume that (2) already holds for $n - 1$. From the dimension theorem, it then follows that

$$\dim(U_1 + \ldots + U_n) = \dim(U_1 + \ldots + U_{n-1}) + \dim(U_n) = \dim(U_1) + \ldots + \dim(U_n).$$

(2) ⇒ (3): The given map is always linear and surjective. Because of

$$\dim(U_1 \times \ldots \times U_n) \overset{7.11}{=} \dim(U_1) + \ldots + \dim(U_n) = \dim(U_1 + \times + U_n)$$

it must also be injective.

(3) ⇒ (4): Follows from the injectivity of $U_1 \times \ldots \times U_n \to U_1 + \ldots + U_n$.

(4) ⇒ (5): The two decompositions of the zero vector $u_1 + \ldots + u_n = 0 + \ldots + 0$ must be identical according to (4), i.e., $u_1 = \ldots = u_n = 0$.

(5) ⇒ (1): For $n = 1$, there is nothing to show. By induction, we can assume $U_1 + \ldots + U_{n-1} = U_1 \oplus \ldots \oplus U_{n-1}$, since the assumption (5) carries over to $U_1, \ldots, U_{n-1}$. Now let $u = u_1 + \ldots + u_{n-1} \in (U_1 + \ldots + U_{n-1}) \cap U_n$. Then

$$0 = u_1 + \ldots + u_{n-1} - u \in U_1 + \ldots + U_n$$

and (5) shows $u = 0$. Thus (1) holds. □

Theorem 8.10. *Let $\lambda_1, \ldots, \lambda_k$ be pairwise distinct eigenvalues of $f \in \mathrm{End}(V)$. Then*

$$E_{\lambda_1}(f) + \ldots + E_{\lambda_k}(f) = E_{\lambda_1}(f) \oplus \ldots \oplus E_{\lambda_k}(f) \leq V. \tag{8.1}$$

In particular, $k \leq \dim V$.

Proof. Induction on k: For $k = 1$, there is nothing to show. So let $k \geq 2$ and assume (8.1) is already proven for $k - 1$. Let $v_i \in E_{\lambda_i}(f)$ with $v_1 + \ldots + v_k = 0$. Then

$$\begin{aligned} 0 &= f(v_1 + \ldots + v_k) - \lambda_k(v_1 + \ldots + v_k) = \lambda_1 v_1 + \ldots + \lambda_k v_k - \lambda_k v_1 - \ldots - \lambda_k v_k \\ &= (\lambda_1 - \lambda_k)v_1 + \ldots + (\lambda_{k-1} - \lambda_k)v_{k-1} \in E_{\lambda_1}(f) \oplus \ldots \oplus E_{\lambda_{k-1}}(f). \end{aligned}$$

From Lemma 8.9 it follows that $(\lambda_i - \lambda_k)v_i = 0$ for $i = 1, \ldots, k-1$. Because $\lambda_i \neq \lambda_k$, it holds that $v_1 = \ldots = v_{k-1} = 0$. Finally, $v_k = v_1 + \ldots + v_k = 0$ as well. Now (8.1) follows from Lemma 8.9. The last assertion follows from $\dim E_{\lambda_i}(f) \geq 1$ for $i = 1, \ldots, k$. □

Remark 8.11. Mnemonic: Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Corollary 8.12. *If $A \in K^{n \times n}$ has exactly n distinct eigenvalues, then A is diagonalizable.*

Proof. For the distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, it holds that

$$\dim\big(E_{\lambda_1}(A) \oplus \ldots \oplus E_{\lambda_n}(A)\big) = \dim(E_{\lambda_1}(A)) + \ldots + \dim(E_{\lambda_n}(A)) \geq n = \dim K^{n \times 1}$$

according to Theorem 8.10. This shows $E_{\lambda_1}(A) \oplus \ldots \oplus E_{\lambda_n}(A) = K^{n \times 1}$. In particular, $K^{n \times 1}$ possesses a basis of eigenvectors of A. □

Example 8.13.

(a) The identity matrix shows that the converse of Corollary 8.12 is false. We will later derive a precise characterization of diagonalizability (Theorem 10.34).

(b) The matrix $A := \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$ from Example 8.7 has the eigenvalues $\pm\sqrt{2}$ and is therefore diagonalizable. We show via the detour of the determinant that the eigenvalues of every matrix are roots of polynomials with coefficients in K (Theorem 10.32).

Definition 8.14. One calls $A = (a_{ij}) \in K^{n\times n}$ an (*upper*[2]) *triangular matrix*, if all entries below the main diagonal vanish, i.e., $a_{ij} = 0$ for all $i > j$:

$$A = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}.$$

If additionally $a_{ii} = 0$ for $i = 1, \ldots, n$, then one speaks of a *strict* (upper) triangular matrix.

Example 8.15. Let $A = (a_{ij}) \in K^{n\times n}$ be an upper triangular matrix. For $\lambda \in \{a_{11}, \ldots, a_{nn}\}$, $A - \lambda 1_n$ does not have full rank, because an offset of the rows occurs during the Gaussian algorithm (Remark 6.20). If, on the other hand, $\lambda \notin \{a_{11}, \ldots, a_{nn}\}$, then the main diagonal entries of $A - \lambda 1_n$ are all non-zero. Therefore, $A - \lambda 1_n$ has full rank. This shows that the eigenvalues of A are exactly the entries on the main diagonal. In particular, A is diagonalizable if $a_{11}, \ldots, a_{nn}$ are pairwise distinct.

[2]Analogously, one defines *lower* triangular matrices.

9 Determinants

9.1 Recursive Definition

Remark 9.1.

(a) Mathematicians often try to replace complicated objects (such as $f \in \mathrm{End}(V)$) with simpler ones (such as $\mathrm{rk}(f)$ or $\mathrm{tr}(f)$) in order to make essential information visible. Thus, we have seen in Lemma 5.15 that $\mathrm{rk}(f)$ provides information about the bijectivity of f, provided that $\dim V$ is known. Such quantities are called *invariants* if they remain unchanged under "natural" transformations (such as a change of basis). In this section, we define the *determinant* $\det(f)$ as a further invariant. We show that f is bijective if and only if $\det(f) \neq 0$ holds (in contrast to $\mathrm{rk}(f)$, this criterion no longer depends on $\dim(V)$.[1])

(b) In measure theory, one tries to assign a "volume" $\mathrm{vol}(S) \in \mathbb{R}_{\geq 0}$ to as many sets $S \subseteq \mathbb{R}^n$ as possible. Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be linear. The number $\det(f)$ described in (a) measures how much the volume changes by applying f, i. e. $\mathrm{vol}(f(S)) = |\det(f)| \,\mathrm{vol}(S)$ holds as long as $\mathrm{vol}(S)$ is defined. The n-dimensional *hypercube*

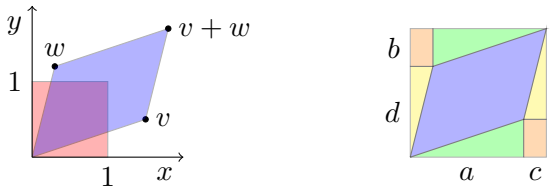
$$H := \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \forall i : 0 \leq x_i \leq 1\big\}$$

is assigned the volume $\mathrm{vol}(H) = 1$. From this it follows

$$|\det(f)| = \mathrm{vol}(f(H)).$$

The sign of $\det(f)$ describes whether f is orientation-preserving (example: rotation) or orientation-reversing (example: reflection). More on this in Example 11.22.

Example 9.2. Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ with $v := f(e_1) = (a, b)$ and $w := f(e_2) = (c, d)$, i. e. $[f] = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$. The image of the (red) square $H = \{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1\}$ is the (blue) parallelogram $f(H)$ spanned by the vectors v and w:



The area of $f(H)$ is

$$\mathrm{vol}(f(H)) = (a + c)(b + d) - 2bc - ab - cd = ad - bc.$$

The area is 0 if and only if v and w lie on a line, i.e., are linearly dependent. This is equivalent to $\mathrm{rk}(f) \leq 1$.

[1] From real life: If you have trouble remembering a person's age, remember their birth year instead, because this invariant does not change every year.

Definition 9.3. Let $A = (a_{ij}) \in K^{n\times n}$ and $1 \le s, t \le n$. By deleting the s-th row and t-th column of A, we obtain the matrix $A_{st} \in K^{(n-1)\times(n-1)}$. The *determinant*[2] of A is defined recursively:

$$\det(A) := \begin{cases} a_{11} & \text{if } n = 1, \\ \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det(A_{i1}) & \text{if } n \ge 2. \end{cases}$$

Example 9.4.

(a) For $n = 2$ one obtains

$$\det \begin{pmatrix} a & c \\ b & d \end{pmatrix} = a \det(A_{11}) - b \det(A_{21}) = ad - bc$$

(cf. Example 9.2).

(b) For every upper triangular matrix $A = (a_{ij})$ it holds that $\det(A) = a_{11} \dots a_{nn}$. This is clear for $n = 1$. Let the assertion for $n - 1$ be already proven by induction. Since A_{11} is also an upper triangular matrix, it follows that

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det(A_{i1}) = a_{11} \det(A_{11}) = a_{11} a_{22} \dots a_{nn}.$$

Since one can transform every matrix into an upper triangular matrix using the Gaussian algorithm, we investigate how the determinant changes under elementary row operations.

Lemma 9.5. *The map* $\det\colon K^{n\times n} \to K$ *is linear in each row, i. e. for* $a_1, \dots, a_n, b \in K^n$, $\lambda \in K$ *and* $1 \le k \le n$ *it holds that*

$$\det \begin{pmatrix} a_1 \\ \vdots \\ \lambda a_k + b \\ \vdots \\ a_n \end{pmatrix} = \lambda \det \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ b \\ \vdots \\ a_n \end{pmatrix}.$$

Proof. Let $a_i = (a_{i1}, \dots, a_{in})$ and $b = (b_1, \dots, b_n)$. For $c \in K^n$ let $M(c)$ be the matrix with rows $a_1, \dots, a_{k-1}, c, a_{k+1}, \dots, a_n$. For $n = 1$ we have $k = 1$ and

$$\det\big(M(\lambda a_1 + b)\big) = \lambda a_{11} + b_1 = \lambda \det(M(a_1)) + \det(M(b))$$

as claimed. Now let $n \ge 2$ and the assertion for $n - 1$ be already proven. Deleting the k-th row yields

$$M(\lambda a_k + b)_{k1} = M(a_k)_{k1} = M(b)_{k1}.$$

It follows that

$$\begin{aligned} \det\big(M(\lambda a_k + b)\big) &= (-1)^{k+1} (\lambda a_{k1} + b_1) \det\big(M(\lambda a_k + b)_{k1}\big) + \sum_{i\ne k} (-1)^{i+1} a_{i1} \det\big(M(\lambda a_k + b)_{i1}\big) \\ &= \lambda (-1)^{k+1} a_{k1} \det(M(a_k)_{k1}) + (-1)^{k+1} b_1 \det(M(b)_{k1}) \end{aligned}$$

[2] In some books, one writes $|A|$ instead of $\det(A)$. We reserve this notation for a matrix norm, see Example 17.61.

$$+\sum_{i\neq k}(-1)^{i+1}a_{i1}\big(\lambda\det(M(a_k)_{i1})+\det(M(b)_{i1})\big)$$
$$=\lambda\sum_{i=1}^{n}(-1)^{i+1}a_{i1}\det(M(a_k)_{i1})+(-1)^{k+1}b_1\det(M(b)_{k1})+\sum_{i\neq k}(-1)^{i+1}a_{i1}\det(M(b)_{i1})$$
$$=\lambda\det(M(a_k))+\det(M(b)).$$ □

Theorem 9.6. *For $A\in K^{n\times n}$ the following holds:*

(a) By multiplying a row of A by $\lambda\in K$, $\det(A)$ is also multiplied by λ.

(b) Swapping two rows of A changes the sign of $\det(A)$.

(c) Adding a multiple of one row to another row of A does not change $\det(A)$.

Proof.

(a) If one first sets $\lambda=1$ and $b=0$ in Lemma 9.5, one sees that the determinant vanishes if A has a zero row. The claim now follows by choosing λ arbitrarily and $b=0$ in Lemma 9.5.

(b) Here $n\geq 2$. Let $a_1,\ldots,a_n$ be the rows of A and $s<t$. Swapping a_s and a_t yields the matrix A'. For $n=2$, we have $(s,t)=(1,2)$ and

$$\det(A')=\det\begin{pmatrix}a_{21}&a_{22}\\a_{11}&a_{12}\end{pmatrix}=a_{21}a_{12}-a_{22}a_{11}=-(a_{11}a_{22}-a_{12}a_{21})=-\det(A)$$

according to Example 9.4. Now let the claim be already proven for $n-1$. For $s\neq i\neq t$, A'_{i1} is obtained by a row swap from A_{i1}. Thus $\det(A'_{i1})=-\det(A_{i1})$. On the other hand, A'_{s1} is obtained from A_{t1} by the swaps $a_s\leftrightarrow a_{s+1}\leftrightarrow a_{s+2}\leftrightarrow\ldots\leftrightarrow a_{t-1}$:

$$A_{t1}=\begin{pmatrix}\vdots\\a_s\\a_{s+1}\\\vdots\\a_{t-1}\\a_{t+1}\\\vdots\end{pmatrix}\sim\begin{pmatrix}\vdots\\a_{s+1}\\a_s\\a_{s+2}\\\vdots\\a_{t-1}\\a_{t+1}\\\vdots\end{pmatrix}\sim\ldots\sim\begin{pmatrix}\vdots\\a_{s+1}\\\vdots\\a_{t-1}\\a_s\\a_{t+1}\\\vdots\end{pmatrix}=A'_{s1}.$$

This shows $\det(A'_{s1})=(-1)^{t-s-1}\det(A_{t1})$. Analogously, $\det(A'_{t1})=(-1)^{t-s-1}\det(A_{s1})$. Because of $(-1)^{t-s-1}=(-1)^{s-t-1}$, it follows that

$$\begin{aligned}\det(A')&=(-1)^{s+1}a_{t1}\det(A'_{s1})+(-1)^{t+1}a_{s1}\det(A'_{t1})+\sum_{i\notin\{s,t\}}(-1)^{i+1}a_{i1}\det(A'_{i1})\\&=(-1)^{t}a_{t1}\det(A_{t1})+(-1)^{s}a_{s1}\det(A_{s1})+\sum_{i\notin\{s,t\}}(-1)^{i}a_{i1}\det(A_{i1})\\&=-\sum_{i=1}^{n}(-1)^{i+1}a_{i1}\det(A_{i1})=-\det(A).\end{aligned}$$

(c) We add λa_k to row a_l with $k \neq l$ and obtain

$$\det \begin{pmatrix} a_1 \\ \vdots \\ a_l + \lambda a_k \\ \vdots \\ a_n \end{pmatrix} \overset{9.5}{=} \det(A) + \lambda \det \begin{pmatrix} \vdots \\ a_k \\ \vdots \\ a_k \\ \vdots \end{pmatrix}.$$

It therefore suffices to show $\det(A) = 0$ if A has two identical rows.[3] For $n = 2$, we have

$$\det(A) = \det \begin{pmatrix} a & b \\ a & b \end{pmatrix} = ab - ba = 0.$$

Now let $n \geq 3$. According to (b), we can assume that the first two rows of A are identical. Then A_{i1} for $i \geq 3$ also has two identical rows. By induction on n, it follows that

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} \det(A_{i1}) = a_{11} \det(A_{11}) - a_{21} \det(A_{21}) = 0. \qquad \square$$

Example 9.7.

$$\det \begin{pmatrix} -4 & -2 & -2 \\ 6 & 3 & 2 \\ 8 & 7 & 6 \end{pmatrix} \; | : (-2) \quad = -2 \det \begin{pmatrix} 2 & 1 & 1 \\ 6 & 3 & 2 \\ 8 & 7 & 6 \end{pmatrix} \; \begin{matrix} \overset{-3}{\leftarrow} + \\ \overset{-4}{\leftarrow} + \end{matrix} \quad = -2 \det \begin{pmatrix} 2 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 3 & 2 \end{pmatrix} \; \begin{matrix} \leftarrow \\ \leftarrow \end{matrix}$$

$$= 2 \det \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 2 \\ 0 & 0 & -1 \end{pmatrix} \overset{9.4}{=} -12$$

Remark 9.8. For $A \in K^{n \times n}$ and $\lambda \in K$, it holds that $\boxed{\det(\lambda A) = \lambda^n \det(A),}$ because each of the n rows is multiplied by λ.

Lemma 9.9. *Let* $A = \left(\begin{smallmatrix} A_1 & A_2 \\ 0 & A_3 \end{smallmatrix} \right) \in K^{(n+m) \times (n+m)}$ *with* $A_1 \in K^{n \times n}$, $A_2 \in K^{n \times m}$ *and* $A_3 \in K^{m \times m}$. *Then* $\det(A) = \det(A_1) \det(A_3)$ *holds.*

Proof. If one performs the Gaussian elimination on A, then first A_1 and then A_3 are transformed into an upper triangular matrix. In the end, A is also an upper triangular matrix and the claim follows. $\square$

9.2 Properties

Theorem 9.10. *For* $A \in K^{n \times n}$ *it holds that*

$$A \text{ invertible} \iff \operatorname{rk}(A) = n \iff \det(A) \neq 0.$$

[3] For $K = \mathbb{Q}$ this follows immediately from (b), but not for $K = \mathbb{F}_2$.

Proof. The first equivalence comes from Lemma 5.15. Since the row echelon form $\widehat{A}$ is an upper triangular matrix, it holds that

$$\det(A) \neq 0 \overset{9.6}{\iff} \det(\widehat{A}) \neq 0 \overset{9.4}{\iff} \widehat{A} = 1_n \iff \operatorname{rk}(A) = \operatorname{rk}(\widehat{A}) = n. \qquad \square$$

Theorem 9.11 (Determinant Theorem)**.** *For $A, B \in K^{n\times n}$ it holds that* $\boxed{\det(AB) = \det(A)\det(B).}$

Proof. If $\det(A) = 0$, then $\operatorname{rk}(AB) \leq \operatorname{rk}(A) < n$ follows from Lemma 5.15. Then $\det(AB) = 0$ also holds by Theorem 9.10. We can therefore assume $A \in \mathrm{GL}(n, K)$. According to Corollary 6.18, A is a product of elementary matrices, say $A = A_1 \dots A_k$. Let $M \in K^{n\times n}$ be arbitrary. For the three types of elementary row operations, it holds respectively that

$$\det(A_i M) = \left\{ \begin{array}{c} \lambda \det(M) \\ -\det(M) \\ \det(M) \end{array} \right\} = \det(A_i 1_n)\det(M) = \det(A_i)\det(M)$$

according to Theorem 9.6. Overall it follows that

$$\begin{aligned} \det(AB) &= \det(A_1 \dots A_k B) = \det(A_1)\det(A_2 \dots A_k B) = \dots = \det(A_1)\dots\det(A_k)\det(B) \\ &= \dots = \det(A_1)\det(A_2 \dots A_k)\det(B) = \det(A)\det(B). \end{aligned} \qquad \square$$

Corollary 9.12.

(a) For $A \in K^{n\times n}$ it holds that $\boxed{\det(A^{\mathrm{t}}) = \det(A).}$

(b) For $A \in \mathrm{GL}(n, K)$ it holds that $\boxed{\det(A^{-1}) = \det(A)^{-1}.}$

(c) Similar matrices have the same determinant.

Proof.

(a) Because of $\operatorname{rk}(A) = \operatorname{rk}(A^{\mathrm{t}})$, we can assume that A is invertible (otherwise $\det(A) = 0 = \det(A^{\mathrm{t}})$). Again, A is a product of elementary matrices $A = A_1 \dots A_k$. For the first two row operations, $A_i^{\mathrm{t}} = A_i$ holds. For the third row operation, $\det(A_i) = 1 = \det(A_i^{\mathrm{t}})$. This shows

$$\begin{aligned} \det(A^{\mathrm{t}}) &\overset{5.8}{=} \det(A_k^{\mathrm{t}} \dots A_1^{\mathrm{t}}) = \det(A_k^{\mathrm{t}})\dots\det(A_1^{\mathrm{t}}) = \det(A_k)\dots\det(A_1) \\ &= \det(A_1)\dots\det(A_k) = \det(A_1 \dots A_k) = \det(A). \end{aligned}$$

(b) The claim follows from $\det(A)\det(A^{-1}) = \det(AA^{-1}) = \det(1_n) = 1$.

(c) For $A \in K^{n\times n}$ and $S \in \mathrm{GL}(n, K)$ we have

$$\det(SAS^{-1}) = \det(S)\det(A)\det(S^{-1}) = \det(S)\det(S)^{-1}\det(A) = \det(A). \qquad \square$$

Definition 9.13. According to the determinant theorem and Corollary 9.12, the matrices with determinant 1 form a subgroup $\mathrm{SL}(n, K) \leq \mathrm{GL}(n, K)$. One calls $\mathrm{SL}(n, K)$ the *special linear group* of degree n over K. It holds that $\mathrm{SL}(n, \mathbb{F}_2) = \mathrm{GL}(n, \mathbb{F}_2)$.

Remark 9.14.

(a) Because of $\det(A^{\mathrm{t}}) = \det(A)$, one may also use elementary column operations when calculating $\det(A)$.

(b) The following statement generalizes the determinant theorem.

Theorem 9.15 (CAUCHY-BINET formula)**.** *For $A, B \in K^{n\times m}$ and $I \subseteq \{1,\dots,m\}$ let $A_I := (a_{ij} : i = 1,\dots,n,\ j \in I)$. Then*

$$\det(AB^{\mathrm{t}}) = \sum_{\substack{I\subseteq\{1,\dots,m\}\\ |I|=n}} \det(A_I)\det(B_I).$$

Proof. In the case $n > m$, the sum is empty and $\det(AB^{\mathrm{t}}) = 0$ according to Lemma 5.15. So let $n \le m$. We decompose the i-th row of A as $a_i = a_i' + a_i''$. If one replaces a_i by a_i' or a_i'', one obtains matrices A' or A'' with $\det(A_I) = \det(A_I') + \det(A_I'')$ for all $I \subseteq \{1,\dots,m\}$ with $|I| = n$ according to Lemma 9.5. The i-th row of AB^{t} is $a_i'B^{\mathrm{t}} + a_i''B^{\mathrm{t}}$. Therefore, $\det(AB^{\mathrm{t}}) = \det(A'B^{\mathrm{t}}) + \det(A''B^{\mathrm{t}})$ also holds. It thus suffices to prove the claim for A' or A'' instead of A. In this way, one achieves that A has at most one non-zero entry in each row. Thus, there exists at most one $I \subseteq \{1,\dots,m\}$ with $|I| = n$ and $\det(A_I) \neq 0$ (all other A_I possess zero columns). It holds that $AB^{\mathrm{t}} = A_I(B_I)^{\mathrm{t}}$ and $\det(AB^{\mathrm{t}}) = \det(A_I)\det(B_I)$ according to Corollary 9.12. □

9.3 Laplace Expansion

Theorem 9.16 (LAPLACE expansion)**.** *Let $n \ge 2$ and $A \in K^{n\times n}$. For $1 \le k \le n$ we have*

$$\boxed{\det(A) = \sum_{i=1}^{n}(-1)^{i+k}a_{ik}\det(A_{ik}) = \sum_{i=1}^{n}(-1)^{i+k}a_{ki}\det(A_{ki}).}$$

Proof. Let $a_1,\dots,a_n$ be the columns of A. According to Remark 9.14, it holds that

$$\begin{aligned}\det\big(\cdots\ \underset{\downarrow}{a_{k-1}}\ \underset{\downarrow}{a_k}\ \cdots\big) &= -\det\big(\cdots\ \underset{\downarrow}{a_{k-2}}\ \underset{\downarrow}{a_k}\ a_{k-1}\ \cdots\big) = \dots\\ &= (-1)^{k-1}\det\big(a_k\ \ a_1\ \ \cdots\ \ a_{k-1}\ \ a_{k+1}\ \ \cdots\big) = (-1)^{k-1}\sum_{i=1}^{n}(-1)^{i+1}a_{ik}\det(A_{ik}).\end{aligned}$$

The second equation follows from the first by using $\det(A^{\mathrm{t}}) = \det(A)$. □

Remark 9.17. The equations in Theorem 9.16 are called *expansion along the k-th column/row*. The signs $(-1)^{i+k}$ are distributed like a checkerboard:

$$\begin{pmatrix} + & - & + & \cdots\\ - & + & - & \cdots\\ + & - & + & \cdots\\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Example 9.18. Like most recursive methods, the Laplace expansion is generally inefficient. However, it is suitable for so-called *sparse* matrices, i.e., when many entries are 0. We first expand along the third row and then along the second column:

$$\det\begin{pmatrix} 1 & 2 & -3 & 0\\ -2 & 0 & 1 & 2\\ 0 & 0 & 1 & 0\\ -2 & 0 & 1 & 0 \end{pmatrix} = \det\begin{pmatrix} 1 & 2 & 0\\ -2 & 0 & 2\\ -2 & 0 & 0 \end{pmatrix} = -2\det\begin{pmatrix} -2 & 2\\ -2 & 0 \end{pmatrix} = -2\cdot 4 = -8$$

Remark 9.19. For a sequence of numbers $a_1, \ldots, a_n \in K$, the product $\prod_{i=1}^{n} a_i = a_1 \cdot \ldots \cdot a_n$ is defined analogously to the summation sign $\sum$.

Theorem 9.20 (VANDERMONDE)**.** *For $x_1, \ldots, x_n \in K$, the matrix*

$$A := (x_i^{j-1}) = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \in K^{n \times n}$$

is called the VANDERMONDE matrix.[4] *It holds that*

$$\det(A) = \prod_{1 \le i < j \le n} (x_j - x_i).$$

In particular, A is invertible if and only if $x_1, \ldots, x_n$ are pairwise distinct.

Proof. Induction on n: In the case $n = 1$, $A = 1_1$ and $\prod_{i<j}(x_j - x_i)$ is the empty product, which is interpreted as 1 (just as the empty sum is interpreted as 0). So let $n \ge 2$. We subtract x_1 times the second to last column from the last column. Subsequently, we subtract x_1 times the $(n-2)$-th column from the second to last column, and so on. This yields the matrix

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 1 & x_2 - x_1 & (x_2 - x_1)x_2 & \cdots & (x_2 - x_1)x_2^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n - x_1 & (x_n - x_1)x_n & \cdots & (x_n - x_1)x_n^{n-2} \end{pmatrix}$$

with the same determinant (Theorem 9.6). By expansion along the first row, one can pass to the smaller matrix

$$\begin{pmatrix} x_2 - x_1 & (x_2 - x_1)x_2 & \cdots & (x_2 - x_1)x_2^{n-2} \\ \vdots & \vdots & & \vdots \\ x_n - x_1 & (x_n - x_1)x_n & \cdots & (x_n - x_1)x_n^{n-2} \end{pmatrix}.$$

The factors $(x_k - x_1)$ can be pulled out of the determinant for $k = 2, \ldots, n$. This gives

$$\det(A) = (x_2 - x_1)(x_3 - x_1) \ldots (x_n - x_1) \det\big((x_{i+1}^{j-1})_{i,j=1}^{n-1}\big).$$

Now the claim follows by induction. □

Definition 9.21. For $A \in K^{n \times n}$, one calls

$$\widetilde{A} := \begin{cases} 1_1 & \text{if } n = 1 \\ \big((-1)^{i+j} \det(A_{ji})\big)_{i,j} & \text{if } n > 1 \end{cases} \in K^{n \times n}$$

the *complementary* matrix to A.[5]

Theorem 9.22. *For all $A \in K^{n \times n}$, it holds that* $\boxed{A\widetilde{A} = \det(A) 1_n = \widetilde{A}A.}$ *In particular,* $\boxed{A^{-1} = \dfrac{1}{\det(A)} \widetilde{A},}$ *if $A \in \mathrm{GL}(n, K)$.*

[4] In some books, the transposed matrix is considered.

[5] also called *adjugate*; risk of confusion with the *adjoint* matrix A^* from Theorem 13.7

Proof. For $n = 1$, the claim is clear. So let $n \geq 2$. Let B_{kl} be the matrix that arises from A by replacing the l-th row with the k-th row. For $k \neq l$, B_{kl} has two identical rows and it follows that $\det(B_{kl}) = 0$. On the other hand, $B_{kk} = A$. Let $A\widetilde{A} = (c_{ij})$. Expansion along the l-th row of B_{kl} yields

$$\delta_{kl} \det(A) = \det(B_{kl}) = \sum_{i=1}^{n} a_{ki}(-1)^{i+l} \det(A_{li}) = c_{kl}.$$

This shows $A\widetilde{A} = \det(A)1_n$. The equation $\widetilde{A}A = \det(A)1_n$ is shown analogously by expansion along a column. □

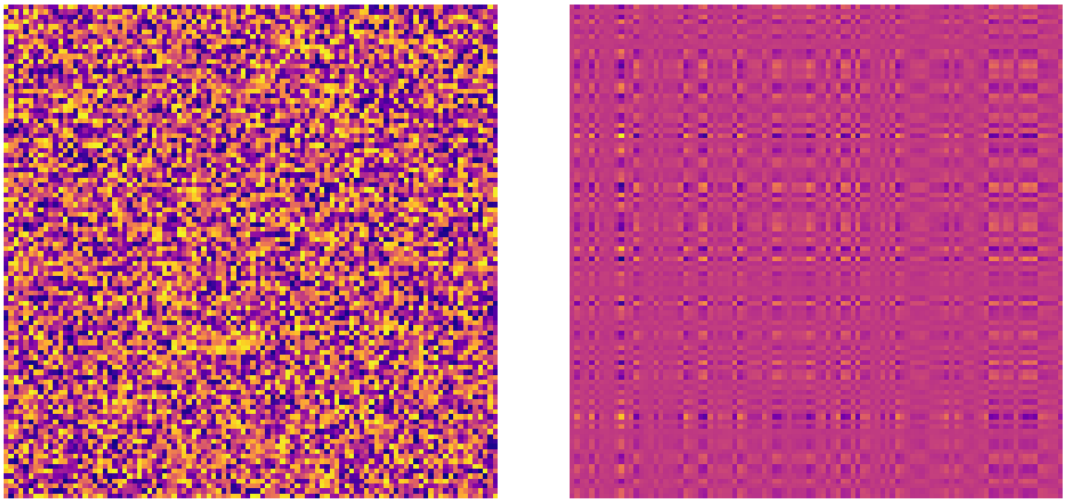
Example 9.23. For every invertible 2×2-matrix $A := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, one obtains

$$A^{-1} = \frac{1}{\det(A)}\widetilde{A} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Remark 9.24.

(a) The formula $A^{-1} = \frac{1}{\det(A)}\widetilde{A}$ shows that the entry of A^{-1} at position (i, j) depends on all a_{st}, except those with either $s = i$ or $t = j$. Therefore, the inverse of a randomly chosen 100×100 matrix exhibits clear structures:[6]

A $\qquad$ A^{-1}



(b) For "large" n, this formula has more theoretical than practical significance (use Theorem 6.17 for the calculation of A^{-1}). For example, if $A \in \mathbb{Z}^{n\times n}$, then $\det(A)A^{-1} = \widetilde{A} \in \mathbb{Z}^{n\times n}$ as well. In particular, $A^{-1} \in \mathbb{Z}^{n\times n}$ if $\det(A) = \pm 1$. This observation is not apparent from the Gaussian algorithm. The next theorem provides a similar statement for systems of equations.

Theorem 9.25 (Cramer's Rule)**.** *Let $A \in \mathrm{GL}(n, K)$ and $b \in K^{n\times 1}$. For $k = 1, \ldots, n$, let A_k be the matrix formed from A by replacing the k-th column with b. For the unique solution $x = (x_1, \ldots, x_n)^{\mathrm{t}}$ of the system of equations $Ax = b$, it then holds that $x_k = \frac{\det(A_k)}{\det(A)}$ for $k = 1, \ldots, n$.*

Proof. We use Theorem 9.22 and expand A_k along the k-th column:

$$(\det(A_k))_k = \Big(\sum_{i=1}^{n}(-1)^{i+k} \det(A_{ik})b_i\Big)_k = \widetilde{A}b = \widetilde{A}Ax = \det(A)x = (\det(A)x_k)_k.$$ □

[6]The graphics were generated with sageMath.

9.4 The Leibniz Formula

Remark 9.26. If one carries out the Laplace expansion for $n \times n$ matrices down to 1×1 matrices, one obtains the determinant as a sum of $n(n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$ terms. We determine these terms explicitly.

Definition 9.27. Let $n \in \mathbb{N}$ and $N := \{1, \ldots, n\}$. A bijection of the form $N \to N$ is called a *permutation* of N. The set of permutations of N is denoted by S_n.

Remark 9.28.

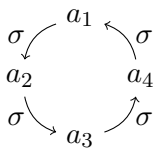
(a) Analogous to $\mathrm{GL}(V)$, S_n is also a group wrt. the composition of mappings. We will therefore often omit the composition symbol $\circ$. S_n is called the *symmetric group of degree n*.

(b) Let $\sigma \in S_n$. For the choice of $\sigma(1) \in \{1, \ldots, n\}$, there are n possibilities. Since σ is injective, $\sigma(2) \neq \sigma(1)$ holds. Thus, for the choice of $\sigma(2)$, there remain $n-1$ possibilities, and so on. In total, there are $n!$ possibilities to define a permutation, i.e., $|S_n| = n!$.

Example 9.29.

(a) For $k \geq 2$, one calls $\sigma \in S_n$ a (k-)*cycle* (or cycle of *length* k), if pairwise distinct $1 \leq a_1, \ldots, a_k \leq n$ exist, such that

$$\sigma(x) = \begin{cases} a_{i+1} & \text{if } x = a_i \text{ with } i < k, \\ a_1 & \text{if } x = a_k, \\ x & \text{otherwise.} \end{cases}$$



One then writes $\sigma = (a_1, \ldots, a_k)$. This notation is unique up to "rotation", i. e.

$$\sigma = (a_2, \ldots, a_k, a_1) = \ldots = (a_k, a_1, \ldots, a_{k-1}).$$

The composition of cycles occurs as usual for mappings from right to left:

$$(1,3,4,5) \circ (3,4,5) \circ (1,3,2) = (1,5,4)(2,3).$$

Furthermore, $(a_1, \ldots, a_k)^{-1} = (a_k, a_{k-1}, \ldots, a_1)$.

(b) Cycles of length 2 are called *transpositions*. A transposition thus swaps two elements and leaves all other elements fixed. Every k-cycle is a composition of $k-1$ transpositions:

$$(a_1, \ldots, a_k) = (a_1, a_2)(a_2, a_3) \ldots (a_{k-1}, a_k).$$

However, there are many possibilities for such a composition.

(c) In S_3, every element is a cycle: $S_3 = \{1, (1,2), (1,3), (2,3), (1,2,3), (1,3,2)\}$.

(d) Cycles $\sigma = (a_1, \ldots, a_k)$ and $\tau = (b_1, \ldots, b_l)$ are called *disjoint*, if

$$\{a_1, \ldots, a_k\} \cap \{b_1, \ldots, b_l\} = \varnothing.$$

In this case, $\sigma\tau = \tau\sigma$ holds.

Theorem 9.30. *Every permutation $\sigma \in S_n$ is a composition of pairwise disjoint cycles $\sigma = \sigma_1 \ldots \sigma_k$. Here, $\sigma_1, \ldots, \sigma_k$ are uniquely determined up to their order.*

Proof. Let $\sigma \in S_n$. In the case $\sigma = \mathrm{id}$, σ is the empty product of cycles. So let $\sigma \neq \mathrm{id}$ and

$$a_1 := \min\{1 \leq i \leq n : \sigma(i) \neq i\}.$$

Let $a_i := \sigma^{i-1}(a_1)$ for $i \geq 2$. Because of $n < \infty$, there exist $i < j$ with $a_i = a_j$, i. e. $\sigma^{j-i}(a_1) = a_1$. Therefore there exists

$$k := \min\{1 \leq i \leq n : \sigma^i(a_1) = a_1\}.$$

Because of $k \leq j - i$, the elements $a_1, \ldots, a_k$ are pairwise distinct. Thus $\sigma_1 := (a_1, \ldots, a_k)$ is a k-cycle. For $\rho := \sigma\sigma_1^{-1}$ it holds that

$$\rho(x) = \begin{cases} x & \text{if } x \in \{a_1, \ldots, a_k\}, \\ \sigma(x) & \text{otherwise.} \end{cases}$$

In the case $\rho = \mathrm{id}$, $\sigma = \sigma_1$ and we are finished. Otherwise, we can repeat the procedure with ρ instead of σ. Since the number of fixed points of σ increases with each repetition, after finitely many steps one reaches $\sigma = \sigma_1 \ldots \sigma_k$ with pairwise disjoint cycles $\sigma_1, \ldots, \sigma_k$.

Let also $\sigma = \tau_1 \ldots \tau_l$ be a composition of pairwise disjoint cycles $\tau_1, \ldots, \tau_l$. Then there exists an i with $\tau_i(a_1) \neq a_1$. Since the cycles are disjoint, it holds that $\tau_i(a_1) = \sigma(a_1) = \sigma_1(a_1) = a_2$, $\tau_i(a_2) = \sigma_1(a_2) = a_3$ etc. Thus $\tau_i = \sigma_1$ and $\sigma_2 \ldots \sigma_k = \tau_1 \ldots \tau_{i-1}\tau_{i+1} \ldots \tau_l$. The uniqueness of the σ_i now follows by induction on k. □

Remark 9.31.

(a) Since according to Example 9.29 every cycle is a composition of transpositions, even every permutation is a composition of (usually not disjoint) transpositions.

(b) One can make the notation in disjoint cycles

$$\sigma = (a_1, \ldots, a_s)(b_1, \ldots, b_t) \ldots$$

completely unique by requiring $a_1 = \min\{a_1, \ldots, a_s\} < b_1 = \min\{b_1, \ldots, b_t\} < \ldots$.

Definition 9.32. For $\sigma \in S_n$, one calls

$$P_\sigma := (\delta_{i\sigma(j)})_{i,j} \in \mathbb{Q}^{n \times n}$$

the *permutation matrix* of σ. Furthermore, $\mathrm{sgn}(\sigma) := \det(P_\sigma)$ is called the *signum* or *sign* of σ.

Remark 9.33. The permutation matrix of σ is formed by permuting the rows of the identity matrix (i.e., the standard basis $e_1, \ldots, e_n$) according to σ. Using the Gaussian algorithm, this permutation can be realized by finitely many row swaps (this corresponds to the sorting algorithm *selection sort*). Because of $\det(1_n) = 1$, $\mathrm{sgn}(\sigma) \in \{\pm 1\}$ is indeed a "sign".

Example 9.34. We determine the permutation matrix and the sign for the permutations in S_3:

σ	id	$(1,2)$	$(1,3)$	$(2,3)$	$(1,2,3)$	$(1,3,2)$
P_σ	1_3	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$
$\mathrm{sgn}(\sigma)$	1	-1	-1	-1	1	1

Theorem 9.35. *For $\sigma, \tau \in S_n$, it holds that $P_{\sigma\circ\tau} = P_\sigma P_\tau$ and* $\boxed{\operatorname{sgn}(\sigma \circ \tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau).}$

Proof. The entry of $P_\sigma P_\tau$ at position (i, j) is

$$\sum_{k=1}^{n} \delta_{i\sigma(k)}\delta_{k\tau(j)} = \delta_{i,\sigma(\tau(j))} = \delta_{i,(\sigma\circ\tau)(j)}.$$

This shows the first equation. The second follows from the determinant theorem. □

Remark 9.36.

(a) The permutation matrix of a transposition is exactly the elementary matrix for the swapping of rows. In particular, every transposition has sign -1. According to Theorem 9.35, the product of an even number of transpositions is never a product of an odd number of transpositions.

(b) According to Example 9.29, every k-cycle has sign $(-1)^{k-1}$. If σ is a product of pairwise disjoint cycles with lengths $l_1, \ldots, l_k$, then

$$\boxed{\operatorname{sgn}(\sigma) = (-1)^{l_1+\ldots+l_k-k}.}$$

For example,

$$\operatorname{sgn}\big((1,2,5,6)(3,7)(4,9,8)\big) = (-1)^{4+2+3-3} = 1.$$

If one counts 1-cycles as well, the formula simplifies to $\operatorname{sgn}(\sigma) = (-1)^{n-k}$.

(c) It follows from Theorem 9.35 that the permutations with sign 1 form a subgroup $A_n \leq S_n$. A_n is called the *alternating* group of degree n. In a certain way, A_n relates to S_n as $\mathrm{SL}(n, K)$ relates to $\mathrm{GL}(n, K)$.

Theorem 9.37 (Leibniz formula)**.** *For $A = (a_{ij}) \in K^{n\times n}$, it holds that*

$$\boxed{\det(A) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)}a_{2\sigma(2)}\ldots a_{n\sigma(n)}.}$$

Proof. The rows $a_1, \ldots, a_n$ of A can be expressed as a linear combination of the standard basis: $a_i = \sum_{j=1}^{n} a_{ij}e_j$. Since det is linear in each row (Lemma 9.5), it holds that

$$\det(A) = \sum_{i_1=1}^{n} a_{1i_1} \det\begin{pmatrix} e_{i_1} \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \sum_{i_1=1}^{n} a_{1i_1} \sum_{i_2=1}^{n} a_{2i_2} \det\begin{pmatrix} e_{i_1} \\ e_{i_2} \\ a_3 \\ \vdots \end{pmatrix} = \ldots$$

$$= \sum_{1\leq i_1,\ldots,i_n\leq n} a_{1i_1}\ldots a_{ni_n} \det\begin{pmatrix} e_{i_1} \\ \vdots \\ e_{i_n} \end{pmatrix}.$$

If there exist $s \neq t$ with $i_s = i_t$, then the corresponding determinant vanishes. Thus, one only needs to sum over the tuples $(i_1, \ldots, i_n)$ with pairwise distinct entries. Each such tuple describes a permutation $\sigma \in S_n$ with $\sigma(j) = i_j$ for $j = 1, \ldots, n$. It follows

$$\det(A) = \sum_{\sigma\in S_n} a_{1\sigma(1)}\ldots a_{n\sigma(n)} \det(P_\sigma) = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)}\ldots a_{n\sigma(n)}.$$ □

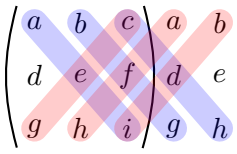
Corollary 9.38 (SARRUS Rule)**.** *For* 3×3 *matrices, it holds that*

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + cdh - gec - hfa - idb.$$

Proof. Use the Leibniz formula with Example 9.34. □

Remark 9.39.

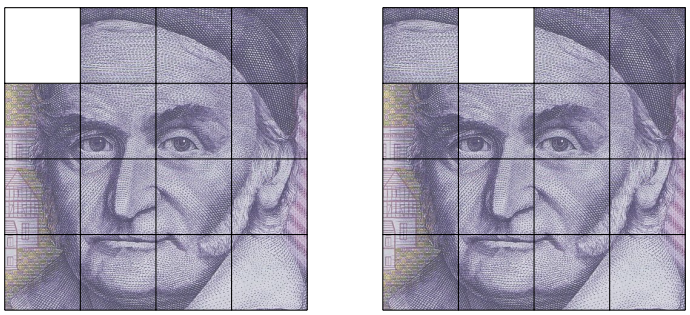
(a) One can remember the Sarrus rule with the following scheme:



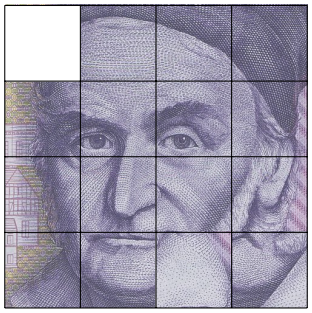
(b) Attention: The Sarrus rule holds *only* for 3×3 matrices (for 4×4 matrices, one needs $4! = 24$ summands).

(c) Let $A \in K^{n \times n}$ with eigenvalue λ. Then $E_\lambda(A) = \mathrm{Ker}(A - \lambda 1_n) \neq \{0\}$ and $\det(A - \lambda 1_n) = 0$. According to the Leibniz formula, $\det(A - \lambda 1_n)$ is a polynomial in λ. In this way, we will calculate all eigenvalues of A.

Example 9.40. The following *sliding puzzle* consists of 15 movable squares and one empty space. A square that is horizontally or vertically adjacent to the empty space may be moved into it (cf. cover):



Sam Loyd offered a prize of 1000 $ to anyone who succeeded in transforming the following configuration into the initial state:[7]



[7] see [D. Slocum and J. Sonneveld, *The 15 puzzle*, The Slocum Puzzle Foundation, Beverly Hills, 2006]

Every move corresponds to a transposition in S_{16}. If one assumes a checkerboard pattern, the empty space moves from black to white or vice versa with every move. Since the empty space in Loyd's configuration is in the initial position, an even number of moves is required for a solution. On the other hand, Loyd's configuration differs from the initial state by only one transposition. According to Remark 9.36, this configuration is therefore unsolvable and Loyd never had to pay out the prize money.

Exercises

Exercise I.1. Which of the following propositions are true in the year 2025?

(a) There is a month with 28 days.

(b) There is a month with exactly 28 days.

(c) There is exactly one month with 28 days.

(d) There is exactly one month with exactly 28 days.

Exercise I.2. Let $1 \leq a \leq b \leq 9$ be natural numbers. The logician (S)iegfried knows only the sum $a + b$, while his colleague (P)etrus knows only the product ab. The two conduct the following dialogue:

S: "I do not know a and b."	P: "I do not know a and b."
S: "I do not know a and b."	P: "I do not know a and b."
S: "I do not know a and b."	P: "I do not know a and b."
S: "I do not know a and b."	P: "I do not know a and b."
S: "I do not know a and b."	P: "Now I know a and b!"

Determine a and b from this.

Exercise I.3. For finite sets A and B, $|A \cup B| = |A| + |B| - |A \cap B|$ holds according to Lemma 1.12. Find and prove an analogous equation for three finite sets.

Exercise I.4. Prove by mathematical induction: The sum of the first n odd numbers is n^2.

Exercise I.5. Prove that $\mathbb{N} \times \mathbb{N}$ is countable.

Exercise I.6. Construct relations with the following properties:

(a) reflexive, but neither symmetric nor transitive.

(b) symmetric, but neither reflexive nor transitive.

(c) transitive, but neither reflexive nor symmetric.

Exercise I.7. Let $U := \{2z + 1 : z \in \mathbb{Z}\} \cup \{0\}$. Investigate whether $(U, +)$ is a group.

Exercise I.8. Let $(G, *)$ and $(H, \circ)$ be groups. Show that $G \times H$ with the operation

$$(g_1, h_1) \cdot (g_2, h_2) := (g_1 * g_2, h_1 \circ h_2)$$

becomes a group.

Exercise I.9. Construct a field with three elements.
Hint: What is $1 + 1$?

Exercise I.10. Show:

(a) A non-empty subset H of a group G is a subgroup if and only if $x, y \in H \Rightarrow xy^{-1} \in H$ holds.

(b) A non-empty subset U of a K-vector space V is a subspace if and only if for all $u, v \in U$ and $\lambda \in K$ it holds: $\lambda u + v \in U$.

Exercise I.11. (Dedekind identity) Let X, Y, Z be subspaces of a vector space V with $X \subseteq Z$. Show: $(X + Y) \cap Z = X + (Y \cap Z)$.

Exercise I.12. Let V be a vector space, $S \subseteq V$ and $U, W \leq V$. Show:

(a) $\langle S \rangle$ is the intersection of all subspaces of V that contain S.

(b) $U + W = \langle U \cup W \rangle$.

(c) $U \cup W \leq V \iff U \cup W \in \{U, W\}$.

Exercise I.13. Let u, v, w be vectors of a vector space V. Prove or disprove: $\{u, v, w\}$ is linearly independent if and only if $\{u, v\}$, $\{u, w\}$ and $\{v, w\}$ are linearly independent.

Exercise I.14. Obviously $\mathbb{R}$ is a $\mathbb{Q}$-vector space in which the scalar multiplication coincides with the usual multiplication in $\mathbb{R}$ (you do not need to check this). Show:

(a) 1 and $\sqrt{2}$ are linearly independent over $\mathbb{Q}$.

(b) $\mathbb{Q}(\sqrt{2}) := \langle 1, \sqrt{2} \rangle = \mathbb{Q} + \mathbb{Q}\sqrt{2}$ is a field with the same operations as in $\mathbb{R}$.

Exercise I.15. Show:

(a) Swapping two rows of a matrix can be realized by the other two elementary row operations.

(b) Every $n \times n$-matrix is a product of matrices of the form $1_n + \lambda E_{ij}$ with $\lambda \in K$ and $1 \leq i, j \leq n$ (the case $i = j$ is allowed).

Exercise I.16. Let $n = n_1 + \ldots + n_k$ and $\lambda_1, \ldots, \lambda_k \in K$ be pairwise distinct. Let

$$A := \operatorname{diag}(\lambda_1 1_{n_1}, \ldots, \lambda_k 1_{n_k}) \in K^{n \times n}$$

and $B \in K^{n \times n}$. Show:

(a) $AB = BA$ if and only if $B = \operatorname{diag}(B_1, \ldots, B_k)$ with $B_i \in K^{n_i \times n_i}$ for $i = 1, \ldots, k$.

(b) $AB = BA$ for all $B \in K^{n \times n}$ if and only if A is a scalar matrix (i. e. $k = 1$).

Exercise I.17. Let $A \in \mathbb{Q}^{n \times m} \subseteq \mathbb{R}^{n \times m}$ and $b \in \mathbb{Q}^{n \times 1} \subseteq \mathbb{R}^{n \times 1}$. Justify:

(a) The rank of A over $\mathbb{Q}$ is the rank of A over $\mathbb{R}$.

(b) If A is invertible over $\mathbb{R}$, then it is also invertible over $\mathbb{Q}$.

(c) If the system of equations $Ax = b$ has a solution in $\mathbb{R}^{m\times 1}$, then there also exists a solution in $\mathbb{Q}^{m\times 1}$.

(d) Give an example in which the solution sets of $Ax = b$ over $\mathbb{Q}$ and $\mathbb{R}$ are different.

Exercise I.18. Let U, V, W be vector spaces and $f\colon U \to V$, $g\colon V \to W$ and $h\colon W \to X$ be linear maps.

(a) Show $\operatorname{rk}(g \circ f) + \operatorname{rk}(h \circ g) \leq \operatorname{rk}(g) + \operatorname{rk}(h \circ g \circ f)$ (Frobenius inequality).
Hint: For $g(V) = g(f(U)) \oplus Y$, it holds that $h(g(V)) = h\big(g(f(U))\big) + h(Y)$.

(b) Deduce Lemma 5.15(a) from part (a).

(c) Show $\operatorname{rk}(A) + \operatorname{rk}(B) \leq \operatorname{rk}(AB) + n$ for $A \in K^{m\times n}$ and $B \in K^{n\times k}$ (Sylvester inequality).

Exercise I.19. Let K be a field and $n \in \mathbb{N}$. Show:

(a) The sum and the product of (strict) upper (resp. lower) triangular matrices in $K^{n\times n}$ are again (strict) upper (resp. lower) triangular matrices (Definition 8.14).

(b) The upper (resp. lower) triangular matrices form a subspace U of $K^{n\times n}$. Calculate $\dim U$.

(c) The set of invertible upper (resp. lower) triangular matrices in $K^{n\times n}$ is a subgroup of $\mathrm{GL}(n, K)$.

Exercise I.20. Let V be a K-vector space. For $\varphi \in \mathrm{GL}(V)$ and $v \in V$ let $f_{\varphi,v}\colon V \to V$, $w \mapsto \varphi(w) + v$. The maps of the form $f_{\mathrm{id}_V,v}$ are called *translations*. Show:

(a)
$$\mathrm{Aff}(V) := \{f_{\varphi,v} : \varphi \in \mathrm{GL}(V), v \in V\} \subseteq \mathrm{Fun}(V, V)$$
is a group wrt. composition of maps. Is it a subgroup of $\mathrm{GL}(V)$?

(b) The translations form a subgroup of $\mathrm{Aff}(V)$.

Remark: $\mathrm{Aff}(V)$ is called the *affine* group of V.

Exercise I.21. Let $A \in \mathrm{GL}(n, K)$. Show that the matrix $\binom{A}{1_n} \in K^{2n\times n}$ can be transformed into the form $\binom{1_n}{B}$ by elementary column operations. In this case, $B = A^{-1}$.

Exercise I.22. Let V, W be vector spaces and $f \in \mathrm{Hom}(V, W)$. Show:

(a) f is injective if and only if there exists a $g \in \mathrm{Hom}(W, V)$ with $g \circ f = \mathrm{id}_V$.

(b) f is surjective if and only if there exists a $g \in \mathrm{Hom}(W, V)$ with $f \circ g = \mathrm{id}_W$.

(c) Are the maps g uniquely determined in each case?

Exercise I.23. Let $\lambda \in K$ be an eigenvalue of $A \in K^{n\times n}$ and $k \in \mathbb{N}$. Show that λ^k is an eigenvalue of A^k. Does the converse also hold? Show that λ^{-1} is an eigenvalue of A^{-1} if A is invertible.

Exercise I.24. Let $A \in K^{n\times n}$ be diagonalizable. Show that A^{t} is also diagonalizable with the same eigenvalues. Do the eigenspaces also coincide?

Exercise I.25. Let $a, b \in K$ and

$$A = \begin{pmatrix} a & b & \cdots & b \\ b & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & b \\ b & \cdots & b & a \end{pmatrix} \in K^{n \times n}.$$

Show:

$$\det(A) = (a-b)^{n-1}(a + (n-1)b).$$

Hint: Eigenvalues.

Exercise I.26. Is the sliding puzzle on the cover solvable?

Exercise I.27. Transpositions of the form $(a, a+1) \in S_n$ are called *basic transpositions*. Show that every permutation is a product of basic transpositions.
Remark: This is the basis of *bubble sort*.

Exercise I.28. Show that $|A_n| = \frac{n!}{2}$ for $n \geq 2$.
Hint: Apply the Leibniz formula to the matrix $(1)_{i,j=1}^{n} \in \mathbb{Q}^{n \times n}$.

Exercise I.29. Let $\sigma, \tau \in S_n$ with $\operatorname{sgn}(\sigma) \neq \operatorname{sgn}(\tau)$. Show that $\det(P_\sigma + P_\tau) = 0$.
Hint: $P_\sigma + P_\tau = P_\sigma(P_\sigma^{-1} + P_\tau^{-1})P_\tau$.

Exercise I.30. Let $\sigma \in S_n$. Show that

$$\operatorname{sgn}(\sigma) = \prod_{1 \leq i < j \leq n} \frac{\sigma(j) - \sigma(i)}{j - i}.$$

Hint: First consider transpositions σ.

Exercise I.31 (General Expansion Theorem)**.** Let $A \in K^{r \times s}$, $B \in K^{s \times t}$, $I \subseteq \{1, \ldots, r\}$ and $J \subseteq \{1, \ldots, t\}$ with $|I| = |J| = k$. For a matrix $M = (m_{ij})$ let $M_{IJ} := (m_{ij})_{i \in I, j \in J}$. Show that:

$$\det\big((AB)_{IJ}\big) = \sum_{\substack{L \subseteq \{1,\ldots,s\} \\ |L| = k}} \det(A_{IL}) \det(B_{LJ}).$$

Remark: This generalizes matrix multiplication, the determinant theorem, and the Cauchy-Binet formula.

Exercise I.32. In the game *Lights Out*, a 5×5 grid of lights is given, which can be on or off. If you touch a light, the light along with its horizontal and vertical neighbors (up/down, right/left) changes state (on $\leftrightarrow$ off). The goal of the game is to turn off all lights of a given state.

(a) Convince yourself that a solution is uniquely given by a vector in $\mathbb{F}_2^{25}$, where each coordinate describes whether the corresponding light must be touched.

(b) Model the solution of the game as a system of equations with a coefficient matrix in $\mathbb{F}_2^{25 \times 25}$.

(c) Check (using a computer) how many of the 2^{25} states are solvable.

(d) How many solutions does a solvable state have and how do the solutions differ?

(e) How many "moves" (light switch touches) are required in the worst case for a solution?

(f) Which states with only one burning light are solvable?

(g) Develop an easy-to-learn algorithm for solving the game that does not require computer calculations.

Hint: Play here: `https://raw.org/research/solving-lightsout-using-linear-algebra`

Linear Algebra II

10 Polynomials

10.1 The Vector Space of Polynomials

Remark 10.1. In Remark 9.39, we hinted that the eigenvalues of a matrix A are solutions of certain (non-linear) polynomial equations. In this chapter, we define polynomials with coefficients in an arbitrary field and investigate their roots. From this, we derive a necessary and sufficient criterion for the diagonalizability of an endomorphism.

Definition 10.2. A (formal) *polynomial* over a field K in the *variable* X is a sum of the form

$$\alpha = \sum_{k=0}^{d} a_k X^k = a_0 + a_1 X + \ldots + a_d X^d$$

with *coefficients* $a_0, \ldots, a_d \in K$.[1]

- a_0 is called the *constant term* of α.
- Unless all coefficients are 0,

  $$\deg(\alpha) := \max\{d \in \mathbb{N}_0 : a_d \neq 0\}$$

  is called the *degree* of α and a_d the *leading* coefficient. In the case $a_d = 1$, α is called *monic*.
- For the *zero polynomial* (all coefficients are 0), one sets $\deg(0) := -\infty$.
- The set of all polynomials over K is denoted by $K[X]$.

Remark 10.3.

(a) If the degree of $\alpha \in K[X]$ is unknown, one writes $\alpha = \sum_{k=0}^{\infty} a_k X^k = \sum a_k X^k$ under the assumption that only finitely many coefficients are non-zero.

(b) Polynomials are considered equal if they have the same coefficients, i.e.

$$\sum_{k=0}^{\infty} a_k X^k = \sum_{k=0}^{\infty} b_k X^k \iff \forall k \in \mathbb{N}_0 : a_k = b_k.$$

(c) The field elements $\lambda \in K$ are identified with the *constant* polynomials $\lambda X^0 \in K[X]$. These are exactly the polynomials of degree ≤ 0. In particular, $0, 1 \in K \subseteq K[X]$ holds.

Example 10.4. The polynomial $\alpha = X^2 - 3X + 1 \in \mathbb{Q}[X]$ is monic of degree 2 with constant term 1.

[1] Formally: A polynomial is a map $\mathbb{N}_0 \to K$, $k \mapsto a_k$ with $|\{k \in \mathbb{N}_0 : a_k \neq 0\}| < \infty$.

Theorem 10.5. *With the operations*

$$\sum a_k X^k + \sum b_k X^k := \sum (a_k + b_k) X^k,$$
$$\lambda \sum a_k X^k := \sum (\lambda a_k) X^k$$

$K[X]$ becomes an infinite-dimensional K-vector space with basis $1, X, X^2, \ldots$.

Proof. Let $\alpha = \sum a_k X^k$ and $\beta = \sum b_k X^k$ with $d := \deg(\alpha) \geq \deg(\beta)$. One can view $(a_0, \ldots, a_d)$ and $(b_0, \ldots, b_d)$ as vectors in K^{d+1}. The operations in $K[X]$ correspond exactly to those in K^{d+1}. Therefore, $K[X]$ satisfies the vector space axioms. By definition, every polynomial is a finite linear combination of $1, X, X^2, \ldots$, i.e. $K[X] = \langle 1, X, X^2, \ldots \rangle$. From the uniqueness of the coefficients (Remark 10.3) follows the linear independence of $\{1, X, X^2, \ldots\}$. □

Remark 10.6.

(a) For $\alpha, \beta \in K[X]$ and $\lambda \in K$, it obviously holds that $\deg(\alpha + \beta) \leq \max\{\deg(\alpha), \deg(\beta)\}$ and $\deg(\lambda\alpha) \leq \deg(\alpha)$. Therefore, the polynomials of degree less than d form a d-dimensional subspace with basis $1, X, \ldots, X^{d-1}$.

(b) You probably know that polynomials can also be multiplied, e.g.

$$\begin{aligned}(2X^3 - X^2 + 5X - 1)(4X^2 + 3) &= 8X^5 - 4X^4 + (6 + 20)X^3 + (-3 - 4)X^2 + 15X - 3 \\ &= 8X^5 - 4X^4 + 26X^3 - 7X^2 + 15X - 3\end{aligned}$$

This can be formalized as follows.

Theorem 10.7. *For polynomials $\alpha = \sum a_k X^k$, $\beta = \sum b_k X^k$,*

$$\alpha \cdot \beta := \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l b_{k-l} \Big) X^k$$

is a polynomial of degree $\deg(\alpha) + \deg(\beta)$. The following calculation rules hold:

$$\alpha\beta = \beta\alpha, \qquad \alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \qquad \alpha(\beta\gamma) = (\alpha\beta)\gamma.$$

Proof. In the case $\alpha = 0$ or $\beta = 0$, we have $\alpha\beta = 0$ and $\deg(\alpha\beta) = -\infty = \deg(\alpha) + \deg(\beta)$. So let $d := \deg(\alpha) \geq 0$ and $e := \deg(\beta) \geq 0$. For $k > d + e$, we have $\sum_{l=0}^{k} a_l b_{k-l} = 0$ and $\deg(\alpha\beta) \leq d + e$. For $k = d + e$, we have $\sum_{l=0}^{k} a_l b_{k-l} = a_d b_e \neq 0$. This shows $\deg(\alpha\beta) = d + e$. In particular, $\alpha\beta \in K[X]$. For $\gamma = \sum c_k X^k$, it holds that

$$\alpha\beta = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l b_{k-l} \Big) X^k = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} b_l a_{k-l} \Big) X^k = \beta\alpha$$

$$\alpha(\beta + \gamma) = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l (b_{k-l} + c_{k-l}) \Big) X^k = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l b_{k-l} \Big) X^k + \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l c_{k-l} \Big) X^k = \alpha\beta + \alpha\gamma.$$

The coefficient of X^k in $\alpha(\beta\gamma)$ is

$$\sum_{l=0}^{k} a_l \sum_{m=0}^{k-l} b_m c_{k-l-m} = \sum_{\substack{r,s,t \in \mathbb{N}_0 \\ r+s+t=k}} a_r b_s c_t = \sum_{l=0}^{k} \Big(\sum_{m=0}^{l} a_m b_{l-m} \Big) c_{k-l}.$$

This is also the coefficient of X^k in $(\alpha\beta)\gamma$. Thus $\alpha(\beta\gamma) = (\alpha\beta)\gamma$. □

Remark 10.8. In contrast to matrix multiplication, the multiplication of polynomials is commutative. The only field axiom that $K[X]$ does not satisfy is the existence of inverses. For example, there exists no $\alpha \in K[X]$ with $X \cdot \alpha = 1$. Nevertheless, the cancellation rule holds: $\alpha\beta = \alpha\gamma \Rightarrow \beta = \gamma$, if $\alpha \neq 0$. This follows from

$$\deg(\beta - \gamma) \leq \deg(\alpha) + \deg(\beta - \gamma) = \deg(\alpha(\beta - \gamma)) = \deg(0) = -\infty.$$

One can therefore calculate in $K[X]$ just as in $\mathbb{Z}$.

Theorem 10.9 (Euclidean division)**.** *For $\alpha, \beta \in K[X]$ with $\beta \neq 0$, there exist uniquely determined polynomials $\gamma, \delta \in K[X]$ with $\alpha = \beta\gamma + \delta$ and $\deg \delta < \deg \beta$.*

Proof. Existence: Choose $\gamma \in K[X]$ such that

$$\delta := \alpha - \beta\gamma = a_d X^d + a_{d-1} X^{d-1} + \ldots + a_0$$

has the smallest possible degree $d \in \mathbb{N}_0 \cup \{-\infty\}$. Let $\beta = b_e X^e + \ldots + b_0$ and $e := \deg \beta$. If $d \geq e$, then

$$a_d b_e^{-1} X^{d-e} \beta = a_d b_e^{-1} (b_e X^d + b_{e-1} X^{d-1} + \ldots + b_0 X^{d-e}) = a_d X^d + \ldots$$

and it follows that

$$\deg\big(\alpha - \beta(\gamma + a_d b_e^{-1} X^{d-e})\big) = \deg\big(\delta - a_d b_e^{-1} X^{d-e} \beta\big) < d.$$

This is a contradiction to the choice of γ. Thus $d < e$ and $\alpha = \beta\gamma + \delta$.

Uniqueness: Now let $\alpha = \beta\tilde{\gamma} + \tilde{\delta}$ with $\tilde{\gamma}, \tilde{\delta} \in K[X]$ and $\deg \tilde{\delta} < e$. According to Theorem 10.7

$$e + \deg(\tilde{\gamma} - \gamma) = \deg(\beta) + \deg(\tilde{\gamma} - \gamma) = \deg(\beta(\tilde{\gamma} - \gamma)) = \deg(\delta - \tilde{\delta}) \leq \max\{\deg(\delta), \deg(\tilde{\delta})\} < e.$$

It follows that $\deg(\tilde{\gamma} - \gamma) = -\infty = \deg(\delta - \tilde{\delta})$. This shows $\tilde{\gamma} = \gamma$ and $\tilde{\delta} = \delta$. □

Definition 10.10. In the situation of Theorem 10.9, δ is called the *remainder* of the division of α by β. In the case $\delta = 0$, β is called a *divisor* of α and we write $\beta \mid \alpha$. If applicable, one also says "β *divides* α" or "α is *divisible* by β".

Example 10.11.

$$\begin{array}{rrrrl} (2X^3 & -X^2 & +5X & +1) : (X^2+3) = 2X - 1 =: \gamma \\ -(2X^3 & & +6X) & \\ & -X^2 & -X & +1 \\ & -(-X^2 & & -3) \\ & & -X & +4 =: \delta \end{array}$$

Thus $\alpha = 2X^3 - X^2 + 5X + 1 = (X^2+3)(2X-1) - X + 4 = \beta\gamma + \delta$ with $\deg \delta = 1 < 2 = \deg \beta$.

Remark 10.12. Division by monic polynomials of degree 1 can be performed efficiently using the HORNER *scheme*[2]. For this, let $\alpha = a_n X^n + a_{n-1} X^{n-1} + \ldots + a_0$ and $\beta = X - b$. We calculate $c_n := 0$, $c_k := a_{k+1} + b c_{k+1}$ for $k = n-1, \ldots, 0$ and $d := a_0 + b c_0$:

$$\begin{array}{cccccc} & a_n & a_{n-1} & a_{n-2} & \cdots & a_0 \\ + & 0 & bc_{n-1} & bc_{n-2} & \cdots & bc_0 \\ \hline & c_{n-1} & c_{n-2} & \cdots & c_0 & d \end{array}$$

[2] also called RUFFINI's *rule*

For $\gamma := c_{n-1}X^{n-1} + \ldots + c_0$, it now holds that

$$\begin{aligned}\beta\gamma + d &= c_{n-1}X^n + (c_{n-2} - bc_{n-1})X^{n-1} + \ldots + (c_0 - bc_1)X - bc_0 + d \\ &= a_nX^n + a_{n-1}X^{n-1} + \ldots + a_0 = \alpha.\end{aligned}$$

Example 10.13. For $\alpha = 2X^3 - X^2 + 3X + 1$ and $\beta = X - 2$ one obtains:

$$\begin{array}{rcccc} & 2 & -1 & 3 & 1 \\ + & 0 & 4 & 6 & 18 \\ \hline & 2 & 3 & 9 & 19 \end{array}$$

This shows $\alpha = \beta(2X^2 + 3X + 9) + 19$.

10.2 Roots

Definition 10.14. Let $\alpha = \sum_{k=0}^{d} a_kX^k \in K[X]$. One can *substitute* an element $x \in K$ for X in α:

$$\alpha(x) := \sum_{k=0}^{d} a_kx^k \in K.$$

One calls x a *root* of α if $\alpha(x) = 0$.

Lemma 10.15. *For $\alpha, \beta \in K[X]$ and $x \in K$, it holds that*

$$\begin{aligned}(\alpha + \beta)(x) &= \alpha(x) + \beta(x), \\ (\alpha\beta)(x) &= \alpha(x)\beta(x).\end{aligned}$$

Proof. Let $\alpha = \sum a_kX^k$ and $\beta = \sum b_kX^k$. Then

$$\begin{aligned}(\alpha + \beta)(x) &= \sum (a_k + b_k)x^k = \sum a_kx^k + \sum b_kx^k = \alpha(x) + \beta(x), \\ (\alpha\beta)(x) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} a_kb_{n-k}x^n = \sum_{n=0}^{\infty}\sum_{k=0}^{n}(a_kx^k)(b_{n-k}x^{n-k}) = \sum a_kx^k \sum b_kx^k = \alpha(x)\beta(x).\end{aligned}$$ □

Remark 10.16. Mnemonic: It does not matter whether you first add/multiply and then evaluate, or first evaluate and then add/multiply. Caution: In general, $\alpha(x+y) \neq \alpha(x)+\alpha(y)$ and $\alpha(xy) \neq \alpha(x)\alpha(y)$ for $\alpha \in K[X]$ and $x, y \in K$.

Theorem 10.17 (Interpolation)**.** *Let $x_1, \ldots, x_n \in K$ be pairwise distinct and $y_1, \ldots, y_n \in K$ be arbitrary. Then there exists exactly one polynomial α of degree $< n$ with $\alpha(x_i) = y_i$ for $i = 1, \ldots, n$.*

Proof. Let $\alpha = a_0 + a_1X + \ldots + a_{n-1}X^{n-1}$. The condition $\alpha(x_i) = y_i$ for $i = 1, \ldots, n$ means:

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The coefficient matrix of this system of equations is a Vandermonde matrix. Since the x_i are pairwise distinct, the matrix is invertible according to Theorem 9.20. Thus, there exists exactly one solution $(a_0, \ldots, a_{n-1})$. □

Remark 10.18. An explicit solution to the interpolation problem is given by the LAGRANGE *polynomial*

$$\alpha := \sum_{i=1}^{n} y_i \prod_{j \neq i} \frac{X - x_j}{x_i - x_j} \in K[X]$$

(verify by calculation).

Example 10.19.

(a) For $n = 2$ and $K = \mathbb{R}$, Theorem 10.17 is the geometric statement that two distinct points in $\mathbb{R}^2$ are connected by exactly one line.

(b) We are looking for a polynomial $\alpha \in \mathbb{R}[X]$ through the points $(-1, 2)$, $(0, 1)$ and $(1, 3)$. The proof of Theorem 10.17 leads to the system of equations

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$

with the unique solution $\alpha = 1 + \frac{1}{2}X + \frac{3}{2}X^2$.

Corollary 10.20.

(a) Every polynomial $\alpha \in K[X]$ of degree $d \geq 0$ has at most d roots.

(b) Let $|K| = \infty$ and $\alpha, \beta \in K[X]$ with $\alpha(x) = \beta(x)$ for all $x \in K$. Then $\alpha = \beta$.

Proof.

(a) Suppose α has pairwise distinct roots $x_1, \ldots, x_{d+1} \in K$. According to Theorem 10.17 with $y_1 = \ldots = y_{d+1} = 0$, α is the unique polynomial of degree $\leq d$ with these roots. On the other hand, the zero polynomial also has these roots. Thus $\alpha = 0$ and $d = -\infty$, contradicting the assumption.

(b) Since $|K| = \infty$, $\alpha - \beta$ has infinitely many roots. From (a) it follows that $\alpha = \beta$. □

Remark 10.21. For $K = \mathbb{R}$, every polynomial $\alpha \in \mathbb{R}[X]$ is uniquely determined by the (continuous) function $\mathbb{R} \to \mathbb{R}$, $x \mapsto \alpha(x)$, because $|K| = \infty$. In analysis, one therefore does not distinguish between a polynomial and a function. Over finite fields K, one would lose information in doing so, because there are only finitely many mappings $K \to K$, but infinitely many polynomials. For example, the polynomials $X, X^2, \ldots \in \mathbb{F}_2[X]$ all correspond to the identity $\mathrm{id}_{\mathbb{F}_2}$.

Lemma 10.22. *$x \in K$ is a root of α if and only if $(X - x) \mid \alpha$.*

Proof. Euclidean division yields $\gamma, \delta \in K[X]$ with $\alpha = (X - x)\gamma + \delta$ and $\deg \delta < \deg(X - x) = 1$, i. e. $\delta \in K$. Now

$$\delta = \delta(x) = \big(\alpha - (X - x)\gamma\big)(x) \overset{10.15}{=} \alpha(x) - (x - x)\gamma(x) = \alpha(x).$$ □

Definition 10.23. Let $x \in K$ be a root of α. One calls $X - x$ a *linear factor* of α. The largest number $e \in \mathbb{N}$ with $(X - x)^e \mid \alpha$ is called the *algebraic multiplicity* of the root x. In the case $e = 1$, one speaks of a *simple* root and otherwise of a *multiple* root.

Lemma 10.24. *Let $x_1, \ldots, x_n \in K$. Then every monic divisor of $(X - x_1) \ldots (X - x_n) \in K[X]$ has the form $(X - x_{i_1}) \ldots (X - x_{i_k})$ with $1 \leq i_1 < \ldots < i_k \leq n$.*

Proof. In the case $n = 1$, 1 (the empty product with $k = 0$) and $X - x_1$ are the only monic divisors. So let $n \geq 2$ and the assertion be already proven for $n-1$. Let $\alpha, \beta \in K[X]$ with $(X-x_1) \ldots (X-x_n) = \alpha\beta$. Then $\alpha(x_n)\beta(x_n) = (\alpha\beta)(x_n) = 0$, wlog. let $\alpha(x_n) = 0$. According to Lemma 10.22, $\alpha = (X - x_n)\gamma$ for some $\gamma \in K[X]$. According to Remark 10.8, one may cancel $X - x_n$ and obtains $(X-x_1) \ldots (X-x_{n-1}) = \gamma\beta$. The assertion now follows by induction. □

Example 10.25.

(a) Let $\alpha = X^3 + X^2 - 5X + 3 \in \mathbb{R}[X]$. A root $x \in \mathbb{R}$ is a solution to the equation

$$x^3 + x^2 - 5x + 3 = 0.$$

Even though there are solution formulas for such equations (of third and fourth degree[3]), these are cumbersome in practice. We will therefore choose our examples (and exercises) such that one can guess "small" integer roots. Suppose there is a root $x \in \mathbb{Z}$. Because of $x(x^2 + x - 5) = -3$, x is a divisor of 3, i.e., $x \in \{\pm 1, \pm 3\}$. One easily checks that $x_1 = 1$ is indeed a root ($1^3 + 1^2 - 5 \cdot 1 + 3 = 0$). Polynomial division (for example with the Horner scheme) yields

$$(X^3 + X^2 - 5X + 3) : (X - 1) = X^2 + 2X - 3 =: \gamma.$$

For every root $y \in \mathbb{R}$ of γ, it now holds that $\alpha(y) = (y - 1)\gamma(y) = 0$, i.e., y is also a root of α. With the p-q-formula $\frac{1}{2}(-p \pm \sqrt{p^2 - 4q})$ for quadratic equations, one obtains the roots of γ:

$$x_2 = \frac{1}{2}\big(-2 + \sqrt{4 + 12}\big) = 1, \qquad x_3 = \frac{1}{2}\big(-2 - \sqrt{4 + 12}\big) = -3.$$

Therefore $x_1 = x_2 = 1$ is a root of α with algebraic multiplicity 2 (a *double* root). Furthermore, α *splits* into linear factors $\alpha = (X - 1)^2(X + 3)$.

(b) Obviously $\alpha(0)$ is the constant term of $\alpha \in K[X]$. Thus $x = 0$ is a root of α if and only if the constant term of α vanishes.

(c) As is well known, $X^2 + 1 \in \mathbb{R}[X]$ has no root. We will later construct a "larger" field over which this polynomial also splits into linear factors (Lemma 11.27).

(d) The polynomial $X^2 + X + 1 \in \mathbb{F}_2[X]$ has no root in $\mathbb{F}_2$, because only 0 and 1 are candidates.

10.3 Characteristic Polynomials

Remark 10.26. In the following, we consider matrices with entries in $K[X]$. Due to the calculation rules for polynomials (Theorem 10.7), one easily sees that the usual calculation rules (Lemma 5.8) for matrices also hold in $K[X]^{n \times n}$. Finally, one can even apply the definition of the determinant to matrices in $K[X]^{n \times n}$ (in doing so, matrix entries are only added and multiplied, but never divided). Likewise, the determinant theorem, the Laplace expansion, the Leibniz formula, and Theorem 9.22 about the complementary matrix remain correct in this greater generality. On the other hand, the Gaussian algorithm does not work in $K[X]^{n \times n}$, because division is required here.

[3] see Algebra notes

Definition 10.27. For $A = (a_{ij}) \in K^{n\times n}$ we consider the matrix

$$X1_n - A = \begin{pmatrix} X - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & X - a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{n-1,n} \\ -a_{n1} & \cdots & -a_{n,n-1} & X - a_{nn} \end{pmatrix} \in K[X]^{n\times n}.$$

One calls $\chi_A := \det(X1_n - A) \in K[X]$ the *characteristic polynomial* of A.[4]

Lemma 10.28. *Similar matrices have the same characteristic polynomial.*

Proof. For $A \in K^{n\times n}$ and $S \in \mathrm{GL}(n, K)$ it holds that

$$\chi_{SAS^{-1}} = \det(X1_n - SAS^{-1}) = \det(S(X1_n - A)S^{-1}) \overset{10.28}{=} \det(X1_n - A) = \chi_A. \qquad \square$$

Definition 10.29. Let V be an n-dimensional K-vector space with basis B. For $f \in \mathrm{End}(V)$ one defines

$$\det(f) := \det({}_B[f]_B), \qquad \chi_f := \det(X1_n - {}_B[f]_B).$$

According to Corollary 7.27, Corollary 9.12 and Lemma 10.28, $\det(f)$ and χ_f do not depend on the choice of B. In the following theorems, one can therefore replace matrices with endomorphisms (and vice versa).

Example 10.30. The characteristic polynomial of $A := \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right) \in \mathbb{Q}[X]$ is

$$\chi_A = \det \begin{pmatrix} X - 1 & -2 \\ -3 & X - 4 \end{pmatrix} = (X - 1)(X - 4) - (-2)(-3) = X^2 - 5X - 2 = X^2 - \mathrm{tr}(A)X + \det(A).$$

Lemma 10.31. *For $A \in K^{n\times n}$ it holds that $\chi_A = X^n - \mathrm{tr}(A)X^{n-1} + \ldots + (-1)^n \det(A)$. In particular, χ_A is monic of degree n.*

Proof. Let $A = (a_{ij})$. According to the Leibniz formula, it holds that

$$\begin{aligned} \chi_A = \det(X1_n - A) &= (X - a_{11})(X - a_{22}) \ldots (X - a_{nn}) \\ &+ \sum_{\sigma \in S_n \setminus \{\mathrm{id}\}} \mathrm{sgn}(\sigma)(\delta_{1\sigma(1)}X - a_{1\sigma(1)}) \ldots (\delta_{n\sigma(n)}X - a_{n\sigma(n)}). \end{aligned}$$

For $\sigma \in S_n \setminus \{\mathrm{id}\}$ there exists a $k \in \{1, \ldots, n\}$ with $l := \sigma(k) \neq k$. Since σ is injective, it holds that $\sigma(l) \neq \sigma(k) = l$. Therefore $\delta_{k\sigma(k)} = 0 = \delta_{l\sigma(l)}$ and

$$(\delta_{1\sigma(1)}X - a_{1\sigma(1)}) \ldots (\delta_{n\sigma(n)}X - a_{n\sigma(n)})$$

is a polynomial of degree $\leq n - 2$. In total,

$$\chi_A = (X - a_{11})(X - a_{22}) \ldots (X - a_{nn}) + \alpha$$

with $\deg(\alpha) \leq n-2$. Multiplying out shows $\chi_A = X^n - (a_{11} + \ldots + a_{nn})X^{n-1} + \ldots = X^n - \mathrm{tr}(A)X^{n-1} + \ldots$. To calculate the constant term, one sets $X = 0$ and obtains $\chi_A(0) \overset{10.15}{=} \det(-A) = (-1)^n \det(A)$ from Remark 9.8. $\square$

[4] In some books, χ_A is defined by $\det(A - X1_n) = (-1)^n \det(X1_n - A)$. This does not make a big difference, but brings the disadvantage that χ_A is not monic if n is odd.

Theorem 10.32. *The eigenvalues of $A \in K^{n\times n}$ are the roots of χ_A.*

Proof. It holds that

$$\operatorname{Ker}(A-\lambda 1_n) \neq \{0\} \iff \det(A-\lambda 1_n) = 0 \overset{9.8}{\iff} \det(\lambda 1_n - A) = 0 \overset{10.15}{\iff} \chi_A(\lambda) = 0. \qquad \square$$

Lemma 10.33. *Let $\lambda \in K$ be an eigenvalue of $f \in \operatorname{End}(V)$. Then the geometric multiplicity of λ is at most as large as the algebraic multiplicity of λ as a root of χ_f.*

Proof. One extends a basis $b_1, \ldots, b_e$ of $E_\lambda(f)$ to a basis $B := \{b_1, \ldots, b_n\}$ of V. Then it holds that

$$\chi_f = \det(X1_n - {}_B[f]_B) = \det \begin{pmatrix} (X-\lambda)1_e & * \\ 0 & * \end{pmatrix}.$$

Lemma 9.9 shows $\chi_f = (X-\lambda)^e \beta$ for some $\beta \in K[X]$. Thus the algebraic multiplicity of λ as a root of χ_f is at least e. $\square$

Theorem 10.34. *$f \in \operatorname{End}(V)$ is diagonalizable if and only if χ_f splits into linear factors and for every root of χ_f the algebraic multiplicity coincides with the geometric multiplicity.*

Proof. Let $\lambda_1, \ldots, \lambda_k \in K$ be the distinct roots of χ_f with algebraic multiplicities $m_1, \ldots, m_k$. Then there exists an $\alpha \in K[X]$ with $\chi_f = (X-\lambda_1)^{m_1} \ldots (X-\lambda_k)^{m_k}\alpha$. Let m_i' be the geometric multiplicity of λ_i as an eigenvalue of f. According to Lemma 10.33 it holds that

$$m_1' + \ldots + m_k' \leq m_1 + \ldots + m_k + \deg(\alpha) \overset{10.7}{=} \deg\big((X-\lambda_1)^{m_1} \ldots (X-\lambda_k)^{m_k}\alpha\big) = \deg(\chi_f) = \dim V.$$

According to Theorem 8.10, V possesses a basis of eigenvectors if and only if $m_1' + \ldots + m_k' = \dim V$. This holds if and only if χ_f splits into linear factors (i.e., $\alpha = 1$) and the algebraic multiplicities coincide with the geometric multiplicities (i.e., $m_i = m_i'$ for $i = 1, \ldots, k$). $\square$

Remark 10.35. If χ_A splits into linear factors, then

$$\chi_A = (X-\lambda_1)\ldots(X-\lambda_n) = X^n - (\lambda_1 + \ldots + \lambda_n)X^{n-1} + \ldots + (-1)^n \lambda_1 \ldots \lambda_n.$$

A comparison with Lemma 10.31 shows:

$$\boxed{\begin{aligned} \operatorname{tr}(A) &= \lambda_1 + \ldots + \lambda_n, \\ \det(A) &= \lambda_1 \ldots \lambda_n, \end{aligned}}$$

i.e., the trace is the sum of the eigenvalues and the determinant is the product of the eigenvalues (provided these exist). If one has already determined $\lambda_1, \ldots, \lambda_{n-1}$, then one obtains $\lambda_n = \operatorname{tr}(A) - \lambda_1 - \ldots - \lambda_{n-1}$.

Theorem 10.36 (Mirsky)**.** *Let $d_1, \ldots, d_n, \lambda_1, \ldots, \lambda_n \in K$ with $d_1 + \ldots + d_n = \lambda_1 + \ldots + \lambda_n$. Then there exists a matrix $A \in K^{n\times n}$ with main diagonal $d_1, \ldots, d_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$.*

Proof. In the case $n = 1$, $A = (d_1) = (\lambda_1)$ satisfies the claim. Let $n \geq 2$ and $A = (a_{ij}) \in K^{n\times n} \setminus K1_n$ be a triangular matrix with main diagonal $\lambda_1, \ldots, \lambda_n$. Then A has eigenvalues $\lambda_1, \ldots, \lambda_n$. According to Fillmore's theorem, A is similar to a matrix with main diagonal $d_1, \ldots, d_n$. This matrix has the same eigenvalues as A. $\square$

Example 10.37.

(a) We are looking for a matrix with eigenvalues $1, 1, 1$ and main diagonal $0, 0, 3$. For this, we apply Fillmore's theorem to

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

The transition to the basis $\{(1,0,0), (1,1,0), (1,2,1)\}$ yields

$$A \approx \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -3 \\ 0 & 1 & 3 \end{pmatrix}.$$

(b) The FIBONACCI *numbers* F_k are defined recursively:

$$F_k := k \quad (k = 0, 1) \qquad F_{k+1} := F_k + F_{k-1} \quad (k \geq 1).$$

k	0	1	2	3	4	5	6	7	8	9	10
F_k	0	1	1	2	3	5	8	13	21	34	55

We are looking for an explicit formula for F_k. For $A := \left(\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$ we have

$$\begin{pmatrix} F_{k+1} \\ F_k \end{pmatrix} = A \begin{pmatrix} F_k \\ F_{k-1} \end{pmatrix} = A^2 \begin{pmatrix} F_{k-1} \\ F_{k-2} \end{pmatrix} = \ldots = A^k \begin{pmatrix} F_1 \\ F_0 \end{pmatrix} = A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

To calculate A^k, we diagonalize A. Because of $\chi_A = (X-1)X - 1 = X^2 - X - 1$, A has the eigenvalues $\varphi := \frac{1+\sqrt{5}}{2}$ and $\psi := \frac{1-\sqrt{5}}{2}$ (one calls $\varphi \approx 1.618$ the *golden ratio*). One calculates

$$E_\varphi(A) = \operatorname{Ker}(A - \varphi 1_2) = \langle \begin{pmatrix} \varphi \\ 1 \end{pmatrix} \rangle,$$

$$E_\psi(A) = \langle \begin{pmatrix} \psi \\ 1 \end{pmatrix} \rangle.$$

For $S := \begin{pmatrix} \varphi & \psi \\ 1 & 1 \end{pmatrix}$ we thus have $S^{-1}AS = \operatorname{diag}(\varphi, \psi)$ and

$$A^k = (S \operatorname{diag}(\varphi, \psi) S^{-1})^k = S \operatorname{diag}(\varphi, \psi)^k S^{-1} = S \operatorname{diag}(\varphi^k, \psi^k) S^{-1}.$$

According to Example 9.23,

$$S^{-1} = \frac{1}{\det(S)} \widetilde{S} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -\psi \\ -1 & \varphi \end{pmatrix}.$$

Overall, one obtains

$$A^k = S \operatorname{diag}(\varphi^k, \psi^k) S^{-1} = \frac{1}{\sqrt{5}} \begin{pmatrix} \varphi^{k+1} & \psi^{k+1} \\ \varphi^k & \psi^k \end{pmatrix} \begin{pmatrix} 1 & -\psi \\ -1 & \varphi \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} * & * \\ \varphi^k - \psi^k & * \end{pmatrix}$$

and

$$\boxed{F_k = \frac{1}{\sqrt{5}} (\varphi^k - \psi^k)} \qquad \text{(BINET } \textit{formula}^5\text{)}$$

Because of $|\psi^k| \approx 0.618^k \to 0$, it holds that $F_k \approx \frac{1}{\sqrt{5}} \varphi^k$, i.e., F_k grows exponentially.

[5] One can also prove the formula by induction, provided one has guessed it beforehand.

Lemma 10.38. *For $A \in K^{m\times n}$ and $B \in K^{n\times m}$, it holds that $X^n\chi_{AB} = X^m\chi_{BA}$. In particular, AB and BA have the same non-zero eigenvalues.*

Proof. According to the rules for block matrices (Remark 5.9, Lemma 9.9) and the determinant theorem, it holds that

$$\begin{aligned} X^n\chi_{AB} &= \det\left(\begin{pmatrix} 1_m & 0 \\ -B & X1_n \end{pmatrix}\begin{pmatrix} X1_m - AB & A \\ 0 & 1_n \end{pmatrix}\begin{pmatrix} 1_m & 0 \\ B & 1_n \end{pmatrix}\right) \\ &= \det\left(\begin{pmatrix} 1_m & 0 \\ -B & X1_n \end{pmatrix}\begin{pmatrix} X1_m & A \\ B & 1_n \end{pmatrix}\right) = \det\begin{pmatrix} X1_m & A \\ 0 & X1_n - BA \end{pmatrix} = X^m\chi_{BA}. \end{aligned}$$

Every eigenvalue $\lambda \in K$ of AB is a root of $X^n\chi_{AB} = X^m\chi_{BA}$. In the case $\lambda \neq 0$, λ must be a root of χ_{BA}. Then λ is also an eigenvalue of BA (and vice versa). □

Remark 10.39. In the case $n = m$, even $\chi_{AB} = \chi_{BA}$ holds in the situation of Lemma 10.38.

10.4 Minimal Polynomials

Remark 10.40. We have already substituted polynomials into matrices. We now conversely substitute matrices into polynomials. For $\alpha = \sum_{k=0}^{d} a_k X^k \in K[X]$ and $A \in K^{n\times n}$, we define

$$\alpha(A) := \sum_{k=0}^{d} a_k A^k \in K^{n\times n}.$$

The rules from Lemma 10.15 also hold in this generality.

Theorem 10.41. *For $A \in K^{n\times n}$, there exists exactly one monic polynomial $\mu_A \in K[X] \setminus \{0\}$ with $\mu_A(A) = 0_n$ and $\deg(\mu_A)$ minimal.*

Proof. Due to $\dim K^{n\times n} = n^2$ (Lemma 5.4), the powers $1_n = A^0, A, A^2, \ldots, A^{n^2}$ are linearly dependent in $K^{n\times n}$. Thus, there exist $a_0, \ldots, a_{n^2} \in K$ (not all 0) with $\sum_{k=0}^{n^2} a_k A^k = 0$. For $\alpha = \sum a_k X^k \in K[X]$, it thus holds that $\alpha(A) = 0$. By dividing by the leading coefficient of α, one can assume that α is monic. This shows that μ_A exists. Let $\tilde{\mu} \in K[X]$ also be monic with $\tilde{\mu}(A) = 0$ and $\deg(\tilde{\mu}) = \deg(\mu_A)$ minimal. Then $(\mu_A - \tilde{\mu})(A) = \mu_A(A) - \tilde{\mu}(A) = 0$ and $\deg(\mu_A - \tilde{\mu}) < \deg(\mu_A)$. The minimality of $\deg(\mu_A)$ shows $\mu_A - \tilde{\mu} = 0$, i.e., μ_A is uniquely determined. □

Definition 10.42. One calls μ_A the *minimal polynomial* of A.

Example 10.43. Let $A := \left(\begin{smallmatrix} 1 & -1 \\ 2 & 0 \end{smallmatrix}\right) \in \mathbb{Q}^{2\times 2}$. Since A is not a scalar matrix, $\deg \mu_A \geq 2$ holds. We use the ansatz

$$A^2 + xA + y1_2 = \begin{pmatrix} -1 & -1 \\ 2 & -2 \end{pmatrix} + x\begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix} + y\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

with $x, y \in \mathbb{Q}$. A comparison of the matrix entries at position $(1, 2)$ shows $x = -1$. Indeed, the equation now holds for $y = 2$. Therefore, $\mu_A = X^2 - X + 2$.

Lemma 10.44. *Let $A \in K^{n\times n}$ and $\alpha \in K[X]$ with $\alpha(A) = 0$. Then $\mu_A \mid \alpha$ holds.*

Proof. We divide with remainder: $\alpha = \mu_A\gamma + \delta$ with $\gamma, \delta \in K[X]$ and $\deg(\delta) < \deg(\mu_A)$. Then

$$\delta(A) = (\alpha - \mu_A\gamma)(A) = \alpha(A) - \mu_A(A)\gamma(A) = 0.$$

From the minimality of $\deg(\mu_A)$ it follows that $\delta = 0$ and $\mu_A \mid \alpha$. □

Lemma 10.45. *Similar matrices have the same minimal polynomial.*

Proof. Let $A \in K^{n\times n}$ with $\mu_A = \sum a_k X^k$. For $S \in \mathrm{GL}(n, K)$ we have

$$\mu_A(SAS^{-1}) = \sum a_k (SAS^{-1})^k = \sum a_k SA^kS^{-1} = S\Big(\sum a_k A^k\Big)S^{-1} = S\mu_A(A)S^{-1} = 0_n.$$

From Lemma 10.44 it follows that $\mu_{SAS^{-1}} \mid \mu_A$. Since similarity is a symmetric relation, $\mu_A \mid \mu_{SAS^{-1}}$ also holds. Since both minimal polynomials are monic, they must be equal. □

Definition 10.46. Let V be a K-vector space with basis B. For $f \in \mathrm{End}(V)$, let as usual $\mu_f := \mu_{{}_B[f]_B}$. According to Lemma 10.45, μ_f does not depend on the choice of B. The following theorems about matrices also apply analogously to endomorphisms.

Remark 10.47. From the proof of Theorem 10.41 one obtains $\deg(\mu_A) \leq n^2$. The next theorem implies $\deg(\mu_A) \leq n$.

Theorem 10.48 (Cayley-Hamilton)**.** *For $A \in K^{n\times n}$, $\chi_A(A) = 0$ and $\mu_A \mid \chi_A$ hold.*

Proof. Let $B := X1_n - A \in K[X]^{n\times n}$ and $\widetilde{B} \in K[X]^{n\times n}$ be the complementary matrix of B. From each entry of $\widetilde{B}$ we extract the coefficient of X^k and form the matrix $B_k \in K^{n\times n}$ from them. It now holds that

$$\widetilde{B} = \sum_{k=0}^{\infty} B_k X^k,$$

where only finitely many of the B_k are non-zero. Let $\chi_A = \sum a_k X^k$. According to Theorem 9.22 we have

$$\sum_{k=0}^{\infty} a_k 1_n X^k = \chi_A 1_n = \det(B)1_n = \widetilde{B}B = \sum_{k=0}^{\infty} B_k X^k (X1_n - A) = \sum_{k=0}^{\infty} (B_{k-1} - B_k A)X^k,$$

where $B_{-1} := 0_n$. A comparison of coefficients yields $a_k 1_n = B_{k-1} - B_k A$ for $k = 0, 1, \ldots$. Therefore

$$\chi_A(A) = \sum_{k=0}^{\infty} a_k A^k = \sum_{k=0}^{\infty} (B_{k-1}A^k - B_k A^{k+1}) = \sum_{k=0}^{\infty} B_{k-1}A^k - \sum_{k=0}^{\infty} B_k A^{k+1} = 0.$$

The second assertion follows from Lemma 10.44. □

Example 10.49.

(a) For $A \in K^{2\times 2}$, $A^2 - \mathrm{tr}(A)A + \det(A)1_2 = 0$ holds according to Lemma 10.31.

(b) Let $A \in \mathrm{GL}(n, K)$ with $\chi_A = \mu_A \gamma$ for some $\gamma \in K[X]$. According to Lemma 10.31,

$$\mu_A(0)\gamma(0) = \chi_A(0) = \det(A) \neq 0.$$

Thus μ_A has the form $\mu_A = X^d + a_{d-1}X^{d-1} + \ldots + a_0$ with $a_0 \neq 0$. One can now multiply the equation $A^d + a_{d-1}A^{d-1} + \ldots + a_0 1_n = 0$ on both sides by A^{-1} and obtains $A^{d-1} + a_{d-1}A^{d-2} + \ldots + a_1 1_n + a_0 A^{-1} = 0$. This yields a formula for the inverse

$$A^{-1} = -\frac{1}{a_0}\big(A^{d-1} + a_{d-1}A^{d-2} + \ldots + a_1 1_n\big).$$

Specifically for $n = 2$:

$$A^{-1} \overset{\text{(a)}}{=} \frac{1}{\det(A)}\big(\mathrm{tr}(A)1_2 - A\big) = \frac{1}{\det(A)}\widetilde{A}$$

(cf. Example 9.23).

Theorem 10.50. *The eigenvalues of $A \in K^{n\times n}$ are the roots of μ_A, i.e., χ_A and μ_A have the same roots (not necessarily with the same multiplicities).*

Proof. By Cayley-Hamilton, every root of μ_A is also a root of χ_A and thus an eigenvalue of A (Theorem 10.32). Conversely, let $\lambda \in K$ be an eigenvalue of A with eigenvector $v \in K^{n\times 1}$. For $k \in \mathbb{N}_0$, $A^k v = A^{k-1}\lambda v = \ldots = \lambda^k v$. Let $\mu_A = \sum a_k X^k$. Then

$$0 = \mu_A(A)v = \sum a_k A^k v = \sum a_k \lambda^k v = \mu_A(\lambda)v.$$

Since $v \neq 0$, $\mu_A(\lambda) = 0$, i.e., λ is a root of μ_A. □

Remark 10.51. Because $\deg \mu_A \leq \deg \chi_A$, Theorem 10.50 simplifies the determination of the eigenvalues. On the other hand, it is not clear how to compute μ_A efficiently. The next theorem improves Corollary 8.12.

Theorem 10.52. *$A \in K^{n\times n}$ is diagonalizable if and only if μ_A splits into pairwise distinct linear factors.*

Proof. Let $A \in K^{n\times n}$ be diagonalizable. Then there exists $S \in \mathrm{GL}(n, K)$ with $S^{-1}AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. The λ_i can be sorted by arranging the columns of S (i. e. the eigenvectors of A) accordingly. According to Lemma 10.45 we can assume

$$A = \begin{pmatrix} \lambda_1 1_{n_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_k 1_{n_k} \end{pmatrix}$$

where $n = n_1 + \ldots + n_k$ and $\lambda_i \neq \lambda_j$ for $i \neq j$. Then

$$(A - \lambda_1 1_n)\ldots(A - \lambda_k 1_n) = \mathrm{diag}(0_{n_1}, *, \ldots, *)\,\mathrm{diag}(*, \ldots, *, 0_{n_2}, *, \ldots, *)\ldots\mathrm{diag}(*, \ldots, *, 0_{n_k}) = 0_n.$$

According to Lemma 10.44, μ_A is a divisor of $(X - \lambda_1)\ldots(X - \lambda_k)$. On the other hand, $\lambda_1, \ldots, \lambda_k$ are eigenvalues and thus roots of μ_A. This shows $\mu_A = (X - \lambda_1)\ldots(X - \lambda_k)$.

Conversely, assume $\mu_A = (X - \lambda_1)\ldots(X - \lambda_k)$ with pairwise distinct $\lambda_1, \ldots, \lambda_k$. Let P_k be the k-dimensional vector space of all polynomials of degree less than k (Remark 10.6). For $i = 1, \ldots, k$ let

$$\gamma_i := (X - \lambda_1)\ldots(X - \lambda_{i-1})(X - \lambda_{i+1})\ldots(X - \lambda_k) \in P_k.$$

Let $a_1, \ldots, a_k \in K$ with $a_1\gamma_1 + \ldots + a_k\gamma_k = 0$. Then

$$a_i(\lambda_i - \lambda_1)\ldots(\lambda_i - \lambda_{i-1})(\lambda_i - \lambda_{i+1})\ldots(\lambda_i - \lambda_n) = a_i\gamma_i(\lambda_i) = (a_1\gamma_1 + \ldots + a_k\gamma_k)(\lambda_i) = 0$$

and it follows that $a_i = 0$ for $i = 1, \ldots, k$. Thus $\gamma_1, \ldots, \gamma_k$ are linearly independent in P_k. Since $\dim P_k = k$, they even form a basis. In particular, there exist $b_1, \ldots, b_k \in K$ with $\gamma := b_1\gamma_1 + \ldots + b_k\gamma_k = X^0 = 1$. For $v \in K^{n\times 1}$ we have

$$(A - \lambda_i 1_n)\gamma_i(A)v = \mu_A(A)v = 0,$$

i. e. $\gamma_i(A)v$ lies in $E_{\lambda_i}(A)$. On the other hand,

$$v = 1_n v = A^0 v = \gamma(A)(v) = b_1\gamma_1(A)(v) + \ldots + b_k\gamma_k(A)(v).$$

This shows $K^{n\times 1} = E_{\lambda_1}(A) + \ldots + E_{\lambda_k}(A)$. According to Remark 8.6, A is diagonalizable. □

Example 10.53. Let $A \in K^{n\times n}$ with exactly two distinct eigenvalues. Suppose we find a vector $v \in K^{n\times 1}$ such that v, Av, A^2v are linearly independent. Then $1_n, A, A^2$ are also linearly independent. This shows $\deg \mu_A \geq 3$. According to Theorem 10.52, A is not diagonalizable.

Theorem 10.54. *For $A \in K^{n\times n}$, it holds that $\chi_A \mid \mu_A^n$.*

Proof. Let $\mu_A = \sum a_i X^i$. For $i \geq 1$ we have

$$(X^i 1_n - A^i) = (X1_n - A)(X^{i-1}1_n + X^{i-2}A + \ldots + XA^{i-2} + A^{i-1}).$$

It follows that

$$\mu_A 1_n = \mu_A(X1_n) - \mu_A(A) = \sum_{i\geq 1} a_i(X^i 1_n - A^i) = (X1_n - A)B$$

for some $B \in K[X]^{n\times n}$. Taking the determinant on both sides yields $\mu_A^n = \chi_A \det(B)$. □

Remark 10.55. The algebraic multiplicity of an eigenvalue of $A \in K^{n\times n}$ is at most n. If $\mu_A = (X - \lambda_1)\ldots(X - \lambda_k)$ splits into linear factors, it follows that $\chi_A \mid (X - \lambda_1)^n \ldots (X - \lambda_k)^n = \mu_A^n$ (the proof of Theorem 10.54 does not require this assumption).

11 Euclidean Geometry

11.1 Scalar Products

Remark 11.1. In this chapter, we consider $K = \mathbb{R}$. In contrast to arbitrary fields, $\mathbb{R}$ can be divided into positive and negative numbers. For $x \in \mathbb{R}$, let as usual

$$|x| := \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0 \end{cases}$$

be the *absolute value* of x. We also use the fact that every positive real number has exactly one positive square root. Thus, $|x| = \sqrt{x^2}$ holds for all $x \in \mathbb{R}$.

Definition 11.2. Let V be an $\mathbb{R}$-vector space. A map $V \times V \to \mathbb{R}$, $(v, w) \mapsto [v, w]$ is called a *scalar product*[1], if the following conditions hold for all $u, v, w \in V$ and $\lambda \in \mathbb{R}$:

- $[v, v] \geq 0$ with equality if and only if $v = 0$ (*positive definite*),
- $[v, w] = [w, v]$ (*symmetric*),
- $[\lambda u + v, w] = \lambda[u, w] + [v, w]$ (*bilinear*).

Together with a scalar product, V becomes a *Euclidean space*. Vectors $v, w \in V$ are called *orthogonal*, if $[v, w] = 0$. One calls $|v| := \sqrt{[v, v]} \geq 0$ the *norm* of v. In the case $|v| = 1$, v is called *normalized*.

Remark 11.3.

(a) The symmetry of the scalar product shows

$$[u, \lambda v + w] = [\lambda v + w, u] = \lambda[v, u] + [w, u] = \lambda[u, v] + [u, w]$$

for all $u, v, w \in V$ and $\lambda \in \mathbb{R}$. For a fixed $x \in V$, the maps $V \to \mathbb{R}$, $v \mapsto [v, x]$ and $V \to \mathbb{R}$, $v \mapsto [x, v]$ are thus linear (this explains the term *bilinear*). In particular, $[v, 0] = 0 = [0, v]$ for all $v \in V$. Nevertheless, the map $V \times V \to \mathbb{R}$, $(v, w) \mapsto [v, w]$ is *not* linear, hence not a functional, because $[v, v] > 0 = [0, v] + [v, 0]$ for $v \neq 0$.

(b) Every subspace of a Euclidean space is itself a Euclidean space with the restricted scalar product.

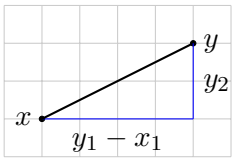
Example 11.4.

(a) The most important example of a Euclidean space is $V = \mathbb{R}^n$ with the *standard inner product*

$$\boxed{[x, y] := xy^{\mathrm{t}} = \sum_{i=1}^{n} x_i y_i} \qquad (x, y \in \mathbb{R}^n).$$

[1] The notation $[v, w]$ is not uniform in the literature. One also finds $\langle v, w \rangle$ (confusion with span), $(v \mid w)$ and similar. In a more general context (Definition 17.48), we use $\|v\|$ instead of $|v|$.

One easily verifies the three properties (positive definite, symmetric, and bilinear). In the case $n = 1$, $|x| = \sqrt{x_1^2}$ is the usual absolute value (this justifies the use of the vertical bars). According to the Pythagorean theorem[2], the norm $|y - x|$ in $\mathbb{R}^2$ corresponds to the geometric distance between x and y:

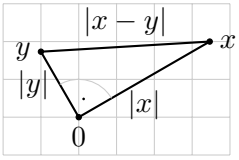


$$|y - x| = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}$$

If x and y are orthogonal, then one obtains

$$|x - y|^2 = [x - y, x - y] = [x, x] - 2[x, y] + [y, y] = |x|^2 + |y|^2.$$

According to the converse of the Pythagorean theorem, x and y form a right angle, i.e., they are perpendicular to each other (one writes $x \perp y$):



In general, the *parallelogram law* holds:

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

(b) Vectors in $\mathbb{R}^n$ are "discrete" functions $\{1, \dots, n\} \to \mathbb{R}$. In analysis, one considers a "continuous" variant: The continuous maps $[0, 1] \to \mathbb{R}$ on the closed interval $[0, 1]$ form an (infinite-dimensional) subspace $V \leq \mathrm{Fun}([0, 1], \mathbb{R})$ with inner product

$$[f, g] := \int_0^1 f(x)g(x)\,\mathrm{d}x \qquad (f, g \in V).$$

Lemma 11.5. *Let V be a Euclidean space, $v, w \in V$ and $\lambda \in \mathbb{R}$. Then:*

(a) $|\lambda v| = |\lambda||v|$ (homogeneity).

(b) $|[v, w]| \leq |v||w|$ *with equality if and only if v and w are linearly dependent* (Cauchy-Schwarz inequality).

(c) $||v| - |w|| \leq |v + w| \leq |v| + |w|$ (triangle inequality).

Proof.

(a) $|\lambda v| = \sqrt{[\lambda v, \lambda v]} = \sqrt{\lambda^2 [v, v]} = \sqrt{\lambda^2}\sqrt{[v, v]} = |\lambda||v|$.

(b) Wlog. let $w \neq 0$. Let $\lambda := \frac{[v,w]}{[w,w]}$. According to the properties of the scalar product, it holds that

$$0 \leq |v - \lambda w|^2 = [v - \lambda w, v - \lambda w] = [v, v] - 2\lambda[v, w] + \lambda^2[w, w] = |v|^2 - \frac{[v, w]^2}{|w|^2}.$$

It follows that $[v, w]^2 \leq |v|^2|w|^2$ and $|[v, w]| \leq |v||w|$. Equality implies $v = \lambda w$, i.e., v and w are linearly dependent. Conversely, if v and w are given as linearly dependent, then there exists a $\mu \in \mathbb{R}$ such that $v = \mu w$ and $|[v, w]| = |\mu||w|^2 \overset{\text{(a)}}{=} |\mu w||w| = |v||w|$.

[2]One could also *define* the distance between x and y by $|y - x|$ and thereby *prove* the Pythagorean theorem.

(c) First,

$$|v+w|^2 = [v+w, v+w] = [v,v] + 2[v,w] + [w,w] \overset{\text{(b)}}{\leq} |v|^2 + 2|v||w| + |w|^2 = (|v| + |w|)^2$$

and $|v+w| \leq |v|+|w|$. From this it follows that $|v| = |v+w-w| \leq |v+w|+|w|$ and $|v|-|w| \leq |v+w|$. Swapping v and w yields $-(|v| - |w|) = |w| - |v| \leq |v+w|$, thus $||v| - |w|| \leq |v+w|$. □

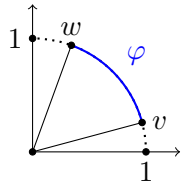
Remark 11.6.

(a) If $v, w \in \mathbb{R}^n$ are linearly independent, then 0, v and $v+w$ form a triangle with sides $|v|$, $|w|$ and $|v+w|$. The triangle inequality states that the sum of any two sides is greater than the third side.

(b) The Cauchy-Schwarz inequality implies $-1 \leq \frac{[v,w]}{|v||w|} \leq 1$ for $v, w \in V \setminus \{0\}$. This fraction does not change by positive scaling of v and w:

$$\frac{[\lambda v, \mu w]}{|\lambda v||\mu w|} = \frac{\lambda\mu[v,w]}{|\lambda||\mu||v||w|} = \frac{[v,w]}{|v||w|} \qquad (\lambda, \mu > 0)$$

So let v and w be normalized. Then one defines the *angle* φ (in radians) between v and w as the length of the arc on the unit circle[3] between v and w:



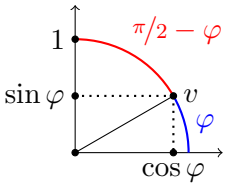
The length of the semicircle arc is called π and is calculated as $\pi \approx 3.14$ (Exercise II.10). The *cosine* of φ is defined by $\cos\varphi := [v, w]$.[4] It holds that

$$\cos 0 = [e_1, e_1] = 1, \qquad \cos(\pi/2) = [e_1, e_2] = 0, \qquad \cos\pi = [e_1, -e_1] = -1.$$

By $\cos(\varphi + 2k\pi) = \cos\varphi$ for $k \in \mathbb{Z}$, the cos is extended periodically to all of $\mathbb{R}$. In this context, $\cos(-\varphi) = \cos\varphi$ and $\cos(\pi - \varphi) = -\cos\varphi$ hold. The "shifted" function

$$\sin\varphi := \cos(\varphi - \pi/2) = \cos(\pi/2 - \varphi)$$

for $\varphi \in \mathbb{R}$ is called the *sine* of φ. For an arbitrary normalized vector $v = (x, y)$, it holds that $x = [v, e_1] = \cos\varphi$ and $y = [v, e_2] = \cos(\pi/2 - \varphi) = \sin\varphi$:



[3] in the plane $\langle v, w\rangle$

[4] One can show that this definition coincides with the analytical definition as a power series.

11.2 Orthonormal Bases

Definition 11.7. Let V be an n-dimensional Euclidean space. Vectors $b_1, \ldots, b_n$ form an *orthonormal basis* of V if they are normalized and pairwise orthogonal, i. e., $[b_i, b_j] = \delta_{ij}$ for $1 \leq i, j \leq n$.

Remark 11.8. An orthonormal basis $b_1, \ldots, b_n$ of V is indeed a basis. For this, it suffices to check the linear independence. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with $\lambda_1 b_1 + \ldots + \lambda_n b_n = 0$. Then

$$\lambda_i = \sum_{j=1}^{n} \lambda_j [b_j, b_i] = \Big[\sum_{j=1}^{n} \lambda_j b_j, b_i\Big] = [0, b_i] = 0$$

for $i = 1, \ldots, n$.

Example 11.9. The standard basis $e_1, \ldots, e_n \in \mathbb{R}^n$ is an orthonormal basis wrt. the standard inner product. Every permutation of an orthonormal basis is again an orthonormal basis.

Theorem 11.10 (Gram-Schmidt Process)**.** *Let $v_1, \ldots, v_k \in V$ be linearly independent. We define recursively:*

$$\boxed{b_s := v_s - \sum_{i=1}^{s-1} \frac{[v_s, b_i]}{[b_i, b_i]} b_i} \qquad (s = 1, \ldots, k).$$

Then $b_1, \ldots, b_k$ are pairwise orthogonal with $\langle v_1, \ldots, v_k \rangle = \langle b_1, \ldots, b_k \rangle$. Consequently, $\frac{1}{|b_1|} b_1, \ldots, \frac{1}{|b_k|} b_k$ is an orthonormal basis of $\langle v_1, \ldots, v_k \rangle$.

Proof. Induction on k: For $k = 1$, $b_1 = v_1 \neq 0$. Now let $k \geq 2$ and assume the statement is already proven for $k - 1$, i. e., $\langle v_1, \ldots, v_{k-1} \rangle = \langle b_1, \ldots, b_{k-1} \rangle$ and $[b_i, b_j] = 0$ for $1 \leq i < j \leq k - 1$. Because $\sum_{i=1}^{k-1} \frac{[v_k, b_i]}{[b_i, b_i]} b_i \in \langle b_1, \ldots, b_{k-1} \rangle$, it holds that

$$\langle v_1, \ldots, v_k \rangle = \langle b_1, \ldots, b_{k-1}, v_k \rangle = \langle b_1, \ldots, b_k \rangle.$$

Furthermore, for $i = 1, \ldots, k - 1$,

$$[b_k, b_i] = [v_k, b_i] - \sum_{j=1}^{k-1} \frac{[v_k, b_j]}{[b_j, b_j]} [b_j, b_i] = [v_k, b_i] - [v_k, b_i] = 0.$$

This proves the first statement. The second statement follows because $|\frac{1}{|b_i|} b_i| = \frac{1}{|b_i|} |b_i| = 1$. □

Corollary 11.11. *Every Euclidean space possesses (at least) one orthonormal basis.*

Proof. Apply the Gram-Schmidt process to an arbitrary basis. □

Example 11.12. Let $v_1 := (1, 0, 1)$, $v_2 := (0, 1, 1)$ and $v_3 := (-1, 2, 0)$ be linearly independent in $\mathbb{R}^3$. wrt. the standard scalar product, one obtains

$$\begin{aligned} b_1 &:= v_1 = (1, 0, 1), \\ b_2 &:= v_2 - \frac{[v_2, b_1]}{[b_1, b_1]} b_1 = (0, 1, 1) - \frac{1}{2}(1, 0, 1) = \frac{1}{2}(-1, 2, 1), \end{aligned}$$

$$b_3 := v_3 - \frac{[v_3, b_1]}{[b_1, b_1]} b_1 - \frac{[v_3, b_2]}{[b_2, b_2]} b_2 = (-1, 2, 0) + \frac{1}{2}(1, 0, 1) - \frac{5}{6}(-1, 2, 1) = \frac{1}{3}(1, 1, -1).$$

Note that scaling factors do not play a role in this calculation. Thus, one can calculate b_3 somewhat more conveniently using $b_2 = (-1, 2, 1)$. After normalization, $\frac{1}{\sqrt{2}}(1, 0, 1)$, $\frac{1}{\sqrt{6}}(-1, 2, 1)$, $\frac{1}{\sqrt{3}}(1, 1, -1)$ is an orthonormal basis of $\mathbb{R}^3$. To minimize the entries of the vectors, it can be useful to first apply the Gaussian algorithm before starting the Gram-Schmidt process. In this example, one would end up with the standard basis of $\mathbb{R}^3$.

Definition 11.13. Let V be a Euclidean space and $S \subseteq V$. Then

$$S^\perp := \big\{ v \in V : \forall s \in S : [v, s] = 0 \big\}$$

is called the *orthogonal complement* of S in V. In the case $S = \{s\}$, one writes $s^\perp := S^\perp$.

Remark 11.14. For $v, w \in S^\perp$ and $\lambda \in \mathbb{R}$, it holds that $[\lambda v + w, s] = \lambda[v, s] + [w, s] = 0$ for all $s \in S$, i. e. $\lambda v + w \in S^\perp$. Therefore, $S^\perp$ is a subspace, even if S is only a subset. Furthermore, $S^\perp = \langle S \rangle^\perp$ holds.

Lemma 11.15. *For subspaces U, W of a Euclidean space V, the following hold:*

(a) $V = U \oplus U^\perp$ *and* $\dim V = \dim U + \dim U^\perp$.

(b) $(U^\perp)^\perp = U$.

(c) $U \subseteq W \iff W^\perp \subseteq U^\perp$.

Proof.

(a) For $v \in U \cap U^\perp$, it holds that $|v|^2 = [v, v] = 0$ and $v = 0$. Thus $U \cap U^\perp = \{0\}$ and $U + U^\perp = U \oplus U^\perp$. We can extend a basis $v_1, \ldots, v_k$ of U to a basis $v_1, \ldots, v_n$ of V. The Gram-Schmidt process yields an orthonormal basis $b_1, \ldots, b_n$ of V with $\langle v_1, \ldots, v_k \rangle = \langle b_1, \ldots, b_k \rangle$. Therefore $b_{k+1}, \ldots, b_n \in U^\perp$ and $V = U \oplus U^\perp$.

(b) By definition, $U \subseteq (U^\perp)^\perp$. According to (a), $\dim (U^\perp)^\perp = \dim V - \dim U^\perp = \dim U$. Thus $U = (U^\perp)^\perp$.

(c) By definition, it holds that

$$U \subseteq W \implies W^\perp \subseteq U^\perp \overset{\text{(b)}}{\implies} U = (U^\perp)^\perp \subseteq (W^\perp)^\perp = W. \qquad \square$$

Example 11.16.

(a) In $V = \mathbb{R}^n$, an orthogonal complement of $U \leq V$ wrt. the standard scalar product can be determined using the Gaussian algorithm: One writes the vectors of a generating system of U as rows into a matrix $A \in \mathbb{R}^{k \times n}$. The solution set L_0 of the homogeneous system of equations $Ax = 0$ is $U^\perp$, because according to Theorem 6.6 it holds that $\dim L_0 = n - \operatorname{rk}(A) = n - \dim U = \dim U^\perp$.

(b) For $v = (x, y) \in \mathbb{R}^2 \setminus \{0\}$, it holds that $v^\perp = \langle (y, -x) \rangle$.

(c) Let $v, w \in \mathbb{R}^3$ be linearly independent. We extend them to a basis u, v, w of $\mathbb{R}^3$, such that the matrix

$$A := \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix}$$

has determinant 1 (this is always possible by scaling u appropriately). There is exactly one vector $x \in \langle v, w\rangle^\perp$ with $Ax^{\mathrm{t}} = e_1$, namely

$$x^{\mathrm{t}} = A^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \overset{9.22}{=} \widetilde{A} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \det(A_{11}) \\ -\det(A_{12}) \\ \det(A_{13}) \end{pmatrix} = \begin{pmatrix} v_2w_3 - v_3w_2 \\ v_3w_1 - v_1w_3 \\ v_1w_2 - v_2w_1 \end{pmatrix} =: v \times w.$$

One calls $v \times w$ the *cross product* of v and w. By construction, $\langle v, w\rangle^\perp = \langle v \times w\rangle$ holds. The direction of $v \times w$ can be determined using the *right-hand rule*: If v points in the direction of the thumb and w in the direction of the index finger, then $v \times w$ points in the direction of the middle finger of the right hand.

11.3 Symmetric and orthogonal maps

Definition 11.17. Let V be a Euclidean space and $f \in \mathrm{End}(V)$. One calls f

- *symmetric*,[5] if $[f(v), w] = [v, f(w)]$ for all $v, w \in V$.
- *orthogonal*,[6] if $[f(v), f(w)] = [v, w]$ for all $v, w \in V$.

Remark 11.18.

(a) The zero map is symmetric. If $f, g \in \mathrm{End}(V)$ are symmetric and $\lambda \in \mathbb{R}$, then $\lambda f + g$ is obviously also symmetric. Thus, the symmetric maps form a subspace of $\mathrm{End}(V)$.

(b) Orthogonal maps $f \in \mathrm{End}(V)$ are isomorphisms, because $v \in \mathrm{Ker}(f)$ implies $|v|^2 = [v, v] = [f(v), f(v)] = 0$, hence $v = 0$. Because of $|f(v - w)| = |v - w|$ and

$$\frac{[f(v), f(w)]}{|f(v)||f(w)|} = \frac{[v, w]}{|v||w|}$$

f preserves distances and angles (Remark 11.6). In particular, f maps orthonormal bases to orthonormal bases. With f, g, the compositions $f \circ g$ and f^{-1} are also orthogonal. The set of orthogonal maps therefore forms a subgroup $\mathrm{O}(V)$ of $\mathrm{GL}(V)$. $\mathrm{O}(V)$ is called the *orthogonal group* of V.

(c) The following theorem shows that length-preserving maps are automatically orthogonal (in particular linear).

Theorem 11.19 (Mazur-Ulam)**.** *Let V be a Euclidean space and $f \in \mathrm{Fun}(V, V)$ with $|f(v)| = |v|$ for all $v \in V$. Then $f \in \mathrm{O}(V)$.*

[5] or *self-adjoint*, see section 13.2
[6] or *isometric*

Proof. For $v, w \in V$ we have

$$[f(v), f(w)] = \frac{1}{2}\big(|f(v)|^2 + |f(w)|^2 - |f(v) - f(w)|^2\big) = \frac{1}{2}\big(|v|^2 + |w|^2 - |v - w|^2\big) = [v, w].$$

From this it follows that

$$\begin{aligned} |f(\lambda v + w) - \lambda f(v) - f(w)|^2 &= [f(\lambda v + w) - \lambda f(v) - f(w),\ f(\lambda v + w) - \lambda f(v) - f(w)] \\ &= [\lambda v + w - \lambda v - w,\ \lambda v + w - \lambda v - w] = 0 \end{aligned}$$

for $\lambda \in \mathbb{R}$. Thus f is linear and orthogonal. □

Lemma 11.20. *Let V be Euclidean with orthonormal basis B, $f \in \mathrm{End}(V)$ and $A := {}_B[f]_B$. Then:*

(a) f is symmetric $\iff$ $A^{\mathrm{t}} = A$.

(b) f is orthogonal $\iff$ $A^{\mathrm{t}} = A^{-1}$.

Proof. Let $B = \{b_1, \ldots, b_n\}$ and $A = (a_{ij})$. Then $f(b_i) = \sum_{j=1}^n a_{ji} b_j$ for $i = 1, \ldots, n$.

(a) If f is symmetric, then $a_{ji} = [f(b_i), b_j] = [b_i, f(b_j)] = a_{ij}$ for $1 \le i, j \le n$. Thus $A = A^{\mathrm{t}}$. Conversely, let $A = A^{\mathrm{t}}$. Then it follows that $[f(b_i), b_j] = [b_i, f(b_j)]$ for $1 \le i, j \le n$. For arbitrary $v = \sum \lambda_i b_i$ and $w = \sum \mu_i b_i$ in V we have

$$[f(v), w] = \sum_{i,j=1}^n \lambda_i \mu_j [f(b_i), b_j] = \sum_{i,j=1}^n \lambda_i \mu_j [b_i, f(b_j)] = [v, f(w)].$$

Thus f is symmetric.

(b) If f is orthogonal, then

$$\sum_{k=1}^n a_{ki} a_{kj} = \left[\sum_{k=1}^n a_{ki} b_k, \sum_{k=1}^n a_{kj} b_k\right] = [f(b_i), f(b_j)] = [b_i, b_j] = \delta_{ij}.$$

This shows $A^{\mathrm{t}} A = 1_n$ and $A^{\mathrm{t}} = A^{-1}$. Conversely, if $A^{\mathrm{t}} A = 1_n$, then $[f(b_i), f(b_j)] = \delta_{ij} = [b_i, b_j]$. As in (a), it follows that $[f(v), f(w)] = [v, w]$ for all $v, w \in V$, i.e., f is orthogonal. □

Remark 11.21.

(a) For an arbitrary field K, matrices $A \in \mathrm{GL}(n, K)$ are called *orthogonal*, if $A^{\mathrm{t}} = A^{-1}$. It is easily shown that the orthogonal matrices form a subgroup $\mathrm{O}(n, K)$ of $\mathrm{GL}(n, K)$. As usual, $\mathrm{O}(n, \mathbb{R})$ corresponds to the group $\mathrm{O}(V)$ through the choice of a basis (just as $\mathrm{GL}(V)$ and $\mathrm{GL}(n, K)$ correspond). For every orthogonal matrix A, it holds that

$$\det(A)^2 \overset{9.12}{=} \det(A)\det(A^{\mathrm{t}}) = \det(AA^{\mathrm{t}}) = \det(1_n) = 1$$

and $\det(A) = \pm 1$. One calls

$$\mathrm{SO}(n, K) := \mathrm{O}(n, K) \cap \mathrm{SL}(n, K) \le \mathrm{O}(n, K) \le \mathrm{GL}(n, K).$$

the *special orthogonal group* of degree n over K.

(b) The equation $A^{\mathrm{t}}A = 1_n = AA^{\mathrm{t}}$ means for orthogonal real matrices that the columns (or rows) of A form an orthonormal basis of $\mathbb{R}^n$ wrt. the standard scalar product. Let v be an eigenvector of A with eigenvalue $\lambda \in \mathbb{R}$. Then

$$|v|^2 = v^{\mathrm{t}}v = v^{\mathrm{t}}A^{\mathrm{t}}Av = (Av)^{\mathrm{t}}(Av) = \lambda^2 v^{\mathrm{t}}v = \lambda^2|v|^2$$

and $\lambda = \pm 1$.

Example 11.22.

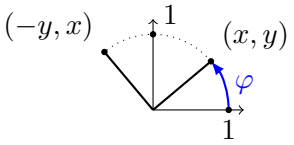
(a) Every permutation matrix is orthogonal, because the rows form an orthonormal basis (namely a permutation of the standard basis).

(b) For $A \in \mathrm{O}(2, \mathbb{R})$ we have

$$A = \begin{pmatrix} x & \mp y \\ y & \pm x \end{pmatrix}$$

with $x^2 + y^2 = 1 = \pm\det(A)$ (cf. Example 11.16).

- In the case $\det(A) = 1$, A describes a *rotation* by the angle φ between e_1 and (x, y):



Then we have

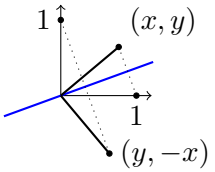
$$A = D(\varphi) := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

For two angles φ and ψ one obtains $D(\varphi + \psi) = D(\varphi)D(\psi)$, from which the well-known *trigonometric identities* follow:

$$\cos(\varphi + \psi) = \cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi),$$
$$\sin(\varphi + \psi) = \sin(\varphi)\cos(\psi) + \sin(\psi)\cos(\varphi).$$

Obviously, $D(\varphi)$ possesses (real) eigenvalues only if $\varphi \in \{0, \pi\}$, i. e. $D(0) = 1_2$ and $D(\pi) = -1_2$.

- In the case $\det(A) = -1$, A describes a *reflection* across the angle bisector between e_1 and (x, y):



Then we have

$$A = S(\varphi) := \begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & -\cos\varphi \end{pmatrix}.$$

The axis of reflection is spanned by an eigenvector for the eigenvalue 1. Orthogonal to it is an eigenvector for the eigenvalue -1 (note: $\det(A)$ is the product of the eigenvalues). According to Corollary 8.12, $S(\varphi)$ is diagonalizable. In fact,

$$D(\varphi/2)^{-1}S(\varphi)D(\varphi/2) = D(-\varphi/2)S(\varphi)D(\varphi/2) = S(0) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For special angles one obtains (cf. Exercise II.9):

φ	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$	π
$D(\varphi)$	$\frac{1}{2}\begin{pmatrix}\sqrt{3} & -1\\ 1 & \sqrt{3}\end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}$	$\frac{1}{2}\begin{pmatrix}1 & -\sqrt{3}\\ \sqrt{3} & 1\end{pmatrix}$	$\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}$	-1_2
$S(\varphi)$	$\frac{1}{2}\begin{pmatrix}\sqrt{3} & 1\\ 1 & -\sqrt{3}\end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$	$\frac{1}{2}\begin{pmatrix}1 & \sqrt{3}\\ \sqrt{3} & -1\end{pmatrix}$	$\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}$	$\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix}$

With this, one can also calculate

$$D(\pi/12) = D(\pi/3 - \pi/4) = D(\pi/3)D(\pi/4)^{-1} = D(\pi/3)D(\pi/4)^{\mathrm{t}}.$$

Remark 11.23. To show that symmetric endomorphisms are diagonalizable, we must temporarily leave the real numbers.

11.4 Complex Numbers

Lemma 11.24. *The $\mathbb{R}$-vector space $\mathbb{C} := \mathbb{R}^2$ becomes a field through the multiplication*

$$(a,b)\cdot(c,d) := (ac - bd,\ ad + bc) \qquad (a,b,c,d \in \mathbb{R}).$$

Proof. As a vector space, $(\mathbb{C}, +)$ is already an abelian group. We trace the multiplication in $\mathbb{C}$ back to matrix multiplication.[7] For this, we consider the injective map

$$\Gamma\colon \mathbb{C} \to \mathbb{R}^{2\times 2}, \quad z = (a,b) \mapsto \Gamma(z) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

For $x, y, z \in \mathbb{C}$, it holds that $\Gamma(x) + \Gamma(y) = \Gamma(x+y)$ and $\Gamma(x)\Gamma(y) = \Gamma(x\cdot y)$ (verify). From Lemma 5.8 it follows that

$$\Gamma(x(yz)) = \Gamma(x)\Gamma(yz) = \Gamma(x)(\Gamma(y)\Gamma(z)) = (\Gamma(x)\Gamma(y))\Gamma(z) = \Gamma(xy)\Gamma(z) = \Gamma((xy)z)$$

and $x(yz) = (xy)z$, since Γ is injective. The commutative and distributive laws are proven analogously. Because $\Gamma(1,0) = 1_2$, $(1,0)$ is the identity element in $\mathbb{C}$. It remains to show that $z \neq 0$ is invertible. It holds that $\det(\Gamma(z)) = a^2 + b^2 > 0$ and

$$\Gamma(z)^{-1} \overset{9.23}{=} \frac{1}{\det(\Gamma(z))}\widetilde{\Gamma(z)} = \frac{1}{a^2+b^2}\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \Gamma\Big(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\Big). \qquad \square$$

Definition 11.25. One calls $\mathbb{C}$ the field of *complex numbers*.

- By means of the map $\mathbb{R} \to \mathbb{C}$, $a \mapsto (a,0)$, we will regard $\mathbb{R}$ as a subset of $\mathbb{C}$. The operations in $\mathbb{R}$ correspond exactly to those in $\mathbb{C}$ with the same neutral elements.
- One calls $\mathrm{i} := (0,1) \in \mathbb{C} \setminus \mathbb{R}$ the *imaginary unit*. It holds that $\mathrm{i}^2 = (-1,0) = -1$. Since $1, \mathrm{i}$ form a basis of $\mathbb{C}$, every complex number can be uniquely written in the form $z = a + b\mathrm{i}$, where $\mathrm{Re}(z) := a \in \mathbb{R}$ is the *real part* and $\mathrm{Im}(z) := b \in \mathbb{R}$ is the *imaginary part* of z.

[7] One can also verify the axioms directly using the definition.

- One calls $|z| := \sqrt{\mathrm{Re}(z)^2 + \mathrm{Im}(z)^2} \geq 0$ the *absolute value* of z (this corresponds to the norm in $\mathbb{R}^2$).

Remark 11.26. In contrast to $\mathbb{R}$, $\mathbb{C}$ is not an *ordered* field, i.e., there is no order relation $\leq$ on $\mathbb{C}$ with

$$\begin{aligned} a \leq b &\implies a + c \leq b + c, \\ a, b \geq 0 &\implies ab \geq 0 \end{aligned}$$

for all $a, b, c \in \mathbb{C}$. For suppose $\mathrm{i} > 0$ holds. Then $-1 = \mathrm{i}^2 > 0$ and $1 = (-1)^2 > 0$. Now one obtains the contradiction $0 = 1 - 1 > 0 + 0 = 0$. If, on the other hand, $\mathrm{i} < 0$, then $0 = \mathrm{i} - \mathrm{i} < -\mathrm{i}$ and $-1 = (-\mathrm{i})^2 > 0$. This leads to the same contradiction.

Lemma 11.27. *For $z \in \mathbb{C} \setminus \{0\}$ and $n \in \mathbb{N}$, there exist pairwise distinct n-th roots $\zeta_1, \ldots, \zeta_n \in \mathbb{C}$ with $\zeta_1^n = \ldots = \zeta_n^n = z$.*

Proof. Let $z_0 := \frac{1}{|z|} z \in \mathbb{C}$ and $\Gamma(z_0) = \left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$ as in the proof of Lemma 11.24. Because of $a^2 + b^2 = |z_0|^2 = 1$, $\Gamma(z_0) = D(\varphi) \in \mathrm{O}(2, \mathbb{R})$ for an angle φ with $a = \cos\varphi$ and $b = \sin\varphi$. We define $\varphi_k := \frac{\varphi + 2k\pi}{n}$ for $k = 1, \ldots, n$. Then

$$D(\varphi_k)^n = D(n\varphi_k) = D(\varphi + 2k\pi) = D(\varphi) = \Gamma(z_0).$$

The numbers $z_k := \cos\varphi_k + \mathrm{i}\sin\varphi_k \in \mathbb{C}$ thus satisfy $z_k^n = z_0$. Because of $|z| > 0$, there exists $\sqrt[n]{|z|} \in \mathbb{R}_{>0}$ (Analysis). For $\zeta_k := \sqrt[n]{|z|} z_k$, one obtains $\zeta_k^n = |z| z_0 = z$ for $k = 1, \ldots, n$. Now let $\zeta_k = \zeta_l$ for $1 \leq k \leq l \leq n$. Then φ_k and φ_l differ only by a multiple of 2π. This shows $2\pi(l - k) = 2\pi cn$ for some $c \in \mathbb{N}_0$. From $0 \leq l - k < n$, it follows that $k = l$. Therefore, the n-th roots $\zeta_1, \ldots, \zeta_n$ are pairwise distinct. □

Example 11.28. The (n-th) roots of 1 are called *roots of unity*. They correspond to rotations by $2\pi k/n$ with $k \in \mathbb{Z}$. The fourth roots of unity are 1, i, -1, $-\mathrm{i}$.

Corollary 11.29. *Let $A \in \mathbb{C}^{n\times n}$ and $k \in \mathbb{N}$ with $A^k = A$. Then A is diagonalizable.*

Proof. The minimal polynomial μ_A divides $X^k - X = X(X^{k-1} - 1)$ according to Lemma 10.44. The roots of $X^k - X$ are the $(k-1)$-th roots of unity and 0. According to Lemma 10.24, μ_A splits into pairwise distinct linear factors. The claim now follows from Theorem 10.52. □

Example 11.30. Let $k \in \mathbb{N}$ and $A := D(2\pi/k)$. Because of $A^{k+1} = AA^k = AD(2\pi) = A$, A is diagonalizable over $\mathbb{C}$, but not necessarily over $\mathbb{R}$.

Definition 11.31. The map $\mathbb{C} \to \mathbb{C}$, $a + b\mathrm{i} \mapsto a - b\mathrm{i} =: \overline{a + b\mathrm{i}}$ is called *complex conjugation*.

Lemma 11.32. *For $z, w \in \mathbb{C}$, it holds that $|z|^2 = z\overline{z}$, $\overline{z + w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z} \cdot \overline{w}$.*

Proof. For $z = a + b\mathrm{i}$ and $w = c + d\mathrm{i}$, it holds that

$$\begin{aligned} |z|^2 &= a^2 + b^2 = (a + b\mathrm{i})(a - b\mathrm{i}) = z\overline{z}, \\ \overline{z + w} &= a + c - (b + d)\mathrm{i} = a - b\mathrm{i} + c - d\mathrm{i} = \overline{z} + \overline{w}, \\ \overline{zw} &= ac - bd - (ad + bc)\mathrm{i} = (a - b\mathrm{i})(c - d\mathrm{i}) = \overline{z} \cdot \overline{w}. \end{aligned}$$

□

Remark 11.33.

(a) For matrices $A = (a_{ij}) \in \mathbb{C}^{n\times m}$ we define $\overline{A} := (\overline{a_{ij}}) \in \mathbb{C}^{n\times m}$. For $B = (b_{ij}) \in \mathbb{C}^{m\times k}$ it holds that

$$\overline{AB} = \Big(\overline{\sum_{l=1}^{m} a_{il}b_{lj}}\Big)_{ij} \overset{11.32}{=} \Big(\sum_{l=1}^{m} \overline{a_{il}b_{lj}}\Big)_{ij} \overset{11.32}{=} \Big(\sum_{l=1}^{m} \overline{a_{il}}\,\overline{b_{lj}}\Big)_{ij} = \overline{A}\cdot\overline{B}.$$

(b) According to Lemma 11.27, the polynomial $X^n - z$ has a root for every $z \in \mathbb{C}$ (in particular, i is a root of X^2+1). Surprisingly, even every non-constant polynomial in $\mathbb{C}[X]$ has a root.[8]

Theorem 11.34 (Fundamental Theorem of Algebra)**.** *Every polynomial $\alpha \in \mathbb{C}[X] \setminus \mathbb{C}$ has a root in $\mathbb{C}$.*

Proof. The proof uses analysis (more precisely the completeness of $\mathbb{R}$) and is too difficult for this lecture. See Algebra notes. □

Corollary 11.35. *Every monic polynomial $\alpha \in \mathbb{C}[X] \setminus \mathbb{C}$ splits into linear factors.*

Proof. Induction on $d := \deg(\alpha) \geq 1$. In the case $d = 1$, α itself is a linear factor, since α is monic. Now let $d \geq 2$. According to the Fundamental Theorem, α has a root $x \in \mathbb{C}$. According to Lemma 10.22, there exists a $\beta \in \mathbb{C}[X]$ with $\alpha = (X-x)\beta$ and $\deg(\beta) = d-1$. Since α is monic, β must also be monic. By induction, β splits into linear factors and thus so does α. □

Example 11.36. Let $\alpha \in \mathbb{R}[X]$ with odd degree. To show that α has a root, we can assume that α is monic. Then $\lim_{x\to\pm\infty} \alpha(x) = \pm\infty$ holds. According to the Intermediate Value Theorem of analysis, the continuous function $\mathbb{R} \to \mathbb{R}$, $x \mapsto \alpha(x)$ has a real root. In practice, however, such a root can only be calculated approximately. Let specifically $\alpha := X^5 - 4X + 2$. Based on the graph of α, we suspect a root near $x_0 := 0.5$. Let $\alpha' = 5X^4 - 4$ be the derivative of α (analysis). In the *Newton's method*, one calculates the recursive sequence

$$x_{n+1} := x_n - \frac{\alpha(x_n)}{\alpha'(x_n)} \qquad (n \geq 0),$$

thus $x_1 = 0.50847\ldots$, $x_2 = 0.50849948\ldots$ etc. If x_0 is chosen "well" (as it is here), the sequence $(x_n)_n$ converges quadratically to a root, i.e., with each iteration the number of correct decimal places doubles. In fact, all given decimal places of x_2 are already correct.

Remark 11.37. Let $x \in \mathbb{C}$ be a root of $\alpha = \sum a_k X^k \in \mathbb{C}[X]$. We define $\overline{\alpha} := \sum \overline{a_k} X^k \in \mathbb{C}[X]$. It then holds that

$$\overline{\alpha}(\overline{x}) = \sum \overline{a_k x^k} = \overline{\alpha(x)} = \overline{0} = 0,$$

i.e., $\overline{x}$ is a root of $\overline{\alpha}$. In the case $\alpha \in \mathbb{R}[X]$, it thus holds: $\alpha(x) = 0 \iff \alpha(\overline{x}) = 0$. If applicable,

$$(X-x)(X-\overline{x}) = X^2 - (x+\overline{x})X + x\overline{x} = X^2 - 2\mathrm{Re}(x)X + |x|^2 \in \mathbb{R}[X]$$

is a divisor of α.

[8] One says: $\mathbb{C}$ is *algebraically closed*.

11.5 The Principal Axis Theorem

Theorem 11.38 (Principal Axis Theorem)**.** *Let V be a Euclidean space and $f \in \mathrm{End}(V)$. f is symmetric if and only if V possesses an orthonormal basis of eigenvectors of f. In particular, symmetric endomorphisms are diagonalizable.*

Proof. Let B be an orthonormal basis of eigenvectors of f. Then ${}_B[f]_B$ is a diagonal matrix and therefore symmetric. According to Lemma 11.20, f is symmetric. Conversely, let f be symmetric. We argue by induction on $n := \dim V$. In the case $n = 1$, any normalized vector can be used for an orthonormal basis of eigenvectors. So let $n \geq 2$ and the claim be already proven for $n-1$. First, let B be an arbitrary orthonormal basis of V. Then we can also consider the symmetric matrix $A := {}_B[f]_B$ as a complex matrix. According to the Fundamental Theorem of Algebra, $\chi_A \in \mathbb{C}[X] \setminus \mathbb{C}$ has a root $\lambda \in \mathbb{C}$. Thus λ is an eigenvalue of A. Let $v = (v_1, \ldots, v_n)^{\mathrm{t}} \in \mathbb{C}^{n \times 1}$ be a corresponding eigenvector. Because of

$$\lambda \sum_{i=1}^{n} |v_i|^2 = \lambda \overline{v}^{\mathrm{t}} v = \overline{v}^{\mathrm{t}} A v = (\overline{v}^{\mathrm{t}} A v)^{\mathrm{t}} = v^{\mathrm{t}} A^{\mathrm{t}} \overline{v} = v^{\mathrm{t}} A \overline{v} = v^{\mathrm{t}} \overline{Av} = \overline{\lambda} v^{\mathrm{t}} \overline{v} = \overline{\lambda} \sum_{i=1}^{n} |v_i|^2$$

it follows that $\lambda = \overline{\lambda} \in \mathbb{R}$. Now λ is also an eigenvalue of f and we can choose a corresponding eigenvector $b_1 \in V$. After normalization, $|b_1| = 1$. For $U := b_1^{\perp}$, it holds that $V = \langle b_1 \rangle \oplus U$ according to Lemma 11.15. For $u \in U$, it holds that

$$[f(u), b_1] = [u, f(b_1)] = \lambda [u, b_1] = 0,$$

i.e., $f(u) \in U$. Therefore, the restriction $g := f_{|U}$ lies in $\mathrm{End}(U)$. Obviously, g is also symmetric. Because of $\dim U = n - 1$, U possesses an orthonormal basis $b_2, \ldots, b_n \in U$ of eigenvectors of g by induction. Because of $f(b_i) = g(b_i)$, $b_2, \ldots, b_n$ are also eigenvectors of f. Overall, $b_1, \ldots, b_n$ is an orthonormal basis of V of eigenvectors of f. □

Corollary 11.39. *$A \in \mathbb{R}^{n \times n}$ is symmetric if and only if there exists a matrix $S \in \mathrm{O}(n, \mathbb{R})$ with $S^{\mathrm{t}} A S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. If applicable, $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.*

Proof. If $D := S^{\mathrm{t}} A S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, then

$$A^{\mathrm{t}} = (S D S^{\mathrm{t}})^{\mathrm{t}} = S D^{\mathrm{t}} S^{\mathrm{t}} = S D S^{\mathrm{t}} = A,$$

i. e. A is symmetric. For the converse, let $V := \mathbb{R}^n$ and $f \in \mathrm{End}(V)$ with $[f] = A$. Let A be symmetric. According to Lemma 11.20, f is symmetric. According to the Principal Axis Theorem, there exists an orthonormal basis B of V consisting of eigenvectors of f. Let E be the standard basis of V and $S := {}_E \Delta_B \in \mathrm{GL}(n, \mathbb{R})$. Since the columns of S consist of the vectors of B, it holds that $S \in \mathrm{O}(n, \mathbb{R})$. From Corollary 7.27 it follows that

$$S^{\mathrm{t}} A S = S^{-1} A S = {}_B[f]_B = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

with the eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ of f. □

Remark 11.40. Let $f \in \mathrm{End}(V)$ be symmetric. Let $v, w \in V$ be eigenvectors of f for distinct eigenvalues λ and μ, respectively. Then $\lambda [v, w] = [f(v), w] = [v, f(w)] = \mu [v, w]$ and it follows that $[v, w] = 0$. Mnemonic: Eigenvectors for distinct eigenvalues are orthogonal. One can therefore calculate the desired orthonormal basis of V by applying the Gram-Schmidt process to each eigenspace.

Example 11.41. In Example 8.4 we had investigated the map $f \in \mathrm{End}(\mathbb{R}^3)$ with symmetric matrix

$$A := [f] = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

It holds that

$$E_1(A) = \langle \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \rangle, \qquad E_4(A) = \langle \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \rangle$$

(note: $\mathrm{tr}(A) = 6 = 1 + 1 + 4$). The Gram-Schmidt process for $E_1(A)$ yields

$$b_1 := (1, 0, -1), \qquad b_2 := (0, 1, -1) - \frac{1}{2}(1, 0, -1) = \frac{1}{2}(-1, 2, -1).$$

The eigenvector for the eigenvalue 4 only needs to be normalized. Overall, one obtains the orthonormal basis $\frac{1}{\sqrt{2}}(1, 0, -1)$, $\frac{1}{\sqrt{6}}(-1, 2, -1)$, $\frac{1}{\sqrt{3}}(1, 1, 1)$ consisting of eigenvectors of f. For the orthogonal matrix

$$S := \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$$

it holds that $S^{\mathrm{t}}AS = S^{-1}AS = \mathrm{diag}(1, 1, 4)$. One can use the result to take roots of matrices. For $\sqrt{A} := S\,\mathrm{diag}(1, 1, 2)S^{\mathrm{t}}$ it holds that

$$\sqrt{A}^2 = S\,\mathrm{diag}(1, 1, 2)S^{\mathrm{t}}S\,\mathrm{diag}(1, 1, 2)S^{\mathrm{t}} = S\,\mathrm{diag}(1, 1, 2)^2S^{\mathrm{t}} = S\,\mathrm{diag}(1, 1, 4)S^{\mathrm{t}} = A.$$

More on this in Theorem 12.46 and Theorem 14.42.

Theorem 11.42 (Euler)**.** *Let V be a 3-dimensional Euclidean space and $f \in \mathrm{O}(V)$. Then there exists an orthonormal basis B of V and an angle φ with*

$${}_B[f]_B = \begin{pmatrix} \pm 1 & 0 \\ 0 & D(\varphi) \end{pmatrix}.$$

Proof. Since 3 is odd, χ_f has a root $\lambda \in \mathbb{R}$ according to Example 11.36 (or Remark 11.37). According to Remark 11.21, $\lambda \in \{\pm 1\}$. Let b_1 be a corresponding normalized eigenvector and $U := b_1^{\perp}$. For $u \in U$, it holds that

$$[f(u), b_1] = [f(u), \lambda^2 b_1] = \lambda[f(u), f(b_1)] = \lambda[u, b_1] = 0$$

and $f(u) \in U$. Therefore, the restriction $g := f_{|U}$ lies in $\mathrm{O}(U)$. For an orthonormal basis $C := \{b_2, b_3\}$ of U, ${}_C[g]_C \in \mathrm{O}(2, \mathbb{R})$ according to Lemma 11.20. In the case $\det(g) = 1$, ${}_C[g]_C = D(\varphi)$ according to Example 11.22 and the assertion follows with $B := \{b_1, b_2, b_3\}$. Now let $\det(g) = -1$. According to Example 11.22, g is a reflection with eigenvalues 1 and -1. Since these are also eigenvalues of f, we can replace λ by $-\lambda$ and choose b_1 accordingly. Then $\det(g) = 1$ and the assertion follows as before. □

Remark 11.43. The matrices in Theorem 11.42 with determinant 1 (i.e., of the form $\mathrm{diag}(1, D(\varphi))$) correspond to rotations by the angle φ, where the axis of rotation is spanned by the first basis vector. According to the determinant theorem, the composition of rotations is again a rotation (in 2-dimensional space this is clear because of $D(\varphi)D(\psi) = D(\varphi + \psi)$). Note: The orthogonal maps with determinant -1 are not necessarily reflections. There are also so-called *rotary reflections*, i.e., compositions of a rotation with a reflection.

Example 11.44. During a football match, there is a point on the surface of the football that is in exactly the same location at two different points in time. Reason: The center of the football lies on the kickoff point at the beginning of the first and second half. In between, the ball performs a rotation (and translation) with every kick. Since the composition of rotations is again a rotation, the transformation at the kickoff point is also described by a rotation f. The axis of rotation of f intersects the surface of the ball at two points. These therefore remain fixed.

Definition 11.45. Let V be a Euclidean space and $v \in V \setminus \{0\}$. One calls

$$S_v \colon V \to V, \qquad w \mapsto w - 2\frac{[w,v]}{[v,v]}v$$

the *reflection* across $v^\perp$.

Remark 11.46. For $\lambda \in \mathbb{R}^\times$, $S_{\lambda v} = S_v$. We can therefore assume that v is normalized. Then

$$S_v(w) = w - 2[w,v]v$$

holds for all $w \in V$. Obviously, S_v is linear and $S_v(w) = w$ holds if and only if $w \in v^\perp$. Furthermore, $S_v(v) = -v$. Geometrically, S_v thus corresponds to a reflection across the hyperplane $v^\perp$. wrt. a suitable basis, S_v has the representation matrix $\operatorname{diag}(-1, 1, \ldots, 1)$ (cf. Exercise II.11). In particular, S_v is orthogonal and $\det S_v = -1$. Furthermore, $S_v \circ S_v = \mathrm{id}_V$.

Theorem 11.47 (Cartan-Dieudonné)**.** *Let V be an n-dimensional Euclidean space and $f \in \mathrm{O}(V)$. Then f is a composition of at most n reflections.*

Proof. In the case $f = \mathrm{id}_V$, f is the composition of 0 reflections. Therefore, let $f \neq \mathrm{id}_V$ and $w \in V$ with $f(w) \neq w$. In the case $n = 1$, $f = -\mathrm{id}_V = S_1$. Now let $n \geq 2$ and assume the claim is already proven for $n-1$. For $v := f(w) - w \neq 0$, it holds that

$$[f(w) + w, v] = [f(w) + w, f(w) - w] = [f(w), f(w)] - [w,w] = 0,$$

i. e. $f(w) + w \in v^\perp$. Thus

$$(S_v \circ f)(w) = \frac{1}{2}\Big(S_v(f(w) + w) + S_v(v)\Big) = \frac{1}{2}\big(f(w) + w - v\big) = w.$$

Let $U := w^\perp$. Because $g := S_v \circ f \in \mathrm{O}(V)$, it follows that $g(U) = U$. By induction, $g_{|U}$ is a composition of at most $n-1$ reflections $S_{w_1}, \ldots, S_{w_k}$ with $w_1, \ldots, w_k \in U$. One can also interpret S_{w_i} as a reflection of V using the same formula. Because $w \in U^\perp$, $S_{w_i}(w) = w$ holds for $i = 1, \ldots, k$. Thus $f = S_v \circ g$ is a product of at most n reflections. □

Remark 11.48. According to Remark 11.46, $f \in \mathrm{SO}(V)$ is a product of an even number of reflections. In particular, in the case $\dim V = 2n+1$, only $2n$ reflections are required.

12 Bilinear Forms

12.1 Gram Matrices

Remark 12.1. In this chapter, we generalize parts of Euclidean geometry to arbitrary fields. Let V always be a finite-dimensional K-vector space.

Definition 12.2.

- A *bilinear form* on V is a map $\beta\colon V \times V \to K$ that is linear in the first and second components, i. e.

$$\beta(\lambda u + v, w) = \lambda\beta(u, w) + \beta(v, w),$$
$$\beta(u, \lambda v + w) = \lambda\beta(u, v) + \beta(u, w)$$

  for all $u, v, w \in V$ and $\lambda \in K$.

- One calls β
  - *symmetric*, if $\beta(v, w) = \beta(w, v)$ holds for all $v, w \in V$.
  - *antisymmetric*, if $\beta(v, w) = -\beta(w, v)$ holds for all $v, w \in V$.
  - *alternating*, if $\beta(v, v) = 0$ holds for all $v \in V$.
  - *degenerate*, if $\exists v \in V \setminus \{0\} : \forall w \in V : \beta(v, w) = 0$.

Example 12.3.

(a) The trivial bilinear form $\beta(v, w) = 0$ for all $v, w \in V$ is symmetric, antisymmetric, and alternating. For $V \neq \{0\}$ it is degenerate. As a rule, we are interested in non-degenerate bilinear forms.

(b) A scalar product $\beta(v, w) = [v, w]$ on a Euclidean space V is a symmetric bilinear form. For $V \neq \{0\}$, β is non-degenerate, because $[v, v] = |v|^2 > 0$ for $v \neq 0$.

(c) For $V = K^n$ and $A \in K^{n\times n}$, $\beta_A(v, w) := vAw^{\mathrm{t}}$ defines a bilinear form. This follows from the calculation rules for matrices. We will see in Theorem 12.9 that every bilinear form has this form.

(d) The choice $A = 1_n$ in (c) leads to the symmetric bilinear form

$$\beta(v, w) = vw^{\mathrm{t}} = \sum_{i=1}^{n} v_i w_i.$$

In the case $K = \mathbb{R}$, one obtains the standard scalar product.

(e) Let $V := K^2$ and $\beta(v, w) := \det\left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right)$ for $v, w \in V$. According to Lemma 9.5, β is an alternating bilinear form. In general, det is a *multilinear form*.

Remark 12.4.

(a) Every alternating bilinear form β is antisymmetric, because from

$$0 = \beta(v+w, v+w) = \beta(v,v) + \beta(v,w) + \beta(w,v) + \beta(w,w) = \beta(v,w) + \beta(w,v)$$

it follows that $\beta(v,w) = -\beta(w,v)$. For $K \in \{\mathbb{Q}, \mathbb{R}, \mathbb{C}\}$, the converse also holds, because from $\beta(v,v) = -\beta(v,v)$ it follows that $\beta(v,v) = 0$. For $K = \mathbb{F}_2$, this reasoning is invalid, because $1+1=0$. Here, symmetric and antisymmetric are even synonymous. In particular, the bilinear form $\beta(v,w) = vw^{\mathrm{t}}$ on $V = K^n$ is (anti)symmetric, but not alternating. To avoid this case, we will often assume $1+1 \neq 0$ in the following.[1]

(b) Every symmetric bilinear form β defines a *quadratic form* $q\colon V \to K$ by $q(v) := \beta(v,v)$. If $1+1 \neq 0$, then β can be recovered from q by *polarization*:

$$\beta(v,w) = \frac{1}{2}\Big(q(v+w) - q(v) - q(w)\Big) \qquad (v, w \in V).$$

Theorem 12.5. *The set* $\mathrm{Bil}(V)$ *of all bilinear forms on* V *is a subspace of* $\mathrm{Fun}(V \times V, K)$.

Proof. The zero map is the trivial bilinear form. Let $\beta, \gamma \in \mathrm{Bil}(V)$ and $\lambda \in K$. As in Theorem 7.13, one shows that $\beta+\gamma$ and $\lambda\beta$ are linear in the first and second components (however, $\mathrm{Bil}(V) \not\subseteq \mathrm{Hom}(V \times V, K)$). This shows $\beta + \gamma, \lambda\beta \in \mathrm{Bil}(V)$. □

Remark 12.6. Obviously, the symmetric (antisymmetric, alternating) bilinear forms each form a subspace of $\mathrm{Bil}(V)$. Theorem 12.9 provides information about the dimension of these subspaces.

Theorem 12.7. *Let* $\beta \in \mathrm{Bil}(V)$. *For* $v \in V$ *let* $F_v\colon V \to K$, $w \mapsto \beta(v,w)$. *Then* $F\colon V \to V^*$, $v \mapsto F_v$ *is a homomorphism.* β *is non-degenerate if and only if* F *is an isomorphism.*

Proof. Due to the bilinearity of β, $F_v \in V^*$ and F is a homomorphism. Apparently, β is non-degenerate if and only if F is injective. Due to $\dim V^* = \dim V$, injective is equivalent to bijective. □

Definition 12.8. Let β be a bilinear form on V. Let $B := \{b_1, \ldots, b_n\}$ be a basis of V. One calls

$${}_B[\beta]_B := \big(\beta(b_i, b_j)\big)_{i,j=1}^n \in K^{n \times n}$$

the *Gram matrix* of β wrt. B. If $V = K^n$ and B is the standard basis, let $[\beta] := {}_B[\beta]_B$.

Theorem 12.9. *Let* B *be a basis of* V *with* $|B| = n$. *Then the map*

$${}_B[.]_B\colon \mathrm{Bil}(V) \to K^{n \times n}, \qquad \beta \mapsto {}_B[\beta]_B$$

is an isomorphism. For $A := {}_B[\beta]_B$ *the following holds:*

(a) β *symmetric* $\iff$ A *symmetric.*

(b) β *antisymmetric* $\iff$ $A^{\mathrm{t}} = -A$.[2]

(c) β *non-degenerate* $\iff$ A *invertible.*

[1] Besides $\mathbb{F}_2$, there are many (even infinite) fields with $1+1=0$.

[2] One calls A *antisymmetric* or *skew-symmetric*.

Proof. For $\beta, \gamma \in \mathrm{Bil}(V)$ and $\lambda \in K$ it holds that

$$ {}_B[\lambda\beta + \gamma]_B = ((\lambda\beta + \gamma)(b_i, b_j)) = \lambda(\beta(b_i, b_j)) + (\gamma(b_i, b_j)) = \lambda\, {}_B[\beta]_B + {}_B[\gamma]_B, $$

i. e. ${}_B[.]_B$ is linear. If ${}_B[\beta]_B = 0$, then

$$ \beta\Big(\sum \lambda_i b_i, \sum \mu_j b_j\Big) = \sum_{i,j} \lambda_i \mu_j \beta(b_i, b_j) = 0 $$

for all $\lambda_i, \mu_j \in K$. Thus $\beta = 0$ and ${}_B[.]_B$ is injective. For surjectivity, let $A \in K^{n\times n}$ be given. From the linearity of the coordinate representation $v \mapsto {}_B[v]$ (see proof of Theorem 7.10), it follows that

$$ \beta(v, w) := {}_B[v] A\, {}_B[v]^{\mathrm{t}} $$

defines a bilinear form. Because of $\beta(b_i, b_j) = e_i A e_j^{\mathrm{t}} = a_{ij}$, we have ${}_B[\beta]_B = A$. Thus ${}_B[.]_B$ is an isomorphism.

(a) If β is symmetric, then $a_{ij} = \beta(b_i, b_j) = \beta(b_j, b_i) = a_{ji}$ for all $1 \le i, j \le n$. Therefore A is symmetric. Conversely, let A be symmetric. Then $\beta(b_i, b_j) = a_{ij} = a_{ji} = \beta(b_j, b_i)$. For ${}_B[v] = (v_1, \ldots, v_n)$ and ${}_B[w] = (w_1, \ldots, w_n)$ it follows that

$$ \beta(v, w) = \sum_{1\le i,j\le n} v_i w_j \beta(b_i, b_j) = \sum_{1\le i,j\le n} v_i w_j \beta(b_j, b_i) = \beta(w, v), $$

i. e. β is symmetric.

(b) If β is antisymmetric, then $a_{ij} = \beta(b_i, b_j) = -\beta(b_j, b_i) = -a_{ji}$ and $A^{\mathrm{t}} = -A$. The converse is shown as in (a).

(c) Let β be degenerate and $v \neq 0$ with $\beta(v, w) = 0$ for all $w \in W$. For $x := {}_B[v]^{\mathrm{t}}$ and $i = 1, \ldots, n$ it then holds that

$$ e_i A x = \sum_{j=1}^{n} \beta(b_i, b_j) x_j = \beta(b_i, v) = 0. $$

This shows $Ax = 0$. According to Theorem 6.6, A is not invertible. Conversely, if A is not invertible, then there exists $x \in K^{n\times 1} \setminus \{0\}$ with $Ax = 0$. For $v = \sum_{i=1}^{n} x_i b_i$ it then holds that $\beta(v, w) = 0$ for all $w \in V$. Therefore β is degenerate. □

Example 12.10. Let $1 + 1 \neq 0$ in K and $n := \dim V$ be odd. Let $\beta \in \mathrm{Bil}(V)$ be antisymmetric with Gram matrix A wrt. an arbitrary basis. According to Theorem 12.9 it holds that

$$ \det(A) = \det(A^{\mathrm{t}}) = \det(-A) \overset{9.8}{=} (-1)^n \det(A) = -\det(A) $$

and $\det(A) = 0$. Therefore β must be degenerate.

Theorem 12.11. *Let B and C be bases of V. For all $\beta \in \mathrm{Bil}(V)$ it holds that*

$$ \boxed{{}_C[\beta]_C = {}_B\Delta_C^{\mathrm{t}}\, {}_B[\beta]_B\, {}_B\Delta_C.} $$

Proof. Let $B = \{b_1, \ldots, b_n\}$, $C = \{c_1, \ldots, c_n\}$ and $S = (s_{ij}) = {}_B\Delta_C$. For $1 \le i, j \le n$ it holds that

$$ \beta(c_i, c_j) = \sum_{k,l=1}^{n} s_{ki} \beta(b_k, b_l) s_{lj} = \sum_{k=1}^{n} s_{ki} ({}_B[\beta]_B S)_{kj} = (S^{\mathrm{t}}\, {}_B[\beta]_B S)_{ij}. $$ □

Definition 12.12. Matrices $A, B \in K^{n\times n}$ are called *congruent*, if there exists an $S \in \mathrm{GL}(n, K)$ with $B = SAS^{\mathrm{t}}$.[3]

Remark 12.13.

(a) As in Remark 7.30, one shows that the congruence of matrices is an equivalence relation. According to Theorem 12.9 and Theorem 12.11, every bilinear form on V determines a congruence class of Gram matrices. We will show in the next section that bilinear forms can be diagonalized similarly to endomorphisms.

(b) According to Lemma 5.15,

$$\mathrm{rk}(B) = \mathrm{rk}(SAS^{\mathrm{t}}) \leq \min\{\mathrm{rk}(S), \mathrm{rk}(AS^{\mathrm{t}})\} = \mathrm{rk}(AS^{\mathrm{t}}) \leq \min\{\mathrm{rk}(A), \mathrm{rk}(S^{\mathrm{t}})\} = \mathrm{rk}(A).$$

By symmetry, it follows that $\mathrm{rk}(B) = \mathrm{rk}(A)$, i.e., congruent matrices have the same rank. On the other hand, A and B do not necessarily have the same determinant, because $\det(B) = \det(SAS^{\mathrm{t}}) = \det(S)^2 \det(A)$. Due to Remark 10.35, A and B generally do not have the same eigenvalues either. Even in the case $n = 1$, one sees that A and B likewise do not have the same trace.

(c) According to the principal axis theorem, every symmetric real matrix is congruent and similar to a diagonal matrix.

12.2 Sylvester's Law of Inertia

Definition 12.14. Let $\beta \in \mathrm{Bil}(V)$ and $S \subseteq V$. As in Euclidean spaces, we define the *orthogonal complement* of S by

$$S^{\perp} := \{v \in V : \forall s \in S : \beta(v, s) = 0\}.$$

Remark 12.15.

(a) In the definition of $S^{\perp}$, one would strictly speaking have to distinguish between "left-orthogonal" and "right-orthogonal" ($\forall s \in S : \beta(s, v) = 0$). Usually, we will assume that β is (anti)symmetric, so that both versions are equivalent.

(b) For $U \leq V$, $U^{\perp} \leq V$. β is degenerate if and only if $V^{\perp} \neq \{0\}$ holds.

Theorem 12.16. *Let $\beta \in \mathrm{Bil}(V)$ be non-degenerate. For $U \leq V$, it holds that* $\dim V = \dim U + \dim U^{\perp}$.

Proof. Let $F\colon V \to V^*$, $v \mapsto F_v$ be the isomorphism from Theorem 12.7. Then

$$F_v \in F(U^{\perp}) \iff \forall u \in U : F_v(u) = \beta(v, u) = 0 \iff F_v \in U^0,$$

i.e., $F(U^{\perp}) = U^0$. The claim follows from Lemma 7.43. □

Remark 12.17. Attention: In contrast to Euclidean spaces, in general neither $U \cap U^{\perp} = \{0\}$ nor $V = U \oplus U^{\perp}$ holds. For example, let $V = \mathbb{F}_2^2$ and $[\beta] = 1_2$. For $U = \langle(1, 1)\rangle$, it holds that $U = U^{\perp}$.

[3] In contrast to similarity, we do not introduce a symbol for the congruence of matrices.

Definition 12.18. Let $\beta \in \mathrm{Bil}(V)$ be symmetric. A basis $B = \{b_1, \ldots, b_n\}$ of V is called an *orthogonal basis* wrt. β if $\beta(b_i, b_j) = 0$ for $i \neq j$. If additionally $\beta(b_i, b_i) = 1$, then B is called an *orthonormal basis* wrt. β.

Remark 12.19.

(a) Obviously, B is an *orthogonal basis* (resp. *orthonormal basis*) wrt. β if and only if ${}_B[\beta]_B$ is a diagonal matrix (resp. ${}_B[\beta]_B = 1_n$).

(b) According to Corollary 11.11, every scalar product on a Euclidean space possesses an orthonormal basis.

Theorem 12.20. *Let $1+1 \neq 0$ in K. Then every symmetric bilinear form on V possesses an orthogonal basis.*

Proof. Let $\beta \in \mathrm{Bil}(V)$ be symmetric. We argue by induction on $n := \dim V$. In the case $n \leq 1$ or $\beta = 0$, every basis is an orthogonal basis. So let $n \geq 2$ and $\beta(v, w) \neq 0$ for certain $v, w \in V$. In the case $\beta(v, v) = \beta(w, w) = 0$, we have

$$\beta(v + w, v + w) = \beta(v, v) + \beta(v, w) + \beta(w, v) + \beta(w, w) = 2\beta(v, w) \neq 0$$

(note $1 + 1 \neq 0$). In any case, there exists a $b_1 \in V \setminus \{0\}$ with $\beta(b_1, b_1) \neq 0$. Let $U := \langle b_1 \rangle$. Because of $U \cap U^\perp = \{0\}$, we have $V = U \oplus U^\perp$. The restriction of β to $U \times U$ is a symmetric bilinear form. By induction, $U^\perp$ possesses an orthogonal basis $b_2, \ldots, b_n$. Now $b_1, \ldots, b_n$ is an orthogonal basis of V. □

Remark 12.21.

(a) The matrix version of Theorem 12.20 states: Every symmetric matrix is congruent to a diagonal matrix (if $1 + 1 \neq 0$). We will see that the diagonal entries can be chosen specifically.

(b) If $1 + 1 = 0$ in K, then Theorem 12.20 becomes false: The bilinear form β with $[\beta] = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has no orthogonal basis, because $\beta(v, v) = 0$ for all $v \in K^2$.

(c) The Gram-Schmidt process for calculating orthogonal bases does not always work in this generality, because one must divide by $\beta(b_i, b_i)$, but this value can be 0. Instead, we modify the Gaussian algorithm as follows:

(1) Let $A := [\beta]$. Let us assume inductively that the rows and columns $1, \ldots, k-1$ of A already have the desired diagonal form (at the beginning let $k = 1$).

(2) Let us first assume $a_{kk} \neq 0$. Then one can achieve $a_{ik} = 0$ for $i = k+1, \ldots, n$ by adding a multiple of the k-th row to the rows below it as usual. This corresponds to multiplication by elementary matrices $S_1, \ldots, S_{n-k}$ from the left. Subsequently, set $a_{ki} = 0$ for $i = k+1, \ldots, n$. This corresponds to the multiplication of $S_1^\mathrm{t}, \ldots, S_{n-k}^\mathrm{t}$ from the right (in any order). Overall, A becomes SAS^t, where $S = S_1 \ldots S_n$. Because $(SAS^\mathrm{t})^\mathrm{t} = (S^\mathrm{t})^\mathrm{t} A^\mathrm{t} S^\mathrm{t} = SAS^\mathrm{t}$, A remains symmetric through this procedure.

(3) Now let $a_{kk} = 0$. If there exists an $l > k$ with $a_{ll} \neq 0$, then swap rows k and l and subsequently columns k and l. This swaps a_{kk} and a_{ll} and A remains symmetric. One is now in situation (2).

(4) Finally, let $a_{ii} = 0$ for $i = k, \dots, n$. If $a_{ij} = 0$ for all $k \le i, j \le n$, then we are finished. Therefore, let $a_{ij} \neq 0$ for certain $k \le i < j \le n$. We add the i-th row to the j-th row and subsequently the i-th column to the j-th column. The entry at position (j, j) thereby becomes $2a_{ij} \neq 0$ (note $1 + 1 \neq 0$). Thus one is in situation (3).

(5) To determine the orthogonal basis B, perform all row operations (but not the column operations) on the identity matrix. This produces the matrix $S \in \mathrm{GL}(n, K)$ such that SAS^{t} is a diagonal matrix (cf. Theorem 6.17). According to Theorem 12.11, the rows of S are the vectors of B.

Example 12.22.

$$(A|1_3) = \left(\begin{array}{ccc|ccc} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 1 & 0 & 0 & 1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 1 & 0 \\ 0 & -1/2 & 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & -1 & -1 & -1 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1/2 & 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & -1 & -1 & -1 & 1 \end{array}\right)$$

$$S = \begin{pmatrix} 1 & 1 & 0 \\ -1/2 & 1/2 & 0 \\ -1 & -1 & 1 \end{pmatrix}$$

Theorem 12.23 (Sylvester's Law of Inertia)**.** *Let β be a symmetric bilinear form on $V = \mathbb{R}^n$. Then there exists a basis B of V with*

$$_B[\beta]_B = \mathrm{diag}(1_r, -1_s, 0_t).$$

The numbers r, s, t do not depend on B.

Proof. According to Theorem 12.20, there exists an orthogonal basis $B = \{b_1, \dots, b_n\}$ of V. After reordering, we can assume

$$\beta(b_i, b_i) \begin{cases} > 0 & \text{for } i = 1, \dots, r, \\ < 0 & \text{for } i = r+1, \dots, r+s, \\ = 0 & \text{for } i = r+s+1, \dots, r+s+t = n \end{cases}$$

For $i = 1, \dots, r+s$, we replace b_i by $\frac{1}{\sqrt{|\beta(b_i, b_i)|}} b_i$. Then $\beta(b_i, b_i) = 1$ for $i = 1, \dots, r$ and $\beta(b_i, b_i) = -1$ for $i = r+1, \dots, r+s$. Thus, the existence of B is shown.

Because $V^\perp = \langle b_{r+s+1}, \dots, b_n \rangle$, $t = \dim V^\perp$ does not depend on B. Let $V_+ := \langle b_1, \dots, b_r \rangle$. For $v = \sum_{i=1}^r \lambda_i b_i \in V_+$ with $\lambda_i \in \mathbb{R}$, it holds that $\beta(v, v) = \sum_{i=1}^n \lambda_i^2 \ge 0$ with equality if and only if $v = 0$. Let $B' = \{b'_1, \dots, b'_n\}$ be another basis of V with

$$_{B'}[\beta]_{B'} = \mathrm{diag}(1_{r'}, -1_{s'}, 0_t).$$

Let $V'_{\le 0} := \langle b'_{r'+1}, \dots, b'_n \rangle$. For $v \in V_+ \cap V'_{\le 0}$, on the one hand $\beta(v, v) \ge 0$ and on the other hand $\beta(v, v) \le 0$. This shows $\beta(v, v) = 0$ and $v = 0$. Thus $V_+ \cap V'_{\le 0} = \{0\}$ and

$$r + s' + t = \dim(V_+ \oplus V'_{\le 0}) \le \dim V = n = r + s + t.$$

It follows that $s' \le s$. By symmetry, $s' = s$ and $r' = r$. □

Definition 12.24. In the situation of Theorem 12.23, $\operatorname{ind}(\beta) := (r, s, t)$ is called the *index* of β.[4] For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, one defines $\operatorname{ind}(A) := \operatorname{ind}(\beta)$ via the bilinear form β with $[\beta] = A$.

Example 12.25.

(a) Every scalar product on $\mathbb{R}^n$ has index $(n, 0, 0)$.

(b) If $\operatorname{ind}(\beta) = (r, s, t)$, then β is degenerate if and only if $t > 0$ holds.

(c) In relativity theory, one considers the *Minkowski space* $\mathbb{R}^4$ wrt. a bilinear form β with index $(3, 1, 0)$. The fourth dimension describes time.

Remark 12.26. To calculate the index of a matrix A, one can first determine an orthogonal basis $\tilde{B}$ using the modified Gaussian algorithm from Remark 12.21 (note: $1 + 1 \neq 0$ in $\mathbb{R}$). Subsequently, one counts the positive and negative entries on the main diagonal. If one divides the vectors in $\tilde{B}$ by the square root of the corresponding diagonal entry, one obtains the basis B as in the proof of Theorem 12.23. The matrix from Example 12.22, for example, has index $(1, -2, 0)$ wrt.

$$B = \Big\{ \frac{1}{\sqrt{2}}(1, 1, 0),\ \frac{1}{\sqrt{2}}(-1, 1, 0),\ (-1 - 1, 1) \Big\}.$$

Theorem 12.27. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with index (r, s, t). Then r is the number of positive eigenvalues and s is the number of negative eigenvalues of A, each counted with (algebraic) multiplicities.*

Proof. According to the matrix version of the principal axis theorem, A is congruent to $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A (the algebraic multiplicity of each eigenvalue coincides with the geometric multiplicity). One can apply the argument from the proof of Theorem 12.23 to this matrix. The positive λ_i are transformed to 1 and the negative ones to -1. □

Corollary 12.28. *For symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, the following statements are equivalent:*

(1) A and B are congruent.

(2) $\operatorname{ind}(A) = \operatorname{ind}(B)$.

(3) A and B have the same number of positive eigenvalues and the same number of negative eigenvalues.[5]

Proof. Since congruent matrices describe the same bilinear form, (1)⇒(2) holds. From Theorem 12.27 follows (2)⇒(3). If (3) holds, then A and B have the same index according to Theorem 12.27. Thus A and B are congruent, i. e., (1) holds. □

Remark 12.29.

(a) For each $m \leq n$ there are exactly $m + 1$ congruence classes of symmetric $n \times n$-matrices with rank m (namely with index $(i, m - i, n - m)$ for $i = 0, \ldots, m$). Therefore, there are

$$\sum_{m=0}^{n}(m + 1) = \sum_{m=1}^{n+1} m = \frac{(n + 1)(n + 2)}{2}$$

[4] This term is not uniform in the literature! Some authors call $r - s$ the *signature* of β. The index can be determined from dimension, rank, and signature.

[5] This describes the *inertia* in Sylvester's theorem.

symmetric bilinear forms on $\mathbb{R}^n$ up to choice of basis (cf. Example 1.16).

(b) If one works over $\mathbb{C}$ instead of $\mathbb{R}$, one can always achieve ${}_B[\beta]_B = \operatorname{diag}(1_r, 0_t)$. Namely, one can replace every basis vector b with $\beta(b,b) < 0$ by $\frac{\mathrm{i}}{\sqrt{|\beta(b,b)|}} b$. Therefore, there are only $n+1$ symmetric bilinear forms on $\mathbb{C}^n$ up to choice of basis. We prove Sylvester's Law of Inertia over $\mathbb{C}$ with a different notion of congruence in Theorem 13.25.

Theorem 12.30. *Let V be an arbitrary K-vector space. Let $\beta \in \operatorname{Bil}(V)$ be alternating and non-degenerate. Then there exists a basis B of V with ${}_B[\beta]_B = \left(\begin{smallmatrix} 0_n & 1_n \\ -1_n & 0_n \end{smallmatrix}\right)$.*

Proof. Induction on $\dim V$.[6] Let $b_1 \in V \setminus \{0\}$. Since β is non-degenerate, there exists $c_1 \in V$ with $\beta(b_1, c_1) \neq 0$. By replacing c_1 with $\beta(b_1, c_1)^{-1} c_1$, one achieves $\beta(b_1, c_1) = 1$. Let $U := \langle b_1, c_1 \rangle$. For $v = \lambda b_1 + \mu c_1 \in U$, it holds that $\beta(v, b_1) = \mu\beta(c_1, b_1) = -\mu$ and $\beta(v, c_1) = \lambda$. This shows $U \cap U^\perp = 0$. According to Theorem 12.16, it holds that $V = U \oplus U^\perp$. For $u \in U^\perp$, there exists a $v \in V$ with $\beta(u, v) \neq 0$. Writing $v = v_1 + v_2$ with $v_1 \in U$ and $v_2 \in U^\perp$, it follows that $\beta(u, v_2) = \beta(u, v) \neq 0$. Therefore, the restriction β' of β to $U^\perp \times U^\perp$ is also non-degenerate and alternating. By induction, $U^\perp$ has a basis $B' = \{b_2, \ldots, b_n, c_2, \ldots, c_n\}$ with

$$ {}_{B'}[\beta']_{B'} = \begin{pmatrix} 0_{n-1} & 1_{n-1} \\ -1_{n-1} & 0_{n-1} \end{pmatrix}. $$

Now $B = \{b_1, \ldots, b_n, c_1, \ldots, c_n\}$ is a basis with the desired property. □

Remark 12.31. A vector space V with an alternating, non-degenerate bilinear form is called a *symplectic space*. We will not go into this further.

12.3 Positive definite matrices

Remark 12.32. In this section, we generalize the positive definiteness of scalar products in Euclidean spaces. Let V always be an $\mathbb{R}$-vector space.

Definition 12.33. A symmetric bilinear form $\beta \in \operatorname{Bil}(V)$ is called

- *positive (semi)definite*, if $\beta(v,v) > 0$ (resp. $\beta(v,v) \geq 0$) for all $v \in V \setminus \{0\}$,
- *negative (semi)definite*, if $\beta(v,v) < 0$ (resp. $\beta(v,v) \leq 0$) for all $v \in V \setminus \{0\}$,
- *indefinite*, if $\exists v, w \in V : \beta(v,v) > 0 > \beta(w,w)$.

These terms transfer to symmetric matrices A by choosing a β with $[\beta] = A$.

Remark 12.34.

(a) Obviously, every positive definite bilinear form is also positive semidefinite. If β is positive (semi)definite, then $-\beta$ is negative (semi)definite. Therefore, one can usually restrict oneself to the positive property.

(b) If β is positive (semi)definite, then the corresponding quadratic form from Remark 12.4 is positive (resp. non-negative).

[6] According to Example 12.10, $\dim V$ is even, but this is not used here.

(c) The sum of positive (semi)definite bilinear forms and matrices is again positive (semi)definite. One can view positive semidefinite matrices as a multidimensional generalization of non-negative real numbers (see Theorem 12.46, Example 18.20 and Exercise III.11). However, the product of positive (semi)definite matrices is generally not symmetric and, according to our definition, cannot be positive (semi)definite.

Example 12.35.

(a) Let

$$A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

For $x \in \mathbb{R}^n \setminus \{0\}$ it holds that

$$xAx^{\mathrm{t}} = 2\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n-1} x_i x_{i+1} = x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + x_n^2.$$

In the case $x_1 \neq 0$ or $x_n \neq 0$ it follows that $xAx^{\mathrm{t}} \geq x_1^2 + x_n^2 > 0$. Otherwise there exists an $i \geq 0$ with $x_i = 0 \neq x_{i+1}$. Then likewise $xAx^{\mathrm{t}} \geq (x_i - x_{i+1})^2 > 0$. Thus A is positive definite.

(b) In analysis, one forms from the second partial derivatives of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ the *Hessian matrix* $(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}) \in \mathbb{R}^{n\times n}$, which at a point $x \in \mathbb{R}^n$ can only be positive definite if x is a local minimum of the function.

Theorem 12.36. *For every symmetric bilinear form $\beta \in \mathrm{Bil}(V)$ with index (r, s, t) it holds that*

(a) β positive semidefinite $\iff s = 0$.

(b) β positive definite $\iff s = t = 0$.

(c) β indefinite $\iff r, s > 0$.

Proof. Let B be the basis from Sylvester's law of inertia. If $s > 0$, then there exists a $b \in B$ with $\beta(b, b) = -1$. Then β cannot be positive semidefinite. If $t > 0$, then there exists $b \in B$ with $\beta(b, b) = 0$. Then β cannot be positive definite. If $s = 0$ (or $s = t = 0$), then $\beta(b, b) \geq 0$ (or $\beta(b, b) > 0$) for all $b \in B$. Since B is an orthogonal basis, it follows easily that β is positive (semi)definite. β is indefinite if and only if there exist $b, c \in B$ with $\beta(b, b) = 1 = -\beta(c, c)$. This means $r, s > 0$. □

Corollary 12.37. *A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive (semi)definite if and only if all eigenvalues of A are positive (or non-negative).*

Proof. As is well known, all eigenvalues of A are real. The claim follows from Theorem 12.27. □

Example 12.38. For

$$A = (1 + \delta_{ij}) = \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 2 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

the trick from Example 12.35 does not work. Because $\operatorname{rk}(A - 1_n) = 1$, 1 is an eigenvalue of A with (geometric) multiplicity $n - 1$ (cf. Exercise I.25). The missing eigenvalue must be $\operatorname{tr}(A) - (n-1) = 2n - n + 1 = n + 1$. Therefore A is positive definite.

Remark 12.39. Since eigenvalues are difficult to calculate in practice, it is useful to know other criteria for positive definiteness. A necessary (but not sufficient) condition is that all diagonal entries are positive, because $a_{ii} = e_i A e_i^{\mathrm{t}}$.

Lemma 12.40. *$A \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if there exists a matrix $S \in \mathbb{R}^{n \times n}$ with $A = SS^{\mathrm{t}}$. A is positive definite if and only if S is invertible.*

Proof. Let $A = SS^{\mathrm{t}}$ for some $S \in \mathbb{R}^{n \times n}$. For $v \in \mathbb{R}^n$, $vAv^{\mathrm{t}} = vS(vS)^{\mathrm{t}} = |vS|^2 \geq 0$ holds. Therefore A is positive semidefinite. If S is invertible, then $|vS| > 0$ for all $v \neq 0$. Then A is positive definite. Conversely, let A be positive semidefinite. By Sylvester's law of inertia and Theorem 12.36, there exists a $U \in \mathrm{GL}(n, \mathbb{R})$ with $A = UDU^{\mathrm{t}}$, where $D = \operatorname{diag}(1_r, 0_t)$. For $S = UD$, $SS^{\mathrm{t}} = A$ holds because $D^2 = D$. If A is positive definite, then $S = U$ is invertible. □

Theorem 12.41 (Sylvester Criterion)**.** *Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ be symmetric and $A_k := (a_{ij})_{i,j=1}^k$ for $1 \leq k \leq n$. A is positive definite if and only if $\det(A_k) > 0$ for $k = 1, \ldots, n$.*

Proof. Induction on n: In the case $n = 1$, A is positive definite if and only if $\det(A) = \det(A_1) = a_{11} > 0$. Now let $n \geq 2$ and $1 \leq k \leq n$. Let A be positive definite and $v \in \mathbb{R}^k \backslash \{0\}$. For $w := (v_1, \ldots, v_k, 0, \ldots, 0) \in \mathbb{R}^n$, it holds that $vA_kv^{\mathrm{t}} = wAw^{\mathrm{t}} > 0$. Thus A_k is positive definite. By Lemma 12.40 there exists an $S \in \mathrm{GL}(k, \mathbb{R})$ with $A_k = SS^{\mathrm{t}}$. It follows that $\det(A_k) = \det(S)^2 > 0$.

Conversely, assume $\det(A_k) > 0$ for $k = 1, \ldots, n$. By induction, A_{n-1} is positive definite. Let $\beta \in \mathrm{Bil}(\mathbb{R}^n)$ with $[\beta] = A$. Then the restriction β_1 of β to $\mathbb{R}^{n-1} \times \mathbb{R}^{n-1}$ is positive definite, because $[\beta_1] = A_{n-1}$. Let $b'_1, \ldots, b'_{n-1} \in \mathbb{R}^{n-1}$ be an orthonormal basis of β_1. By appending a 0, one obtains the vectors $b_i := (b'_i, 0) \in \mathbb{R}^n$ with

$$\beta(b_i, b_j) = b_i A b_j^{\mathrm{t}} = b'_i A_{n-1} b'_j = \delta_{ij}$$

for $i = 1, \ldots, n-1$. Let $V_1 := \langle b_1, \ldots, b_{n-1} \rangle$. Since β is positive definite on V_1, it holds that $V_1 \cap V_1^{\perp} = \{0\}$. With $b_n \in V_1^{\perp} \setminus \{0\}$, $B = \{b_1, \ldots, b_n\}$ is an orthogonal basis of $\mathbb{R}^n$ and

$$D := {}_B[\beta]_B = \operatorname{diag}(1, \ldots, 1, \lambda)$$

for some $\lambda \in \mathbb{R}$. Since A and D are congruent, there exists an $S \in \mathrm{GL}(n, \mathbb{R})$ with $D = SAS^{\mathrm{t}}$. It follows that $\lambda = \det(D) = \det(S)^2 \det(A) > 0$. Since D is positive definite, A is now also positive definite. □

Example 12.42.

(a) The matrix $A = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ is positive definite if and only if $a > 0$ and $ac > b^2$.

(b) Let

$$A := \begin{pmatrix} 2 & -1 & . & . & -1 \\ -1 & 2 & -1 & . & . \\ . & -1 & 2 & -1 & . \\ . & . & -1 & 2 & -1 \\ -1 & . & . & -1 & 2 \end{pmatrix}$$

According to Example 12.35, $\det(A_k) > 0$ holds for $k = 1, \ldots, 4$. Since the row sums of A vanish, $(1, \ldots, 1)^{\mathrm{t}}$ is an eigenvector for the eigenvalue 0. This shows $\det(A) = 0$. The proof of Theorem 12.41 shows that A is positive semidefinite, but not positive definite.

Remark 12.43.

(a) One calls $\det(A_k)$ the *principal minors* of A.

(b) Caution: From $\det(A_k) \geq 0$ for $k = 1, \ldots, n$ it does not follow that A is positive semidefinite (consider for example $A = \left(\begin{smallmatrix} 0 & 0 \\ 0 & -1 \end{smallmatrix}\right)$).

(c) Likewise, negative definiteness is not equivalent to $\det(A_k) < 0$. As is well known, A is negative definite if and only if $-A$ is positive definite. This is equivalent to $(-1)^k \det(A_k) > 0$ for $k = 1, \ldots, n$.

Theorem 12.44. *A symmetric matrix $A \in \mathbb{R}^{n \times n}$ with characteristic polynomial*

$$\chi_A = X^n + a_1 X^{n-1} + \ldots + a_n \in \mathbb{R}[X]$$

is positive definite if and only if the coefficients of χ_A have alternating signs, i. e. $a_i(-1)^i > 0$ for $i = 1, \ldots, n$.

Proof. Let $a_i(-1)^i > 0$ for $i = 1, \ldots, n$. For $\lambda < 0$ it then holds that

$$\chi_A(\lambda) = (-1)^n \big(|\lambda|^n + |a_1||\lambda|^{n-1} + \ldots + |a_n|\big) \neq 0.$$

Thus A possesses only positive eigenvalues. According to Corollary 12.37, A is positive definite.

Conversely, let A be positive definite with eigenvalues $\lambda_1, \ldots, \lambda_n > 0$. Then it holds that

$$\chi_A = \prod_{i=1}^{n} (X - \lambda_i) = X^n - \Big(\sum_{i=1}^{n} \lambda_i\Big) X^{n-1} + \Big(\sum_{i<j} \lambda_i \lambda_j\Big) X^{n-2} + \ldots + (-1)^n \lambda_1 \ldots \lambda_n$$

and

$$a_k (-1)^k = \sum_{i_1 < \ldots < i_k} \lambda_{i_1} \ldots \lambda_{i_n} > 0$$

for $k = 1, \ldots, n$. □

Example 12.45. For

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

it holds that

$$\chi_A = (X-2)^2(X-1) + 2 - 2(X-2) - (X-1) = X^3 - 5X^2 + 5X + 3.$$

Since the last two coefficients have the same sign, A cannot be positive definite.

Theorem 12.46. *Let $A \in \mathbb{R}^{n\times n}$ be positive (semi)definite and $k \in \mathbb{N}$. Then A possesses exactly one positive (semi)definite k-th root $W \in \mathbb{R}^{n\times n}$, i. e. $W^k = A$.*

Proof. According to the principal axis theorem, there exists an $S \in \mathrm{O}(n, \mathbb{R})$ with $S^{\mathrm{t}}AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Since A is positive semidefinite, $\lambda_1, \ldots, \lambda_n \geq 0$ holds according to Corollary 12.37. Now $W := S\,\mathrm{diag}(\sqrt[k]{\lambda_1}, \ldots, \sqrt[k]{\lambda_n})S^{\mathrm{t}}$ is positive (semi)definite with $W^k = A$. Let $\alpha \in \mathbb{R}[X]$ with $\alpha(\lambda_i) = \sqrt[k]{\lambda_i}$ for $i = 1, \ldots, n$ (interpolation). Because $(S^{\mathrm{t}}AS)^i = S^{\mathrm{t}}A^iS$ for $i \in \mathbb{N}$, it holds that

$$\alpha(A) = S\alpha(S^{\mathrm{t}}AS)S^{\mathrm{t}} = W.$$

Let also $B \in \mathbb{R}^{n\times n}$ be positive (semi)definite with $B^k = A$. The principal axis theorem yields a $T \in \mathrm{O}(n, \mathbb{R})$ with $T^{\mathrm{t}}BT = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ and $\mu_1, \ldots, \mu_n \geq 0$. Here $\mu_1^k, \ldots, \mu_n^k$ are the eigenvalues of $B^k = A$. Thus there exists a permutation matrix P with $P^{\mathrm{t}}T^{\mathrm{t}}BTP = S^{\mathrm{t}}WS$ and $P^{\mathrm{t}}T^{\mathrm{t}}ATP = S^{\mathrm{t}}AS$. Therefore TPS^{t} commutes with A and with $\alpha(A) = W$. This shows $B = TPS^{\mathrm{t}}WSP^{\mathrm{t}}T^{\mathrm{t}} = W$. □

13 Unitary Spaces

13.1 Sesquilinear Forms

Remark 13.1. In this chapter, we develop a geometry for the complex numbers. Instead of the Euclidean scalar product, a "skew-symmetric" map appears, which is linear only in the first coordinate. The counterpart of the principal axis theorem is the spectral theorem.

Definition 13.2. A *scalar product* on a $\mathbb{C}$-vector space V is a map $V \times V \to \mathbb{C}$ with the following properties ($u, v, w \in V$, $\lambda \in \mathbb{C}$):

- $[v, v] \in \mathbb{R}$ and $[v, v] \geq 0$ with equality if and only if $v = 0$ (*positive definite*),
- $[v, w] = \overline{[w, v]}$ (*skew-symmetric*),
- $[\lambda u + v, w] = \lambda[u, w] + [v, w]$.

Together with a scalar product, V becomes a *unitary space*. Vectors $v, w \in V$ are called *orthogonal* if $[v, w] = 0$. One calls $|v| := \sqrt{[v, v]} \geq 0$ the *norm* of v. In the case $|v| = 1$, v is called *normalized*.

Example 13.3. The *standard scalar product* on $V = \mathbb{C}^n$ is defined by

$$[v, w] := v\overline{w}^{\mathrm{t}} = \sum_{i=1}^{n} v_i \overline{w_i}$$

for $v, w \in V$. It is easy to check that the properties of the scalar product are satisfied.

Remark 13.4.

(a) For $v, w \in V$ and $\lambda \in \mathbb{C}$ it holds that

$$[u, \lambda v + w] = \overline{[\lambda v + w, u]} = \overline{\lambda[v, u] + [w, u]} = \overline{\lambda}[u, v] + [u, w],$$

i.e., the scalar product is not linear in the second component. One therefore speaks of a *sesquilinear form*.[1]

(b) Most of the theorems proven in section 11.1 do not depend significantly on the bilinearity of the scalar product and can therefore be transferred without problems.

Lemma 13.5. *Let V be a unitary space, $v, w \in V$ and $\lambda \in \mathbb{C}$. Then:*

(a) $|\lambda v| = |\lambda||v|$ (homogeneity).

(b) $|[v, w]| \leq |v||w|$ *with equality if and only if v and w are linearly dependent* (Cauchy-Schwarz inequality).

[1] *sesqui* is Latin for one and a half.

(c) $||v| - |w|| \leq |v + w| \leq |v| + |w|$ (triangle inequality).

Proof.

(a) $|\lambda v| = \sqrt{[\lambda v, \lambda v]} = \sqrt{\lambda\overline{\lambda}[v, v]} = \sqrt{|\lambda|^2}\sqrt{[v, v]} = |\lambda||v|$.

(b) Wlog. let $w \neq 0$ and $\lambda := \frac{[v,w]}{[w,w]}$. Due to $\overline{\lambda} = \frac{[w,v]}{[w,w]}$ it holds that

$$0 \leq |v - \lambda w|^2 = [v - \lambda w, v - \lambda w] = [v, v] - \lambda[w, v] - \overline{\lambda}[v, w] + |\lambda|^2[w, w] = |v|^2 - \frac{|[v, w]|^2}{|w|^2}.$$

It follows that $|[v, w]|^2 \leq |v|^2|w|^2$ and $|[v, w]| \leq |v||w|$. Equality implies $v = \lambda w$, i. e. v and w are linearly dependent. Conversely, if v and w are given as linearly dependent, then there exists a $\mu \in \mathbb{C}$ with $v = \mu w$ and $|[v, w]| = |\mu||w|^2 \overset{\text{(a)}}{=} |\mu w||w| = |v||w|$.

(c) Let $[v, w] = a + b\mathrm{i}$. Then it holds that

$$[v, w] + [w, v] = [v, w] + \overline{[v, w]} = 2a \leq 2\sqrt{a^2 + b^2} = 2|[v, w]|.$$

It follows that

$$|v + w|^2 = [v + w, v + w] \leq [v, v] + 2|[v, w]| + [w, w] \overset{\text{(b)}}{\leq} |v|^2 + 2|v||w| + |w|^2 = (|v| + |w|)^2$$

and $|v + w| \leq |v| + |w|$. Thus $|v| = |v + w - w| \leq |v + w| + |w|$ and $|v| - |w| \leq |v + w|$. Swapping v and w yields $-(|v| - |w|) = |w| - |v| \leq |v + w|$, i. e. $||v| - |w|| \leq |v + w|$. □

Remark 13.6. Let V be a unitary space.

(a) A basis $b_1, \ldots, b_n$ of V is called an *orthonormal basis* if $[b_i, b_j] = \delta_{ij}$ for $1 \leq i, j \leq n$ holds. The Gram-Schmidt process for calculating an orthonormal basis works unchanged for unitary spaces. In particular, every unitary space possesses an orthonormal basis.

(b) For $U \leq V$ one defines as usual $U^\perp := \{v \in V : \forall u \in U : [v, u] = 0\} \leq V$. The rules from Lemma 11.15

- $V = U \oplus U^\perp$,
- $(U^\perp)^\perp = U$,
- $U \subseteq W \iff W^\perp \subseteq U^\perp$,

also hold for unitary spaces.

13.2 Adjoint Maps

Theorem 13.7. *Let V be a unitary space and $f \in \mathrm{End}(V)$. Then there exists exactly one $f^* \in \mathrm{End}(V)$ with $[f(v), w] = [v, f^*(w)]$ for all $v, w \in V$.*

Proof. Let $b_1, \dots, b_n$ be an orthonormal basis of V. We define $f^* \in \mathrm{End}(V)$ by

$$f^*(b_j) := \sum_{i=1}^{n} [f(b_i), b_j] b_i$$

for $j = 1, \dots, n$. For $v = \sum \lambda_i b_i$ and $w = \sum \mu_j b_j$ it holds that

$$[f(v), w] = \sum_{i,j=1}^{n} \lambda_i \overline{\mu_j} [f(b_i), b_j] = \sum_{i,j=1}^{n} \lambda_i \overline{\mu_j} [b_i, f^*(b_j)] = [v, f^*(w)].$$

Let also $f_1 \in \mathrm{End}(V)$ with $[f(v), w] = [v, f_1(w)]$ for all $v, w \in V$. Then $[v, f^*(w) - f_1(w)] = 0$. For $v := f^*(w) - f_1(w)$ it follows that $f^*(w) = f_1(w)$ for all $w \in V$. This shows $f_1 = f^*$. □

Definition 13.8. In the situation of Theorem 13.7, f^* is called the *adjoint* map to f.

Remark 13.9.

(a) For $v, w \in V$ we have $[v, f(w)] = \overline{[f(w), v]} = \overline{[w, f^*(v)]} = [f^*(v), w]$. In particular, $(f^*)^* = f$.

(b) For $f, g \in \mathrm{End}(V)$ and $\lambda \in \mathbb{C}$ we have $(\lambda f + g)^* = \overline{\lambda} f^* + g^*$.

(c) Attention: We use the same notation for the adjoint map of f as for the dual map of f. However, the dual map lies in $\mathrm{End}(V^*)$.

Definition 13.10. Let V be a unitary space. We call $f \in \mathrm{End}(V)$

- *Hermitian*, if $f = f^*$, i. e. $[f(v), w] = [v, f(w)]$ for all $v, w \in V$.
- *unitary*, if $[f(v), f(w)] = [v, w]$ for all $v, w \in V$.
- *normal*, if $f \circ f^* = f^* \circ f$.

Remark 13.11.

(a) Let $f \in \mathrm{End}(V)$ be Hermitian. Let $\lambda \in \mathbb{C}$ be an eigenvalue of f with eigenvector $v \in V$. Then

$$\lambda |v|^2 = [\lambda v, v] = [f(v), v] = [v, f(v)] = [v, \lambda v] = \overline{\lambda} |v|^2$$

and $\lambda = \overline{\lambda} \in \mathbb{R}$.

(b) Let $f \in \mathrm{End}(V)$ be unitary. For $v \in \mathrm{Ker}(f)$ we have $|v|^2 = [v, v] = [f(v), f(v)] = 0$, i. e. $v = 0$. Therefore f is an isomorphism. Because of $[f(v), w] = [v, f^{-1}(w)]$ for all $v, w \in V$ it also follows that $f^* = f^{-1}$. If $f, g \in \mathrm{End}(V)$ are unitary, then so are $f \circ g$ and f^{-1}. Therefore the unitary maps form a subgroup $\mathrm{U}(V)$ of $\mathrm{GL}(V)$. We call $\mathrm{U}(V)$ the *unitary group* of degree n.

(c) Let λ be an eigenvalue of a unitary map f with eigenvector $v \in V$. Then $|\lambda|^2 |v|^2 = [\lambda v, \lambda v] = [f(v), f(v)] = [v, v] = |v|^2$ and $|\lambda| = 1$.

(d) Hermitian and unitary maps are obviously normal.

Lemma 13.12. *Let V be unitary with orthonormal basis B and $f \in \mathrm{End}(V)$. Then ${}_B[f^*]_B = \overline{{}_B[f]_B}^{\mathrm{t}}$.*

Proof. Let $B = \{b_1, \ldots, b_n\}$. Let $f(b_i) = \sum_{j=1}^n a_{ji} b_j$ and $f^*(b_i) = \sum_{j=1}^n a^*_{ji} b_j$ for $i = 1, \ldots, n$. The claim follows from

$$\overline{a^*_{ji}} = [b_j, f^*(b_i)] = [f(b_j), b_i] = a_{ij}. \qquad \square$$

Definition 13.13. For $A \in \mathbb{C}^{n \times m}$ let $A^* := \overline{A}^{\mathrm{t}} = \overline{A^{\mathrm{t}}}$ be the matrix *adjoint* to A. For $A \in \mathrm{GL}(n, \mathbb{C})$ we use the abbreviation $A^{-*} := (A^{-1})^* = (A^*)^{-1}$.

Corollary 13.14. *Let V be unitary with orthonormal basis B, $f \in \mathrm{End}(V)$ and $A := {}_B[f]_B$. Then*

(a) f Hermitian $\iff A^* = A \iff A^{\mathrm{t}} = \overline{A}$.

(b) f unitary $\iff A^* = A^{-1} \iff A^{\mathrm{t}} = \overline{A}^{-1}$.

(c) f normal $\iff A^*A = AA^* \iff A^{\mathrm{t}}\overline{A} = \overline{A}A^{\mathrm{t}}$.

Proof. It holds that

$$f \text{ hermitian} \iff f^* = f \overset{13.12}{\iff} A^* = A \iff \overline{A^*} = \overline{A} \iff A^{\mathrm{t}} = \overline{A}.$$

The other statements are proven analogously. $\square$

Definition 13.15. A matrix $A \in \mathbb{C}^{n \times n}$ is called

- *hermitian*, if $A^* = A$.
- *unitary*, if $A^* = A^{-1}$.
- *normal*, if $AA^* = A^*A$.

Example 13.16.

(a) Hermitian and unitary matrices are normal. Diagonal matrices are normal, but not necessarily hermitian or unitary.

(b) Real matrices are hermitian (or unitary) if and only if they are symmetric (or orthogonal).

(c) Inverting, transposing, and complex-conjugating are commuting operators on $\mathbb{C}^{n \times n}$ (cf. Remark 5.9). If A is hermitian (or unitary, normal), then so are $\overline{A}$, A^{t} and A^* (verify).

Remark 13.17. The unitary matrices form a subgroup $\mathrm{U}(n, \mathbb{C})$ of $\mathrm{GL}(n, \mathbb{C})$, which as usual corresponds to $\mathrm{U}(V)$. For $A \in \mathrm{U}(n, \mathbb{C})$ it holds that $|\det(A)|^2 = \det(A)\overline{\det(A)} = \det(A^{\mathrm{t}}\overline{A}) = 1$. According to Remark 13.11, all eigenvalues of A also have absolute value 1. One calls

$$\mathrm{SU}(n, \mathbb{C}) := \mathrm{U}(n, \mathbb{C}) \cap \mathrm{SL}(n, \mathbb{C}) \leq \mathrm{U}(n, \mathbb{C})$$

the *special unitary group* of degree n.

13.3 The Spectral Theorem

Theorem 13.18 (Spectral Theorem[2]). *Let V be unitary and $f \in \mathrm{End}(V)$. f is normal if and only if V has an orthonormal basis of eigenvectors of f. In particular, normal endomorphisms are diagonalizable.*

Proof. Let B be an orthonormal basis of eigenvectors of f. Then $A := {}_B[f]_B$ is a diagonal matrix. Certainly A^* is also a diagonal matrix. In particular, A and A^* commute. According to Corollary 13.14, f is normal. Conversely, let f be normal. We argue by induction on $n := \dim V$. In the case $n = 1$, every normalized vector provides an orthonormal basis of eigenvectors. Let $n \geq 2$. According to the Fundamental Theorem of Algebra, f has an eigenvalue $\lambda \in \mathbb{C}$ with eigenvector $b_1 \in V$. Wlog. let $|b_1| = 1$. Because of

$$\begin{aligned}|f^*(b_1) - \overline{\lambda} b_1|^2 &= [f^*(b_1), f^*(b_1)] - \lambda[f^*(b_1), b_1] - \overline{\lambda}[b_1, f^*(b_1)] + |\lambda|^2 [b_1, b_1] \\ &= [f(f^*(b_1)), b_1] - \lambda[b_1, f(b_1)] - \overline{\lambda}[f(b_1), b_1] + |\lambda|^2 = [f^*(f(b_1)), b_1] - |\lambda|^2 \\ &= \lambda[f^*(b_1), b_1] - |\lambda|^2 = \lambda[b_1, f(b_1)] - |\lambda|^2 = 0\end{aligned}$$

$\overline{\lambda}$ is an eigenvalue of f^* for the eigenvector b_1. For $U := \langle b_1 \rangle$, it holds that $V = U \oplus U^\perp$. For $u \in U^\perp$, we have $[f(u), b_1] = [u, f^*(b_1)] = \lambda[u, b_1] = 0$ and $f(u) \in U^\perp$. Therefore, the restriction of f to $U^\perp$ is a normal endomorphism. By induction, there exists an orthonormal basis $b_2, \ldots, b_n$ of $U^\perp$ consisting of eigenvectors of f. Now $b_1, \ldots, b_n$ is an orthonormal basis of V consisting of eigenvectors of f. □

Remark 13.19. If $A \in \mathbb{R}^{n \times n}$ is symmetric, then the Spectral Theorem yields an $S \in \mathrm{U}(n, \mathbb{C})$ with $S^* A S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. In contrast to the Principal Axis Theorem, one does not obtain that S is orthogonal (i.e., real).

Corollary 13.20. *Let V be unitary and $f \in \mathrm{End}(V)$ be normal with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Then:*

(a) f is Hermitian $\iff$ $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.

(b) f is unitary $\iff$ $|\lambda_1| = \ldots = |\lambda_n| = 1$.

Proof. According to the Spectral Theorem, there exists an orthonormal basis B of V with $D := {}_B[f]_B = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. According to Corollary 13.14, f is Hermitian (resp. unitary) if and only if $\overline{D} = D^{\mathrm{t}} = D$ (resp. $1_n = D\overline{D} = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$) holds. This proves the claim. □

Remark 13.21. Let $f \in \mathrm{End}\, V$ be normal. Let $v, w \in V$ be eigenvectors for distinct eigenvalues λ and μ, respectively. As in the proof of the Spectral Theorem, w is an eigenvector of f^* for the eigenvalue $\overline{\mu}$. From

$$\lambda[v, w] = [f(v), w] = [v, f^*(w)] = [v, \overline{\mu} w] = \mu[v, w]$$

it follows that $[v, w] = 0$. Therefore, eigenvectors for distinct eigenvalues are orthogonal (as with symmetric matrices, see Remark 11.40). The orthonormal basis in the Spectral Theorem can thus be determined using the Gram-Schmidt process, just as in the Principal Axis Theorem.

[2]The set of eigenvalues of $f \in \mathrm{End}(V)$ is called the *spectrum* of f.

Example 13.22. Suppose that

$$A := \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 4 & 5\mathrm{i} & 2\mathrm{i} \\ -2 & 2\mathrm{i} & 8\mathrm{i} \end{pmatrix}$$

is normal (otherwise a contradiction will arise). We calculate

$$\chi_A = (X - 5\mathrm{i})^2(X - 8\mathrm{i}) - 32\mathrm{i} + 8(X - 5\mathrm{i}) + 16(X - 8\mathrm{i}) = X^3 - 18\mathrm{i}X^2 - 81X = X(X - 9\mathrm{i})^2.$$

Due to

$$A \sim \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 0 & 9\mathrm{i}/5 & 18\mathrm{i}/5 \\ 0 & 18\mathrm{i}/5 & 36\mathrm{i}/5 \end{pmatrix} \sim \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 0 & 9\mathrm{i}/5 & 18\mathrm{i}/5 \\ 0 & 0 & 0 \end{pmatrix}$$

$(2\mathrm{i}, -2, 1)$ is an eigenvector for the eigenvalue 0. After normalization, let $b_1 := \frac{1}{3}(2\mathrm{i}, -2, 1)$. The eigenspace for the eigenvalue 9i is spanned by $c_2 := (0, 1, 2)$ and $c_3 := (\mathrm{i}, 1, 0)$. We observe that these vectors are indeed orthogonal to b_1. Therefore, A must be normal. The Gram-Schmidt process yields:

$$\begin{aligned} b_2 &:= \frac{1}{\sqrt{5}}(0, 1, 2), \\ c_3' &:= c_2 - [c_2, b_2]b_2 = (\mathrm{i}, 4/5, -2/5), \\ b_3 &:= \frac{1}{3\sqrt{5}}(5\mathrm{i}, 4, -2). \end{aligned}$$

Now $B = \{b_1, b_2, b_3\}$ is an orthonormal basis of eigenvectors of A.

Theorem 13.23 (Schur Decomposition)**.** *For $A \in \mathbb{C}^{n\times n}$ there exists an $S \in \mathrm{U}(n, \mathbb{C})$ such that S^*AS is a triangular matrix. A is normal if and only if S^*AS is a diagonal matrix.*

Proof. Wlog. let $n \geq 2$. Let $\lambda \in \mathbb{C}$ be an eigenvalue of A and $v \in \mathbb{C}^n$ a normalized eigenvector for λ. We extend v with Gram-Schmidt to an orthonormal basis of $\mathbb{C}^n$. After changing to this basis, $A = \left(\begin{smallmatrix} \lambda & * \\ 0 & A_1 \end{smallmatrix}\right)$ holds with $A_1 \in \mathbb{C}^{(n-1)\times(n-1)}$. By induction, there exists an $S_1 \in \mathrm{U}(n-1, \mathbb{C})$ such that $S_1^*A_1S_1$ is an upper triangular matrix. Now $S := \mathrm{diag}(1, S_1) \in \mathrm{U}(n, \mathbb{C})$ and S^*AS is an upper triangular matrix. A lower triangular matrix is obtained by transforming A^{t} into an upper triangular matrix and then transposing.

If A is normal, then $D := (d_{ij}) = S^*AS$ is a normal upper triangular matrix. Because of

$$|d_{ii}|^2 = (DD^*)_{ii} = (D^*D)_{ii} = |d_{1i}|^2 + \ldots + |d_{ii}|^2$$

$a_{ji} = 0$ for $j = 1, \ldots, i-1$ and $i = 1, \ldots, n$. This shows that D is a diagonal matrix. Conversely, if D is a diagonal matrix, then A is normal according to the spectral theorem. □

Example 13.24. The matrix

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

obviously has the eigenvector e_3 for the eigenvalue 1. Transition to the orthonormal basis $\{e_3, e_2, e_1\}$ yields

$$A \approx \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}.$$

The matrix $A_1 = \left(\begin{smallmatrix} 1 & 2 \\ -1 & -1 \end{smallmatrix}\right)$ in the proof of Theorem 13.23 has the eigenvector $(1+\mathrm{i}, -1)$ for the eigenvalue i. Orthogonal to it is $(1, 1-\mathrm{i})$. After normalization, one obtains

$$A \approx \frac{1}{3}\begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1-\mathrm{i} & -1 \\ 0 & 1 & 1+\mathrm{i} \end{pmatrix}\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}\begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1+\mathrm{i} & 1 \\ 0 & -1 & 1-\mathrm{i} \end{pmatrix} = \begin{pmatrix} 1 & (2+2\mathrm{i})/\sqrt{3} & 2/\sqrt{3} \\ 0 & \mathrm{i} & 1-2\mathrm{i} \\ 0 & 0 & -\mathrm{i} \end{pmatrix}.$$

Theorem 13.25 (Inertia Theorem for Hermitian Matrices)**.** *For every Hermitian matrix $A \in \mathbb{C}^{n\times n}$, there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with*

$$S^* A S = \mathrm{diag}(1_r, -1_s, 0_t).$$

The numbers r, s, t are uniquely determined.

Proof. According to the spectral theorem, we can assume $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. According to Corollary 13.20, the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are real. Let wlog. $\lambda_1, \ldots, \lambda_r > 0$, $\lambda_{r+1}, \ldots, \lambda_{r+s} < 0$ and $\lambda_{r+s+1} = \ldots = \lambda_n = 0$. The equation now holds with

$$S := \mathrm{diag}\big(\sqrt{|\lambda_1|}^{-1}, \ldots, \sqrt{|\lambda_{r+s}|}^{-1}, 1_t\big) \in \mathrm{GL}(n, \mathbb{C}).$$

With $\mathrm{rk}(A) = r + s$, t is uniquely determined. For all $v \in \langle e_1, \ldots, e_r\rangle \setminus \{0\}$, it holds that $\overline{v} A v^{\mathrm{t}} > 0$. Conversely, let $U \leq \mathbb{C}^n$ with $\overline{u} A u^{\mathrm{t}} > 0$ for all $u \in U \setminus \{0\}$. For $u \in U \cap \langle e_{r+1}, \ldots, e_n\rangle$, on the one hand $\overline{u} A u^{\mathrm{t}} \geq 0$ and on the other hand $\overline{u} A u^{\mathrm{t}} \leq 0$. This shows $u = 0$ and $U \cap \langle e_{r+1}, \ldots, e_n\rangle = \{0\}$. It follows that $\dim U \leq r$. Thus, r is the maximum dimension of a subspace $U \leq \mathbb{C}^n$ with $\overline{u} A u^{\mathrm{t}} > 0$ for all $u \in U \setminus \{0\}$. In this way, r is uniquely determined by A. Now $s = n - r - t$ is also uniquely determined. □

Example 13.26. The numbers r, s, t in Theorem 13.25 can be determined using the modified Gaussian method from Remark 12.21 (note: $1 + 1 \neq 0$ in $\mathbb{C}$). For example:

$$\begin{aligned}
(A|1_3) &= \left(\begin{array}{ccc|ccc} 1 & \mathrm{i} & 2 & 1 & 0 & 0 \\ -\mathrm{i} & 1 & -1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & \mathrm{i} & 2 & 1 & 0 & 0 \\ 0 & 0 & 2\mathrm{i}-1 & \mathrm{i} & 1 & 0 \\ 0 & -2\mathrm{i}-1 & -1 & -2 & 0 & 1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -2\mathrm{i}-1 & -2 & 0 & 1 \\ 0 & 2\mathrm{i}-1 & 0 & \mathrm{i} & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & -2 & 0 & 1 \\ 0 & 0 & 5 & 2-3\mathrm{i} & 1 & 2\mathrm{i}-1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & (2-3\mathrm{i})/\sqrt{5} & 1/\sqrt{5} & (2\mathrm{i}-1)/\sqrt{5} \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array}\right) \\
S^* &= \frac{1}{\sqrt{5}}\begin{pmatrix} \sqrt{5} & 0 & 0 \\ 2-3\mathrm{i} & 1 & 2\mathrm{i}-1 \\ -2\sqrt{5} & 0 & \sqrt{5} \end{pmatrix}
\end{aligned}$$

Remark 13.27. Mirsky's theorem gives a necessary and sufficient condition for the existence of matrices with prescribed eigenvalues and main diagonal elements (their sum must be equal). For Hermitian matrices, stronger restrictions apply to these numbers.

Theorem 13.28 (Schur-Horn)**.**

(a) If $A \in \mathbb{C}^{n\times n}$ is Hermitian with main diagonal $d_1 \geq \ldots \geq d_n$ and eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$, then

$$\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{k} \lambda_i$$

for $k = 1, \ldots, n$ with equality in the case $k = n$.

(b) If real numbers $d_1 \geq \ldots \geq d_n$ and $\lambda_1 \geq \ldots \geq \lambda_n$ are given with $\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{k} \lambda_i$ for $k = 1, \ldots, n$ and $\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \lambda_i$, then there exists a real symmetric matrix with main diagonal $d_1, \ldots, d_n$ and eigenvalues $\lambda_1, \ldots, \lambda_n$.

Proof (Chan-Li).

(a) By the spectral theorem, there exists a unitary matrix $S := (s_{ij})$ with

$$A = (a_{ij}) = S^* \operatorname{diag}(\lambda_1, \ldots, \lambda_n) S.$$

For $1 \leq k \leq n$ it holds that

$$\sum_{i=1}^{k} d_i = \sum_{i=1}^{k} a_{ii} = \sum_{i=1}^{k} \sum_{j=1}^{n} \overline{s_{ji}} s_{ji} \lambda_j = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{k} |s_{ji}|^2. \tag{13.1}$$

Since S is unitary, $t_j := \sum_{i=1}^{k} |s_{ji}|^2 \leq 1$ holds with equality if $k = n$. It follows that

$$\begin{aligned}
\sum_{j=1}^{n} t_j &= \sum_{i=1}^{k} \sum_{j=1}^{n} |s_{ji}|^2 = k, \\
\sum_{i=1}^{k} (d_i - \lambda_i) &= \sum_{j=1}^{n} \lambda_j t_j - \sum_{j=1}^{k} \lambda_j + \lambda_k \overbrace{\Big(k - \sum_{i=1}^{n} t_i\Big)}^{=0} \\
&= \sum_{j=1}^{k} \underbrace{(\lambda_j - \lambda_k)}_{\geq 0} \underbrace{(t_j - 1)}_{\leq 0} + \sum_{i=k+1}^{n} t_i \underbrace{(\lambda_i - \lambda_k)}_{\leq 0} \leq 0.
\end{aligned}$$

Thus $\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{k} \lambda_i$ and $\sum_{i=1}^{n} d_i = \operatorname{tr}(A) = \sum_{i=1}^{n} \lambda_i$.

(b) Induction on n: In the case $n = 1$, $A := (d_1) = (\lambda_1)$ satisfies the claim. Let $n = 2$. Then $\lambda_1 \geq d_1 \geq d_2 = \lambda_1 + \lambda_2 - d_1 \geq \lambda_2$ holds. In the case $\lambda_1 = \lambda_2$ we can choose $A := d_1 1_2$. In the case $\lambda_1 > \lambda_2$ we have

$$S := \frac{1}{\sqrt{\lambda_1 - \lambda_2}} \begin{pmatrix} \sqrt{d_1 - \lambda_2} & -\sqrt{\lambda_1 - d_1} \\ \sqrt{\lambda_1 - d_1} & \sqrt{d_1 - \lambda_2} \end{pmatrix} \in \mathrm{O}(2, \mathbb{R}).$$

For $A = (a_{ij}) = S^{\mathrm{t}} \operatorname{diag}(\lambda_1, \lambda_2) S$ it holds that

$$a_{11} = \frac{1}{\lambda_1 - \lambda_2} \big((d_1 - \lambda_2)\lambda_1 + (\lambda_1 - d_1)\lambda_2 \big) = d_1$$

according to (13.1). It follows that $a_{22} = \lambda_1 + \lambda_2 - d_1 = d_2$. Now let $n \geq 3$ and

$$\lambda_2' := \lambda_1 + \lambda_2 - d_1 \geq \lambda_2.$$

By induction, there exists $S \in \mathrm{O}(2,\mathbb{R})$ such that $S^{\mathrm{t}} \operatorname{diag}(\lambda_1, \lambda_2) S$ has main diagonal (d_1, λ_2'). Since the sequences $(\lambda_2', \lambda_3, \ldots, \lambda_n)$ and $(d_2, \ldots, d_n)$ satisfy the same assumptions, there exists by induction a $T \in \mathrm{O}(n-1,\mathbb{R})$ such that $T^{\mathrm{t}} \operatorname{diag}(\lambda_2', \ldots, \lambda_n) T$ has main diagonal $d_2, \ldots, d_n$. For $U := \operatorname{diag}(S, 1_{n-2}) \operatorname{diag}(1_1, T) \in \mathrm{O}(n,\mathbb{R})$,

$$A := U^{\mathrm{t}} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U = \begin{pmatrix} 1 & 0 \\ 0 & T^{\mathrm{t}} \end{pmatrix} \begin{pmatrix} d_1 & * & & & 0 \\ * & \lambda_2' & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ 0 & & & & \lambda_n \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & T \end{pmatrix}$$

has main diagonal $(d_1, \ldots, d_n)$ and eigenvalues $\lambda_1, \ldots, \lambda_n$. □

Example 13.29. We construct a symmetric matrix with eigenvalues $3, 2, 1$ and main diagonal $2, 2, 2$. Using the notation from the above proof, $\lambda_2' = 3$. One obtains

$$\begin{aligned} A &= \begin{pmatrix} 1 & . & . \\ . & 1/\sqrt{2} & 1/\sqrt{2} \\ . & -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} . & 1 & . \\ -1 & . & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} 3 & . & . \\ . & 2 & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} . & -1 & . \\ 1 & . & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} 1 & . & . \\ . & 1/\sqrt{2} & -1/\sqrt{2} \\ . & 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} \sqrt{2} & . & . \\ . & 1 & 1 \\ . & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & . & . \\ . & 3 & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & . & . \\ . & 1 & -1 \\ . & 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & . & . \\ . & 2 & -1 \\ . & -1 & 2 \end{pmatrix}. \end{aligned}$$

14 The Jordan Normal Form

14.1 Generalized Eigenspaces

Remark 14.1. According to Theorem 10.34, there are two reasons why an endomorphism is not diagonalizable:

- The characteristic polynomial does not split into linear factors.
- The eigenspaces are too "small", i.e., the geometric multiplicity of an eigenvalue is smaller than the corresponding algebraic multiplicity.

Over $\mathbb{C}$, the first problem does not occur according to the Fundamental Theorem of Algebra. To circumvent the second problem, we replace eigenspaces with larger subspaces. In the following, let V be a finite-dimensional K-vector space.

Definition 14.2. Let $f \in \mathrm{End}(V)$. A subspace $U \leq V$ is called *f-invariant*, if $f(U) \subseteq U$ holds.

Example 14.3. For $f \in \mathrm{End}(V)$, $\mathrm{Ker}(f)$ and $f(V)$ are always f-invariant subspaces, because $f(\mathrm{Ker}(f)) = \{0\} \subseteq \mathrm{Ker}(f)$ and $f(f(V)) \subseteq f(V)$.

Remark 14.4. Let $U \leq V$ be an f-invariant subspace.

(a) For $v, w \in V$, it holds that

$$v + U = w + U \implies v - w \in U \implies f(v) - f(w) = f(v - w) \in U \implies f(v) + U = f(w) + U.$$

Therefore, $\overline{f}\colon V/U \to V/U$, $v + U \mapsto f(v) + U$ is a well-defined endomorphism.

(b) Let B_1 be a basis of U and $B = B_1 \dot\cup B_2$ a basis of V. For dimensional reasons, $\{b + U : b \in B_2\}$ is then a basis of V/U. Furthermore, ${}_B[f]_B = \left(\begin{smallmatrix} A & C \\ 0 & D \end{smallmatrix}\right)$, where $A = {}_{B_1}[f_{|U}]_{B_1}$ and $D = {}_{B_2}[\overline{f}]_{B_2}$. If U possesses an f-invariant complement W (i.e., $V = U \oplus W$), then $C = 0$ can be achieved by a suitable choice of basis. We therefore attempt to decompose V into the smallest possible f-invariant subspaces.

Definition 14.5. A map $f \in \mathrm{End}(V)$ is called *trigonalizable*, if there exists a basis B of V such that ${}_B[f]_B$ is a triangular matrix. Analogously, $A \in K^{n\times n}$ is called *trigonalizable*, if A is similar to a triangular matrix.

Remark 14.6. Let $B = \{b_1, \ldots, b_n\}$ be a basis of V such that ${}_B[f]_B$ is an upper triangular matrix. Then ${}_{B'}[f]_{B'}$ with $B' = \{b_n, b_{n-1}, \ldots, b_1\}$ is a lower triangular matrix. Thus, in Definition 14.5 one does not need to specify whether upper or lower triangular matrices are considered (cf. Remark 15.26).

Example 14.7. According to the Schur decomposition, every complex square matrix is triangulable. The next theorem extends Theorem 10.52.

Theorem 14.8. *A map $f \in \mathrm{End}(V)$ is triangulable if and only if μ_f (or χ_f) splits into linear factors.*

Proof. If ${}_B[f]_B$ (for a basis B of V) is a triangular matrix with diagonal $\lambda_1, \ldots, \lambda_n$, then $\chi_f = (X - \lambda_1) \ldots (X - \lambda_n)$ splits into linear factors. By Cayley-Hamilton and Lemma 10.24, μ_f also splits into linear factors.

Conversely, assume that μ_f splits into linear factors. By Theorem 10.54, χ_f also splits into linear factors. Wlog. let $V \neq \{0\}$. Then there exists an eigenvector $v \in V$ of f. Obviously, $U := \langle v \rangle$ is f-invariant. Let $\overline{V} := V/U$ and $\overline{f}$ as in Remark 14.4. Then $\mu_{\overline{f}} \mid \mu_f$. By Lemma 10.24, $\mu_{\overline{f}}$ also splits into linear factors. By induction on $\dim V$, we obtain a basis $\overline{B} = \{\overline{b_1}, \ldots, \overline{b_{n-1}}\}$ of $\overline{V}$ such that ${}_{\overline{B}}[\overline{f}]_{\overline{B}}$ is a lower triangular matrix. We choose $b_i \in \overline{b_i}$ for $i = 1, \ldots, n-1$. Then $f(b_i) \in \langle b_i, \ldots, b_{n-1}, v \rangle$ for $i = 1, \ldots, n-1$. Therefore, $B := \{b_1, \ldots, b_{n-1}, v\}$ is a basis of V such that ${}_B[f]_B$ is a lower triangular matrix. □

Lemma 14.9. *Let $f \in \mathrm{End}(V)$.*

(a) If $U \leq V$ is f-invariant, then $U^0 \leq V^*$ is f^*-invariant.

(b) If $U \leq V^*$ is f^*-invariant, then $U_0 \leq V$ is f-invariant.

Proof. The dual complements U^0, U_0 and the dual map f^* were defined in section 7.3.

(a) Let $\gamma \in U^0$. For all $u \in U$, it holds that $f^*(\gamma)(u) = \gamma(f(u)) = 0$, i. e. $f^*(\gamma) \in U^0$.

(b) Let $v \in U_0$. For all $\gamma \in U$, it holds that $\gamma(f(v)) = f^*(\gamma)(v) = 0$, i. e. $f(v) \in U_0$. □

Lemma 14.10. *Let $f \in \mathrm{End}(V)$ be diagonalizable and $U \leq V$ be an f-invariant subspace. Then the restriction $f_{|U}$ is also diagonalizable.*

Proof. For the minimal polynomial μ_1 of $f_{|U}$, we have $\mu_1 \mid \mu_f$ according to Lemma 10.44. According to Theorem 10.52, μ_f splits into pairwise distinct linear factors. According to Lemma 10.24, this also holds for μ_1. □

Lemma 14.11 (Simultaneous Diagonalization)**.** *Let $f, g \in \mathrm{End}(V)$ be diagonalizable. Then $f \circ g = g \circ f$ if and only if f and g are* simultaneously *diagonalizable, i. e. there exists a basis B of V such that ${}_B[f]_B$ and ${}_B[g]_B$ are diagonal matrices.*

Proof. Let B be a basis of V such that $D_f := {}_B[f]_B$ and $D_g := {}_B[g]_B$ are diagonal matrices. From $D_f D_g = D_g D_f$ it then follows that $f \circ g = g \circ f$. Conversely, let $f \circ g = g \circ f$. For the eigenvalues $\lambda_1, \ldots, \lambda_k \in K$ of f, we have $V = E_{\lambda_1}(f) \oplus \ldots \oplus E_{\lambda_k}(f)$, since f is diagonalizable. For $v \in U := E_{\lambda_i}(f)$, we have

$$f(g(v)) = g(f(v)) = \lambda_i g(v)$$

and $g(v) \in U$. Therefore, U is a g-invariant subspace. According to Lemma 14.10, $g_{|U}$ is diagonalizable. Thus, one can choose a basis of V consisting of common eigenvectors of f and g. □

Lemma 14.12 (Simultaneous Trigonalization)**.** *Let $f, g \in \mathrm{End}(V)$ be commuting and trigonalizable. Then f and g are* simultaneously *trigonalizable, i. e. there exists a basis B of V such that ${}_B[f]_B$ and ${}_B[g]_B$ are lower triangular matrices.*

Proof. Wlog. let $V \neq \{0\}$. According to Theorem 14.8, f has an eigenvalue $\lambda \in K$. As in the proof of Lemma 14.11, $U := E_\lambda(f)$ is a g-invariant subspace. The minimal polynomial of $g_{|U}$ divides μ_g and therefore splits into linear factors. Thus, there exists an eigenvector $u \in U$ of g. Now $W := \langle u \rangle$ is both f-invariant and g-invariant. Let $\overline{V} := V/W$, $\overline{f} \in \mathrm{End}(\overline{V})$ and $\overline{g} \in \mathrm{End}(\overline{V})$ as in the proof of Theorem 14.8. Because $\mu_{\overline{f}} \mid \mu_f$ and $\mu_{\overline{g}} \mid \mu_g$, one can inductively assume that $\overline{f}$ and $\overline{g}$ are simultaneously trigonalizable. The claim now follows as in the proof of Theorem 14.8. □

Example 14.13. In contrast to simultaneous diagonalization, the converse of Lemma 14.12 is false: $\mathrm{diag}(1,0)$ and $\left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$ are trivially trigonalizable, but do not commute.

Definition 14.14. Let $\dim V = n$ and $f \in \mathrm{End}(V)$ with eigenvalue $\lambda \in K$. One calls

$$H_\lambda(f) := \mathrm{Ker}\big((f - \lambda \,\mathrm{id}_V)^n\big) \leq V$$

the *generalized eigenspace* of f for λ. If λ is an eigenvalue of a matrix $A \in K^{n\times n}$, one defines analogously $H_\lambda(A) := \mathrm{Ker}\big((A - \lambda 1_n)^n\big)$.

Example 14.15. Let $f \in \mathrm{End}(K^2)$ with $A := [f] = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Then $E_1(f) = \langle e_1 \rangle$. Because of $(A - 1_2)^2 = 0$, $H_1(f) = K^2$, i. e. the generalized eigenspace is larger than the eigenspace.

Remark 14.16. For $v \in E_\lambda(f)$ it holds that

$$(f - \lambda \,\mathrm{id}_V)^n(v) = (f - \lambda \,\mathrm{id}_V)^{n-1}\big((f - \lambda \,\mathrm{id}_V)(v)\big) = (f - \lambda \,\mathrm{id}_V)^{n-1}(0) = 0.$$

This shows $E_\lambda(f) \subseteq H_\lambda(f)$. Obviously f commutes with $f - \lambda \,\mathrm{id}$ and $(f - \lambda \,\mathrm{id})^n$. For $v \in H_\lambda(f)$ it therefore holds that

$$(f - \lambda \,\mathrm{id}_V)^n(f(v)) = \big((f - \lambda \,\mathrm{id}_V)^n \circ f\big)(v) = f\big((f - \lambda \,\mathrm{id}_V)^n(v)\big) = f(0) = 0.$$

Thus $f(v) \in H_\lambda(f)$ and $H_\lambda(f)$ is f-invariant. We show next that $H_\lambda(f)$ possesses an f-invariant complement.

Lemma 14.17 (Fitting)**.** *Let* $\dim V = n$ *and* $f \in \mathrm{End}(V)$*. Then*

$$\boxed{V = f^n(V) \oplus \mathrm{Ker}(f^n)}$$

is a decomposition into f-invariant subspaces.

Proof. For $v \in \mathrm{Ker}(f^k)$ it holds that $f^{k+1}(v) = f(f^k(v)) = f(0) = 0$. This shows $\mathrm{Ker}(f) \leq \mathrm{Ker}(f^2) \leq \ldots$. Since the dimension of these subspaces is bounded by n, there exists a $k \leq n$ with $\mathrm{Ker}(f^k) = \mathrm{Ker}(f^{k+1})$. For $v \in \mathrm{Ker}(f^{k+2})$ it holds that $f^{k+1}(f(v)) = f^{k+2}(v) = 0$, so $f(v) \in \mathrm{Ker}(f^{k+1}) = \mathrm{Ker}(f^k)$. This shows $f^{k+1}(v) = 0$ and $v \in \mathrm{Ker}(f^{k+1})$. Inductively one obtains

$$\mathrm{Ker}(f^k) = \mathrm{Ker}(f^{k+1}) = \ldots = \mathrm{Ker}(f^n) = \mathrm{Ker}(f^{n+1}) = \ldots.$$

For $v \in \mathrm{Ker}(f^n) \cap f^n(V)$ there exists a $w \in V$ with $v = f^n(w)$. Because of $f^{2n}(w) = f^n(v) = 0$, $w \in \mathrm{Ker}(f^{2n}) = \mathrm{Ker}(f^n)$ and it follows that $v = f^n(w) = 0$. Thus $\mathrm{Ker}(f^n) \cap f^n(V) = \{0\}$. The homomorphism theorem shows $V = \mathrm{Ker}(f^n) \oplus f^n(V)$. Because of $f(\mathrm{Ker}(f^n)) \subseteq \mathrm{Ker}(f^{n-1}) \subseteq \mathrm{Ker}(f^n)$ and $f(f^n(V)) = f^n(f(V)) \subseteq f^n(V)$, both subspaces are f-invariant. □

Example 14.18. Let $\lambda, \lambda' \in K$ be distinct eigenvalues of $f \in \operatorname{End}(V)$. For $U := H_\lambda(f) \cap H_{\lambda'}(f)$ it holds that

$$(f - \lambda \operatorname{id})^n(U) = \{0\} = (f - \lambda' \operatorname{id})^n(U).$$

The minimal polynomial of the restriction $g := f_{|U} \in \operatorname{End}(U)$ therefore divides $(X - \lambda)^n$ and $(X - \lambda')^n$ according to Lemma 10.44. From Lemma 10.24 it follows that $\mu_g = 1$ and $U = \{0\}$.

Theorem 14.19 (Generalized Eigenspace Decomposition)**.** *Let $f \in \operatorname{End}(V)$ such that μ_f splits into linear factors. Then*

$$V = H_{\lambda_1}(f) \oplus \ldots \oplus H_{\lambda_k}(f)$$

for the distinct eigenvalues $\lambda_1, \ldots, \lambda_k \in K$ of f.

Proof. We argue by induction on $n := \dim V$. The case $n = 1$ is trivial. So let $n \geq 2$ and assume the claim is already proven for $n - 1$. Since μ_f splits into linear factors, f has an eigenvalue $\lambda = \lambda_1 \in K$ according to Theorem 10.50. For the map $g := f - \lambda \operatorname{id}_V$, it holds that

$$V = g^n(V) \oplus \operatorname{Ker}(g^n) = g^n(V) \oplus H_\lambda(f)$$

according to Fitting. For $U := g^n(V)$, we have $f(U) = (g + \lambda \operatorname{id})(U) \subseteq g(U) + U \subseteq U$, i. e. U is f-invariant. We now consider the restriction $h := f_{|U} \in \operatorname{End}(U)$. According to Lemma 10.44, $\mu_h \mid \mu_f$ and μ_h splits into linear factors (Lemma 10.24). Because $\dim H_\lambda(f) \geq \dim E_\lambda(f) \geq 1$, we have $\dim U < n$. By induction, it therefore holds that

$$U = H_{\lambda_2}(h) \oplus \ldots \oplus H_{\lambda_k}(h)$$

for the distinct eigenvalues $\lambda_2, \ldots, \lambda_k$ of h. We show $H_{\lambda_i}(h) = H_{\lambda_i}(f)$ for $i = 2, \ldots, k$. Because $E_{\lambda_i}(h) \subseteq E_{\lambda_i}(f)$, $\lambda_2, \ldots, \lambda_k$ are also eigenvalues of f. Because $E_{\lambda_i}(h) \cap H_\lambda(f) \leq U \cap H_\lambda(f) = \{0\}$, we have $\lambda \neq \lambda_i$ for $i = 2, \ldots, k$. Certainly $H_{\lambda_i}(h) \leq H_{\lambda_i}(f)$. Conversely, let $v \in H_{\lambda_i}(f)$ for some $i \geq 2$. Then there exist $u \in U$ and $w \in H_\lambda(f)$ with $v = u + w$. It follows

$$0 = (f - \lambda_i \operatorname{id})^n(v) = (f - \lambda_i \operatorname{id})^n(u) + (f - \lambda_i \operatorname{id})^n(w) \in U \oplus H_\lambda(f).$$

From Lemma 8.9 one obtains $(f - \lambda_i \operatorname{id})^n(u) = 0 = (f - \lambda_i \operatorname{id})^n(w)$. According to Example 14.18, $w \in H_\lambda(f) \cap H_{\lambda_i}(f) = \{0\}$ and $v = u \in H_{\lambda_i}(f) \cap U = H_{\lambda_i}(h)$. This shows

$$V = H_\lambda(f) \oplus U = H_\lambda(f) \oplus H_{\lambda_2}(h) \oplus \ldots \oplus H_{\lambda_k}(h) = H_{\lambda_1}(f) \oplus \ldots \oplus H_{\lambda_k}(f). \qquad \square$$

Remark 14.20. If μ_f splits into pairwise distinct linear factors, then f is diagonalizable (Theorem 10.52). The generalized eigenspace decomposition is then exactly the decomposition into eigenspaces, since $E_\lambda(f) \subseteq H_\lambda(f)$.

Example 14.21. Let

$$A := \begin{pmatrix} . & 1 & 1 & 1 \\ 1 & 1 & 1 & . \\ 1 & 1 & . & 1 \\ . & . & 1 & 1 \end{pmatrix} \in \mathbb{F}_2^{4\times 4}.$$

Since 0 and 1 are the only possible eigenvalues, we calculate as a trial

$$A^2 = \begin{pmatrix} . & . & . & . \\ . & 1 & . & . \\ 1 & . & 1 & . \\ 1 & 1 & 1 & . \end{pmatrix}, \qquad (A - 1_4)^2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & . & 1 & . \\ 1 & 1 & 1 & 1 \\ . & . & 1 & . \end{pmatrix}^2 = \begin{pmatrix} 1 & . & . & . \\ . & . & . & . \\ 1 & . & . & . \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

One sees $(0,0,0,1),(1,0,1,0) \in H_0(A)$ and $(0,1,1,0),(0,1,0,1) \in H_1(A)$. For dimensional reasons, it follows

$$\mathbb{F}_2^4 = H_0(A) \oplus H_1(A) = \langle (0,0,0,1),(1,0,1,0)\rangle \oplus \langle (0,1,1,0),(0,1,0,1)\rangle.$$

14.2 Jordan Blocks

Remark 14.22. Following the generalized eigenspace decomposition, we are interested in bases of generalized eigenspaces.

Definition 14.23. For $\lambda \in K$ and $n \geq 1$, one calls

$$J_n(\lambda) := \begin{pmatrix} \lambda & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & \lambda \end{pmatrix} \in K^{n\times n}$$

a *Jordan block* for the eigenvalue λ.[1]

Remark 14.24. Obviously, $A := J_n(\lambda)$ has the characteristic polynomial $\chi_A = (X-\lambda)^n$, i.e., λ has algebraic multiplicity n. On the other hand, λ has geometric multiplicity $\dim E_\lambda(A) = 1$. Thus, A is particularly far from being diagonalizable (Theorem 10.34). For the standard basis vector $e_1 \in K^n$, we have $Ae_1 = (*,1,0,\ldots,0)^{\mathrm{t}}$ and inductively

$$A^k e_1 = A(A^{k-1}e_1) = A(\underbrace{*,\ldots,*}_{k-1},1,0,\ldots,0)^{\mathrm{t}} = (\underbrace{*,\ldots,*}_{k},1,0,\ldots,0)^{\mathrm{t}}.$$

Thus $e_1, Ae_1, \ldots, A^{n-1}e_1$ are linearly independent. Therefore, the matrices $1_n, A, \ldots, A^{n-1}$ must also be linearly independent. This shows $\deg \mu_A \geq n$. From Cayley-Hamilton, it follows that $\mu_A = \chi_A = (X-\lambda)^n$.

Definition 14.25. One calls $f \in \mathrm{End}(V)$ (resp. $A \in K^{n\times n}$) *nilpotent*, if $f^m = f \circ \ldots \circ f = 0$ (resp. $A^m = 0_n$) holds for some $m \in \mathbb{N}$.

Example 14.26. Let A be a strict (upper or lower) triangular matrix. Then $\chi_A = X^n$. By Cayley-Hamilton, $\mu_A \mid X^n$ and therefore $A^n = 0_n$. Thus A is nilpotent. In particular, Jordan blocks for the eigenvalue 0 are nilpotent.

Remark 14.27. If $A \in K^{n\times n}$ is nilpotent, then $\mu_A \mid X^m$ for some $m \in \mathbb{N}$. Because $\deg \mu_A \leq n$, we have $\mu_A \mid X^n$ and $A^n = 0$. The following theorem provides a canonical system of representatives for the similarity classes of nilpotent matrices. The number of similarity classes is the number $p(n)$ of *partitions* of n, i.e., decompositions of the form $n = n_1 + \ldots + n_k$ with $n_1 \geq \ldots \geq n_k$. For example, $p(5) = 7$, because

$$5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1.$$

No simple formula for $p(n)$ is known.

[1] In some books, $J_n(\lambda)^{\mathrm{t}}$ is used. This makes no essential difference (Theorem 14.44).

Theorem 14.28. *Let $f \in \mathrm{End}(V)$ be nilpotent. Then there exists a basis B of V such that*

$$ {}_B[f]_B = \mathrm{diag}(J_{n_1}(0), \ldots, J_{n_s}(0)). $$

The numbers $n_1 \geq \ldots \geq n_s \geq 1$ are uniquely determined by the equations

$$ \boxed{\big|\{1 \leq i \leq s : n_i \geq k\}\big| = \mathrm{rk}(f^{k-1}) - \mathrm{rk}(f^k)} \qquad (k = 1, \ldots, m) \tag{14.1} $$

In particular, $s = \dim \mathrm{Ker}(f)$.

Proof. Let $m \in \mathbb{N}$ with $f^m = 0 \neq f^{m-1}$. For $k \in \mathbb{N}$, it holds that $f(\mathrm{Ker}(f^k)) \subseteq \mathrm{Ker}(f^{k-1})$. According to Corollary 4.16, there exist subspaces $U_1, \ldots, U_m$ with

$$ \begin{aligned} V = \mathrm{Ker}(f^m) &= \mathrm{Ker}(f^{m-1}) \oplus U_1, \\ \mathrm{Ker}(f^{m-1}) &= \big(\mathrm{Ker}(f^{m-2}) + f(U_1)\big) \oplus U_2, \\ &\vdots \\ \mathrm{Ker}(f) &= \big(f^{m-1}(U_1) + \ldots + f(U_{m-1})\big) \oplus U_m. \end{aligned} $$

We show that all sums are direct. This is given for the first sum. Suppose inductively it has already been shown:

$$ \mathrm{Ker}(f^{m-k+1}) = \mathrm{Ker}(f^{m-k}) \oplus f^{k-1}(U_1) \oplus f^{k-2}(U_2) \oplus \ldots \oplus f(U_{k-1}) \oplus U_k. \tag{14.2} $$

Let $w + f^k(u_1) + \ldots + f(u_k) = 0$ with $w \in \mathrm{Ker}(f^{m-k-1})$ and $u_i \in U_i$ for $i = 1, \ldots, k$. Then it follows

$$ \begin{aligned} f^{m-k}\big(f^{k-1}(u_1) + \ldots + u_k\big) &= f^{m-k-1}\big(f^k(u_1) + \ldots + f(u_k)\big) \\ &= f^{m-k-1}\big(w + f^k(u_1) + \ldots + f(u_k)\big) = f^{m-k-1}(0) = 0 \end{aligned} $$

and

$$ f^{k-1}(u_1) + \ldots + u_k \in \mathrm{Ker}(f^{m-k}) \cap \big(f^{k-1}(U_1) \oplus \ldots \oplus f(U_{k-1}) \oplus U_k\big) \overset{(14.2)}{=} \{0\}. $$

This shows $f^{k-1}(u_1) = \ldots = u_k = 0$ (Lemma 8.9) and it follows $w = 0$. Thus

$$ \mathrm{Ker}(f^{m-k}) = \mathrm{Ker}(f^{m-k-1}) \oplus f^k(U_1) \oplus f^{k-1}(U_2) \oplus \ldots \oplus f(U_k) \oplus U_{k+1} $$

as desired.

Overall,

$$ V = U_1 \oplus f(U_1) \oplus \ldots \oplus f^{m-1}(U_1) \oplus U_2 \oplus f(U_2) \oplus \ldots \oplus f^{m-2}(U_2) \oplus \ldots \oplus U_m. $$

Let $b_{i1}, \ldots, b_{ik_i}$ be a basis of U_i for $i = 1, \ldots, m$ (the case $U_i = \{0\}$ with $k_i = 0$ is allowed). Because $\mathrm{Ker}(f^j) \cap U_i = \{0\}$, the restriction $f^j_{|U_i}$ is injective for $j = 1, \ldots, m - i$ (Lemma 7.7). In particular, $f^j(b_{i1}), \ldots, f^j(b_{ik_i})$ is a basis of $f^j(U_i)$. Thus

$$ B := \bigcup_{i=1}^{m} \bigcup_{j=1}^{k_i} \big\{b_{ij}, f(b_{ij}), \ldots, f^{m-i}(b_{ij})\big\} $$

is a basis of V. Because $U_i \subseteq \mathrm{Ker}(f^{m-i+1})$, it holds that $f^{m-i+1}(b_{ij}) = 0$. Thus the elements $b_{ij}, f(b_{ij}), \ldots, f^{m-i}(b_{ij})$ correspond to the Jordan block $J_{m-i+1}(0)$ in ${}_B[f]_B$. Overall, ${}_B[f]_B$ now has the desired form. Furthermore,

$$ \begin{aligned} k_1 + \ldots + k_l &= \dim\big(f^{l-1}(U_1) \oplus f^{l-2}(U_2) \oplus \ldots \oplus U_l\big) = \dim \mathrm{Ker}(f^{m-l+1}) - \dim \mathrm{Ker}(f^{m-l}) \\ &\overset{7.12}{=} \mathrm{rk}(f^{m-l}) - \mathrm{rk}(f^{m-l+1}) \end{aligned} $$

is the number of Jordan blocks $J_{n_i}(0)$ with $n_i \geq m - l + 1$. By replacing $m - l + 1$ with k, (14.1) follows. The last claim is obtained with $k = 1$ in (14.1). □

Remark 14.29. In the situation of Theorem 14.28, it holds that

$$\begin{aligned}\big|\{1 \le i \le s : n_i = k\}\big| &= \big|\{1 \le i \le s : n_i \ge k\}\big| - \big|\{1 \le i \le s : n_i \ge k+1\}\big| \\ &= \operatorname{rk}(f^{k-1}) + \operatorname{rk}(f^{k+1}) - 2\operatorname{rk}(f^k)\end{aligned}$$

for $k = 1, \dots, n$. In particular, $2\operatorname{rk}(f^k) \le \operatorname{rk}(f^{k+1}) + \operatorname{rk}(f^{k-1})$, i.e., the sequence $\operatorname{rk}(f)$, $\operatorname{rk}(f^2), \dots$ cannot fall too "fast".

Example 14.30. Let $f \in \operatorname{End}(\mathbb{R}^5)$ with

$$A := {}_B[f]_B = \begin{pmatrix} 2 & -3 & -1 & -1 & 2 \\ 1 & -2 & 0 & 0 & 1 \\ 0 & 2 & -1 & -1 & 0 \\ 1 & -4 & 1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 0 \end{pmatrix}.$$

One calculates

$$A^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 & 1 \\ -1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad A^3 = 0.$$

Thus $\operatorname{rk}(f^2) = 1$. As in the proof of Theorem 14.28, we can choose $b_1 := e_1 \notin \operatorname{Ker}(f^2)$ and $U_1 := \langle b_1 \rangle$ with $\mathbb{R}^5 = \operatorname{Ker}(f^2) \oplus \langle b_1 \rangle$. Furthermore,

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

$$\operatorname{Ker}(f) \oplus f(U_1) = \left\langle \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle.$$

With $b_2 := e_3 \in \operatorname{Ker}(f^2) \setminus \big(\operatorname{Ker}(f) \oplus f(U_1)\big)$ and $U_2 := \langle b_2 \rangle$, it holds that $\operatorname{Ker}(f^2) = \operatorname{Ker}(f) \oplus f(U_1) \oplus U_2$. Finally, $\operatorname{Ker}(f) = f^2(U_1) \oplus f(U_2)$ (thus $U_3 := \{0\}$). wrt. the basis $B := \{b_1, f(b_1), f^2(b_1), b_2, f(b_2)\}$, f has the representation matrix $\operatorname{diag}\big(J_3(0), J_2(0)\big)$.

Theorem 14.31. *Let $f \in \operatorname{End}(V)$ such that μ_f splits into linear factors. Then there exists a basis B of V with*

$${}_B[f]_B = \operatorname{diag}(J_{n_1}(\lambda_1), \dots, J_{n_s}(\lambda_s)), \qquad \text{(JORDAN normal form)}$$

where $\lambda_1, \dots, \lambda_s \in K$ are the eigenvalues of f (possibly with multiplicities). The Jordan blocks $J_{n_i}(\lambda_i)$ are uniquely determined up to their order.

Proof. We put all the puzzle pieces together: Let $\lambda_1, \ldots, \lambda_k \in K$ be the distinct eigenvalues of f and $H_i := H_{\lambda_i}(f)$ for $i = 1, \ldots, k$. According to the generalized eigenspace decomposition, $V = H_1 \oplus \ldots \oplus H_k$ holds. For $g_i := (f - \lambda_i \operatorname{id}_V)_{|H_i}$, we have $g_i^n = 0$. By Theorem 14.28, there exists a basis B_i of H_i with

$$ {}_{B_i}[g_i]_{B_i} = \operatorname{diag}\big(J_{m_1}(0), \ldots, J_{m_t}(0)\big). $$

By Theorem 7.18, it follows that

$$ \begin{aligned} {}_{B_i}[f_{|H_i}]_{B_i} &= {}_{B_i}[g_i + \lambda_i \operatorname{id}_{H_i}]_{B_i} = {}_{B_i}[g_i]_{B_i} + \lambda_i {}_{B_i}[\operatorname{id}_{H_i}]_{B_i} = \operatorname{diag}\big(J_{m_1}(0), \ldots, J_{m_t}(0)\big) + \lambda_i 1 \\ &= \operatorname{diag}\big(J_{m_1}(\lambda_i), \ldots, J_{m_t}(\lambda_i)\big). \end{aligned} $$

Thus $B := B_1 \cup \ldots \cup B_k$ is a suitable basis. For a fixed pair (n_i, λ_i), the number of blocks $J_{n_i}(\lambda_i)$ is obtained from Remark 14.29 applied to g_i. □

Corollary 14.32. *Every matrix $A \in \mathbb{C}^{n \times n}$ possesses a Jordan normal form (more precisely: A is similar to a Jordan normal form).*

Proof. According to Corollary 11.35, μ_A splits into linear factors. □

Remark 14.33.

(a) Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ cannot generally be ordered in a meaningful way (cf. Remark 11.26), there is, in contrast to Theorem 14.28, no canonical order of the Jordan blocks. However, one can define the *lexicographical* order

$$ x <_l y \;:\Longleftrightarrow\; \operatorname{Re}(x) < \operatorname{Re}(y) \vee \big(\operatorname{Re}(x) = \operatorname{Re}(y) \wedge \operatorname{Im}(x) < \operatorname{Im}(y)\big) $$

on $\mathbb{C}$ and sort the Jordan blocks first by size and then by eigenvalue wrt. $<_l$. In the following, we speak somewhat loosely of *the* Jordan normal form of f.

(b) Let $A \in K^{n \times n}$. In Theorem 15.39, we construct a field $L \supseteq K$ such that μ_A splits into linear factors in $L[X]$. One can then regard A as a matrix in $L^{n \times n}$ and determine the Jordan normal form there.

Example 14.34. Let $f \in \operatorname{End}(\mathbb{C}^3)$ with

$$ A := [f] = \begin{pmatrix} 5 & 0 & 1 \\ -5 - \mathrm{i} & -\mathrm{i} & -1 \\ -9 & 0 & -1 \end{pmatrix}. $$

By Laplace expansion along the second column, one obtains

$$ \chi_f = \chi_A = (X + \mathrm{i})\big((X - 5)(X + 1) + 9\big) = (X + \mathrm{i})(X^2 - 4X + 4) = (X + \mathrm{i})(X - 2)^2. $$

Thus $\lambda_1 = 2$ and $\lambda_2 = -\mathrm{i}$ are the eigenvalues of f. Obviously $b_3 := e_2$ is an eigenvector for λ_2. Because of

$$ A - 2 \cdot 1_3 = \begin{pmatrix} 3 & 0 & 1 \\ -5 - \mathrm{i} & -2 - \mathrm{i} & -1 \\ -9 & 0 & -3 \end{pmatrix} \sim \begin{pmatrix} 3 & 0 & 1 \\ -5 - \mathrm{i} & -2 - \mathrm{i} & -1 \\ 0 & 0 & 0 \end{pmatrix} $$

we have $\operatorname{rk}(f - 2\,\mathrm{id}) = 2$, i.e., $\lambda_1 = 2$ has geometric multiplicity 1. Thus f is not diagonalizable and the Jordan normal form of f must be $\operatorname{diag}\big(J_2(2), J_1(-\mathrm{i})\big)$. Because of

$$(A - 2 \cdot 1_3)^2 = \begin{pmatrix} 0 & 0 & 0 \\ 3+4\mathrm{i} & 3+4\mathrm{i} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

we can choose $b_1 := e_3 \in \operatorname{Ker}\big((f - 2\,\mathrm{id})^2\big) \setminus \operatorname{Ker}(f - 2\,\mathrm{id})$ as in Example 14.30. For

$$b_2 := (f - 2\,\mathrm{id})(b_1) = (1, -1, -3)$$

it holds that $f(b_1) = 2b_1 + b_2$ and $f(b_2) = (f - 2\,\mathrm{id})(b_2) + 2b_2 = 2b_2$. For the basis $B := \{b_1, b_2, b_3\}$, one obtains

$${}_B[f]_B = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & -\mathrm{i} \end{pmatrix} = \operatorname{diag}\big(J_2(2), J_1(-\mathrm{i})\big).$$

14.3 Applications

Theorem 14.35. *Let $J := \operatorname{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$ be the Jordan normal form of $A \in K^{n\times n}$. Let $\rho_1, \ldots, \rho_k$ be the distinct eigenvalues of A. For $i = 1, \ldots, k$, let a_i, m_i and s_i be the number, the maximum and the sum of the n_j with $\lambda_j = \rho_i$. Then a_i is the geometric multiplicity of ρ_i and*

$$\chi_A = (X - \rho_1)^{s_1} \ldots (X - \rho_k)^{s_k} \qquad \mu_A = (X - \rho_1)^{m_1} \ldots (X - \rho_k)^{m_k}.$$

In particular, A is diagonalizable if and only if J is a diagonal matrix.

Proof. One easily sees $\operatorname{rk}(J - \mu_i 1_n) = n - a_i$, i.e., a_i is the geometric multiplicity of ρ_i (cf. Theorem 14.28). Since J is a lower triangular matrix, s_i is the algebraic multiplicity of ρ_i. Since μ_A does not depend on the choice of basis, it holds that

$$0 = \mu_A(J) = \operatorname{diag}\big(\mu_A(J_{n_1}(\lambda_1)), \ldots, \mu_A(J_{n_s}(\lambda_s))\big).$$

From Remark 14.24 it follows that $\mu_A = (X - \rho_1)^{m_1} \ldots (X - \rho_k)^{m_k}$. The second assertion holds by Theorem 10.52. □

Example 14.36. For $A = \operatorname{diag}(J_3(1), J_2(1), J_4(\mathrm{i}))$, it holds that $\chi_A = (X - 1)^5(X - \mathrm{i})^4$ and $\mu_A = (X - 1)^3(X - \mathrm{i})^4$.

Corollary 14.37. *Let $A, B \in \mathbb{C}^{n\times n}$. In the case $n \leq 3$, it holds that*

$$A \approx B \iff \chi_A = \chi_B,\ \mu_A = \mu_B.$$

In the case $n \leq 6$, the following statements are equivalent:

(1) $A \approx B$.

(2) $\chi_A = \chi_B$, $\mu_A = \mu_B$ *and the geometric multiplicities of the eigenvalues of A coincide with those of B.*

Proof. As is well known, $A \approx B$ implies the respective other statement. Now let $n \leq 3$, $\chi_A = \chi_B$ and $\mu_A = \mu_B$. Then A and B have the same eigenvalues with the same algebraic multiplicities. Let λ be an eigenvalue with algebraic multiplicity $r \leq 3$. For the corresponding sizes of the Jordan blocks $n_1 \geq \ldots \geq n_k$ of A, it holds that $r = n_1 + \ldots + n_k$. Since n_1 is determined by μ_A, k and $n_2, \ldots, n_k$ are also uniquely determined by r. Thus A and B have the same Jordan normal form. It follows that $A \approx B$.

Now let $n \leq 6$ and (2) be satisfied. Besides r and n_1, the geometric multiplicity k of λ is now also uniquely determined. There are the following possibilities:

- $k = 1$: Here $n_1 = r$.
- $k = 2$: Here $n_2 = r - n_1$.
- $k = 3$: Because $r \leq 6$, $(n_1, n_2, n_3) \in \{(1,1,1), (2,1,1), (2,2,1), (2,2,2), (3,1,1), (3,2,1), (4,1,1)\}$. These triples are uniquely determined by r and n_1.
- $k = 4$: Here $(n_1, n_2, n_3, n_4) \in \{(1,1,1,1), (2,1,1,1), (2,2,1,1), (3,1,1,1)\}$.
- $k = 5$: Here $n_1 = r - 4$ and $n_2 = \ldots = n_5 = 1$.
- $k = 6$: Here $n_1 = \ldots = n_6 = 1$.

In all cases, $A \approx B$ follows. □

Example 14.38. The matrices

$$\operatorname{diag}(J_2(0), J_2(0)) \not\approx \operatorname{diag}(J_2(0), J_1(0), J_1(0)),$$
$$\operatorname{diag}(J_3(0), J_2(0), J_2(0)) \not\approx \operatorname{diag}(J_3(0), J_3(0), J_1(0))$$

show that Corollary 14.37 cannot be extended to $n = 4$ or $n = 7$.

Theorem 14.39. *A matrix $A \in \mathbb{C}^{n \times n}$ is nilpotent if and only if $\operatorname{tr}(A^k) = 0$ for $k = 1, \ldots, n$.*

Proof. Wlog. let A be in Jordan normal form with pairwise distinct eigenvalues $\lambda_1, \ldots, \lambda_s \in \mathbb{C}$. Let m_i be the algebraic multiplicity of λ_i. The eigenvalues of A^k are then $\lambda_1^k, \ldots, \lambda_s^k$ with the corresponding multiplicities. According to Remark 10.35, it holds that

$$\operatorname{tr}(A^k) = m_1\lambda_1^k + \ldots + m_s\lambda_s^k = 0.$$

If A is nilpotent, then $s = 1$, $\lambda_1 = 0$ and $\operatorname{tr}(A^k) = 0$ for $k = 1, \ldots, n$.

Conversely, let $\operatorname{tr}(A^k) = m_1\lambda_1^k + \ldots + m_s\lambda_s^k = 0$ for $k = 1, \ldots, n$. Suppose indirectly that A is not nilpotent. Wlog. let $\lambda_i \neq 0$ for $i = 1, \ldots, s$. Then $(m_1, \ldots, m_s)^{\mathrm{t}}$ is a solution to the homogeneous linear system of equations with coefficient matrix $V = (\lambda_j^i) \in K^{s \times s}$. However, according to Vandermonde, $\det(V) = \lambda_1 \ldots \lambda_s \det(\lambda_j^{i-1}) \neq 0$. Contradiction. □

Remark 14.40. Theorem 14.39 does not hold over arbitrary fields. For example, $\operatorname{tr}(1_2^k) = \operatorname{tr}(1_2) = 0$ in $\mathbb{F}_2$ for all $k \in \mathbb{N}$.

Lemma 14.41. *For $\lambda \in \mathbb{C}$, $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$ it holds that*

$$J_n(\lambda)^k = \begin{pmatrix} \lambda^k & & & & 0 \\ k\lambda^{k-1} & \ddots & & & \\ \binom{k}{2}\lambda^{k-2} & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ \binom{k}{n-1}\lambda^{k-n+1} & \cdots & \binom{k}{2}\lambda^{k-2} & k\lambda^{k-1} & \lambda^k \end{pmatrix},$$

where $\binom{k}{l} = \frac{k(k-1)\ldots(k-l+1)}{l!}$ for $l = 1, \ldots, n-1$.

Proof. In the case $\lambda = 0$ one easily sees:

$$J_n(0)^2 = \begin{pmatrix} 0 & & & & 0 \\ 0 & \ddots & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 0 & 0 \end{pmatrix}, \ldots, J_n(0)^{n-1} = \begin{pmatrix} 0 & & & 0 \\ \vdots & \ddots & & \\ 0 & & \ddots & \\ 1 & 0 & \cdots & 0 \end{pmatrix}, J_n(0)^n = 0.$$

The general case follows from the binomial formula[2]

$$J_n(\lambda)^k = (\lambda 1_n + J_n(0))^k = \sum_{l=0}^{k} \binom{k}{l} \lambda^{k-l} J_n(0)^k. \qquad \square$$

Theorem 14.42. *For $k \in \mathbb{N}$ and $A \in \mathrm{GL}(n, \mathbb{C})$ there exists a $W \in \mathbb{C}^{n \times n}$ with $W^k = A$.*

Proof. Because of $(SWS^{-1})^k = SW^kS^{-1}$ for $S \in \mathrm{GL}(n, \mathbb{C})$, we can assume that A is a Jordan normal form. It suffices to construct a k-th root for each Jordan block. So let $A = J_n(\lambda)$ with $\lambda \in \mathbb{C}$. Since A is invertible, $\lambda \neq 0$ holds. According to Lemma 11.27, there exists a $\mu \in \mathbb{C}$ with $\mu^k = \lambda$. Let $J := J_n(\mu)$. According to Lemma 14.41, J^k has entries $k\mu^{k-1} \neq 0$ directly below the main diagonal. Thus $\mu^k = \lambda$ is the only eigenvalue of J^k and the geometric multiplicity is 1. According to Theorem 14.35, A is the Jordan normal form of J^k. In particular, $A \approx J^k$. $\square$

Example 14.43.

(a) The proof shows $J_n(1)^k \approx J_n(1)$ for all $k, n \in \mathbb{N}$. We are looking for a third root of $J_3(1)$. For

$$S := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 9 \end{pmatrix}$$

it holds that

$$S^{-1}J_3(1)^3S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & -1/9 & 1/9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 9 \end{pmatrix} = J_3(1).$$

[2]The formula is applicable because 1_n and $J_n(0)$ commute.

For

$$W := S^{-1}J_3(1)S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & -1/9 & 1/9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ -1/9 & 1/3 & 1 \end{pmatrix}$$

it holds that $W^3 = J_3(1)$. In Theorem 15.35, we determine the number of k-th roots of $J_n(\lambda)$ for all $\lambda \in \mathbb{C}$ and $n, k \in \mathbb{N}$.

(b) The nilpotent matrix $J_2(0)$ has no square root, because for $A \in K^{2\times 2}$ with $A^2 = J_2(0)$, it would follow that $A^4 = 0$ and thus $A^2 = 0$ (Remark 14.27).

(c) One can easily verify that $J_2(1) \in \mathrm{GL}(2, \mathbb{F}_2)$ also has no square root.

Theorem 14.44. *For all $A \in \mathbb{C}^{n\times n}$, $A \approx A^{\mathrm{t}}$ holds.*

Proof. As usual, we can assume $A = J_n(\lambda)$. Since λ is the only eigenvalue of A^{t} and the geometric multiplicity is 1, A is the Jordan normal form of A^{t}. □

Remark 14.45. One can also verify Theorem 14.44 directly:

$$\begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix}^{(-1)} J_n(\lambda) \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} \begin{pmatrix} 0 & & & \lambda \\ & & \iddots & 1 \\ & \iddots & \iddots & \\ \lambda & 1 & & 0 \end{pmatrix} = J_n(\lambda)^{\mathrm{t}}.$$

In Theorem 15.25, we prove Theorem 14.44 over arbitrary fields.

15 The Frobenius Normal Form

15.1 Irreducible Polynomials

Remark 15.1. In this chapter, we construct a normal form for endomorphisms that, in contrast to the Jordan normal form, exists over every field. Since in general not every polynomial splits into linear factors, one cannot expect the normal form to be a triangular matrix (Theorem 14.8). Nevertheless, we achieve the same number of zero entries as in the Jordan normal form. The basic idea is to factorize the minimal polynomial into "small" polynomials as much as possible. As always, let V be a finite-dimensional K-vector space.

Definition 15.2.

- A monic polynomial $\alpha \in K[X] \setminus K$ is called *irreducible*, if there is no factorization $\alpha = \beta\gamma$ with $\beta, \gamma \in K[X] \setminus K$ (cf. prime number).
- We call $\alpha, \beta \in K[X]$ *coprime*, if every common divisor of α and β is constant, i.e., lies in K.

Example 15.3.

(a) Monic polynomials of degree 1 are irreducible, because $\deg(\alpha\beta) = \deg(\alpha)+\deg(\beta)$ for $\alpha, \beta \in K[X]$. In $\mathbb{C}[X]$, conversely, every irreducible polynomial has degree 1 according to the Fundamental Theorem of Algebra and Lemma 10.22. In general, monic polynomials of degree ≤ 3 are irreducible if and only if they have no root. For example, $X^2 + X + 1 \in \mathbb{F}_2[X]$ is irreducible.

(b) The polynomial X^2-2 is irreducible in $\mathbb{Q}[X]$, but not in $\mathbb{R}[X]$, because $X^2-2 = (X+\sqrt{2})(X-\sqrt{2})$. Similarly, $X^2 + 1$ is irreducible in $\mathbb{R}[X]$, but not in $\mathbb{C}[X]$ due to $X^2 + 1 = (X + \mathrm{i})(X - \mathrm{i})$.

(c) Two different irreducible polynomials are coprime. If $\gamma \in K[X] \setminus K$ is a common divisor of α and β, then every irreducible divisor of γ is also a common divisor of α and β. Thus, α and β are coprime if and only if they have no irreducible common divisors.

Theorem 15.4 (Euclidean Algorithm)**.** *Let $\alpha, \beta \in K[X]$ with $\deg \alpha \geq \deg \beta$. The following algorithm determines whether α and β are coprime:*

(1) Set $\alpha_0 := \alpha$, $\alpha_1 := \beta$ and $i := 1$.

(2) Repeat:

- *Euclidean division: $\alpha_{i-1} = \alpha_i\gamma_i + \alpha_{i+1}$ with $\deg(\alpha_{i+1}) < \deg(\alpha_i)$.*
- *If $\alpha_{i+1} = 0$, then terminate.*
- *Increase i by 1.*

(3) α and β are coprime if and only if $\alpha_i \in K$ holds.

Proof. Every common divisor of α_{i-1} and α_i also divides $\alpha_{i-1} - \alpha_i\gamma_i = \alpha_{i+1}$ for $i = 1, 2, \ldots$. Since $\deg(\alpha_i)$ decreases with each Euclidean division, the algorithm must terminate after finitely many steps. At the end, $\alpha_{i+1} = 0 \neq \alpha_i$, i. e. $\alpha_{i-1} = \alpha_i\gamma_i$. Therefore, α and β are coprime if and only if α_i is constant. □

Example 15.5. Let $\alpha = X^5 + X^4 + X^3 + 1 \in \mathbb{F}_2[X]$ and $\beta = X^4 + X^3 + X + 1 \in \mathbb{F}_2[X]$. We have

$$\begin{aligned} X^5 + X^4 + X^3 + 1 &= (X^4 + X^3 + X + 1)X + X^3 + X^2 + X + 1 && \Longrightarrow \alpha_2 = X^3 + X^2 + X + 1, \\ X^4 + X^3 + X + 1 &= (X^3 + X^2 + X + 1)X + X^2 + 1 && \Longrightarrow \alpha_3 = X^2 + 1 \notin \mathbb{F}_2, \\ X^3 + X^2 + X + 1 &= (X^2 + 1)(X + 1) && \Longrightarrow \alpha_4 = 0. \end{aligned}$$

Thus $X^2 + 1$ is a common divisor of α and β.

Lemma 15.6 (Bézout)**.** *If $\alpha, \beta \in K[X]$ are coprime, then there exist $\tilde{\alpha}, \tilde{\beta} \in K[X]$ with $\alpha\tilde{\alpha} + \beta\tilde{\beta} = 1$.*

Proof. By assumption, α and β are not both 0. Thus there exist $\tilde{\alpha}, \tilde{\beta} \in K[X]$ such that $\rho := \alpha\tilde{\alpha} + \beta\tilde{\beta} \neq 0$ has minimal degree. Euclidean division yields $\gamma, \delta \in K[X]$ with $\alpha = \gamma\rho + \delta$ and $\deg\delta < \deg\rho$. Thus

$$\delta = \alpha - \gamma\rho = \alpha - \gamma\alpha\tilde{\alpha} - \gamma\beta\tilde{\beta} = \alpha(1 - \gamma\tilde{\alpha}) - \beta(\gamma\tilde{\beta}).$$

The choice of ρ shows $\delta = 0$. It follows that $\rho \mid \alpha$. Analogously, we obtain $\rho \mid \beta$. Since α and β are coprime, $\rho \in K^\times$. After normalization, one can assume $\rho = 1$. □

Remark 15.7. The polynomials $\tilde{\alpha}$ and $\tilde{\beta}$ in Lemma 15.6 can be calculated by reading the Euclidean algorithm backwards

$$\alpha_i = \alpha_{i-2} - \alpha_{i-1}\gamma_{i-1} = \alpha_{i-2} - (\alpha_{i-3} - \alpha_{i-2}\gamma_{i-2})\gamma_{i-1} = \alpha_{i-3}\gamma_{i-2} + \alpha_{i-2}(1 - \gamma_{i-2}\gamma_{i-1}) = \ldots$$

and dividing by α_i.

Example 15.8. Let $\alpha := X^3 + X^2 + 1$ and $\beta := X^2 - X$. The Euclidean algorithm yields

$$\begin{aligned} X^3 + X^2 + 1 &= (X^2 - X)(X + 2) + 2X + 1 \\ X^2 - X &= (2X + 1)\frac{1}{4}(2X - 3) + \frac{3}{4}. \end{aligned}$$

Thus

$$\begin{aligned} \frac{3}{4} &= \beta - \frac{1}{4}(2X + 1)(2X - 3) = \beta - \frac{1}{4}\Big(\alpha - \beta(X + 2)\Big)(2X - 3) \\ &= -\frac{1}{4}(2X - 3)\alpha + \frac{1}{4}\Big(4 + (X + 2)(2X - 3)\Big)\beta \end{aligned}$$

and

$$-\frac{1}{3}(2X - 3)\alpha + \frac{1}{3}(2X^2 + X - 2)\beta = 1.$$

Theorem 15.9 (Prime factorization in $K[X]$)**.** *Every monic polynomial in $K[X] \setminus K$ is a product of irreducible factors, which are uniquely determined up to their order.*

Proof. Let $\alpha \in K[X]$ be monic. Induction on $d := \deg(\alpha)$. If α is irreducible (for example $d = 1$), then we are done. Otherwise $\alpha = \beta\gamma$ with $\deg(\beta), \deg(\gamma) < d$. We can assume that β and γ are monic. By induction, β and γ are products of irreducible factors and therefore so is α.

Now let $\alpha = \sigma_1 \dots \sigma_n = \tau_1 \dots \tau_m$ with irreducible polynomials $\sigma_1, \dots, \sigma_n, \tau_1, \dots, \tau_m \in K[X]$. Induction on m. In the case $m = 1$, $n = 1$ and $\sigma_1 = \tau_1$. Now let $m \geq 2$. In the case $\sigma_1 \neq \tau_1$, σ_1 and τ_1 are coprime. By Bézout there exist $\tilde{\sigma}, \tilde{\tau} \in K[X]$ with $\sigma_1\tilde{\sigma} + \tau_1\tilde{\tau} = 1$. It follows

$$\sigma_1(\tilde{\sigma}\tau_2 \dots \tau_m + \sigma_2 \dots \sigma_n\tilde{\tau}) = \tau_2 \dots \tau_m.$$

Inductively one obtains $\sigma_1 = \tau_i$ for some $i \in \{1, \dots, m\}$. Then $\sigma_2 \dots \sigma_n = \tau_1 \dots \tau_{i-1}\tau_{i+1} \dots \tau_m$ and induction yields $\{\sigma_1, \dots, \sigma_n\} = \{\tau_1, \dots, \tau_m\}$. □

Example 15.10. As with natural numbers, we will group identical irreducible factors into powers in the prime factorization of polynomials. For example,

$$X^6 + X^5 - X^4 - X^3 = X^3(X + 1)^2(X - 1).$$

From an algorithmic point of view, determining the prime factorization (of numbers as well as polynomials) is a very difficult problem.

Corollary 15.11. *Let $\alpha, \beta \in K[X]$ and γ be an irreducible divisor of $\alpha\beta$. Then $\gamma \mid \alpha$ or $\gamma \mid \beta$ holds.*

Proof. Because $\gamma \mid \alpha\beta$, γ must appear in the prime factorization of $\alpha\beta$. This prime factorization is obtained by combining the prime factorizations of α and β. Thus γ must appear in the factorization of α or β. □

Theorem 15.12 (Primary Decomposition)**.** *Let $f \in \mathrm{End}(V)$ and $\mu_f = \gamma_1^{a_1} \dots \gamma_k^{a_k}$ be the prime factorization of μ_f in $K[X]$. Then*

$$V = \mathrm{Ker}\big(\gamma_1^{a_1}(f)\big) \oplus \dots \oplus \mathrm{Ker}\big(\gamma_k^{a_k}(f)\big)$$

is a decomposition into f-invariant subspaces. Furthermore, $\gamma_i^{a_i}$ is the minimal polynomial of the restriction of f to $\mathrm{Ker}\big(\gamma_i^{a_i}(f)\big)$ for $i = 1, \dots, k$.

Proof. For $1 \leq i \leq k$ let $V_i := \mathrm{Ker}(\gamma_i^{a_i}(f))$. For $v \in V_i$ we have

$$\gamma_i^{a_i}(f)(f(v)) = f(\gamma_i^{a_i}(f)(v)) = f(0) = 0,$$

i. e. $f(v) \in V_i$. Thus V_i is f-invariant. To prove the direct decomposition, we argue by induction on k. For $k = 1$ we have $V_1 = \mathrm{Ker}(\mu_f(f)) = \mathrm{Ker}(0) = V$. So let $k \geq 2$. The polynomials $\alpha := \gamma_1^{a_1}$ and $\beta := \gamma_2^{a_2} \dots \gamma_k^{a_k}$ are coprime, since they have no irreducible common divisor. By Bézout there exist $\tilde{\alpha}, \tilde{\beta} \in K[X]$ with $\alpha\tilde{\alpha} + \beta\tilde{\beta} = 1$. Let $V_\beta := \mathrm{Ker}(\beta(f))$. Because $(\alpha\beta)(f) = 0 = (\beta\alpha)(f)$, it follows that $\beta(f)(V) \subseteq V_1$ and $\alpha(f)(V) \subseteq V_\beta$. It follows that

$$V = \mathrm{id}(V) = (\alpha\tilde{\alpha} + \beta\tilde{\beta})(f)(V) \subseteq \alpha(f)(V) + \beta(f)(V) \subseteq V_\beta + V_1.$$

For $v \in V_1 \cap V_\beta$, on the other hand,

$$v = (\alpha\tilde{\alpha} + \beta\tilde{\beta})(f)(v) = \tilde{\alpha}(f)\big(\alpha(f)(v)\big) + \tilde{\beta}(f)\big(\beta(f)(v)\big) = 0.$$

This shows $V = V_1 \oplus V_\beta$.

The minimal polynomial α_1 of the restriction $f_{|V_1}$ divides α, because $\alpha(f)(V_1) = 0$. In particular, $\deg(\alpha_1) \leq \deg(\alpha)$. Likewise, β is divisible by the minimal polynomial β_1 of $f_\beta := f_{|V_\beta}$. Because $V = V_1 \oplus V_\beta$, on the other hand, $(\alpha_1\beta_1)(f) = 0$ and $\mu_f \mid \alpha_1\beta_1$. From

$$\deg(\alpha) + \deg(\beta) = \deg(\mu_f) \leq \deg(\alpha_1\beta_1) = \deg(\alpha_1) + \deg(\beta_1)$$

it follows that $\alpha_1 = \alpha$ and $\beta_1 = \beta$. We can now apply the induction hypothesis to $f_\beta \in \mathrm{End}(V_\beta)$. For $i = 2, \ldots, k$ we have $\mathrm{Ker}(\gamma_i^{a_i}(f_\beta)) = V_i \cap V_\beta = V_i$. This shows

$$V_\beta = V_2 \oplus \ldots \oplus V_k.$$

The claim follows from Lemma 8.9. □

Remark 15.13. Suppose that μ_f splits into linear factors, i. e. $\gamma_i = X - \lambda_i$ for $i = 1, \ldots, k$, where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of f. Because

$$\mathrm{Ker}(\gamma_i^{a_i}(f)) = \mathrm{Ker}\big((f - \lambda_i \,\mathrm{id})^{a_i}\big) \subseteq \mathrm{Ker}\big((f - \lambda_i \,\mathrm{id})^{n}\big) = H_{\lambda_i}(f)$$

the primary decomposition coincides with the generalized eigenspace decomposition from Theorem 14.19.

Definition 15.14. For $v \in V$ and $f \in \mathrm{End}(V)$, one calls $U := \langle f^i(v) : i \in \mathbb{N}_0 \rangle \leq V$ a *cyclic* subspace. Obviously U is f-invariant. We denote the minimal polynomial of $f_{|U}$ by μ_v.

Remark 15.15. Let $f \in \mathrm{End}(V)$ and $v \in V \setminus \{0\}$. Let $d \in \mathbb{N}$ be minimal such that $v, f(v), \ldots, f^d(v)$ are linearly dependent. Then there exist $a_0, \ldots, a_{d-1} \in K$ with $f^d(v) + a_{d-1}f^{d-1}(v) + \ldots + a_1 f(v) + a_0 v = 0$. Let $\alpha := X^d + a_{d-1}X^{d-1} + \ldots + a_0 \in K[X]$. Then

$$\alpha(f)(f^i(v)) = f^i(\alpha(f)(v)) = f^i(0) = 0$$

for all $i \in \mathbb{N}_0$. Thus $\mu_v \mid \alpha$. Since $v, f(v), \ldots, f^{d-1}(v)$ are linearly independent, on the other hand $\deg(\mu_v) \geq d$. This shows $\mu_v = \alpha$.

Lemma 15.16. *For $f \in \mathrm{End}(V)$ there exists a $v \in V$ with $\mu_v = \mu_f$.*

Proof. Let $V_i := \mathrm{Ker}(\gamma_i^{a_i}(f))$ for $i = 1, \ldots, k$ as in Theorem 15.12. For $v \in V_i$, $\langle f^j(v) : j \in \mathbb{N}_0 \rangle \subseteq V_i$. Therefore μ_v is a divisor of $\gamma_i^{a_i}$. By the unique prime factorization, $\mu_v = \gamma_i^{b_i}$ for some $b_i \leq a_i$. Since $\gamma_i^{a_i}$ is the minimal polynomial of $f_{|V_i}$, there must exist a $v_i \in V_i$ with $\mu_{v_i} = \gamma_i^{a_i}$. We set $v := v_1 + \ldots + v_k$. Since V_i is f-invariant, $\mu_v(f)(v_i) = 0$ for $i = 1, \ldots, k$. This shows $\gamma_i^{a_i} = \mu_{v_i} \mid \mu_v$ and it follows that $\mu_v = \mu_f$. □

15.2 Companion Matrices

Remark 15.17. We define the counterpart to the Jordan blocks over arbitrary fields.

Definition 15.18. For $\alpha := X^n + a_{n-1}X^{n-1} + \ldots + a_0 \in K[X] \setminus K$, one calls

$$B(\alpha) := \begin{pmatrix} 0 & & & -a_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & -a_{n-2} \\ 0 & & 1 & -a_{n-1} \end{pmatrix} \in K^{n\times n}$$

the *companion matrix* of α.[1]

Example 15.19. Obviously $B_{X^n} = J_n(0)$ is a Jordan block and $B_{X^n-1} = P_\sigma$ is the permutation matrix of the n-cycle $\sigma = (1, \ldots, n)$.

Lemma 15.20. *Let $\alpha \in K[X] \setminus K$ be monic and $k \in \mathbb{N}$. Then:*

(a) $\chi_{B(\alpha)} = \mu_{B(\alpha)} = \alpha$.

(b) $\alpha(B(\alpha^k)) \approx \operatorname{diag}\big(J_k(0), \ldots, J_k(0)\big)$.

Proof. Let $\alpha = X^n + a_{n-1}X^{n-1} + \ldots + a_0$.

(a) Let $B := B(\alpha)$. For $i = 1 \ldots, n-1$ it holds that $Be_i = e_{i+1}$. Therefore $e_1, Be_1, \ldots, B^{n-1}e_1$ is the standard basis of K^n and $\deg(\mu_B) \geq \deg(\mu_{e_1}) \geq n$. On the other hand,

$$\alpha(B)e_i = B^{i-1}\alpha(B)e_1 = B^{i-1}(Be_n + a_{n-1}e_n + \ldots + a_1e_2 + a_0e_1) = 0$$

for $i = 1, \ldots, n$. This shows $\mu_B \mid \alpha$. Overall, $\mu_B = \alpha$. By Cayley-Hamilton, $\chi_B = \mu_B$ because of $\deg(\chi_B) = n$.

(b) Let $d := \deg(\alpha^k) = k\deg(\alpha) = kn$. For $A := B(\alpha^k)$ and $N := \alpha(A)$, it holds that $N^k = \alpha^k(A) = 0$ according to (a). According to Theorem 14.28, N has a Jordan normal form J consisting of blocks of the form $J_l(0)$ with $l \leq k$. As in (a), $e_{i+1} = Ae_i$ for $i = 1, \ldots, d-1$. Therefore the vectors

$$Ne_i = A^ne_i + a_{n-1}A^{n-1}e_i + \ldots + e_i \in e_{n+i} + \langle e_1, \ldots, e_{n+i-1}\rangle$$

are linearly independent for $i = 1, \ldots, d-n$. This shows $\operatorname{rk}(N) \geq d-n$ and $\dim \operatorname{Ker}(N) \leq n$ by the homomorphism theorem. Thus J has at most n Jordan blocks. Consequently, these must all have size $k \times k$. □

Theorem 15.21. *Let $f \in \operatorname{End}(V)$ with $\chi_f = \mu_f$. Then there exists a basis B of V with ${}_B[f]_B = B(\mu_f)$.*

Proof. Let $n := \dim V$. According to Lemma 15.16, there exists a $v \in V$ with $\mu_f = \mu_v$. For $U := \langle f^i(v) : i \in \mathbb{N}_0\rangle \leq V$ it holds that

$$n = \deg(\chi_f) = \deg(\mu_f) = \deg(\mu_v) \leq \dim U.$$

Thus $U = V$. According to Remark 15.15, $B := \{v, f(v), \ldots, f^{n-1}(v)\}$ is a basis of V and ${}_B[f]_B = B(\mu_f)$. □

[1] The letter B derives from the German term *Begleitmatrix*.

Theorem 15.22. *For every matrix $A \in K^{n\times n}$ the following holds:*

(a) There exist uniquely determined monic polynomials $\alpha_1, \dots, \alpha_k \in K[X] \setminus K$ with $\alpha_k \mid \alpha_{k-1} \mid \dots \mid \alpha_1$ and

$$A \approx \operatorname{diag}\big(B(\alpha_1), \dots, B(\alpha_k)\big). \qquad \text{(Frobenius normal form}^2\text{)}$$

In this case, $\mu_A = \alpha_1$ and $\chi_A = \alpha_1 \dots \alpha_k$.

(b) There exist irreducible polynomials $\gamma_1, \dots, \gamma_s \in K[X]$ and $a_1, \dots, a_s \in \mathbb{N}$ with

$$A \approx \operatorname{diag}\big(B({\gamma_1}^{a_1}), \dots, B({\gamma_s}^{a_s})\big). \qquad \text{(Weierstrass normal form)}$$

In this case, the powers $\gamma_1^{a_1}, \dots, \gamma_s^{a_s}$ are uniquely determined up to their order and $\chi_A = \gamma_1^{a_1} \dots \gamma_s^{a_s}$.

Proof (Jacob). Let $V := K^n$ and $f \in \operatorname{End}(V)$ with $[f] = A$. According to Lemma 15.16, there exists $v \in V$ with $\alpha_1 := \mu_v = \mu_f$. Let $d := \deg(\alpha_1)$ and

$$U := \langle f^i(v) : i \in \mathbb{N}_0 \rangle \leq V.$$

According to Remark 15.15, $B(\alpha_1)$ is the representation matrix of $f_{|U}$ wrt. $\{b_i := f^{i-1}(v) : i = 1, \dots, d\}$. In the case $U = V$, we are finished. So let $U < V$. We extend the b_i to a basis $b_1, \dots, b_n$ of V. Let $b_1^*, \dots, b_n^*$ be the dual basis of V^* and $f^* \in \operatorname{End}(V^*)$ be the dual map to f. According to Lemma 14.9, $U^0 \leq V^*$ is f^*-invariant. The cyclic subspace

$$L := \langle (f^*)^i(b_d^*) : i \in \mathbb{N}_0 \rangle \leq V^*$$

is also f^*-invariant. Because $[f^*] = A^{\mathrm{t}}$ (Theorem 7.49), $\mu_{b_d^*} \mid \mu_{f^*} = \mu_f$. In particular, $\dim L \leq d$. Suppose there exist $\lambda_1, \dots, \lambda_t \in K$ with $t \leq d$, $\lambda_t \neq 0$ and

$$v^* := \sum_{i=1}^{t} \lambda_i (f^*)^{i-1}(b_d^*) \in L \cap U^0.$$

According to Remark 7.50, $(f^*)^i = (f^i)^*$ holds for $i \in \mathbb{N}_0$. The contradiction follows:

$$0 = v^*(b_{d-t+1}) = b_d^*\Big(\sum_{i=1}^{t} \lambda_i f^{i-1}(b_{d-t+1})\Big) = b_d^*(\lambda_t b_d) = \lambda_t.$$

Therefore, $\{(f^*)^i(b_d^*) : i = 0, \dots, d-1\}$ is a basis of L and $L \cap U^0 = \{0\}$. From $\dim U^0 = n - \dim U = n - d$ it follows that $V^* = L \oplus U^0$. According to Lemma 7.43,

$$V = L_0 \oplus U,$$

where L_0 is f-invariant according to Lemma 14.9. The minimal polynomial of $f_1 := f_{|L_0}$ divides α_1. By induction on n, f_1 possesses a Frobenius normal form $\operatorname{diag}(B(\alpha_2), \dots, B(\alpha_k))$ with $\alpha_k \mid \alpha_{k-1} \mid \dots \mid \alpha_2 = \mu_{f_1} \mid \alpha_1$. Altogether, a Frobenius normal form exists for f. From the block diagonal form and Lemma 15.20, it follows that $\chi_A = \alpha_1 \dots \alpha_k$.

For uniqueness, we first construct the Weierstrass normal form. For this, let $\mu_f = \gamma_1^{c_1} \dots \gamma_l^{c_l}$ be the prime factorization, $V_i := \operatorname{Ker}(\gamma_i^{c_i}(f))$ and $f_i := f_{|V_i}$ for $i = 1, \dots, l$. According to the primary decomposition,

$$V = V_1 \oplus \dots \oplus V_l$$

[2] or *rational canonical form*

and $\mu_{f_i} = \gamma_i^{c_i}$. A Frobenius normal form of f_i thus has the form $\operatorname{diag}(B(\gamma_i^{a_1}), \ldots, B(\gamma_i^{a_s}))$ with $1 \leq a_1, \ldots, a_s \leq c_i$. From this, one obtains a Weierstrass normal form for f. According to Lemma 15.20,

$$\operatorname{diag}\big(\underbrace{J_{a_1}(0), \ldots, J_{a_1}(0)}_{\deg(\gamma_i)}, \ldots, \underbrace{J_{a_s}(0), \ldots, J_{a_s}(0)}_{\deg(\gamma_i)}\big)$$

is the Jordan normal form of $\gamma_i(f_i)$. Therefore, the numbers $a_1, \ldots, a_s$ are uniquely determined. Conversely, let $B(\rho)$ with $\rho \in K[X]$ be a block of any Weierstrass normal form W of f. Then there exists an f-invariant subspace $U \leq V$ such that ρ is the minimal polynomial of $f_{|U}$. It follows that $\rho \mid \mu_f$. According to Corollary 15.11, $\rho \mid \gamma_i^{c_i}$ for some i and $U \leq V_i$. The corresponding blocks of W thus also provide a Weierstrass normal form of f_i. Since the numbers $a_1, \ldots, a_s$ are uniquely determined, the Weierstrass normal form of f is also uniquely determined (up to the order of the blocks).

Finally, we consider the f-invariant decomposition $V = U_1 \oplus \ldots \oplus U_k$ of a Frobenius normal form of f as above. Let $\alpha_i = \gamma_{i1}^{a_{i1}} \ldots \gamma_{is}^{a_{is}}$ be the prime factorization of the minimal polynomial of $f_{|U_i}$. In the Weierstrass normal form of $f_{|U_i}$, the blocks $B(\gamma_{ij}^{a_{ij}})$ with $j = 1, \ldots, s$ must then occur. Since α_i is also the characteristic polynomial of $f_{|U_i}$ (Lemma 15.20), no further blocks can occur. Together, all $B(\gamma_{ij}^{a_{ij}})$ yield the unique Weierstrass normal form of f. In this way, $\alpha_1, \ldots, \alpha_k$ are also uniquely determined (see Example 15.24). □

Remark 15.23.

(a) In contrast to the Jordan normal form, the blocks $B(\alpha_i)$ of the Frobenius normal form are in a fixed order. Furthermore, the Frobenius normal form does not depend on the factorization of the minimal polynomial. In particular, the Frobenius normal form does not change if K is replaced by a larger field (for example $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$).

(b) The polynomials $\alpha_1, \ldots, \alpha_k$ in the Frobenius normal form can alternatively be described by the sequence $\beta_1 := \alpha_1/\alpha_2$, $\beta_2 := \alpha_2/\alpha_3, \ldots, \beta_k := \alpha_k$. Conversely, if arbitrary monic polynomials $\beta_1, \ldots, \beta_k$ are given, one obtains the Frobenius normal form

$$\operatorname{diag}(\beta_1 \ldots \beta_k, \beta_2 \ldots \beta_k, \ldots, \beta_k) \in K^{n \times n}$$

with

$$n = \deg(\beta_1) + 2\deg(\beta_2) + \ldots + k\deg(\beta_k).$$

In this way, the similarity classes of matrices can be systematically enumerated. If one only wants to count invertible matrices, all β_i must have a constant term $\neq 0$ (otherwise X would be a divisor of the characteristic polynomial).

(c) If $q := |K| < \infty$, then there are exactly q^d monic polynomials of degree $d \geq 0$. Among these, $q^d - q^{d-1}$ have a constant term $\neq 0$. Let $n = 4$. With the notation from (b), there are the following possibilities:

$\deg(\beta_1)$	$\deg(\beta_2)$	$\deg(\beta_3)$	$\deg(\beta_4)$	Number	invertible
0	0	0	1	q	$q-1$
1	0	1		q^2	$(q-1)^2$
2	1			q^3	$(q^2-q)(q-1)$
0	2			q^2	q^2-q
4				q^4	q^4-q^3

In total, there are $q^4 + q^3 + 2q^2 + q$ similarity classes of matrices in $K^{4\times 4}$. Of these, $q^4 - q$ consist of invertible matrices.

(d) If A is nilpotent, then the Jordan, Frobenius, and Weierstraß normal forms coincide.

(e) A disadvantage of the Frobenius normal form is that companion matrices are less easy to multiply than Jordan blocks. Another disadvantage is that the Frobenius normal form F is a diagonal matrix only if $A = F$ is a scalar matrix (note $\deg(\mu_A) = \deg(\alpha_1) = 1$). However, one can see from the first block $B(\alpha_1)$ whether A is diagonalizable (Theorem 10.52).

(f) Note that the irreducible polynomials γ_i in the Weierstraß normal form are not necessarily distinct. Since these polynomials arise from the prime factorization of the α_i, the Weierstraß normal form, in contrast to the Frobenius normal form, depends on K (in a larger field, the γ_i might decompose). A is diagonalizable if and only if the Weierstraß normal form is a diagonal matrix. In this case, the Weierstraß normal form thus coincides with the Jordan normal form.

Example 15.24.

(a) Let $\alpha, \beta, \gamma \in K[X]$ be irreducible. The conversion between Frobenius normal form and Weierstrass normal form works as follows:

$$\operatorname{diag}\big(B(\alpha^2\beta^3\gamma), B(\alpha^2\beta^2), B(\alpha)\big) \approx \operatorname{diag}\big(B(\alpha), B(\alpha^2), B(\alpha^2), B(\beta^2), B(\beta^3), B(\gamma)\big).$$

(b) The calculation of the Frobenius/Weierstrass normal form (by hand) is naturally very laborious. Our proof of Lemma 15.16, for example, is not constructive, but there are corresponding algorithms.[3] Often one can use ad hoc arguments. The matrix

$$A := \begin{pmatrix} 7 & -1 & -3 \\ 30 & -4 & -15 \\ 0 & 0 & 1 \end{pmatrix} \in \mathbb{Q}^{3\times 3}$$

has the characteristic polynomial

$$\chi_A = \big((X-7)(X+4)+30\big)(X-1) = (X^2 - 3X + 2)(X-1) = (X-1)^2(X-2).$$

Because of

$$(A - 1_2)(A - 2\cdot 1_2) = \begin{pmatrix} 6 & -1 & -3 \\ 30 & -5 & -15 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 5 & -1 & -3 \\ 30 & -6 & -15 \\ 0 & 0 & -1 \end{pmatrix} = 0$$

we have $\mu_A = (X-1)(X-2)$. Thus

$$\begin{pmatrix} B(X-1) & 0 \\ 0 & B(\mu) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 1 & 3 \end{pmatrix}$$

is the Frobenius normal form of A. The Weierstrass normal form, on the other hand, is $\operatorname{diag}(1,1,2)$.

(c) We describe the similarity classes of matrices in $\mathbb{F}_2^{3\times 3}$ with the sequence $(\beta_1, \ldots, \beta_k)$ from Remark 15.23:

$$\begin{array}{ccc} (1,1,X) = 0_3, & (1,1,X+1) = 1_3, & (X,X), \\ (X,X+1) \approx \operatorname{diag}(1,1,0), & (X+1,X) \approx \operatorname{diag}(1,0,0), & (X+1,X+1) = P_{(1,2)}, \\ (X^3) = J_3(0), & (X^3+1) = P_{(1,2,3)}, & (X^3+X), \\ (X^3+X+1), & (X^3+X^2), & (X^3+X^2+1), \\ (X^3+X^2+X), & (X^3+X^2+X+1). & \end{array}$$

[3] see [M. Geck, *On Jacob's construction of the rational canonical form of a matrix*, Electron. J. Linear Algebra 36 (2020), 177–182]

Thus there are exactly 14 similarity classes, six of which belong to invertible matrices (which ones?).

Theorem 15.25. *For all $A \in K^{n\times n}$, it holds that $A \approx A^{\mathrm{t}}$.*

Proof. According to the Frobenius normal form, one can assume $A = B(\alpha)$ for a monic polynomial $\alpha \in K[X] \setminus K$. By Lemma 15.20 and Exercise II.4, we have $\chi_{A^{\mathrm{t}}} = \chi_A = \mu_A = \mu_{A^{\mathrm{t}}}$. By Theorem 15.21, $A^{\mathrm{t}} \approx A$. □

Remark 15.26. A direct proof of Theorem 15.25 can be found in Exercise II.27.

15.3 Centralizers

Definition 15.27. For $f \in \mathrm{End}(V)$ and $A \in K^{n\times n}$, let

$$\mathrm{C}(f) := \{g \in \mathrm{End}(V) : f \circ g = g \circ f\} \subseteq \mathrm{End}(V),$$
$$\mathrm{C}(A) := \{B \in K^{n\times n} : AB = BA\} \subseteq K^{n\times n}$$

be the *centralizer* of f and A, respectively.

Remark 15.28. Obviously, centralizers are subspaces. For $C, D \in \mathrm{C}(A)$, it also holds that $CD \in \mathrm{C}(A)$. For $S \in \mathrm{GL}(n, K)$, we have $\mathrm{C}(SAS^{-1}) = S\,\mathrm{C}(A)S^{-1}$. Therefore, $\dim \mathrm{C}(A)$ is an invariant of the similarity class of A. It holds that

$$\{\alpha(A) : \alpha \in K[X]\} = \langle A^i : i \in \mathbb{N}_0\rangle \subseteq \mathrm{C}(A).$$

We investigate when equality holds.

Example 15.29.

(a) Let $A \in K^{n\times n}$ be a diagonal matrix. After permutation of the basis vectors, we can assume $A = \mathrm{diag}(\lambda_1 1_{d_1}, \ldots, \lambda_k 1_{d_k})$ with pairwise distinct $\lambda_1, \ldots, \lambda_k \in K$. According to Exercise I.16, $\mathrm{C}(A) = \{\mathrm{diag}(A_1, \ldots, A_k) : \forall i : A_i \in K^{d_i\times d_i}\}$. In particular, $\dim \mathrm{C}(A) = \sum_{i=1}^k d_i^2$. This will be generalized in Exercise II.30 and Remark 15.33.

(b) Let $A \in K^{d\times d}$ and $D := \mathrm{diag}(A, \ldots, A) \in K^{dk\times dk}$. Let $B = (B_{ij}) \in K^{dk\times dk}$ with blocks $B_{ij} \in K^{d\times d}$. Then

$$\begin{pmatrix} AB_{11} & \cdots & AB_{1k} \\ \vdots & & \vdots \\ AB_{k1} & \cdots & AB_{kk} \end{pmatrix} = DB = BD = \begin{pmatrix} B_{11}A & \cdots & B_{1k}A \\ \vdots & & \vdots \\ B_{k1}A & \cdots & B_{kk}A \end{pmatrix} \iff \forall i, j : B_{ij} \in \mathrm{C}(A).$$

In particular, $\dim \mathrm{C}(D) = k^2 \dim \mathrm{C}(A)$.

(c) According to Exercise II.22, $\dim \mathrm{C}(J_n(\lambda)) = n$ for all $\lambda \in K$. This will be generalized in Corollary 15.34.

Lemma 15.30. *Let V, W be vector spaces and $f \in \mathrm{End}(V)$, $g \in \mathrm{End}(W)$ with $\chi_f = \mu_f \mid \chi_g = \mu_g$. Then*

$$H := \{h \in \mathrm{Hom}(V,W) : h \circ f = g \circ h\}$$

is a vector space with $\dim H = \dim V$.

Proof. Let $d := \dim V$ and $e := \dim W$. Because $\chi_f = \mu_f$ and $\chi_g = \mu_g$, there exist $v \in V$ and $w \in W$ with

$$V = \langle f^i(v) : i = 0, \ldots, d-1\rangle, \qquad W = \langle g^i(w) : i = 0, \ldots, e-1\rangle.$$

Let $h \in H$. Then there exists exactly one polynomial $\alpha \in K[X]$ with $h(v) = \alpha(g)(w)$ and $\deg(\alpha) < e$. For $i = 1, \ldots, d-1$ it follows that $h(f^i(v)) = g^i(h(v))$. Thus h is uniquely determined by α. According to Lemma 10.15, the map $H \to K[X]$, $h \mapsto \alpha$ is linear and injective. Because

$$0 = h\big(\mu_f(f)(v)\big) = \mu_f(g)(h(v)) = \mu_f(g)\big(\alpha(g)(w)\big) = (\alpha\mu_f)(g)(w)$$

it also holds that $\mu_g \mid \alpha\mu_f$, i.e. $\tau := \frac{\mu_g}{\mu_f} \mid \alpha$. Let $P_d \subseteq K[X]$ be the vector space of all polynomials of degree $< d$. Then

$$\Gamma\colon H \to P_d, \qquad h \mapsto \alpha/\tau$$

is an injective linear map.

Conversely, let $\beta \in P_d$ be given and $\alpha := \beta\tau$. Then $\mu_g \mid \alpha\mu_f$ holds. We define $h \in \mathrm{Hom}(V,W)$ by $h(v) = \alpha(g)(w)$ and $h(f^i(v)) = g^i(h(v))$ for $i = 1, \ldots, d-1$. For $i = 0, \ldots, d-2$ it then holds that

$$(h \circ f)(f^i(v)) = g^{i+1}(h(v)) = g(g^i(h(v))) = (g \circ h)(f^i(v)).$$

Let $\mu_f = X^d + a_{d-1}X^{d-1} + \ldots + a_0$. Because $\mu_f(f) = 0$ it holds that

$$\begin{aligned}(h \circ f)(f^{d-1}(v)) &= h(f^d(v)) = h\big(-a_{d-1}f^{d-1}(v) - \ldots - a_0 v\big) = -a_{d-1}g^{d-1}(h(v)) - \ldots - a_0 h(v)\\ &= g^d(h(v)) - \mu_f(g)(h(v)) = g^d(h(v)) - (\mu_f\alpha)(g)(w) = g^d(h(v)) = (g \circ h)(f^{d-1}(v)).\end{aligned}$$

This shows $h \in H$ with $\Gamma(h) = \beta$. Thus Γ is an isomorphism and $\dim H = \dim P_d = d = \dim V$. □

Example 15.31. Let $V = K$ and $f = \lambda\,\mathrm{id}_V$ for some $\lambda \in K$. For $h \in H$ it holds that $\lambda h(1) = h(f(1)) = g(h(1))$, i.e. $h(1)$ is an eigenvector of g for the eigenvalue λ. Furthermore, $H \to E_\lambda(g)$, $h \mapsto h(1)$ is an isomorphism. Because $\dim H = 1$, λ has geometric multiplicity 1. This can of course also be derived directly from the condition $\chi_f = X - \lambda \mid \chi_g = \mu_g$.

Theorem 15.32 (Frobenius)**.** *Let $A \in K^{n\times n}$ with Frobenius normal form $\mathrm{diag}(B(\alpha_1), \ldots, B(\alpha_k))$. Then*

$$\boxed{\dim \mathrm{C}(A) = \sum_{i=1}^{k} (2i-1)\deg(\alpha_i).}$$

In particular, $\dim \mathrm{C}(A) \geq n$ with equality if and only if $\chi_A = \mu_A$.

Proof. Let $d_i := \deg(\alpha_i)$ and $A_i := B(\alpha_i) \in K^{d_i \times d_i}$ for $i = 1, \ldots, k$. According to Remark 15.28, we can assume $A = \mathrm{diag}(A_1, \ldots, A_k)$. Let $C = (C_{ij}) \in K^{n\times n}$ be a block matrix with $C_{ij} \in K^{d_i \times d_j}$ for $1 \leq i, j \leq k$. Then

$$C \in \mathrm{C}(A) \iff CA = AC \iff \forall i,j : C_{ij}A_j = A_iC_{ij}.$$

For $i \leq j$, it holds that $\chi_{A_j} = \mu_{A_j} = \alpha_j \mid \alpha_i = \chi_{A_i} = \mu_{A_i}$ according to Lemma 15.20. According to Lemma 15.30, there are d_j linearly independent possibilities for the choice of C_{ij}. In the case $i > j$, we can consider

$$C_{ij}^{\mathrm{t}} A_i^{\mathrm{t}} = (A_i C_{ij})^{\mathrm{t}} = (C_{ij} A_j)^{\mathrm{t}} = A_j^{\mathrm{t}} C_{ij}^{\mathrm{t}}.$$

Because of $\chi_{A_i^{\mathrm{t}}} = \alpha_i = \mu_{A_i^{\mathrm{t}}}$ (Exercise II.4), there are d_i linearly independent possibilities for C_{ij}^{t} (and thus also for C_{ij}). Since the blocks C_{ij} in C can be chosen independently of each other, one obtains

$$\dim \mathrm{C}(A) = \sum_{i,j=1}^{k} \min\{d_i, d_j\} = d_1 + 3d_2 + 5d_3 + \ldots + (2k-1)d_k \geq d_1 + \ldots + d_k = n$$

with equality if and only if $k = 1$ and $\mu_A = \alpha_1 = \chi_A$ (Theorem 15.22). □

Remark 15.33. In Remark 15.23, we described the Frobenius normal form using the polynomials $\beta_1 := \alpha_1/\alpha_2$, $\beta_2 := \alpha_2/\alpha_3, \ldots, \beta_k := \alpha_k$. It holds that $\deg(\alpha_i) = \deg(\beta_i \beta_{i+1} \ldots \beta_k) = \sum_{j=i}^{k} \deg(\beta_j)$ and $\sum_{i=1}^{m}(2i-1) = m^2$ according to Exercise I.4. This shows

$$\boxed{\dim \mathrm{C}(A) = \sum_{l=1}^{k} l^2 \deg(\beta_l).}$$

Corollary 15.34. *Let $A \in K^{n\times n}$ with $\chi_A = \mu_A$. Then $\mathrm{C}(A) = \langle A^i : i = 0, \ldots, n-1\rangle$.*

Proof. From $\deg(\mu_A) = \deg(\chi_A) = n$, it follows that $\dim \mathrm{C}(A) \geq \dim\langle A^i : i \in \mathbb{N}_0\rangle = n$. According to Frobenius, on the other hand, $\dim \mathrm{C}(A) = n$. □

15.4 Splitting Fields

Theorem 15.35. *For $n, k \in \mathbb{N}$ and $\lambda \in \mathbb{C}$, there exist exactly k matrices $W \in \mathbb{C}^{n\times n}$ with $W^k = J_n(\lambda)$.*

Proof. Let $J := J_n(\lambda)$. By Theorem 14.42 there exists at least one k-th root W with $W^k = J$. For every k-th root of unity $\zeta \in \mathbb{C}$, $(\zeta W)^k = J$ also holds. By Lemma 11.27 there exist at least k roots of J. By Corollary 15.34 it holds that $W \in \mathrm{C}(J) = \langle J^i : i \in \mathbb{N}_0\rangle$. In particular, W is a lower triangular matrix with diagonal $(\mu, \ldots, \mu)$ for a k-th root of unity μ (cf. Exercise II.22). By replacing W with $\mu^{-1}W$, one can assume $\mu = 1$.

Conversely, let $A \in \mathbb{C}^{n\times n}$ with $A^k = J$. Then A is also a lower triangular matrix with diagonal $(\zeta, \ldots, \zeta)$ for a k-th root of unity ζ. Since $B := \zeta W \in \mathrm{C}(J)$ is a polynomial in J, A and B commute. It follows that

$$0 = A^k - B^k = (A-B)(\underbrace{A^{k-1} + A^{k-2}B + \ldots + AB^{k-2} + B^{k-1}}_{=:C}).$$

Obviously C is a lower triangular matrix with main diagonal $k\zeta^{k-1}(1, \ldots, 1) \neq 0$. In particular, C is invertible and $A - B = 0 \cdot C^{-1} = 0$, i. e. $A = B$. □

Theorem 15.36 (Schur's Lemma)**.** *Let $f \in \mathrm{End}(V)$. Then χ_f is irreducible if and only if $\{0\}$ and V are the only f-invariant subspaces.*

Proof. Let $\{0\}$ and V be the only f-invariant subspaces of V. According to the Frobenius normal form, $\chi_f = \mu_f$. Let γ be an irreducible divisor of μ_f. Then $U := \operatorname{Ker}(\gamma(f)) \leq V$ is an f-invariant subspace. For $v \in V$ with $\mu_v = \mu_f$ (Lemma 15.16), $0 \neq \frac{\mu_f}{\gamma}(f)(v) \in U$ holds. This shows $U = V$ and $\chi_f = \mu_f = \gamma$ is irreducible.

Conversely, let χ_f be irreducible and $U < V$ be f-invariant. We extend a basis B_1 of U to a basis B of V. Then

$$_B[f]_B = \begin{pmatrix} {}_{B_1}[f_{|U}]_{B_1} & * \\ 0 & * \end{pmatrix}.$$

This shows $\chi_{f_{|U}} \mid \chi_f$ and $\chi_{f_{|U}} \in K$. It follows that $U = \{0\}$. □

Theorem 15.37. *For every irreducible polynomial $\gamma \in K[X]$, $L := \mathrm{C}(B(\gamma))$ is a field with $\dim_K L = \deg(\gamma)$.*

Proof. Let $n := \deg(\gamma)$, $V = K^n$ and $f \in \operatorname{End}(V)$ with $A := B(\gamma) = [f]$. By Lemma 15.20, $\chi_f = \mu_f = \gamma$. By Corollary 15.34, $L = \langle A^i : i \in \mathbb{N}_0 \rangle \subseteq K^{n \times n}$. In particular, the (matrix) multiplication in L is commutative. For $g \in \mathrm{C}(f) \setminus \{0\}$, $\operatorname{Ker}(g) \leq V$ is, as usual, an f-invariant subspace. By Schur's Lemma, $\operatorname{Ker}(g) = \{0\}$ and $g \in \mathrm{GL}(V)$. Since g is in $\mathrm{C}(f)$, g^{-1} is also in $\mathrm{C}(f)$. Thus every non-trivial element in L has an inverse wrt. multiplication. The remaining field axioms for $\mathrm{C}(A)$ follow from the calculation rules in $K^{n \times n}$. By Corollary 15.34, $\dim L = n$. □

Example 15.38.

(a) As is well known, $\gamma := X^2 + 1 \in \mathbb{R}[X]$ is irreducible. Let $A := B(\gamma) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$. According to Theorem 15.37, $\mathrm{C}(A) = \left\{ \left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right) : a, b \in \mathbb{R} \right\}$ is a field. In fact, $\mathrm{C}(A)$ coincides with the construction of $\mathbb{C}$ in the proof of Lemma 11.24.

(b) For $\gamma = X^2 - 2 \in \mathbb{Q}[X]$, $\mathrm{C}(B(\gamma)) = \mathbb{Q}(\sqrt{2})$ is the field from Exercise I.14.

(c) For $\gamma := X^2 + X + 1 \in \mathbb{F}_2[X]$,

$$\mathrm{C}(B(\gamma)) = \left\{ 0_2, 1_2, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$

is a field with four elements.

Theorem 15.39. *For every monic polynomial $\alpha \in K[X]$, there exists a field $L \supseteq K$ such that α splits into linear factors in $L[X]$.*

Proof. In the case $\deg(\alpha) = 1$, α itself is a linear factor. So let $\deg(\alpha) \geq 2$. Let γ be an irreducible divisor of α with $d := \deg(\gamma)$. According to Theorem 15.37, $L_1 := \mathrm{C}(B(\gamma_1))$ is a field. We can identify K with the scalar matrices $\{\lambda 1_d \in L_1 : \lambda \in K\}$ (the calculation rules for scalar matrices correspond to those in K). For $x := B(\gamma) \in L_1$, we have $\gamma(x) = 0$, i.e., x is a root of γ. According to Lemma 10.22, $\alpha = (X - x)\beta$ holds for some $\beta \in L_1[X]$ with $\deg(\beta) = \deg(\alpha) - 1$. By induction on $\deg(\alpha)$, there exists a field $L \supseteq L_1$ in which β splits into linear factors. Obviously, α also splits into linear factors in $L[X]$. □

Definition 15.40. In the situation of Theorem 15.39, L is called a *splitting field* of α (not uniquely determined).

16 The Jordan-Chevalley Decomposition

16.1 The Chinese Remainder Theorem

Remark 16.1. We extend our knowledge about polynomials. With this tool, we decompose a matrix in a unique way into a diagonalizable part (over a splitting field) and a nilpotent part.

Definition 16.2. Let $\alpha, \beta, \delta \in K[X]$ with $\delta \neq 0$. We say α is *congruent* to β *modulo* δ if $\delta \mid \alpha - \beta$. If applicable, we write $\alpha \equiv \beta \pmod{\delta}$.

Lemma 16.3. *Congruence modulo $\delta \in K[X]\setminus\{0\}$ is an equivalence relation on $K[X]$. For $\alpha_i, \beta_i \in K[X]$ with $\alpha_i \equiv \beta_i \pmod{\delta}$ $(i = 1, 2)$, it holds that $\alpha_1 + \alpha_2 \equiv \beta_1 + \beta_2 \pmod{\delta}$ and $\alpha_1\alpha_2 \equiv \beta_1\beta_2 \pmod{\delta}$.*

Proof. Because $\delta \mid 0 = \alpha - \alpha$, we have $\alpha \equiv \alpha \pmod{\delta}$. From $\alpha \equiv \beta \pmod{\delta}$ it follows that $\beta \equiv \alpha \pmod{\delta}$. Now let $\alpha \equiv \beta \pmod{\delta}$ and $\beta \equiv \gamma \pmod{\delta}$. Then $\delta \mid (\alpha - \beta) + (\beta - \gamma) = \alpha - \gamma$ and $\alpha \equiv \gamma \pmod{\delta}$ holds. Therefore, $\equiv$ is an equivalence relation. From $\alpha_i \equiv \beta_i \pmod{\delta}$ it follows that

$$\begin{aligned}\delta \mid (\alpha_1 - \beta_1) + (\alpha_2 - \beta_2) &= (\alpha_1 + \alpha_2) - (\beta_1 + \beta_2),\\ \delta \mid (\alpha_1 - \beta_1)\alpha_2 + (\alpha_2 - \beta_2)\beta_1 &= \alpha_1\alpha_2 - \beta_1\beta_2.\end{aligned}$$

This shows the second assertion. □

Example 16.4.

(a) For $\delta \in K^\times$, we have $\alpha \equiv \beta \pmod{\delta}$ for all $\alpha, \beta \in K[X]$, i.e., $\equiv$ is the trivial relation.

(b) It holds that $\alpha \equiv \beta \pmod{X}$ if and only if α and β have the same constant term.

(c) For all $\alpha \in K[X]$ there exists a $\beta \in K[X]$ with $\alpha \equiv \beta \pmod{\delta}$ and $\deg(\beta) < \deg(\delta)$. This follows from Euclidean division.

(d) Lemma 16.3 simplifies many divisibility considerations: To check whether

$$\alpha = (X^2 + X + 2)^4 + (X^5 - 1)(X^3 + X) \in \mathbb{Q}[X]$$

is divisible by $\delta = X + 1 \in \mathbb{Q}[X]$, one can perform the congruence with each individual term. Because

$$\begin{aligned}X^2 + X + 2 = X(X + 1) + 2 &\equiv 2 \pmod{\delta},\\ X^5 - 1 = (X + 1)(X^4 - X^3 + X^2 - X + 1) - 2 &\equiv -2 \pmod{\delta},\\ X^3 + X + 10 = (X + 1)(X^2 - X + 2) + 8 &\equiv 8 \pmod{\delta}\end{aligned}$$

it holds that $\alpha \equiv 2^4 + (-2)8 \equiv 0 \pmod{\delta}$, i.e., $\delta \mid \alpha$.

Theorem 16.5 (Chinese Remainder Theorem)**.** *Let $\alpha_1, \ldots, \alpha_n \in K[X] \setminus \{0\}$ be pairwise coprime and $\beta_1, \ldots, \beta_n \in K[X]$ be arbitrary. Then there exists a $\gamma \in K[X]$ with $\gamma \equiv \beta_i \pmod{\alpha_i}$ for $i = 1, \ldots, n$.*

Proof. For $i = 1, \ldots, n$ let $\alpha_i' := \prod_{j \neq i} \alpha_j$. According to Corollary 15.11, every irreducible divisor of α_i' must divide an α_j with $j \neq i$. Since α_i and α_j are coprime, α_i and α_i' must also be coprime. By Bézout, there exist $\gamma_1, \ldots, \gamma_n \in K[X]$ with $\alpha_i' \gamma_i \equiv 1 \pmod{\alpha_i}$. Let

$$\gamma := \sum_{i=1}^{n} \alpha_i' \beta_i \gamma_i.$$

Then $\gamma \equiv \alpha_i' \beta_i \gamma_i \equiv \beta_i \pmod{\alpha_i}$ holds for $i = 1, \ldots, n$. □

Remark 16.6. Along with γ, $\gamma + \rho \prod_{i=1}^{n} \alpha_i$ also satisfies the congruences from the Chinese Remainder Theorem for all $\rho \in K[X]$. Therefore, there are infinitely many solutions.

Example 16.7. Since the polynomials $X^2 - 2, X^2 + 1 \in \mathbb{Q}[X]$ have no common roots, they are coprime. Suppose we are looking for $\gamma \in \mathbb{Q}[X]$ with

$$\gamma \equiv X \pmod{X^2 - 2}, \qquad \gamma \equiv 3 \pmod{X^2 + 1}.$$

Bézout yields $\frac{1}{3}(X^2 + 1) \equiv 1 \pmod{X^2 - 2}$ and $-\frac{1}{3}(X^2 - 2) \equiv 1 \pmod{X^2 + 1}$ (if in doubt, one must use the Euclidean algorithm). As in the proof of Theorem 16.5,

$$\gamma = \frac{1}{3}(X^2 + 1)X - \frac{1}{3}(X^2 - 2)3 = \frac{1}{3}X^3 - X^2 + \frac{1}{3}X + 2$$

is a solution to the congruences.

Definition 16.8.

- For $\alpha = \sum_{k=0}^{\infty} a_k X^k \in K[X]$ we define the (formal) *derivative*

$$\alpha' = \sum_{k=1}^{\infty} k a_k X^{k-1}$$

  of α as in analysis.

- An irreducible polynomial α is called *separable*, if $\alpha' \neq 0$. An arbitrary polynomial α is called *separable*, if all irreducible divisors of α are separable (the opposite is *inseparable*).

Example 16.9.

(a) The constant polynomials are separable, since they have no irreducible divisors.

(b) If the map $\mathbb{N} \to K$, $n \mapsto n \cdot 1_K$ is injective, then $\alpha' \neq 0$ holds for all $\alpha \in K[X] \setminus K$. In particular, every polynomial in $\mathbb{C}[X]$ is separable.

(c) Over finite fields, the condition $\alpha' \neq 0$ is less obvious. For example, $(X^2)' = 0$ holds in $\mathbb{F}_2[X]$. Let generally $\alpha = \sum a_k X^k \in \mathbb{F}_2[X]$ with $\alpha' = \sum k a_k X^{k-1} = 0$. For odd k it follows that $a_k X^{k-1} = k a_k X^{k-1} = 0$, thus $a_k = 0$. This shows

$$\alpha = a_0 + a_2 X^2 + \ldots + a_{2n} X^{2n}$$

for some $n \in \mathbb{N}$. Because of $(x+y)^2 = x^2 + 2xy + y^2 = x + y$ in $\mathbb{F}_2$ one obtains inductively

$$\alpha = (a_0 + a_2 X + \ldots + a_{2n} X^n)^2.$$

In particular, α is reducible (i.e., not irreducible). Thus every polynomial in $\mathbb{F}_2[X]$ is separable.

(d) One constructs $\mathbb{Q}$ from $\mathbb{Z}$ by introducing fractions. In the same way, one can obtain the field of *rational functions*

$$K(X) := \Big\{ \frac{\alpha}{\beta} : \alpha, \beta \in K[X], \beta \neq 0 \Big\}$$

from $K[X]$ by introducing fractions of polynomials. Specifically, let $K := \mathbb{F}_2(X)$. One can show that $\alpha := X^2 + Y \in K[Y]$ is irreducible and inseparable.

Lemma 16.10. *For $\alpha, \beta \in K[X]$ we have*

$$(\alpha + \beta)' = \alpha' + \beta', \qquad \text{(Sum rule)}$$
$$(\alpha\beta)' = \alpha'\beta + \alpha\beta'. \qquad \text{(Product rule)}$$

Proof. For $\alpha = \sum a_k X^k$ and $\beta = \sum b_k X^k$ we have

$$\begin{aligned}(\alpha + \beta)' &= \Big(\sum (a_k + b_k) X^k\Big)' = \sum k(a_k + b_k) X^{k-1} \\ &= \sum k a_k X^{k-1} + \sum k b_k X^{k-1} = \alpha' + \beta'.\end{aligned}$$

From this it follows that

$$\begin{aligned}(\alpha\beta)' &= \Big(\sum_{k,l=0}^{\infty} a_k b_l X^{k+l}\Big)' = \sum_{k,l} a_k b_l \big(X^{k+l}\big)' = \sum_{k,l} (k+l) a_k b_l X^{k+l-1} \\ &= \sum_{k,l} k a_k b_l X^{k-1+l} + \sum_{k,l} l a_k b_l X^{l-1+k} = \alpha'\beta + \alpha\beta'.\end{aligned}$$

□

Lemma 16.11. *Let $\gamma \in K[X]$ be irreducible. Then γ is separable if and only if γ has no multiple roots in a splitting field.*

Proof. Let $\lambda \in L$ be a root of γ in a splitting field L. Then there exists a $\rho \in L[X]$ with $\gamma = (X - \lambda)\rho$. From the product rule it follows that

$$\gamma' = \rho + (X - \lambda)\rho'$$

and $\gamma'(\lambda) = \rho(\lambda)$. If γ is inseparable, then $\rho(\lambda) = \gamma'(\lambda) = 0$, i.e., λ is a multiple root. Conversely, let γ be separable, i.e., $\gamma' \neq 0$. Then $0 \leq \deg(\gamma') < \deg(\gamma)$. Since γ is irreducible, γ' and γ are coprime. By Bézout, there exist $\alpha, \beta \in K[X]$ with $\alpha\gamma + \beta\gamma' = 1$. It follows that $\beta(\lambda)\gamma'(\lambda) = 1$ and $\rho(\lambda) = \gamma'(\lambda) \neq 0$. Thus λ is a simple root. Since λ was arbitrary, all roots are simple. □

16.2 Separable and semisimple maps

Definition 16.12. We call $f \in \mathrm{End}(V)$ (resp. $A \in K^{n\times n}$)

- *separable*, if μ_f (resp. μ_A) is separable.
- *semisimple*, if μ_f (resp. μ_A) factors into pairwise distinct irreducible polynomials.

Example 16.13.

(a) Over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{F}_2$, all endomorphisms are separable according to Example 16.9.

(b) Let $K = \mathbb{C}$. According to the Fundamental Theorem of Algebra and Theorem 10.52, $f \in \mathrm{End}(V)$ is semisimple if and only if f is diagonalizable. This is generalized in Lemma 16.15. On the other hand, $B(X^2 - 2) \in \mathbb{Q}^{2\times 2}$ and $B(X^2 + X + 1) \in \mathbb{F}_2^{2\times 2}$ are semisimple, but not diagonalizable.

Theorem 16.14. *$f \in \mathrm{End}(V)$ is semisimple if and only if every f-invariant subspace $U \leq V$ has an f-invariant complement.*

Proof. Let f be semisimple and $\mu_f = \gamma_1 \dots \gamma_k$ with pairwise distinct irreducible polynomials $\gamma_1, \dots, \gamma_k$. The Weierstrass normal form of f provides a decomposition $V = V_1 \oplus \dots \oplus V_s$ into f-invariant subspaces, such that the representation matrix of f on V_i is given by $B(\gamma_j)$ for some $1 \leq j \leq k$ (note $k \leq s$). According to Lemma 15.20 and Schur's Lemma, V_i has no proper non-trivial f-invariant subspaces for $i = 1, \dots, s$. Let $U \leq V$ be f-invariant. Let $W \leq V$ be f-invariant with $U \cap W = \{0\}$, such that $\dim W$ is as large as possible (if necessary $W = \{0\}$). Suppose $U \oplus W < V$ holds. Then there exists an i with $V_i \not\subseteq U + W$. Obviously,

$$L := V_i \cap (U + W) < V_i$$

is f-invariant. From Schur's Lemma, it follows that $L = \{0\}$. Let $u = w + v \in U \cap (W + V_i)$ with $w \in W$ and $v \in V_i$. Then $v = u - w \in V_i \cap (U + W) = \{0\}$ and $u = w \in U \cap W = \{0\}$. Thus $U \cap (W + V_i) = \{0\}$, contradicting the choice of W. This shows $V = U \oplus W$.

Conversely, assume that every f-invariant subspace of V has an f-invariant complement. We assume indirectly $\mu_f = \gamma^2\delta$ for an irreducible polynomial γ. By assumption, $U := \mathrm{Ker}(\gamma(f)) \leq V$ has an f-invariant complement W. For $w \in W$, it holds that

$$(\gamma\delta)(f)(w) \in U \cap W = \{0\}.$$

Because $(\gamma\delta)(f)(U) = \{0\}$, it even holds that $(\gamma\delta)(f)(v) = 0$ for all $v \in V$. But then $\mu_f \mid \gamma\delta$. Contradiction. □

Lemma 16.15. *Let $A \in K^{n\times n}$ be separable. A is semisimple if and only if there exists a field $L \supseteq K$ such that A is diagonalizable in $L^{n\times n}$.*

Proof. Let A be semisimple and $\mu_A = \gamma_1 \dots \gamma_k$ with pairwise distinct (separable) irreducible polynomials $\gamma_1, \dots, \gamma_k$. Let L be a splitting field of μ_A. According to Lemma 16.11, each γ_i has no multiple roots in L. For $i \neq j$, there exist $\alpha, \beta \in K[X]$ with $\alpha\gamma_i + \beta\gamma_j = 1$ by Bézout. Therefore, γ_i and γ_j have no common roots in L. Overall, μ_A splits into pairwise distinct linear factors in $L[X]$. According to Theorem 10.52, A is diagonalizable in $L^{n\times n}$. If A is not semisimple, then μ_A obviously has multiple roots in $L[X]$. Thus A cannot be diagonalizable. □

Lemma 16.16. *If $A, B \in K^{n\times n}$ are commuting, separable, and semisimple, then $A + B$ is also semisimple.*

Proof. Let L be a splitting field of $\mu_A\mu_B$. According to Lemma 16.15 and Lemma 14.11, A and B are simultaneously diagonalizable in $L^{n\times n}$. Thus, $A + B$ is also diagonalizable in $L^{n\times n}$. According to Theorem 10.52, μ_{A+B} has no multiple roots in L. According to Lemma 16.11, $A + B$ is separable and according to Lemma 16.15 semisimple. □

16.3 Generalized Jordan Blocks

Remark 16.17. If $\gamma \in K[X]$ is irreducible, then one can interpret $\lambda := B(\gamma)$ as an element of the field $L := \mathrm{C}(B(\gamma))$. In the following lemma, we study the Jordan block $J_k(\lambda) \in L^{k\times k}$ as a matrix in $K^{n\times n}$, where $n := k \deg(\gamma)$. For $\gamma = X - \mu$, one obtains $J_k(\gamma) = J_k(\mu)$.

Theorem 16.18 (Generalized Jordan Block)**.** *Let $\gamma \in K[X]$ be irreducible and separable of degree d and $k \in \mathbb{N}$. Then*

$$J_k(\gamma) := \begin{pmatrix} B(\gamma) & & & 0 \\ 1_d & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1_d & B(\gamma) \end{pmatrix} \approx B(\gamma^k).$$

Proof. Let $J := J_k(\gamma)$. A simple induction shows

$$J^m = \begin{pmatrix} B(\gamma)^m & & & 0 \\ mB(\gamma)^{m-1} & \ddots & & \\ & \ddots & \ddots & \\ * & & mB(\gamma)^{m-1} & B(\gamma)^m \end{pmatrix}$$

for $m \in \mathbb{N}_0$ (cf. Theorem 14.42). It follows

$$\gamma(J) = \begin{pmatrix} 0 & & & 0 \\ \gamma'(B(\gamma)) & \ddots & & \\ & \ddots & \ddots & \\ * & & \gamma'(B(\gamma)) & 0 \end{pmatrix}.$$

In particular, $\gamma(J)$ is nilpotent and $\mu_J \mid \gamma^k$. A further induction shows

$$\gamma(J)^2 = \begin{pmatrix} 0 & & & & 0 \\ 0 & \ddots & & & \\ \gamma'(B(\gamma))^2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ * & & \gamma'(B(\gamma))^2 & 0 & 0 \end{pmatrix}, \ldots, \gamma(J)^{k-1} = \begin{pmatrix} 0 & & & 0 \\ \vdots & \ddots & & \\ 0 & & \ddots & \\ \gamma'(B(\gamma))^{k-1} & 0 & \cdots & 0 \end{pmatrix}.$$

Since γ is separable, $0 \leq \deg(\gamma') < \deg(\gamma)$ holds. Since γ is irreducible, γ and γ' must be coprime. By Bézout, there exist $\alpha, \beta \in K[X]$ with $\alpha\gamma + \beta\gamma' = 1$. It follows

$$1 = \alpha(B(\gamma))\gamma(B(\gamma)) + \beta(B(\gamma))\gamma'(B(\gamma)) = \beta(B(\gamma))\gamma'(B(\gamma)).$$

Thus, $\gamma'(B(\gamma))$ is invertible and $\gamma'(B(\gamma))^{k-1} \neq 0$. This shows $\mu_J = \gamma^k$. The claim follows from Theorem 15.21. □

Theorem 16.19 (Jordan-Chevalley Decomposition)**.** *For every separable matrix $A \in K^{n\times n}$, there exist uniquely determined matrices $D, N \in K^{n\times n}$ with the following properties:*

(a) $A = D + N$ and $DN = ND$.

(b) D is semisimple and N is nilpotent.

If applicable, there exists an $\alpha \in K[X]$ with $\alpha(A) = D$.

Proof. We first transform A into the Weierstrass normal form. Since μ_A is separable, we can subsequently transform each block $B(\gamma^a)$ in the Weierstrass normal form into $J_a(\gamma)$ (Theorem 16.18). So let $S \in \mathrm{GL}(n, K)$ with

$$W := SAS^{-1} = \operatorname{diag}(J_{a_1}(\gamma_1), \ldots, J_{a_s}(\gamma_s))$$

For $i = 1, \ldots, s$ let $d_i := \deg(\gamma_i)$ and

$$D_i := \operatorname{diag}(B(\gamma_i), \ldots, B(\gamma_i)) \in K^{a_i d_i \times a_i d_i}, \qquad N_i := J_{a_i}(\gamma_i) - D_i,$$
$$\check{D} := \operatorname{diag}(D_1, \ldots, D_s), \qquad \check{N} := \operatorname{diag}(N_1, \ldots, N_s).$$

Then $W = \check{D} + \check{N}$ holds. Wlog. let $\gamma_1, \ldots, \gamma_k$ be the *distinct* prime divisors of μ_A. For $1 \leq i \leq k$ and $1 \leq j \leq s$ with $\gamma_i = \gamma_j$ let furthermore $a_i \geq a_j$. According to Lemma 15.20, $\mu_{D_i} = \gamma_i$ and $\mu_{\check{D}} = \gamma_1 \ldots \gamma_k$, i.e., $\check{D}$ is semisimple. On the other hand, $\check{N}$ is a strictly lower triangular matrix and therefore nilpotent (Example 14.26). A calculation as in Theorem 16.18 shows $D_i N_i = N_i D_i$ and $\check{D}\check{N} = \check{N}\check{D}$. Since the minimal polynomial does not depend on the choice of basis, $D := S^{-1}\check{D}S$ is semisimple and $N := S^{-1}\check{N}S$ is nilpotent. Furthermore, $DN = ND$ and $A = S^{-1}WS = D + N$ hold.

Because of $J_i := J_{a_i}(\gamma_i) = D_i + N_i$, we have $D_i \in \mathrm{C}(J_i)$. According to Theorem 16.18 and Corollary 15.34, there exists an $\alpha_i \in K[X]$ with $\alpha_i(J_i) = D_i$ for $i = 1, \ldots, k$. Let $1 \leq j \leq s$ with $\gamma_j = \gamma_i$. Then $a_j \leq a_i$ and $J_i = \left(\begin{smallmatrix} J_j & 0 \\ * & * \end{smallmatrix}\right)$. It follows that

$$\begin{pmatrix} D_j & 0 \\ * & * \end{pmatrix} = D_i = \alpha_i(J_i) = \begin{pmatrix} \alpha_i(J_j) & 0 \\ * & * \end{pmatrix}$$

and $\alpha_i(J_j) = D_j$. According to the Chinese Remainder Theorem, there exists an $\alpha \in K[X]$ with $\alpha \equiv \alpha_i \pmod{\gamma_i^{a_i}}$ for $i = 1, \ldots, k$. From $\gamma_i^{a_i}(J_i) = 0$ it follows that

$$\alpha(W) = \operatorname{diag}\big(\alpha_1(J_1), \ldots, \alpha_s(J_s)\big) = \operatorname{diag}(D_1, \ldots, D_s) = \check{D}$$

and $\alpha(A) = S^{-1}\alpha(W)S = D$. Thus the existence statement is proven. Since μ_D has the same prime divisors as μ_A, D is separable.

Now let $A = \grave{D} + \grave{N}$ with the same properties. We first show that $\grave{D}$ is separable. Over a splitting field, $\grave{D}$ and $\grave{N}$ are simultaneously trigonalizable according to Theorem 14.8 and Lemma 14.12. With respect to a suitable basis, A is then a triangular matrix. As a nilpotent matrix, $\grave{N}$ must be a strictly triangular matrix wrt. this basis. Therefore, A and $\grave{D}$ have the same main diagonal and thus the same characteristic polynomial (note: χ_A does not depend on the splitting field). Since A is separable, $\grave{D}$ must

also be separable. By assumption, $\grave{D}$ commutes with A and with $\alpha(A) = D$. According to Lemma 16.16, $D - \grave{D}$ is semisimple. Analogously, $\grave{N}$ and N also commute. Multiplying out $(\grave{N} - N)^{2n}$, one obtains summands of the form $\grave{N}^i N^{2n-i}$ with $0 \leq i \leq 2n$. In the case $i \geq n$, $\grave{N}^i = 0$. Otherwise, $2n - i \geq n$ and $N^{2n-i} = 0$. In any case, $(\grave{N} - N)^{2n} = 0$. Overall, $D - \grave{D} = \grave{N} - N$ is semisimple and nilpotent. This is only possible with the minimal polynomial X, i.e., $\grave{D} = D$ and $\grave{N} = N$. □

Corollary 16.20. *For every matrix $A \in \mathbb{C}^{n\times n}$, there exist uniquely determined matrices $D, N \in \mathbb{C}^{n\times n}$ with the following properties:*

(a) $A = D + N$ and $DN = ND$.

(b) D is diagonalizable and N is nilpotent.

Proof. According to Example 16.13, A is separable and D is semisimple if and only if D is diagonalizable. □

Remark 16.21. One can compute the Jordan-Chevalley decomposition of $A \in K^{n\times n}$ by considering A over a splitting field of μ_A and using the Jordan normal form there. Since the Jordan-Chevalley decomposition is uniquely determined, D and N (nevertheless) lie in $K^{n\times n}$.

Example 16.22. Let $\gamma := X^2 + 1 \in \mathbb{Q}[X]$ and

$$A := B(\gamma^2) = \begin{pmatrix} . & . & . & -1 \\ 1 & . & . & . \\ . & 1 & . & -2 \\ . & . & 1 & . \end{pmatrix} \in \mathbb{Q}[X].$$

We determine the Jordan-Chevalley decomposition of A without the detour via $\mathbb{C}$ (or $\mathbb{Q}(\mathrm{i})$, cf. Exercise I.14) described in Remark 16.21. According to Theorem 16.18, $A \approx J_2(\gamma)$ holds. To realize this change of basis, we need a basis vector $b_3 \in \mathbb{Q}^4$ with $\mu_{b_3} = \gamma$. For this,

$$b_3 := \gamma(A)e_1 = e_1 + e_3 = (1, 0, 1, 0)^{\mathrm{t}}$$

is suitable. Then $b_4 := Ab_3 = (0, 1, 0, 1)^{\mathrm{t}}$. For b_1 and b_2, the following should hold: $Ab_1 = b_2 + b_3$ and $Ab_2 = -b_1 + b_4$. This yields the system of equations

$$\gamma(A)b_1 = (A^2 + 1_4)b_1 = 2b_4 = (0, 2, 0, 2)^{\mathrm{t}}$$

with the solution $b_1 = 2e_2$. Finally, $b_2 = Ab_1 - b_3 = (-1, 0, 1, 0)^{\mathrm{t}}$. For

$$S := \begin{pmatrix} . & -1 & 1 & . \\ 2 & . & . & 1 \\ . & 1 & 1 & . \\ . & . & . & 1 \end{pmatrix} \in \mathrm{GL}(4, \mathbb{Q})$$

it holds that $S^{-1}AS = J_2(\gamma)$. This results in the Jordan-Chevalley decomposition $A = D + N$ with

$$D := S\,\mathrm{diag}(B(\gamma), B(\gamma))S^{-1} = \frac{1}{2}\begin{pmatrix} . & -1 & . & -1 \\ 3 & . & -1 & . \\ . & 1 & . & -3 \\ 1 & . & 1 & . \end{pmatrix},$$

$$N := S \begin{pmatrix} 0_2 & 0_2 \\ 1_2 & 0_2 \end{pmatrix} S^{-1} = \frac{1}{2} \begin{pmatrix} . & 1 & . & -1 \\ -1 & . & 1 & . \\ . & 1 & . & -1 \\ -1 & . & 1 & . \end{pmatrix}.$$

Theorem 16.23. *Let $A \in K^{n\times n}$ be separable with Jordan-Chevalley decomposition $A = D + N$. Then $\mathrm{C}(A) = \mathrm{C}(D) \cap \mathrm{C}(N)$ holds.*

Proof. For $B \in \mathrm{C}(D) \cap \mathrm{C}(N)$, it holds that $BA = BD + BN = DB + NB = AB$ and $B \in \mathrm{C}(A)$. Let $\alpha \in K[X]$ with $\alpha(A) = D$. For $B \in \mathrm{C}(A)$, it then holds that $BD = B\alpha(A) = \alpha(A)B = DB$ and $BN = B(A - \alpha(A)) = (A - \alpha(A))B = NB$. This shows $B \in \mathrm{C}(D) \cap \mathrm{C}(N)$. □

Exercises

Exercise II.1. Let $\alpha, \beta \in K[X]$ with $\alpha \mid \beta \mid \alpha$. Show that there exists a $c \in K^\times$ with $\alpha = c\beta$.

Exercise II.2. Let $n \in \mathbb{N}$. Show:

(a) $A \in \mathbb{F}_2^{n\times n}$ is diagonalizable if and only if $A^2 = A$.

(b) $A \in \mathrm{GL}(n, \mathbb{F}_2)$ is diagonalizable if and only if $A = 1_n$.
Hint: Theorem 10.52.

(c) $|\mathrm{GL}(n, \mathbb{F}_2)| = (2^n - 1)(2^n - 2)\dots(2^n - 2^{n-1})$.
Hint: Theorem 7.45.

Exercise II.3. Let $A \in K^{n\times m}$ and $B \in K^{m\times n}$. Prove Sylvester's *determinant formula*

$$\det(1_n + AB) = \det(1_m + BA).$$

Hint: Lemma 10.38.

Exercise II.4. Let $A, B \in K^{n\times n}$.

(a) Prove or disprove:

$$\chi_A = \chi_{A^\mathrm{t}}, \qquad \mu_A = \mu_{A^\mathrm{t}}, \qquad \mu_{AB} = \mu_{BA}.$$

(b) How can $\chi_{A^{-1}}$ and $\mu_{A^{-1}}$ be calculated from χ_A and μ_A, if A is invertible?

Exercise II.5. Show $\chi_A = X^3 - \mathrm{tr}(A)X^2 + \frac{1}{2}\big(\mathrm{tr}(A)^2 - \mathrm{tr}(A^2)\big)X - \det(A)$ for all $A \in K^{3\times 3}$.

Exercise II.6. Let V be a Euclidean space and $U, W \leq V$. Show:

(a) $(U + W)^\perp = U^\perp \cap W^\perp$.

(b) $(U \cap W)^\perp = U^\perp + W^\perp$.

Hint: One can use Lemma 16.3.

Exercise II.7. Let K be a field, $n \in \mathbb{N}$ and $S \in K^{n\times n}$. Show that

$$\{A \in \mathrm{GL}(n, K) : ASA^\mathrm{t} = S\}$$

is a subgroup of $\mathrm{GL}(n, K)$. For $S = 1_n$ one obtains $\mathrm{O}(n, K)$.

Exercise II.8. Let $v, w \in \mathbb{R}^2$ be linearly independent. Then $0, v, w$ form the vertices of a triangle with side lengths $A := |v|$, $B := |w|$ and $C := |v - w|$. Let α, β, γ be the angles opposite to A, B, C. Show:

(a) (Law of sines) $\frac{\sin\alpha}{A} = \frac{\sin\beta}{B} = \frac{\sin\gamma}{C}$.

(b) (Law of cosines) $C^2 = A^2 + B^2 - 2AB\cos\gamma$.

(c) (Trigonometric Pythagoras) $\sin(\alpha)^2 + \cos(\alpha)^2 = 1$.

Exercise II.9. For $\zeta := \cos(\pi/5) + \mathrm{i}\sin(\pi/5) \in \mathbb{C}$ it holds that $\zeta^5 = -1$ (see proof of Lemma 11.27). Let $\omega := \zeta + \zeta^{-1} = 2\mathrm{Re}(\zeta) \in \mathbb{R}$. Show:

(a) $\omega^2 - \omega - 1 = 0$.
Hint: $X^5 + 1 = (X+1)(X^4 - X^3 + X^2 - X + 1)$.

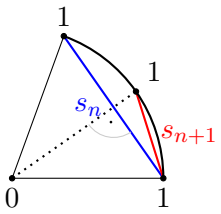
(b) $\omega = \frac{1}{2}(\sqrt{5} + 1)$.

(c)
$$D(\pi/5) = \frac{1}{4}\begin{pmatrix} \sqrt{5}+1 & -\sqrt{10-2\sqrt{5}} \\ \sqrt{10-2\sqrt{5}} & \sqrt{5}+1 \end{pmatrix}.$$
Hint: $\cos(\varphi)^2 + \sin(\varphi)^2 = 1$.

Exercise II.10. We approximate the semicircular arc length π by half the circumference of a regular 2^n-gon with "radius" 1. For this, let s_n be the side length of the regular 2^n-gon.

(a) Show $s_2 = \sqrt{2}$.

(b) Show $s_{n+1} = \sqrt{2 - \sqrt{4 - s_n^2}}$ by applying Pythagoras twice:



(c) Show
$$2^n \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \ldots \sqrt{2}}}}}_{n \text{ roots}} = \lim_{n\to\infty} 2^n s_{n+1} = \pi.$$

Exercise II.11. Let $V := \mathbb{R}^n$ be the Euclidean space wrt. the standard scalar product. Let $v \in V$ be normalized and $S_v \in \mathrm{O}(V)$ be the reflection across $v^\perp$. Show $[S_v] = 1_n - 2v^{\mathrm{t}}v$.
Remark: Such matrices are called *Householder transformations*.

Exercise II.12. Let $\alpha \in \mathbb{R}[X]$. Show that polynomials $\gamma_1, \ldots, \gamma_k \in \mathbb{R}[X]$ with $\alpha = \gamma_1 \ldots \gamma_k$ and $\deg(\gamma_i) \leq 2$ for $i = 1, \ldots, k$ exist.
Hint: Remark 11.37.

Exercise II.13. Show that in the Principal Axis Theorem and in the Spectral Theorem, the transformation matrix S can be chosen with $\det(S) = 1$.

Exercise II.14. Let V be an $\mathbb{R}$-vector space and $\beta \in \mathrm{Bil}(V)$ symmetric with $\mathrm{ind}(\beta) = (r, s, t)$. Show that r (resp. s) is the maximum dimension of a subspace $U \leq V$ such that the restriction of β to $U \times U$ is positive (resp. negative) definite.

Exercise II.15. Let $A \in \mathbb{R}^{n\times n}$ symmetric with $\mathrm{ind}(A) = (r, s, t)$ and $\lambda \in \mathbb{R}$. Show

$$\mathrm{ind}(\lambda A) = \begin{cases} (r, s, t) & \text{if } \lambda > 0, \\ (s, r, t) & \text{if } \lambda < 0, \\ (0, 0, n) & \text{if } \lambda = 0. \end{cases}$$

Exercise II.16. Let $A \in \mathbb{R}^{n\times n}$ symmetric. How can one read from χ_A whether A is positive semidefinite? Prove a criterion in analogy to Theorem 12.44.

Exercise II.17. Show for all $A \in \mathbb{C}^{n\times m}$:

(a) $\mathrm{Ker}(A) = \mathrm{Ker}(A^*A)$.

(b) $\mathrm{rk}(A) = \mathrm{rk}(A^*A)$.

Exercise II.18. A matrix $A \in \mathbb{C}^{n\times n}$ is called *positive (semi)definite*, if $vAv^* > 0$ (resp. $vAv^* \geq 0$) holds for all $v \in \mathbb{C}^n \setminus \{0\}$. Show that:

(a) Every positive (semi)definite matrix is Hermitian.
Remark: The analogous statement for real matrices is false.

(b) A Hermitian matrix A is positive (semi)definite if and only if all eigenvalues of A are positive (resp. non-negative).

(c) For $k \in \mathbb{N}$, every positive (semi)definite matrix has exactly one positive (semi)definite k-th root.

Exercise II.19. Let $A_1, \ldots, A_k \in K^{n\times n}$ be diagonalizable and pairwise commuting. Show that $A_1, \ldots, A_k$ are simultaneously diagonalizable.
Hint: The induction on k is non-trivial! One can use Exercise I.16.

Exercise II.20. Let $A_1, \ldots, A_k \in K^{n\times n}$ be pairwise commuting trigonalizable matrices. Show that $A_1, \ldots, A_k$ are simultaneously trigonalizable.

Exercise II.21 (Real Schur Decomposition)**.** Show that for every matrix $A \in \mathbb{R}^{n\times n}$ there exists an orthogonal matrix $Q \in \mathrm{O}(n, \mathbb{R})$ such that

$$Q^{\mathrm{t}}AQ = \begin{pmatrix} R_{11} & & * \\ & \ddots & \\ 0 & & R_{kk} \end{pmatrix}$$

with $R_{ii} \in \mathbb{R} \cup \mathbb{R}^{2\times 2}$ for $i = 1, \ldots, k$. In the case $R_{ii} \in \mathbb{R}^{2\times 2}$, R_{ii} has two complex conjugate (non-real) eigenvalues. In particular, $Q^{\mathrm{t}}AQ$ is an upper triangular matrix if A has only real eigenvalues.

Exercise II.22. Let $\lambda \in K$ and $A := J_n(\lambda)$ be a Jordan block. Show that $C \in K^{n\times n}$ commutes with A if and only if there exist $c_1, \ldots, c_n \in K$ with

$$C = \begin{pmatrix} c_1 & & & 0 \\ c_2 & \ddots & & \\ \vdots & \ddots & \ddots & \\ c_n & \cdots & c_2 & c_1 \end{pmatrix}.$$

Exercise II.23. Let V be a K-vector space and $f \in \mathrm{End}(V)$. Let $b_1, \ldots, b_n$ be a basis of V. Show that μ_f is the least common multiple of $\mu_{b_1}, \ldots, \mu_{b_n}$, i.e., there is no monic polynomial of smaller degree that is divisible by $\mu_{b_1}, \ldots, \mu_{b_n}$.

Exercise II.24. Let V be a K-vector space and $f \in \mathrm{End}(V)$. Let $\lambda \in K$ be an eigenvalue of f that occurs with multiplicity k as a root of μ_f. Show that

$$E_\lambda(f) \subsetneq \mathrm{Ker}((f - \lambda\,\mathrm{id})^2) \subsetneq \ldots \subsetneq \mathrm{Ker}((f - \lambda\,\mathrm{id})^k) = H_\lambda(f).$$

Hint: One can use the Weierstrass normal form.

Exercise II.25. Let $\alpha \in K[X] \setminus K$ be monic. Show that $\chi_{B(\alpha)} = \alpha$ using the definition of the characteristic polynomial (and not via the minimal polynomial as in Lemma 15.20).

Exercise II.26. Let $\alpha = (X - \lambda_1) \ldots (X - \lambda_n) \in K[X]$ with pairwise distinct $\lambda_1, \ldots, \lambda_n$. Let $V := (\lambda_i^{j-1}) \in K^{n\times n}$ be the Vandermonde matrix. Show that $VB(\alpha)V^{-1} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Exercise II.27. Let $\alpha = X^n + a_{n-1}X^{n-1} + \ldots + a_0$ and

$$S = \begin{pmatrix} a_1 & a_2 & \cdots & a_{n-1} & 1 \\ a_2 & a_3 & \iddots & 1 & \\ \vdots & \iddots & \iddots & & \\ a_{n-1} & 1 & & & \\ 1 & & & & 0 \end{pmatrix} \in \mathrm{GL}(n, K)$$

(a) Show that $S^{-1}B(\alpha)S = B(\alpha)^{\mathrm{t}}$.

(b) (Taussky) Conclude that every square matrix is the product of two symmetric matrices.

Exercise II.28. Let $A \in \mathbb{C}^{n\times n}$ with only real eigenvalues. Show that A is similar to a real matrix. *Hint:* Frobenius normal form.

Exercise II.29. Prove Corollary 14.37 over an arbitrary field K instead of $\mathbb{C}$.

Exercise II.30.

(a) Let $A \in K^{n\times n}$ and $B \in K^{m\times m}$ such that μ_A and μ_B are coprime. Show that

$$\mathrm{C}(\mathrm{diag}(A, B)) = \big\{\mathrm{diag}(C, D) : C \in \mathrm{C}(A),\ D \in \mathrm{C}(B)\big\} \cong \mathrm{C}(A) \times \mathrm{C}(B).$$

Hint: Bézout's identity.

(b) Let A be semisimple with $\chi_A = \gamma_1^{a_1} \dots \gamma_k^{a_k}$ (prime factorization). Show that

$$\dim \mathrm{C}(A) = \sum_{i=1}^{k} a_i^2 \deg(\gamma_i)$$

using (a). Compare with Remark 15.33.

Exercise II.31. Let V be a K-vector space and $F \subseteq \mathrm{End}(V)$. A subspace $U \leq V$ is called *F-invariant*, if $f(U) \subseteq U$ holds for all $f \in F$. Let

$$\mathrm{C}(F) := \bigcap_{f \in F} \mathrm{C}(f)$$

be the *centralizer* of F (analogously for matrices).

(a) Suppose $\{0\}$ and V are the only F-invariant subspaces. Show that all $g \in \mathrm{C}(F) \setminus \{0\}$ are invertible.

(b) Let $F = \{f_1, f_2\}$ with

$$f_1 := \mathrm{diag}\big(B(X^2+1), B(X^2+1)\big) \in \mathbb{R}^{4\times 4}, \qquad f_2 := \begin{pmatrix} 0_2 & B(X^2-1) \\ -B(X^2-1) & 0_2 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

Show that $\mathbb{H} := \mathrm{C}(F) \subseteq \mathbb{R}^{4\times 4}$ is a 4-dimensional $\mathbb{R}$-vector space in which every non-zero element is invertible.
Remark: In contrast to Theorem 15.39, the multiplication in $\mathbb{H}$ is not commutative. $\mathbb{H}$ is called the HAMILTONian *skew field*.

Linear Algebra III

17 Numerical Methods

17.1 Efficient Arithmetic

Remark 17.1. In Linear Algebra I and II, we investigated subjects mainly from a theoretical perspective. For example, eigenvalues were calculated as roots of the characteristic polynomial, although this is impractical for larger matrices (cf. Example 10.25). In this chapter, we describe algorithms with which one solves problems of linear algebra efficiently and robustly (against rounding errors) in practice. Eigenvalues are calculated more or less via the definition (Theorem 17.70).

Example 17.2. On computers, the multiplication of two numbers is generally more expensive than addition (there are, however, exceptions: multiplication by 2 corresponds to a shift of the binary sequence by one digit to the left). In the following runtime analyses, we will therefore neglect additions. The multiplication of two n-digit decimal numbers using the "school method" requires n^2 digit-multiplications (the multiplication table can be persistently stored on the chip):

$$\begin{array}{rr} 123 \cdot 567 & = \quad 69741 \\ \hline 861 & \\ + \quad 738 & \\ + \quad 615 & \\ \hline 69741 & \end{array}$$

This can be done faster.

Theorem 17.3 (Karatsuba Algorithm)**.** *Let $x, y \in \mathbb{N}$ be decimal numbers with n digits.*

(1) Divide x and y into two halves of length $m \approx n/2$:

$$x = x_1 10^m + x_0, \qquad y = y_1 10^m + y_0 \qquad (x_0, y_0 < 10^m)$$

(2) Recursively multiply the m-digit numbers:

$$z_0 := x_0 y_0, \qquad z_2 := x_1 y_1, \qquad z_1 := (x_1 - x_0)(y_0 - y_1) + z_0 + z_2.$$

(3) Then $xy = 10^{2m} z_2 + 10^m z_1 + z_0$ holds.

This algorithm requires approx. $n^{\log_2(3)} \approx n^{1.58} < n^2$ digit-multiplications.

Proof. It holds that $z_1 = x_1 y_0 + x_0 y_1 - x_1 y_1 - x_0 y_0 + z_0 + z_2 = x_1 y_0 + x_0 y_1$ and

$$xy = (x_1 10^m + x_0)(y_1 10^m + y_0) = 10^{2m} x_1 y_1 + 10^m (x_1 y_0 + x_0 y_1) + x_0 y_0 = 10^{2m} z_2 + 10^m z_1 + z_0.$$

For the second assertion, we argue by induction on n: For $n = 1$, one needs $1 = 1^{\log_2(3)}$ digit-multiplication. The calculation of z_0, z_1, z_2 for $n \geq 2$ requires 3 multiplications of m-digit numbers. Inductively, one obtains

$$3m^{\log_2(3)} \approx 3(n/2)^{\log_2(3)} = n^{\log_2(3)}$$

digit-multiplications. □

Example 17.4. For $x = 87 = 80 + 7$ and $y = 91 = 90 + 1$, one obtains

$$z_0 = 7 \cdot 1 = 7, \qquad z_2 = 8 \cdot 9 = 72, \qquad z_1 = (8-7)(1-9) + z_0 + z_2 = -8 + 7 + 72 = 71,$$
$$xy = 7200 + 710 + 7 = 7917.$$

Remark 17.5. Before the introduction of electronic calculators, the multiplication of numbers $x, y \in \mathbb{R}$ was reduced to the simpler addition $\log(x) + \log(y)$ by means of *logarithm tables*. The basis for this is the functional equation $\log(xy) = \log(x) + \log(y)$. A similar reduction with the help of the FOURIER transform is still relevant today.

Definition 17.6. For $n \in \mathbb{N}$ let

$$\zeta_n := \zeta := \cos(2\pi/n) + \mathrm{i}\sin(2\pi/n)$$

be an n-th root of unity (Example 11.28). The symmetric Vandermonde matrix

$$W_n := W := (\zeta^{ij})_{i,j=0}^{n-1} \in \mathbb{C}^{n\times n}$$

is called the n-th *Fourier matrix*. The mapping $\mathcal{F}_n \colon \mathbb{C}^n \to \mathbb{C}^n$, $x \mapsto xW =: \hat{x}$ is called the *discrete Fourier transform*.

Example 17.7. According to Example 11.28, it holds that

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}.$$

Lemma 17.8. *For $x \in \mathbb{C}^n$ it holds that $\mathcal{F}_n^{-1}(x) = \frac{1}{n}\overline{\mathcal{F}_n(\overline{x})}$.*

Proof. Due to $\zeta\overline{\zeta} = |\zeta|^2 = 1$, we have $\zeta^{-1} = \overline{\zeta}$. For $1 \le i, j \le n$, $\sigma := \zeta^{i-j}$ is also an n-th root of unity. In the case $i = j$, $\sigma = 1$ and otherwise $\sum_{k=0}^{n-1} \sigma^k = \frac{\sigma^n - 1}{\sigma - 1} = 0$ according to the formula for the geometric series. Therefore,

$$(W_n\overline{W_n})_{ij} = \sum_{k=0}^{n-1} \zeta^{ik}\zeta^{-kj} = \sum_{k=0}^{n-1} \sigma^k = \begin{cases} n & \text{if } i = j, \\ 0 & \text{if } i \ne j. \end{cases}$$

This shows $W_n^{-1} = \frac{1}{n}\overline{W_n}$ and $n\mathcal{F}_n^{-1}(x) = x\overline{W_n} = \overline{\mathcal{F}_n(\overline{x})}$. □

Theorem 17.9 (Convolution Theorem)**.** *For $\alpha = \sum a_k X^{k-1} \in \mathbb{C}[X]$ let $[\alpha] := (a_1, \ldots, a_n)$. For all $\alpha, \beta \in \mathbb{C}[X]$ with $\deg(\alpha) + \deg(\beta) < n$ it holds that*

$$\mathcal{F}_n([\alpha\beta]) = (\hat{a}_1\hat{b}_1, \ldots, \hat{a}_n\hat{b}_n).$$

Proof. Due to

$$\alpha\beta = \sum_{k=2}^{n+1}\Big(\sum_{i+j=k} a_i b_j\Big)X^{k-2} = \sum_{k=1}^{n}\Big(\sum_{i+j=k+1} a_i b_j\Big)X^{k-1}$$

it holds that $[\alpha\beta]_k = \sum_{i+j=k+1} a_i b_j$ for $1 \le k \le n$. It follows that

$$\hat{a}_k\hat{b}_k = \sum_{i,j=1}^{n} a_i b_j \zeta^{(k-1)(i+j-2)} = \sum_{l=2}^{2n}\Big(\zeta^{(k-1)(l-2)} \sum_{i+j=l} a_i b_j\Big) = \sum_{l=1}^{n} [\alpha\beta]_l \zeta^{(k-1)(l-1)} = \mathcal{F}_n([\alpha\beta])_k. \qquad \square$$

Remark 17.10.

(a) To multiply two numbers $a, b \in \mathbb{N}$ (with $ab < 10^n$), one considers them as polynomials in $X = 10$ as $a = \sum a_k 10^{k-1}$, $b = \sum b_k 10^{k-1}$ and applies the convolution theorem:

$$ab = \mathcal{F}_n^{-1}(\hat{a}_1\hat{b}_1, \ldots, \hat{a}_n\hat{b}_n) \cdot (1, 10, \ldots, 10^{n-1})^{\mathrm{t}}.$$

Note that carries, as in school multiplication, are not resolved here. Carries can generally be avoided by representing a and b *binarily*, i.e., as a polynomial in $X = 2$. However, this type of calculation does not yet provide any speedup.

(b) Let $n = 2m$ be even. With the notation $y_k := x_{2k-1}$ and $z_k := x_{2k}$, it holds that

$$\begin{aligned} \mathcal{F}_n(x)_i &= \sum_{k=1}^{n} x_k \zeta_n^{(i-1)(k-1)} = \sum_{k=1}^{m} y_k \zeta_m^{(i-1)(k-1)} + \zeta_n^{i-1} \sum_{k=1}^{m} z_k \zeta_m^{(i-1)(k-1)} \\ &= \mathcal{F}_m(y)_i + \zeta_n^{i-1} \mathcal{F}_m(z)_i \end{aligned} \tag{17.1}$$

for $i = 1, \ldots, n$, where $\mathcal{F}_m(y)_i = \mathcal{F}_m(y)_{i-m}$ for $i > m$. With Lemma 17.8, analogously

$$\mathcal{F}_n^{-1}(x)_i = \frac{1}{2}\Big(\mathcal{F}_m^{-1}(y)_i + \zeta_n^{1-i} \mathcal{F}_m^{-1}(z)_i\Big).$$

In the *fast* Fourier transform (FFT), one assumes that n is a power of 2 (which is always possible by appending zeros). As with the Karatsuba algorithm, one can reduce $\mathcal{F}_n$ down to $\mathcal{F}_1(x) = x$. For a fixed n, the n-th roots of unity can be calculated once and stored. The addition of such numbers can be performed efficiently with limited precision.

(c) The widespread Schönhage-Strassen algorithm performs the multiplication of n-digit numbers using FFT in an asymptotic runtime of $n \log(n) \log\log(n)$. In 2019, Harvey-van der Hoeven improved the runtime to $n \log(n)$, although this is a *galactic* algorithm, i.e., the time savings only become relevant for extremely large numbers beyond any practical significance. It is conjectured that $n \log(n)$ is asymptotically the best possible bound.

(d) The *continuous* Fourier transform assigns to an integrable function $f\colon \mathbb{R}^n \to \mathbb{R}$ the integral

$$\mathcal{F}(y) = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} f(x) e^{2\pi \mathrm{i}[x,y]} \mathrm{d}x.$$

Example 17.11. We multiply $342 \cdot 87$ using the fast Fourier transform. According to (17.1), we have

$$\mathcal{F}_4(2,4,3,0)^{\mathrm{t}} = \begin{pmatrix} \mathcal{F}_2(2,3)_1 + \mathcal{F}_2(4,0)_1 \\ \mathcal{F}_2(2,3)_2 + \mathrm{i}\mathcal{F}_2(4,0)_2 \\ \mathcal{F}_2(2,3)_1 - \mathcal{F}_2(4,0)_1 \\ \mathcal{F}_2(2,3)_2 - \mathrm{i}\mathcal{F}_2(4,0)_2 \end{pmatrix} = \begin{pmatrix} (2+3)+4 \\ (2-3)+4\mathrm{i} \\ (2+3)-4 \\ (2-4)-4\mathrm{i} \end{pmatrix} = \begin{pmatrix} 9 \\ -1+4\mathrm{i} \\ 1 \\ -1-4\mathrm{i} \end{pmatrix},$$

$$\mathcal{F}_4(7,8,0,0) = (15, 7+8\mathrm{i}, -1, 7-8\mathrm{i}),$$

$$\mathcal{F}_4^{-1}(135, -39+20\mathrm{i}, -1, -39-20\mathrm{i})^{\mathrm{t}} = \frac{1}{4} \begin{pmatrix} 135-1-2\cdot 39 \\ 135+1-\mathrm{i}40\mathrm{i} \\ 135-1+2\cdot 39 \\ 135+1+\mathrm{i}40\mathrm{i} \end{pmatrix} = \big(14, 44, 53, 24\big),$$

$$342 \cdot 87 = \mathcal{F}_4^{-1}(\ldots)(1, 10, 100, 1000)^{\mathrm{t}} = 14 + 440 + 5.300 + 24.000 = 29.754.$$

Remark 17.12. The direct multiplication of two $n \times n$ matrices A, B requires n^3 number multiplications $a_{ij}b_{jk}$ with $1 \leq i, j, k \leq n$, which in turn can be performed using one of the above algorithms. This can also be improved.

Theorem 17.13 (STRASSEN algorithm)**.** *Let $A, B \in K^{n\times n}$.*

(1) By appending a zero row and column, one can assume $n = 2m$.

(2) Divide A and B into $m \times m$ blocks:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

(3) Compute recursively:

$$\begin{aligned} C_1 &:= (A_{11} + A_{22})(B_{11} + B_{22}), & C_2 &:= (A_{21} + A_{22})B_{11}, \\ C_3 &:= A_{11}(B_{12} - B_{22}), & C_4 &:= A_{22}(B_{21} - B_{11}), \\ C_5 &:= (A_{11} + A_{12})B_{22}, & C_6 &:= (A_{21} - A_{11})(B_{11} + B_{12}), \\ C_7 &:= (A_{12} - A_{22})(B_{21} + B_{22}). \end{aligned}$$

(4) Then

$$AB = \begin{pmatrix} C_1 + C_4 - C_5 + C_7 & C_3 + C_5 \\ C_2 + C_4 & C_1 - C_2 + C_3 + C_6 \end{pmatrix}.$$

This algorithm requires approximately $n^{\log_2(7)} \approx n^{2.81}$ number multiplications.

Proof. The formula for AB follows from

$$\begin{aligned} C_1 + C_4 - C_5 + C_7 &= (A_{11} + A_{22})(B_{11} + B_{22}) + A_{22}(B_{21} - B_{11}) - (A_{11} + A_{12})B_{22} \\ &\quad + (A_{12} - A_{22})(B_{21} + B_{22}) = A_{11}B_{11} + A_{12}B_{21}, \\ C_3 + C_5 &= A_{11}(B_{12} - B_{22}) + (A_{11} + A_{12})B_{22} = A_{11}B_{12} + A_{12}B_{22}, \\ C_2 + C_4 &= (A_{21} + A_{22})B_{11} + A_{22}(B_{21} - B_{11}) = A_{21}B_{11} + A_{22}B_{21}, \\ C_1 - C_2 + C_3 + C_6 &= (A_{11} + A_{22})(B_{11} + B_{22}) - (A_{21} + A_{22})B_{11} + A_{11}(B_{12} - B_{22}) \\ &\quad + (A_{21} - A_{11})(B_{11} + B_{12}) = A_{21}B_{12} + A_{22}B_{22}. \end{aligned}$$

The calculation of each C_i inductively requires approximately $m^{\log_2(7)} = \frac{1}{7}n^{\log_2(7)}$ number multiplications. Therefore, one needs approximately $n^{\log_2(7)}$ number multiplications for AB. □

Remark 17.14.

(a) Strassen's algorithm has also been further improved. Most recently, in 2025, an algorithm with $n^{2.371339}$ scalar multiplications was discovered.[1] Currently, *Large Language Models* such as AlphaEvolve are being used. The asymptotically best possible complexity is an open problem in computer science.

(b) Matrix powers can be calculated efficiently by iterated squaring: $A^{11} = A(A^5)^2 = A(A(A^2)^2)^2$ requires only five multiplications instead of ten (Exercise III.1).

[1] See More Asymmetry Yields Faster Matrix Multiplication

(c) The multiplication of arbitrary (not necessarily square) matrices $A \in K^{n\times m}$ and $B \in K^{m\times k}$ requires nmk scalar multiplications. If more than two such matrices are multiplied, the placement of parentheses has a significant influence on the number of scalar multiplications. This is particularly evident in the extreme case $n = k > 1 = m = l$:

$$(AB)C = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{11}b_{1n} \\ \vdots & & \vdots \\ a_{n1}b_{11} & \cdots & a_{n1}b_{1n} \end{pmatrix} \begin{pmatrix} c_{11} \\ \vdots \\ c_{n1} \end{pmatrix} \qquad (2n^2 \text{ multiplications}),$$

$$A(BC) = \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} (b_{11}c_{11} + \ldots + b_{1n}c_{11}) \qquad (2n \text{ multiplications}).$$

Theorem 17.15. *Let $A \in K^{n\times m}$, $B \in K^{m\times k}$ and $C \in K^{k\times l}$. The calculation of $A(BC)$ requires fewer scalar multiplications than $(AB)C$ if and only if $\frac{1}{n} + \frac{1}{k} < \frac{1}{m} + \frac{1}{l}$.*

Proof. The calculation of $A(BC)$ and $(AB)C$ requires $mkl+nml = ml(n+k)$ and $nmk+nkl = nk(m+l)$ scalar multiplications, respectively. It holds that

$$ml(n+k) < nk(m+l) \iff \frac{n+k}{nk} < \frac{m+l}{ml} \iff \frac{1}{n} + \frac{1}{k} < \frac{1}{m} + \frac{1}{l}. \qquad \square$$

Remark 17.16. For the concrete implementation, it is recommended to use well-established libraries such as BLAS, LAPACK, Armadillo, NumPy and SciPy, programming languages such as julia or programs such as MATLAB, Scilab and Octave.

17.2 The Condition Number

Example 17.17. Even though linear maps are continuous in the sense of analysis, small changes in the arguments can still have large effects on the values (in the ϵ-δ-definition of continuity, $\delta \ll \epsilon$ must be chosen):

$$Ax = \begin{pmatrix} 9 & 8 \\ 8 & 7 \end{pmatrix} x = \begin{pmatrix} 17 \\ 15 \end{pmatrix} \Longrightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$
$$\begin{pmatrix} 9 & 8 \\ 8 & 7 \end{pmatrix} x = \begin{pmatrix} 17 \\ 15.\mathbf{1} \end{pmatrix} \Longrightarrow x = \begin{pmatrix} 1.8 \\ 0.1 \end{pmatrix}.$$

Definition 17.18. For $A \in \mathbb{C}^{n\times m}$ one calls

$$\kappa(A) := \frac{\max\{|Ax| : x \in \mathbb{C}^{m\times 1}, |x| = 1\}}{\min\{|Ax| : x \in \mathbb{C}^{m\times 1}, |x| = 1\}} \geq 1$$

the *condition number* of A, where $|.|$ denotes the norm of the standard inner product (Example 13.3). If $\kappa(A)$ is "small" (or "large"), then A is *well* (or *ill*) *conditioned*.

Remark 17.19.

(a) We want to measure how small changes to a vector b affect the solution of the system $Ax = b$. For this, let $\tilde{b} \approx b$ and $A\tilde{x} = \tilde{b}$. The relative error of x is

$$\frac{|x - \tilde{x}|}{|x|} = \frac{|x - \tilde{x}|}{|A(x - \tilde{x})|} \frac{|Ax|}{|x|} \frac{|b - \tilde{b}|}{|b|} \leq \frac{\max\{|Ax|/|x| : 0 \neq x \in \mathbb{C}^{m\times 1}\}}{\min\{|Ay|/|y| : 0 \neq y \in \mathbb{C}^{m\times 1}\}} \frac{|b - \tilde{b}|}{|b|} = \kappa(A) \frac{|b - \tilde{b}|}{|b|}.$$

The condition number is thus the maximum factor by which the relative error of b can be amplified to x.

(b) In analysis, one shows that the map $\mathbb{C}^m \to \mathbb{C}$, $x \mapsto |Ax|$ is continuous. Since the set $\{x \in \mathbb{C}^m : |x| = 1\}$ is compact, the maximum/minimum over $\{|Ax| : |x| = 1\}$ is actually attained (i.e., one does not need a supremum/infimum). For $\mathbb{R}$ instead of $\mathbb{C}$, the condition number describes how much a matrix deforms the n-dimensional unit ball.

Example 17.20.

(a) If A does not have full rank, then there exists a normalized $x \in \mathbb{C}^{n\times 1}$ with $Ax = 0$. In this case, we interpret $\kappa(A) = \infty$, i.e., A is particularly ill-conditioned.

(b) For unitary matrices $S \in \mathrm{U}(n, \mathbb{C})$, it is well known that $|Sx| = |x|$ for all $x \in \mathbb{C}^{n\times 1}$. For arbitrary $A \in \mathbb{C}^{n\times m}$, it follows that $\kappa(SA) = \kappa(A)$ and analogously $\kappa(AS) = \kappa(A)$, if $S \in \mathrm{U}(m, \mathbb{C})$. In particular, $\kappa(S) = \kappa(1_n) = 1$, i.e., unitary matrices are well-conditioned.

(c) For arbitrary matrices $A, S \in \mathbb{C}^{n\times n}$, in general $\kappa(SA) \neq \kappa(A)$ holds. In this way, one can improve the condition number of the coefficient matrix of a system of equations $Ax = b$. However, the necessary replacement of b by Sb can be error-prone for the same reason.

(d) A classic example of an ill-conditioned matrix is the symmetric *Hilbert matrix*

$$H_n := \left(\frac{1}{i + j - 1}\right)_{i,j} = \begin{pmatrix} 1 & 1/2 & \cdots & 1/n \\ 1/2 & 1/3 & \cdots & 1/n+1 \\ \vdots & \vdots & & \vdots \\ 1/n & 1/n+1 & \cdots & 1/2n-1 \end{pmatrix}.$$

One can show that $\kappa(H_n)$ grows exponentially in n. For example, $\kappa(H_4) \approx 15.514$.

(e) For normal matrices A, one can use the spectral theorem and (b) to reduce the calculation of $\kappa(A)$ to a diagonal matrix. The following theorem allows this for arbitrary (especially non-square) matrices.

Theorem 17.21 (Singular Value Decomposition)**.** *For $A \in \mathbb{C}^{n\times m}$, there exist $U \in \mathrm{U}(n, \mathbb{C})$ and $V \in \mathrm{U}(m, \mathbb{C})$ such that UAV is a real diagonal matrix[2] with non-negative entries. The positive entries (on the main diagonal) are called* singular values *of A. They are uniquely determined up to their order.*

Proof. Existence: The positive semidefinite matrix $B := A^*A \in \mathbb{C}^{m\times m}$ has, according to Exercise II.18, only non-negative eigenvalues

$$\lambda_1 \geq \ldots \geq \lambda_k > \lambda_{k+1} = \ldots = \lambda_m = 0,$$

[2]not necessarily square

where $k = \operatorname{rk}(B) = \operatorname{rk}(A) \leq \min\{n, m\}$ according to Exercise II.17. By the spectral theorem, there exists a $V \in \mathrm{U}(m, \mathbb{C})$ with $V^*BV = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. We write $V = (V_1, V_2)$ with $V_1 \in \mathbb{C}^{m \times k}$. From $(AV_2)^*(AV_2) = V_2^*BV_2 = 0$ it follows that $AV_2 = 0$. We set

$$U_1 := AV_1 \operatorname{diag}\bigl(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}\bigr) \in \mathbb{C}^{n \times k}.$$

Because $U_1^*U_1 = 1_k$, one can extend U_1 to $U = (U_1, U_2) \in \mathrm{U}(n, \mathbb{C})$ using Gram-Schmidt. Now

$$\begin{aligned} U_2^*AV_1 &= U_2^*U_1 \operatorname{diag}\bigl(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\bigr) = 0, \\ U^*AV &= \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix} A \begin{pmatrix} V_1 & V_2 \end{pmatrix} = \begin{pmatrix} U_1^*AV_1 & U_1^*AV_2 \\ U_2^*AV_1 & U_2^*AV_2 \end{pmatrix} = (\delta_{ij}\sqrt{\lambda_i}), \end{aligned}$$

where $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}$ are the singular values of A.

Uniqueness: Let SAT with $S \in \mathrm{U}(n, \mathbb{C})$ and $T \in \mathrm{U}(m, \mathbb{C})$ also be a real diagonal matrix with non-negative entries. Let $\mu_1, \ldots, \mu_l$ be the positive entries. Then $\mu_1^2, \ldots, \mu_l^2$ are the positive eigenvalues of

$$(SAT)^2 = (SAT)^*(SAT) = T^*BT.$$

With appropriate numbering, it follows that $k = l$ and $\mu_1^2 = \lambda_1, \ldots, \mu_k^2 = \lambda_k$. From $\mu_i > 0$ it follows that $\mu_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, k$. Thus, the singular values are uniquely determined up to their order. □

Remark 17.22. The proof shows: If $\lambda_1, \ldots, \lambda_k$ are the non-zero eigenvalues of the positive semidefinite matrix A^*A, then $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}$ are the singular values of A.

Corollary 17.23. *Let $A \in \mathbb{C}^{n \times m}$ have full rank and singular values $\sigma_1 \geq \ldots \geq \sigma_k$. Then $\kappa(A) = \sigma_1/\sigma_k$ holds.*

Proof. Let $D := (\delta_{ij}\sigma_i) \in \mathbb{R}^{n \times m}$ be the matrix from the singular value decomposition of A. According to Example 17.20, $\kappa(A) = \kappa(D) = \kappa(\operatorname{diag}(\sigma_1, \ldots, \sigma_k))$ holds. For $x \in \mathbb{C}^k$ with $|x| = 1$, we have

$$|(\sigma_1 x_1, \ldots, \sigma_k x_k)|^2 = \sigma_1^2|x_1|^2 + \ldots + \sigma_k^2|x_k|^2 \leq \sigma_1^2|x|^2 = \sigma_1^2$$

with equality for $x = e_1$. This shows $\kappa(\operatorname{diag}(\sigma_1, \ldots, \sigma_k)) = \sigma_1/\sigma_k$. □

Corollary 17.24. *For all $A \in \mathbb{C}^{n \times m}$, $\kappa(A) = \kappa(A^*)$ holds.*

Proof. According to Lemma 10.38, A^*A and AA^* have the same non-zero eigenvalues. Therefore, the assertion follows from Remark 17.22. □

Theorem 17.25. *If $A \in \mathbb{C}^{n \times n}$ is normal, then the singular values of A are the absolute values of the non-zero eigenvalues.*

Proof. According to the spectral theorem, one can assume $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with the eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Let

$$\tilde{\lambda}_i := \begin{cases} \overline{\lambda_i}/|\lambda_i| & \text{if } \lambda_i \neq 0, \\ 1 & \text{if } \lambda_i = 0. \end{cases}$$

Then $U := \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n) \in \mathrm{U}(n, \mathbb{C})$ and $UA = \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_n|)$. The claim follows from the uniqueness of the singular values. □

Example 17.26. For the matrix from Example 17.17, $\kappa(A) = \frac{8+\sqrt{65}}{8-\sqrt{65}} \approx 258$ holds.

Theorem 17.27 (Moore-Penrose)**.** *For $A \in \mathbb{C}^{n\times m}$ there exists exactly one $A^+ \in \mathbb{C}^{m\times n}$ with the following properties:*

(a) $AA^+A = A$ *and* $A^+AA^+ = A^+$,

(b) $(AA^+)^* = AA^+$ *and* $(A^+A)^* = A^+A$.

One calls A^+ the pseudoinverse[3] *of A.*

Proof. Let $P := UAV = (\delta_{ij}\sigma_i)$ be a singular value decomposition of A with singular values $\sigma_1, \ldots, \sigma_k > 0$. For $i = 1, \ldots, m$ we define

$$\tilde{\sigma}_i := \begin{cases} \sigma^{-1} & \text{if } i \le k, \\ 0 & \text{if } i > k \end{cases}$$

and $Q := (\delta_{ij}\tilde{\sigma}_i) \in \mathbb{C}^{m\times n}$. Set $A^+ := VQU \in \mathbb{C}^{m\times n}$. The four stated properties reduce to $PQP = P$, $QPQ = Q$, $(PQ)^* = PQ$ and $(QP)^* = QP$. This obviously holds. Now let B and C be pseudoinverses of A. Then

$$B = BAB = B(ACA)B = (BA)^*(CA)^*B = (ABA)^*C^*B = (CA)^*B = CAB = \ldots = CAC = C. \quad \square$$

Example 17.28.

(a) Trivially, $0^+_{n\times m} = 0_{m\times n}$.

(b) If A is invertible, then A^{-1} satisfies the conditions of the pseudoinverse and it follows that $A^+ = A^{-1}$.

(c) For $A \in \mathbb{C}^{n\times m}$ and $\lambda \in \mathbb{C}^\times$, we have $(\lambda A)^+ = \lambda^{-1}A^+$.

(d) The proof of Theorem 17.27 reduces the calculation of A^+ to the singular value decomposition and thus to the spectral theorem. For $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, the principal axis theorem yields $S := \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$ with $S^{-1} = S^{\mathrm{t}} = S$ and $SAS = \operatorname{diag}(2, 0)$. Thus $A^+ = S \operatorname{diag}(1/2, 0)S = \frac{1}{4}A$.

(e) Apparently, $(A^{\mathrm{t}})^+ = (A^+)^{\mathrm{t}}$ and $\overline{A}^+ = \overline{A^+}$. In particular, A^+ is symmetric (resp. real, Hermitian) if A is symmetric (resp. real, Hermitian).

(f) If $A \in \mathbb{C}^{n\times m}$ has full rank, then A^*A (if $n \ge m$) or AA^* (if $n \le m$) is invertible. In these cases, $A^+ = (A^*A)^{-1}A^*$ or $A^+ = A^*(AA^*)^{-1}$ holds. For $n \ge m$, one can calculate an (approximate) solution of the system of equations $Ax = b$ in this way: $x \approx A^+b$. Example:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 3 \\ 5 \\ 7 \end{pmatrix}, \qquad A^+ = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} A^* = \frac{1}{3}\begin{pmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{pmatrix}, \qquad x \approx \frac{1}{3}\begin{pmatrix} 8 \\ 14 \end{pmatrix}.$$

More on this in Theorem 17.42.

[3] or *Moore-Penrose inverse*

17.3 Stable Variants of the Gaussian Elimination

Remark 17.29. On computers, real numbers can only be represented approximately by *floating-point numbers*. Even finite decimal fractions cannot be stored exactly if the internal binary representation is infinite:

$$0.1 = 2^{-4} + 2^{-5} + 2^{-8} + 2^{-9} + \ldots \approx 0.0996.$$

The IEEE Standard 754 defines the 32-bit data type `float` (simple precision) by the following division: 1 bit for the sign, 23 bits for the *mantissa* (binary expansion), and 8 bits for the exponent. This allows numbers of the form

$$\pm\Big(2^e + \sum_{i=1}^{23} a_i 2^{e-i}\Big) \qquad (a_i \in \{0,1\},\ -126 \le e \le 127)$$

to be represented ($e = -127$ and $e = 128$ are reserved for ∞ and `NaN`). In the following, we consider a simpler model with only three decimal digits, i.e., numbers of the form

$$\pm \sum_{i=0}^{2} a_i 10^{e-i} \qquad (0 \le a_i \le 9)$$

with $a_0 \neq 0$ (since overflows and underflows occur rarely, we do not restrict the exponent e). Arbitrary real numbers are converted as follows:

$$1001 \rightsquigarrow 1.00\text{e}4, \qquad 0.00123 \rightsquigarrow 1.23\text{e}{-3}.$$

Subtraction of numbers of almost the same size leads to a loss of precision (*cancellation*):

$$1.23\text{e}0 - 1.22\text{e}0 = 1.??\text{e}{-3}.$$

The digits marked with a question mark contain no meaningful information. Furthermore, the addition of floating-point numbers is not associative:

$$(1.00\text{e}{-3} + 1.00\text{e}1) - 1.00\text{e}1 = 0.00\text{e}0 \neq 1.00\text{e}{-3} = 1.00\text{e}{-3} + (1.00\text{e}1 - 1.00\text{e}1).$$

Equally problematic is multiplication by large numbers, because existing rounding errors are amplified as a result.

Example 17.30. The system of equations

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 10^4 \\ 1 \end{pmatrix}$$

has the unique solution $x = (-10^4, 10^4 + 1)$. However, the Gaussian elimination with floating-point numbers yields

$$\left(\begin{array}{cc|c} 1\text{e}{-4} & 1 & 1\text{e}4 \\ 1 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1\text{e}4 & 1\text{e}8 \\ 1 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1\text{e}4 & 1\text{e}8 \\ 0 & -1\text{e}4 & -1\text{e}8 \end{array}\right) \quad \Longrightarrow \quad x = (0, 1\text{e}4).$$

The Gaussian elimination in its pure form is thus numerically *unstable*.

Remark 17.31. One approach to improve the stability of a system of equations is to perform the singular value decomposition of the coefficient matrix A. In this process, A is multiplied by unitary matrices U and V from the left/right. As a first approximation, one can choose permutation matrices for U and V. This realizes permutations of the rows and columns of A.

Theorem 17.32 (Pivoting)**.** *Given the linear system of equations $Ax = b$ with $A \in \mathbb{C}^{n\times m}$ and $b \in \mathbb{C}^{n\times 1}$. The following modification of the Gaussian elimination on $(A|b)$ reduces rounding errors:*

(1) Set $z := 1$ (row index).

(2) For $s = 1, \dots, m$ (column index):

- *Determine an element a_{ij} with the largest absolute value* (pivot) *with $i \geq z$ and $j \geq s$.*
- *If $a_{ij} = 0$, then we are finished. Otherwise, swap the i-th with the z-th row and the j-th with the s-th column. Swap x_j with x_s in the solution vector.*
- *Divide the z-th row by a_{zs}.*
- *Eliminate as usual $a_{ws} = 0$ for $w \neq z$.*
- *Increase z by 1.*

Proof. If one applies the specified procedure as in Theorem 6.15, one obtains the particularly simple augmented matrix

$$M = \begin{pmatrix} 1 & & & s_{11} & \cdots & s_{1,m-k} & c_1 \\ & \ddots & & \vdots & & \vdots & \vdots \\ & & 1 & s_{k1} & \cdots & s_{k,m-k} & c_k \\ & & & -1 & & & 0 \\ & & & & \ddots & & \vdots \\ & & & & & -1 & 0 \end{pmatrix}$$

with the solution set

$$L = \begin{pmatrix} c_1 \\ \vdots \\ c_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \langle \begin{pmatrix} s_{11} \\ \vdots \\ s_{k1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} s_{1,m-k} \\ \vdots \\ s_{k,m-k} \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \rangle.$$

The solution of the original system is obtained by permuting the coordinates according to the chosen swaps $j \leftrightarrow s$. The choice of the pivot guarantees that during the algorithm, one does not multiply by large numbers that would amplify potential rounding errors. □

Example 17.33. In Example 17.30 we swap the columns:

$$\left(\begin{array}{cc|c} 1 & 1\text{e}{-4} & 1\text{e}4 \\ 1 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1\text{e}{-4} & 1\text{e}4 \\ 0 & 1 & -1\text{e}4 \end{array}\right) \quad \Longrightarrow \quad \begin{cases} y = (1\text{e}4, -1\text{e}4), \\ x = (-1\text{e}4, 1\text{e}4) \approx (-10^4, 10^4 + 1). \end{cases}$$

Remark 17.34. For arbitrary fields, the following theorem states that pivoting theoretically only needs to be performed once at the beginning.

Theorem 17.35 (LU decomposition[4])**.** *For every matrix $A \in K^{n\times n}$, there exist a permutation matrix P, a lower triangular matrix L with ones on the main diagonal, and an upper triangular matrix U with $A = PLU$.*

[4] lower-upper.

Proof. Induction on n: For $n = 1$, one can choose $P = L = 1_1$ and $U = A$. Let $n \geq 2$. If the first column of A is the zero vector, we set $P_1 = L_1 = 1_n$. Otherwise, there exists a permutation matrix P_1 that swaps the first row of A with another row, such that subsequently $a_{11} \neq 0$ holds. The elimination of a_{i1} for $i > 1$ is achieved by multiplication with elementary matrices from the left. The product of these elementary matrices has the form $L_1 = \left(\begin{smallmatrix} 1 & 0 \\ * & 1_{n-1} \end{smallmatrix}\right)$. Now it holds that

$$L_1 P_1 A = \begin{pmatrix} a_{11} & * \\ 0 & A_2 \end{pmatrix}$$

with $A_2 \in K^{(n-1)\times(n-1)}$. By induction, there exist P_2, L_2 and U_2 with $A_2 = P_2 L_2 U_2$. Let $\hat{P}_2 := \operatorname{diag}(1, P_2)$ and $\hat{L}_2 := \operatorname{diag}(1, L_2)$. We can also extend U_2 to an upper triangular matrix $\hat{U}_2$ such that $L_1 P_1 A = \hat{P}_2 \hat{L}_2 \hat{U}_2$ holds. The inverse L_1^{-1} has the same form as L_1 (only the signs of the entries below the main diagonal are inverted). Since $\hat{P}_2$ does not swap the first row, $L_1^{-1} \hat{P}_2 = \hat{P}_2 L_1'$ holds for a lower triangular matrix L_1' with ones on the main diagonal. Overall, it holds that

$$A = P_1^{-1} L_1^{-1} \hat{P}_2 \hat{L}_2 \hat{U}_2 = P_1 \hat{P}_2 L_1' \hat{L}_2 \hat{U}_2 = PLU. \qquad \square$$

Example 17.36.

$$\begin{pmatrix} 0 & 0 & 2 \\ 2 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix} = P_{(1,2)} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 2 & 1 & -1 \end{pmatrix} = P_{(1,2)} \begin{pmatrix} 1 & . & . \\ . & 1 & . \\ 1 & . & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 1 & -2 \end{pmatrix}$$
$$= P_{(1,2)} L_1 P_{(2,3)} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 2 \end{pmatrix} = P_{(1,2)} P_{(2,3)} \begin{pmatrix} 1 & . & . \\ 1 & 1 & . \\ . & . & 1 \end{pmatrix} R = P_{(1,2,3)} LU$$

Remark 17.37.

(a) Since there are usually many possibilities for pivoting in the Gaussian algorithm, the LU decomposition is not unique. However, if one can dispense with P, the decomposition becomes unique (Exercise III.9).

(b) Let $A \in \mathrm{GL}(n, K)$. The following algorithm provides a variant of the LU decomposition where P stands between L and U: Choose the smallest j_1 with $a_{1,j_1} \neq 0$ (exists since A is invertible). By left multiplication with a lower triangular matrix L_1 with ones on the main diagonal, one achieves $a_{i,j_1} = 0$ for $i \geq 2$. Analogously, one obtains $a_{1,j_1} = 1$ and $a_{1j} = 0$ for $j > j_1$ by right multiplication with an upper triangular matrix. Now choose the minimal j_2 with $a_{2,j_2} \neq 0$ and proceed as before. In the end, one obtains a lower triangular matrix $L' := L_n \dots L_1$ with ones on the main diagonal and an upper triangular matrix $U' := U_1 \dots U_n$, such that $L'AU' = P_\sigma$ is the permutation matrix with $\sigma^{-1}(k) = j_k$ for $k = 1, \dots, n$. Thus, one has a decomposition of the form $A = LPU$ (cf. Exercise III.10).

(c) The next theorem provides a further approximation to the singular value decomposition.

Theorem 17.38 (QR decomposition)**.** *For every $A \in \mathbb{C}^{n\times m}$ there exist $Q \in \mathrm{U}(n, \mathbb{C})$ and an upper triangular matrix $R \in \mathbb{C}^{n\times m}$ with real, non-negative diagonal entries and $A = QR$.*

(a) If A has full rank, then R is uniquely determined.

(b) If A is real, then one can choose $Q \in \mathrm{O}(n, \mathbb{R})$ and $R \in \mathbb{R}^{n\times m}$.

Proof. Let $a_1, \dots, a_m$ be the columns of A and $s := \mathrm{rk}(A)$. We choose from left to right the first s linearly independent columns $a_{i_1}, \dots, a_{i_s}$ of A. Every further column a_j can then be represented as a linear combination of the a_{i_k} with $i_k < j$. We complement $a_{i_1}, \dots, a_{i_s}$ with real vectors to a basis of $\mathbb{C}^n$. The Gram-Schmidt process (Remark 13.6) transforms this basis into an orthonormal basis $q_1, \dots, q_n$ with $q_1 = \frac{1}{|a_{i_1}|} a_{i_1}$, $q_2 = \lambda a_{i_1} + \mu a_{i_2}$ with $\mu \in \mathbb{R}_{>0}$ etc. By rearranging, one obtains

$$a_{i_k} = \sum_{j=1}^{k} \lambda_j q_j \qquad (\lambda_k \in \mathbb{R}_{>0}).$$

By representing the remaining columns of A also wrt. $q_1, \dots, q_n$, one obtains an upper triangular matrix $R = (r_{ij}) \in \mathbb{C}^{n \times m}$ with $a_k = \sum_{i=1}^{k} r_{ik} q_i$ for $k = 1, \dots, m$ and $r_{ii} \in \mathbb{R}_{\geq 0}$ for $i = 1, \dots, n$. For $Q = (q_1, \dots, q_n) \in \mathrm{U}(n, \mathbb{C})$, it now holds that $A = QR$.

(a) Wlog. let A itself be an upper triangular matrix with positive diagonal entries. For the first column, $a_{11} e_1 = r_{11} q_1$ holds. From $a_{11}, r_{11} > 0$ and $|q_1| = 1$, it follows that $q_1 = e_1$ and $a_{11} = r_{11}$. Inductively, let $r_{ij} = a_{ij}$ and $q_j = e_j$ for $i = 1, \dots, n$ and $j = 1, \dots, k-1$ already be proven. For the k-th column of A, it then holds that

$$\sum_{i=1}^{k} a_{ik} e_i = r_{kk} q_k + \sum_{i=1}^{k-1} r_{ik} e_i.$$

This shows $q_{ik} = 0$ for $i > k$. Because $[e_i, q_k] = [q_i, q_k] = 0$ for $i < k$, it follows that $q_k = e_k$ and $r_{ik} = a_{ik}$ for $i = 1, \dots, n$. Inductively, one obtains $A = R$ (but not necessarily $Q = 1_n$ if $n > m$).

(b) If A is real, then the Gram-Schmidt process can be performed in $\mathbb{R}^n$ (note that we complemented $a_{i_1}, \dots, a_{i_s}$ with real vectors to a basis). □

Remark 17.39.

(i) If $\mathrm{rk}(A) = n \leq m$, then the proof shows that Q and R in the QR decomposition are uniquely determined.

(ii) In the case $n = m = 1$, one obtains the *polar representation* of a complex number $z = e^{\varphi \mathrm{i}} |z|$. In general, the polar decomposition of a matrix has a different structure (Exercise III.11).

(iii) We will see in section 17.7 that the direct application of the Gram-Schmidt process is unstable.

Example 17.40.

$$\begin{pmatrix} 1 & \mathrm{i} \\ 1 & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sqrt{2} & \mathrm{i} - 1 \\ \sqrt{2} & 1 - \mathrm{i} \end{pmatrix} \begin{pmatrix} \sqrt{2} & {}^{(1+\mathrm{i})}/_{\sqrt{2}} \\ 0 & 1 \end{pmatrix}$$

Remark 17.41. In practice, overdetermined systems of equations $Ax = b$ without an exact solution often arise due to imprecise measurements. The quality of an approximate solution $\tilde{x}$ can be quantified by $|\tilde{x} - x|$ or $|\tilde{x} - x|^2 = \sum (\tilde{x}_i - x_i)^2$.

Theorem 17.42 (Least Squares Method)**.** *Let $A \in \mathbb{C}^{n \times m}$ with full rank $m \leq n$ and $b \in \mathbb{C}^{n \times 1}$. Then the system of equations $A^* A x = A^* b$ has a unique solution $\tilde{x}$ and for all $x \neq \tilde{x}$ it holds that $|A\tilde{x} - b| < |Ax - b|$.*

Proof. Since A has full rank $m \leq n$, A^*A is positive definite, thus in particular invertible. This shows the existence and uniqueness of $\tilde{x}$. For $y \neq 0$ it holds that

$$\begin{aligned}[Ay, A\tilde{x} - b] &= y^*A^*(A\tilde{x} - b) = 0 = [A\tilde{x} - b, Ay], \\ |A(\tilde{x} + y) - b|^2 &= [(A\tilde{x} - b) + Ay, (A\tilde{x} - b) + Ay] = |A\tilde{x} - b|^2 + |Ay|^2 > |A\tilde{x} - b|^2.\end{aligned}$$ □

Remark 17.43.

(a) On New Year's Eve 1801, the astronomer Piazzi discovered the dwarf planet *Ceres*. The position was recorded for 40 days until Ceres disappeared from the field of view. Using the least squares method, Gauss extrapolated the orbit of Ceres with the available measurement data and was thus able to successfully predict the position. He gained international fame through this.

(b) According to Example 17.28, $\tilde{x} = A^+b$ holds in the situation of Theorem 17.42.

(c) The least squares method leads to the so-called *normal equation* $A^*Ax = A^*b$ with a positive definite coefficient matrix. There are special methods for this. The following decomposition specifies Lemma 12.40.

Theorem 17.44 (Cholesky Decomposition)**.** *$A \in \mathbb{C}^{n\times n}$ is positive definite[5] if and only if there exists an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ with real, positive main diagonal entries and $A = R^*R$. If it exists, R is uniquely determined.*

Proof. For $R \in \mathrm{GL}(n, \mathbb{C})$, R^*R is known to be positive definite. Conversely, let A be positive definite and $\sqrt{A} = QR$ be the (unique) QR decomposition of the square root of A (Exercise II.18). Then it holds that

$$A = \sqrt{A}^2 = \sqrt{A}^*\sqrt{A} = R^*Q^*QR = R^*R.$$

Let $P \in \mathbb{C}^{n\times n}$ be another upper triangular matrix with positive main diagonal and $A = P^*P$. Then $P^{-*}R^* = P^{-*}AR^{-1} = PR^{-1}$ is an upper and lower triangular matrix, thus a diagonal matrix with positive main diagonal entries. Because of

$$(PR^{-1})^*PR^{-1} = R^{-*}AR^{-1} = 1_n$$

PR^{-1} is unitary. It follows that $P = R$. □

Remark 17.45.

(a) The matrix $R = (r_{ij})$ can be calculated iteratively without the detour via the QR decomposition: Because $a_{11} = e_1Ae_1^{\mathrm{t}} > 0$, we have $r_{11} = \sqrt{a_{11}}$. Suppose $R_1 \in \mathbb{C}^{(n-1)\times(n-1)}$ with $A_{nn} = R_1^*R_1$ has already been determined. With the ansatz $A = \left(\begin{smallmatrix} A_{nn} & a \\ a^* & a_{nn} \end{smallmatrix}\right)$ and $R = \left(\begin{smallmatrix} R_1 & v \\ 0 & r_{nn} \end{smallmatrix}\right)$, one obtains $R_1^*v = a$ and $r_{nn} = \sqrt{a_{nn} - v^*v}$.

(b) The system $Ax = b$ with $A = R^*R$ can be conveniently solved in two steps: First, one solves $R^*y = b$ by *forward substitution*:

$$\begin{pmatrix} \overline{r_{11}} & & 0 \\ \vdots & \ddots & \\ \overline{r_{1n}} & \cdots & \overline{r_{nn}} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Longrightarrow \begin{cases} y_1 = \frac{b_1}{\overline{r_{11}}} \\ y_2 = \frac{b_2 - \overline{r_{12}}y_1}{\overline{r_{22}}} \\ \quad \vdots \end{cases}$$

[5] in the sense of Exercise II.18

and subsequently $Rx = y$ by *backward substitution*:

$$\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \Rightarrow \begin{cases} x_n & = \frac{y_n}{r_{nn}} \\ x_{n-1} & = \frac{y_{n-1} - r_{n-1,n} x_n}{r_{n-1,n-1}} \\ & \vdots \end{cases}$$

Example 17.46. We consider the overdetermined system

$$Ax = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 0 & -1 \end{pmatrix} x = \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix}.$$

The system of normal equations is $\left(\begin{smallmatrix} 2 & 3 \\ 3 & 6 \end{smallmatrix}\right) x = -\left(\begin{smallmatrix} 1 \\ 4 \end{smallmatrix}\right)$. The matrix R is obtained by $r_{11} = \sqrt{2}$, $r_{12} = \frac{3}{\sqrt{2}}$ and $r_{22} = \sqrt{6 - 9/2} = \sqrt{3/2}$. Now $y = -\frac{1}{\sqrt{2}}(1, 5/\sqrt{3})^{\mathrm{t}}$ is the solution of

$$\begin{pmatrix} \sqrt{2} & 0 \\ 3/\sqrt{2} & \sqrt{3/2} \end{pmatrix} y = R^{\mathrm{t}} y = -\begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

and $x = (2, -5/3)^{\mathrm{t}}$ is the solution of

$$\begin{pmatrix} \sqrt{2} & 3/\sqrt{2} \\ 0 & \sqrt{3/2} \end{pmatrix} x = Rx = y = -\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 5/\sqrt{3} \end{pmatrix}.$$

For the original system, one obtains the approximation $Ax = \frac{1}{3}(1, -4, 5)^{\mathrm{t}}$. In practice, the radical expressions are approximated by floating-point numbers.

17.4 Iterative Methods

Remark 17.47.

(a) For large matrices, factorization methods such as the QR decomposition are slow and memory-intensive. To solve linear systems of equations in these cases, one can switch to faster iterative methods. For convergence proofs, we need some tools from analysis.

(b) So far, norms have always originated from scalar products (Definition 11.2 and Definition 13.2). However, they can also be defined directly via the properties from Lemma 13.5.

Definition 17.48. Let V be a $\mathbb{C}$-vector space. A map $V \to \mathbb{R}$, $v \mapsto \|v\|$ is called a *norm*, if for all $v, w \in V$ and $\lambda \in \mathbb{C}$ the following holds:

(a) $\|v\| \geq 0$ with equality if and only if $v = 0$ (*positive definite*).

(b) $\|\lambda v\| = |\lambda| \|v\|$ (*homogeneity*).

(c) $\|v + w\| \leq \|v\| + \|w\|$ (*triangle inequality*).

Example 17.49.

(a) If $[.,.]$ is a scalar product on V, then $\|v\| := \sqrt{[v, v]}$ defines a norm according to Lemma 13.5.

(b) On $V = \mathbb{C}^n$, $\|x\|_\infty := \max\{|x_1|, \ldots, |x_n|\}$ defines a norm. This norm cannot be obtained from a scalar product, because the parallelogram law

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

from Example 11.4 is not satisfied for $x = (1,0)$ and $y = (0,1)$.

(c) For $V = \mathbb{C}^n$ and $p \geq 1$,

$$\|x\|_p := \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

defines a norm (Exercise III.13). For $p = 1$, one obtains $\|x\|_1 = |x_1| + \ldots + |x_n|$. For $p = 2$, one obtains the "Euclidean" norm $|x| = \sqrt{[x,x]}$. As $p \to \infty$, $\|.\|_p$ approaches the norm $\|.\|_\infty$.

(d) If $\|.\|$ is a norm on $\mathbb{C}^n$ and $A \in \mathrm{GL}(n, \mathbb{C})$, then $\|x\|_A := \|Ax^{\mathrm{t}}\|$ also defines a norm on $\mathbb{C}^n$.

Definition 17.50. Norms $\|.\|$ and $\|.\|'$ on a $\mathbb{C}$-vector space V are called *equivalent*, if there exist $\lambda, \mu > 0$ with

$$\lambda\|v\| \leq \|v\|' \leq \mu\|v\|$$

for all $v \in V$.

Lemma 17.51. *Any two norms on a finite-dimensional $\mathbb{C}$-vector space are equivalent.*

Proof. Wlog. let $V = \mathbb{C}^n$. It suffices to show that every norm $\|.\|$ is equivalent to $\|.\|_1$. For $\lambda := \max\{\|e_1\|, \ldots, \|e_n\|\} > 0$ it holds that

$$\|x\| = \Big\|\sum_{i=1}^{n} x_i e_i\Big\| \leq \sum_{i=1}^{n} |x_i|\|e_i\| \leq \lambda\|x\|_1$$

for all $x \in \mathbb{R}^n$. Let $y_i := \overline{x_i}/|x_i|$ for $i = 1, \ldots, n$. By the Cauchy-Schwarz inequality, $\|x\|_1 = [x,y] \leq |x||y| = \sqrt{n}|x|$ holds. From the (reverse) triangle inequality it follows that

$$|\|x\| - \|y\|| \leq \|x-y\| \leq \lambda\|x-y\|_1 \leq \lambda\sqrt{n}|x-y|.$$

This shows that the map $f\colon \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|x\|$ is continuous wrt. the Euclidean norm. Therefore, f attains its minimum $\mu > 0$ on the compact set $\{x \in \mathbb{R}^n : \|x\|_1 = 1\} \neq \varnothing$. For $x \neq 0$ it now holds that

$$\|x\| = \|x\|_1 f(\|x\|_1^{-1} x) \geq \|x\|_1 \mu. \qquad \square$$

Remark 17.52. As is well known, the Euclidean norm on $\mathbb{C}^n$ is complete (i.e., every Cauchy sequence converges). According to Lemma 17.51, every norm on a finite-dimensional $\mathbb{C}$-vector space V is therefore complete, i.e., V is a *Banach space*.

Theorem 17.53 (Banach's Fixed Point Theorem)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a* contraction *wrt. a norm $\|.\|$, i.e., there exists a constant $c < 1$ with $\|f(x) - f(y)\| \leq c\|x-y\|$ for all $x, y \in \mathbb{R}^n$. Then the sequence $x_{k+1} := f(x_k)$ converges for all $x_0 \in \mathbb{R}^n$ to the unique fixed point of f.*

Proof. For $k, l \in \mathbb{N}$ it holds that

$$\|x_{k+l} - x_k\| = \Big\|\sum_{i=0}^{l-1}(x_{k+i+1} - x_{k+i})\Big\| \leq \sum_{i=0}^{l-1}\|x_{k+i+1} - x_{k+i}\| = \sum_{i=0}^{l-1}\|f^{k+i}(x_1) - f^{k+i}(x_0)\|$$
$$\leq \|x_1 - x_0\|\sum_{i=0}^{l-1} c^{k+i} = \|x_1 - x_0\|c^k\frac{1-c^l}{1-c} \xrightarrow{k\to\infty} 0,$$

i.e., $(x_k)_k$ is a Cauchy sequence. By Remark 17.52, the limit $\tilde{x} := \lim_{k\to\infty} x_k$ exists. Since f as a contraction is (uniformly) continuous (set $\delta := \frac{\epsilon}{c}$), it holds that $f(\tilde{x}) = \lim_{k\to\infty} f(x_k) = \tilde{x}$. If $y \neq \tilde{x}$ were also a fixed point of f, one would have the contradiction

$$\|\tilde{x} - y\| = \|f(\tilde{x}) - f(y)\| \leq c\|\tilde{x} - y\| < \|\tilde{x} - y\|. \qquad \square$$

Remark 17.54.

(a) A linear map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is a contraction if and only if there exists a $c < 1$ with $\|f(x)\| \leq c\|x\|$ for all $x \in \mathbb{R}^n$.

(b) The next lemma provides an explicit equivalence between the Euclidean norm and the norm $|.|_A$ introduced in Example 17.49.

Lemma 17.55. *Let $A \in \mathbb{C}^{n\times n}$ be normal with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Let*

$$\lambda_{\min} := \min\{|\lambda_i| : i = 1, \ldots, n\}, \qquad \lambda_{\max} := \max\{|\lambda_i| : i = 1, \ldots, n\}.$$

For $x \in \mathbb{C}^n$, it holds that $\lambda_{\min}|x| \leq |Ax| \leq \lambda_{\max}|x|$ and $|x^*Ax| \leq \lambda_{\max}|x|^2$. If A is positive definite, then $|x^*Ax| \geq \lambda_{\min}|x|^2$ holds.

Proof. Let $S \in \mathrm{U}(n, \mathbb{C})$ with $S^*AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ (Spectral Theorem). Because of $|x| = |Sx|$ and $|Ax| = |S^*Ax|$, we can assume $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then

$$|Ax|^2 = (Ax)^*Ax = \sum_{i=1}^n |\lambda_i|^2|x_i|^2 \leq \lambda_{\max}^2\sum_{i=1}^n |x_i|^2 = \lambda_{\max}^2|x|^2$$

and analogously $|Ax| \geq \lambda_{\min}|x|$. Likewise,

$$|x^*Ax| = \Big|\sum_{i=1}^n \lambda_i|x_i|^2\Big| \leq \sum_{i=1}^n |\lambda_i||x_i|^2 \leq \lambda_{\max}|x|^2.$$

If A is positive definite, then $\lambda_1, \ldots, \lambda_n$ are real and positive. Then it follows as before that $|x^{\mathrm{t}}Ax| \geq \lambda_{\min}|x|^2$. $\square$

Theorem 17.56 (Gauss-Seidel Method)**.** *Let $A = L + R \in \mathbb{C}^{n\times n}$ be positive definite, where L is a lower triangular matrix and R is a strict upper triangular matrix. Let $b \in \mathbb{C}^{n\times 1}$. Then the sequence*

$$x_{k+1} := L^{-1}b - L^{-1}Rx_k$$

converges for all starting values $x_0 \in \mathbb{C}^{n\times 1}$ to the solution of $Ax = b$.

Proof. Since A has real, positive diagonal entries (Remark 12.39) and R is a strict upper triangular matrix, L must be invertible. For $A\tilde{x} = b$, it holds that

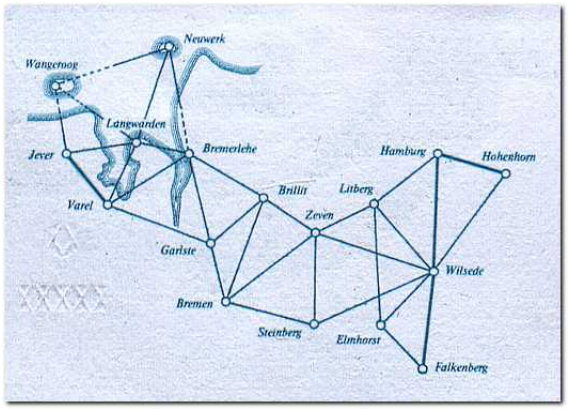
$$\tilde{x} = L^{-1}L\tilde{x} = L^{-1}(b - R\tilde{x}) = L^{-1}b - L^{-1}R\tilde{x},$$

i. e. $\tilde{x}$ is a fixed point of the mapping $f(x) := L^{-1}b - L^{-1}Rx$. Since A is positive definite, $[x, y] := y^*Ax$ defines an inner product with norm $\|x\| := \sqrt{[x, x]}$. By Banach's Fixed Point Theorem, it suffices to show that f is a contraction wrt. $\|.\|$. Since $L^{-1}b$ is constant, we can consider the linear mapping $g(x) := L^{-1}Rx$ instead of f. According to Remark 17.54 and Lemma 17.55, we must show that every eigenvalue $\lambda \in \mathbb{C}$ of $L^{-1}R$ is smaller than 1 in absolute value. Let $x \in \mathbb{C}^{n\times 1}$ be a corresponding eigenvector and $y := x + g(x) = L^{-1}Ax \neq 0$. By assumption, $D := L^* - R = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ is positive definite. Thus, it holds that

$$\begin{aligned} |\lambda|\|x\| = \|g(x)\|^2 = \|x - y\|^2 &= \|x\|^2 - x^*A^*y - y^*Ax + y^*Ay \\ &= \|x\|^2 - y^*L^*y - y^*Ly + y^*(L + R)y = \|x\|^2 - y^*Dy < \|x\|. \end{aligned}$$

This shows $|\lambda| < 1$. □

Remark 17.57. Gauss surveyed the Kingdom of Hanover using this procedure, as evidenced by a drawing on the 10-DM note:



He wrote in a letter:

> "I recommend this mode to you for imitation. You will hardly ever eliminate directly again, at least not if you have more than 2 unknowns. The indirect procedure can be carried out half in one's sleep, or one can think of other things during it."

Example 17.58. The system $H_4x = (1, 1, 1, 1)^{\mathrm{t}}$ with the ill-conditioned Hilbert matrix H_4 has the solution $x = (-4, 60, -180, 140)^{\mathrm{t}}$. The Gauss-Seidel method converges only slowly:

$$\begin{array}{rcl} x_0 & := & (\ 0, \quad 0, \quad 0, \quad 0\ ) \\ x_{10} & \approx & (\ -1.90, \quad -2.97, \quad 4.45, \quad 8.06\ ) \\ x_{100} & \approx & (\ 2.84, \quad -14.20, \quad -5.17, \quad 27.94\ ) \\ x_{1000} & \approx & (\ -1.10, \quad 28.86, \quad -106.96, \quad 93.31\ ) \\ x_{10000} & \approx & (\ -4.000, \quad 59.995, \quad -179.988, \quad 139.992\ ) \end{array}$$

17.5 Matrix Norms

Remark 17.59. The set of $n \times m$ matrices is known to form a vector space of dimension nm. It is therefore natural to introduce the norms defined in Definition 17.48 for matrices as well.

Definition 17.60. A mapping $\mathbb{C}^{n\times m} \to \mathbb{R}$, $A \mapsto \|A\|$ is called a *matrix norm*, if for $A, B \in \mathbb{C}^{n\times m}$ and $\lambda \in \mathbb{C}$ the following hold:

- $\|A\| \geq 0$ with equality if and only if $A = 0_{n\times m}$.
- $\|\lambda A\| = |\lambda|\|A\|$.
- $\|A + B\| \leq \|A\| + \|B\|$.

In the case $n = m$, we call the norm *submultiplicative*, if $\|AB\| \leq \|A\|\|B\|$ holds.

Example 17.61.

(a) The "Euclidean" norm on $\mathbb{C}^{n\times m}$ is called the *Frobenius norm*

$$|A| := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2} = \sqrt{\operatorname{tr}(A^*A)}.$$

For $A, B \in \mathbb{C}^{n\times n}$, it follows from the Cauchy-Schwarz inequality that

$$|AB|^2 = \sum_{i,j}\Big|\sum_{k=1}^{n} a_{ik}b_{kj}\Big|^2 \leq \sum_{i,j}\Big(\sum_{k=1}^{n} |a_{ik}||b_{kj}|\Big)^2 \leq \sum_{i,j}\Big(\sum_{k=1}^{n} |a_{ik}|^2\Big)\Big(\sum_{s=1}^{n} |b_{sj}|^2\Big) = |A|^2|B|^2,$$

i. e. $|.|$ is submultiplicative. If $\sigma_1, \ldots, \sigma_k$ are the singular values of A, then

$$|A| = \sqrt{\operatorname{tr}(A^*A)} = \sqrt{\sigma_1^2 + \ldots + \sigma_k^2}$$

according to Remark 17.22.

(b) Every "natural"[6] vector norm $\|.\|$ *induces* via[7]

$$\|A\| = \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

a matrix norm (Exercise III.14). In the case $m = 1$, the norms coincide. For all $x \in \mathbb{C}^{m\times 1}$, it holds that $\|Ax\| \leq \|A\|\|x\|$. For $A, B \in \mathbb{C}^{n\times n}$, it follows that

$$\|AB\| = \max_{\|x\|=1} \|ABx\| \leq \max_{\|x\|=1} \|A\|\|Bx\| = \|A\|\|B\|,$$

i. e. $\|.\|$ is submultiplicative.

(c) According to (the proof of) Corollary 17.23, $\|A\|_2$ is the largest singular value of A (or 0). In particular, $\|A\|_2 \leq |A|$ with equality if and only if $\operatorname{rk}(A) \leq 1$.

(d) Not every matrix norm is submultiplicative: For $\|A\|_{\max} := \max_{i,j} |a_{ij}|$ and $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, it holds that $\|A^2\|_{\max} = 2 > 1 = \|A\|_{\max}^2$.

[6] defined for arbitrary dimension

[7] As in Remark 17.19, the maximum is actually attained.

Lemma 17.62. *For $A \in \mathbb{C}^{n\times m}$, it holds that*

(a) $$\|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^{n} |a_{ij}| \qquad \text{(column sum norm)}.$$

(b) $$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}| = \|A^{\mathrm{t}}\|_1 \qquad \text{(row sum norm)}.$$

(c) $$\|A\|_2 \le |A| \le \sqrt{\min\{m,n\}}\|A\|_2.$$

(d) $$\frac{1}{\sqrt{m}}\|A\|_\infty \le \|A\|_2 \le \sqrt{n}\|A\|_\infty.$$

(e) $$\frac{1}{\sqrt{n}}\|A\|_1 \le \|A\|_2 \le \frac{1}{\sqrt{m}}\|A\|_1.$$

Proof.

(a) By the triangle inequality, it holds that

$$\|A\|_1 = \max_{\|x\|_1=1} \sum_{i=1}^{n} \Big|\sum_{j=1}^{m} a_{ij}x_j\Big| \le \max_{\sum|x_i|=1} \sum_{j=1}^{m} |x_j| \sum_{i=1}^{n} |a_{ij}| \le \max_{1\le j\le m} \sum_{i=1}^{n} |a_{ij}|.$$

If the maximum on the right is attained for j, then equality is obtained with $x = e_j$.

(b) It holds that

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \max_{1\le i\le n} \Big|\sum_{j=1}^{m} a_{ij}x_j\Big| \le \max_{\max_i|x_i|=1} \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}||x_j| \le \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}|.$$

If the maximum on the right is attained for i, then equality is obtained with $x_j = \overline{a_{ij}}/|a_{ij}|$ for $a_{ij} \neq 0$ and $x_j = 0$ otherwise.

(c) If $\sigma_1 \ge \ldots \ge \sigma_k$ are the singular values of A, then it holds that

$$\|A\|_2 = \sigma_1 \le \sqrt{\sigma_1^2 + \ldots + \sigma_k^2} = |A| \le \sqrt{k}\sigma_1 \le \sqrt{\min\{n,m\}}\|A\|_2.$$

(d) For $x \in \mathbb{C}^n$ it holds that

$$\|x\|_\infty = \max_{1\le i\le n} |x_i| \le \sqrt{|x_1|^2 + \ldots + |x_n|^2} = |x| \le \sqrt{n}\|x\|_\infty.$$

This shows

$$\|A\|_\infty = \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{\|Ax\|_\infty}{\|x\|_\infty} \le \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{|Ax|}{|x|/\sqrt{m}} = \sqrt{m}\|A\|_2,$$
$$\|A\|_2 = \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{|Ax|}{|x|} \le \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{\sqrt{n}\|Ax\|_\infty}{\|x\|_\infty} = \sqrt{n}\|A\|_\infty.$$

(e) From the singular value decomposition, it follows that $\|A^{\mathrm{t}}\|_2 = \|A\|_2$. Therefore, the claim follows from (b) and (d). □

Lemma 17.63. *The condition number of $A \in \mathrm{GL}(n,\mathbb{C})$ is $\kappa(A) = \|A\|_2\|A^{-1}\|_2$.*

Proof. By definition,

$$\|A\|_2 = \max_{x \in \mathbb{C}^{n\times 1}\setminus\{0\}} |Ax|/|x|,$$
$$\|A^{-1}\|_2 = \max_x |A^{-1}x|/|x| = \max_y |y|/|Ay| = (\min_y |Ay|/|y|)^{-1}.$$

The claim follows from the definition of the condition number. □

Definition 17.64. A sequence of matrices $(A_k)_k \in \mathbb{C}^{n\times m}$ *converges* to a matrix $A \in \mathbb{C}^{n\times m}$ if $\lim_{k\to\infty} \|A - A_k\| = 0$ holds for a matrix norm. Since all norms are equivalent according to Lemma 17.51, this definition does not depend on the choice of the norm. If applicable, we write $A = \lim_{k\to\infty} A_k$

Remark 17.65.

(a) As in analysis, one shows that the sequence $A_k = (a_{ij}^{(k)})$ converges to $A = (a_{ij})$ if and only if $\lim_{k\to\infty} a_{ij}^{(k)} = a_{ij}$ holds for all i, j.

(b) Obviously, the maps $A \to A^{\mathrm{t}}$, $A \to \overline{A}$ and tr are continuous (wrt. any matrix norm). As is well known, addition and multiplication of complex numbers are continuous maps. Therefore, matrix multiplication is also continuous. For $A = \lim_{k\to\infty} A_k$ and $B = \lim_{k\to\infty} B_k$, it thus holds that $\lim_{k\to\infty} A_k B_k = AB$. The Leibniz formula shows that det is also continuous. The representation $A^{-1} = \det(A)^{-1}\widetilde{A}$ (Theorem 9.22) implies that the inversion $A \mapsto A^{-1}$ is continuous.

(c) If the sequence of partial sums $B_k := \sum_{i=1}^k A_i$ converges, we write $\sum_{k=1}^\infty A_k := \lim_{k\to\infty} B_k$ as usual.

Theorem 17.66. *For $A \in \mathbb{C}^{n\times n}$, the sequence $(A^k)_k$ converges if and only if the following two statements hold:*

(a) For every eigenvalue λ of A, it holds that $|\lambda| < 1$ or $\lambda = 1$.

(b) If 1 is an eigenvalue, then the algebraic multiplicity coincides with the geometric multiplicity.

Proof. Wlog. let $A = \operatorname{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$ be in Jordan normal form. If $|\lambda_i| > 1$, then A^k has the entry $\lambda_i^k \xrightarrow{k\to\infty} \infty$. Now let $|\lambda_i| < 1$. The absolute values of the entries of $J_{n_i}(\lambda_i)^k$ have, according to Lemma 14.41, the form

$$\binom{k}{l}|\lambda_i|^{k-l} \le k^n |\lambda_i|^{k-n} \xrightarrow{k\to\infty} 0 \qquad (0 \le l < n).$$

This shows $J_{n_i}(\lambda_i)^k \to 0$. Now let $|\lambda_i| = 1$. For A^k to converge, there must exist a k with $\lambda_i^k = \lambda_i^{k+1} = \ldots$, i. e. $\lambda_i = 1$. If $n_i > 1$, then $J_{n_i}(\lambda_i)_{21} = k \to \infty$. Thus $n_i = 1$ must hold for all i with $\lambda_i = 1$. Wlog. let $\lambda_1 = \ldots = \lambda_t = 1 > |\lambda_j|$ for $j > t$. Obviously, (A^k) then converges to $\operatorname{diag}(1_t, 0_{n-t})$. □

Corollary 17.67. *Let the conditions from Theorem 17.66 be satisfied for $A \in \mathbb{C}^{n\times n}$ with $\dim E_1(A) = 1$. Let $v, w \in \mathbb{C}^{n\times 1}$ be eigenvectors of A and A^*, respectively, for the eigenvalue 1 with $[v, w] = 1$. Then $\lim_{k\to\infty} A^k = vw^*$ holds.*

Proof. Because of $\chi_{A^*} = \overline{\chi_A}$, 1 is indeed an eigenvalue of A^*. As is well known (for example, by the Jordan normal form), there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with first column v, such that $S^{-1}AS = \mathrm{diag}(1, B)$ with $\lim_{k\to\infty} B^k = 0$. Let u be the first row of S^{-1}. Then

$$C := \lim_{k\to\infty} A^k = \lim_{k\to\infty} S\,\mathrm{diag}(1, B^k)S^{-1} = S\,\mathrm{diag}(1, 0_{n-1})S^{-1} = vu.$$

On the other hand, $w = C^*w = u^*v^*w = u^*[v, w] = u^*$. □

17.6 Eigenvalue Computation

Remark 17.68.

(a) The determinant is in general ill-conditioned (i.e., sensitive to small changes):

$$\det\begin{pmatrix} 1 & 33 \\ 3 & 100 \end{pmatrix} = 1 \qquad\qquad \det\begin{pmatrix} 1.\mathbf{1} & 33 \\ 3 & 100 \end{pmatrix} = 11$$

Since $\pm\det(A)$ is the constant term of χ_A, χ_A is also ill-conditioned. Even worse conditioned is μ_A, because here small changes can even change $\deg \mu_A$.

(b) The eigenvalues as roots of χ_A (or μ_A) are likewise ill-conditioned:

$X^2 - 21X + 110$	Roots: $10, 11$
$X^2 - 21.\mathbf{1}X + 110$	Roots: $\approx 9.41, 11.69$
$X^2 - 20.\mathbf{9}X + 110$	Roots: $\approx 10.45 \pm 0.89\mathrm{i}$

Deviations in A thus have a "double" effect on the eigenvalue computation. Furthermore, there is no explicit formula for the roots of a polynomial of degree ≥ 5. Nevertheless, the determination of the eigenvalues of *normal* matrices – without the detour via χ_A – is well-conditioned according to the following theorem.

Theorem 17.69 (Bauer-Fike)**.** *Let $A \in \mathbb{C}^{n\times n}$ be diagonalizable with $S^{-1}AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Let $X \in \mathbb{C}^{n\times n}$ and $\mu \in \mathbb{C}$ be an eigenvalue of $A + X$. Then*

$$\min_{i=1,\ldots,n} |\mu - \lambda_i| \leq \kappa(S)\|X\|_2.$$

In particular, $\min_{i=1,\ldots,n} |\mu - \lambda_i| \leq \|X\|_2$ *if A is normal.*

Proof. Wlog. let $\mu \notin \{\lambda_1, \ldots, \lambda_n\}$. Let $D := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then $D - \mu 1_n$ is invertible. By assumption, $A + X - \mu 1_n$ is not invertible. Therefore,

$$(D - \mu 1_n)^{-1}S^{-1}(A - \mu 1_n + X)S = 1_n + (D - \mu 1_n)^{-1}S^{-1}XS$$

is also not invertible. From Theorem 18.8 and Lemma 17.63 it follows that

$$1 \leq \rho\big((D - \mu 1_n)^{-1}S^{-1}XS\big) \leq \|(D - \mu 1_n)^{-1}S^{-1}XS\|_2 \leq \|(D - \mu 1_n)^{-1}\|_2\kappa(S)\|X\|_2.$$

According to Example 17.61, $\|(D - \mu 1_n)^{-1}\|_2 = \max_{i=1,\ldots,n}(|\lambda_i - \mu|)^{-1}$. This shows

$$\min_{i=1,\ldots,n} |\mu - \lambda_i| = \|(D - \mu 1_n)^{-1}\|_2^{-1} \leq \kappa(S)\|X\|_2.$$

If A is normal, then one can choose $S \in \mathrm{U}(n, \mathbb{C})$ according to the spectral theorem. Then $\kappa(S) = 1$. □

Theorem 17.70 (*Power Method*). *Let $A \in \mathbb{C}^{n\times n}$ with eigenvalues $\lambda = \lambda_1, \ldots, \lambda_n \in \mathbb{C}$, such that $|\lambda| > |\lambda_i|$ for $i \geq 2$ holds. Then the sequence*

$$x_{k+1} := \frac{Ax_k}{|Ax_k|}$$

converges for "almost all" starting vectors $x_0 \in \mathbb{C}^{n\times 1}$ to an eigenvector for λ. If applicable,

$$\lim_{k\to\infty} x_k^* A x_k = \lambda.$$

Proof. By assumption, the Jordan normal form of A has the form

$$J = \operatorname{diag}(\lambda, J_{n_2}(\lambda_2), \ldots, J_{n_s}(\lambda_s)),$$

provided that $\lambda_2, \ldots, \lambda_n$ are suitably arranged. Let $b_1, \ldots, b_n \in \mathbb{C}^{n\times 1}$ be a corresponding basis with $Ab_1 = \lambda b_1$. For $S := (b_1, \ldots, b_n) \in \mathrm{GL}(n, \mathbb{C})$, it holds that $S^{-1}AS = J$. Let $x_0 = \alpha_1 b_1 + \ldots + \alpha_n b_n$ be chosen randomly with $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$. In almost all cases, $\alpha_1 \neq 0$. We assume this from now on. Then

$$A^k x_0 = A^k S(\alpha_1, \ldots, \alpha_n)^{\mathrm{t}} = SJ^k(\alpha_1, \ldots, \alpha_n)^{\mathrm{t}} = \lambda^k(\alpha_1 b_1 + \beta_2 b_2 + \ldots + \beta_n b_n).$$

For $i \geq 2$, according to Lemma 14.41, there exist $2 \leq j \leq s$ and $0 \leq t \leq i-2$ with

$$\beta_i = \sum_{l=0}^{t} \binom{k}{l} \lambda_j^{-l} \alpha_{i-l} \Big(\frac{\lambda_j}{\lambda}\Big)^k \xrightarrow{k\to\infty} 0$$

because $|\lambda_j| < |\lambda|$. As $A^k x_0$ does, x_k also tends towards a multiple of b_1. The second assertion follows from $|x_k| = 1$ for all $k \geq 1$. □

Remark 17.71. If one has (approximately) determined an eigenvector b_1 for the eigenvalue λ of A with the largest absolute value, then one can extend b_1 to a basis $b_1, \ldots, b_n$ and form the matrix S from these columns. It holds that $S^{-1}AS = \left(\begin{smallmatrix} \lambda & * \\ 0 & A_1 \end{smallmatrix}\right)$, where every eigenvalue of $A_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ is also an eigenvalue of A. If all eigenvalues of A have pairwise distinct absolute values, then one can apply Theorem 17.70 to A_1 and iterate. We will see that this can be accomplished with only one iteration.

Lemma 17.72. *Let (A_k) be a sequence of invertible matrices with $\lim_{k\to\infty} A_k = 1_n$. Let $A_k = Q_k R_k$ be the QR decomposition. Then $\lim_{k\to\infty} Q_k = 1_n = \lim_{k\to\infty} R_k$.*

Proof. Because $|Q_k| = \sqrt{\operatorname{tr}(Q_k^* Q_k)} = \sqrt{\operatorname{tr}(1_n)} = \sqrt{n}$, the sequence $(Q_k)_k$ is bounded. By Bolzano-Weierstraß, there exists a convergent subsequence $(Q_{k_i})_i$ with $Q := \lim_{i\to\infty} Q_{k_i}$. Because

$$Q^*Q = \lim_{i\to\infty} Q_{k_i}^* Q_{k_i} = 1_n$$

Q is unitary. From the continuity of matrix multiplication, one obtains

$$\lim_{i\to\infty} R_{k_i} = \lim_{i\to\infty} Q_{k_i}^*(Q_{k_i} R_{k_i}) = Q^* \lim_{i\to\infty} A_{k_i} = Q^*.$$

Since the set of upper triangular matrices with non-negative main diagonal entries is closed, Q^* must also belong to this set. This is obviously only possible if $Q = 1_n$. In particular, every convergent subsequence of $(Q_k)_k$ converges to the same limit. Therefore, (Q_k) must also converge to 1_n and consequently so must (R_k). □

Definition 17.73. A sequence (A_k) of matrices *quasi-converges* to a matrix A if there exists a sequence of unitary diagonal matrices (D_k) with $\lim_{k\to\infty} D_k^* A_k D_k = A$.

Remark 17.74. In the situation of Definition 17.73, $D_k = \operatorname{diag}(d_1, \ldots, d_n)$ holds with $|d_i| = 1$ for $i = 1, \ldots, n$. The main diagonals of A_k and $D_k^* A_k A_k$ are identical for all k. Therefore, the main diagonal of A_k converges to a vector in $\mathbb{C}^n$. The entries of A_k outside the main diagonal, however, can "oscillate" in each iteration by factors of absolute value 1.

Theorem 17.75 (FRANCIS algorithm[8])**.** *Let $A_1 := A \in \mathbb{C}^{n\times n}$ be invertible, such that the eigenvalues of A have pairwise distinct absolute values. For $k = 1, 2, \ldots$ let $A_k = Q_k R_k$ be the QR decomposition and $A_{k+1} := R_k Q_k$. Then $(A_k)_k$ quasi-converges to an upper triangular matrix with the eigenvalues of A on the main diagonal.*

Proof (WILKINSON). Let $\mathcal{Q}_k := Q_1 \ldots Q_k$ and $\mathcal{R}_k := R_k \ldots R_1$. From $Q_k^* A_k Q_k = R_k Q_k = A_{k+1}$ it follows that $\mathcal{Q}_k^* A \mathcal{Q}_k = A_{k+1}$. We show $\mathcal{Q}_k \mathcal{R}_k = A^k$ for all $k \geq 1$. This is clear for $k = 1$. For $k \geq 2$ it holds inductively

$$\mathcal{Q}_k \mathcal{R}_k = \mathcal{Q}_{k-1} Q_k R_k \mathcal{R}_{k-1} = \mathcal{Q}_{k-1} A_k \mathcal{R}_{k-1} = \mathcal{Q}_{k-1} \mathcal{Q}_{k-1}^* A \mathcal{Q}_{k-1} \mathcal{R}_{k-1} = A A^{k-1} = A^k.$$

For the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, $|\lambda_1| > \ldots > |\lambda_n| > 0$ holds by assumption. By Corollary 8.12, there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with $S^{-1} A S = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) =: D$. Let $S^{-1} = L_{S'} P U_{S'}$ be the modified LU decomposition of S^{-1} from Remark 17.37. Let $SP = Q_S R_S$ be the QR decomposition of SP. For $L_{S'} = (x_{ij})$ we have

$$D^k L_{S'} D^{-k} = \big((\lambda_i/\lambda_j)^k x_{ij}\big)_{ij} \xrightarrow{k\to\infty} 1_n$$

because of $x_{ij} = \delta_{ij}$ for $i \leq j$ and $|\lambda_i| < |\lambda_j|$ for $i > j$. Thus $M_k := R_S P^{\mathrm{t}} D^k L_{S'} D^{-k} P R_S^{-1}$ also tends towards 1_n. For the QR decomposition $M_k = Q_{M_k} R_{M_k}$, $\lim_{k\to\infty} Q_{M_k} = 1_n$ holds by Lemma 17.72. In total,

$$\begin{aligned} \mathcal{Q}_k \mathcal{R}_k = A^k = S D^k S^{-1} &= S P P^{\mathrm{t}} D^k L_{S'} P U_{S'} = Q_S R_S P^{\mathrm{t}} D^k L_{S'} D^{-k} D^k P U_{S'} \\ &= Q_S M_k R_S P^{\mathrm{t}} D^k P U_{S'} = Q_S Q_{M_k} R_{M_k} P^{\mathrm{t}} D^k P U_{S'}. \end{aligned}$$

Like D^k, $P^{\mathrm{t}} D^k P$ is also a diagonal matrix. Consequently, $T_k := (t_{ij}) = R_{M_k} P^{\mathrm{t}} D^k P U_{S'}$ is an upper triangular matrix. For

$$U_k := \operatorname{diag}\big(\overline{t_{11}}/|t_{11}|, \ldots, \overline{t_{nn}}/|t_{nn}|\big) \in \mathrm{U}(n, \mathbb{C})$$

$U_k T_k$ is an upper triangular matrix with positive main diagonal and $Q_S Q_{M_k} U_k^* \in \mathrm{U}(n, \mathbb{C})$. From the uniqueness of the QR decomposition, it follows that $\mathcal{Q}_k = Q_S Q_{M_k} U_k^*$. Thus

$$\lim_{k\to\infty} U_k^* A_{k+1} U_k = \lim_{k\to\infty} U_k^* \mathcal{Q}_k^* A \mathcal{Q}_k U_k = \lim_{k\to\infty} Q_{M_k}^* Q_S^* S D S^{-1} Q_S Q_{M_k} = R_S P^{\mathrm{t}} D P R_S^{-1},$$

i. e. (A_k) quasi-converges to the upper triangular matrix $R_S P^{\mathrm{t}} D P R_S^{-1} \approx D$. The eigenvalues of D and A, respectively, are on the main diagonal. □

Remark 17.76.

(a) Apparently (A_k) quasi-converges towards a Schur decomposition (Theorem 13.23) of A. Let A be real. Then it follows from the assumption that the eigenvalues are also real, since $|\lambda| = |\overline{\lambda}|$ for $\lambda \in \mathbb{C}$. If this assumption is not satisfied, one can achieve that (A_k) converges towards a real block diagonal matrix with 2×2 blocks (Exercise II.21).

[8] also called the *QR algorithm*

(b) In its pure form, Francis' algorithm converges only slowly. In practice, the procedure is accelerated by first transforming A into a *Hessenberg matrix* (i.e., $a_{ij} = 0$ for $i > j + 1$) and shifting A_k by a suitable multiple of the identity matrix in each iteration. This algorithm is counted (alongside the FFT and the simplex algorithm) among the most important algorithms of the 20th century.[9]

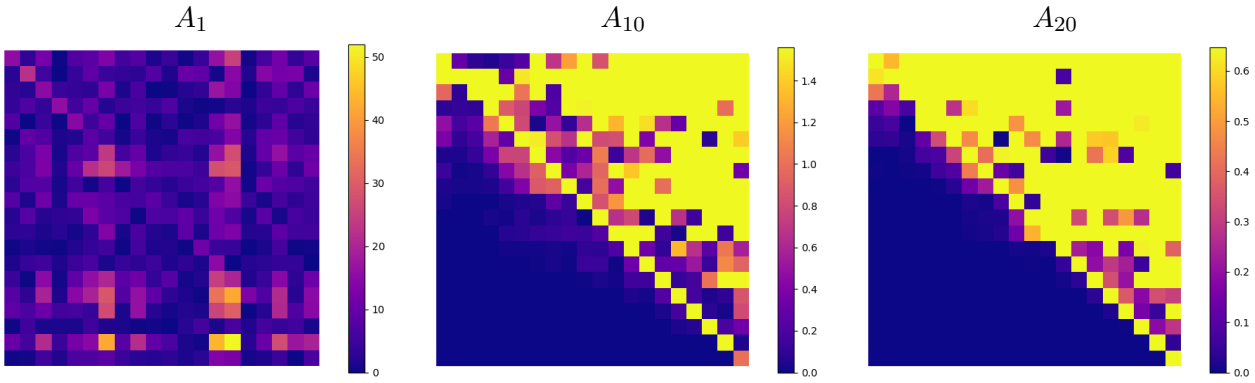
Example 17.77.

(a) For

$$A = \begin{pmatrix} 2 & -1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} = Q_1 R_1$$

it holds that $A_2 = R_1 Q_1 = \left(\begin{smallmatrix} 2 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ and inductively $A_k = \left(\begin{smallmatrix} 2 & (-1)^k \\ 0 & -1 \end{smallmatrix}\right)$. This shows the phenomenon of quasi-convergence.

(b) Let $A = S^{-1} \operatorname{diag}(1, \ldots, 20) S \in \mathbb{R}^{20 \times 20}$ for a randomly chosen matrix $S \in \mathrm{GL}(20, \mathbb{R})$. The following graphic illustrates Francis' algorithm for A. The color scale represents the absolute values of entries below the main diagonal (the structure of A was justified in Remark 9.24).



Remark 17.78. For positive definite matrices, the QR decomposition in Francis' algorithm can be replaced by the Cholesky decomposition.

Theorem 17.79 (Cholesky method)**.** *Let $A = A_1 \in \mathbb{C}^{n \times n}$ be positive definite. For $k = 1, 2, \ldots$ let $A_k = R_k^* R_k$ be the Cholesky decomposition and $A_{k+1} := R_k R_k^*$. Then $(A_k)_k$ converges towards a diagonal matrix with the eigenvalues of A on the main diagonal.*

Proof (Schatzman). Because $A_{k+1} = R_k A_k R_k^{-1}$, all A_k have the same eigenvalues. Let $A_k = (a_{ij}^{(k)})$ and $R_k = (r_{ij}^{(k)})$. Because

$$\delta_k(s) := \sum_{i=1}^{s} a_{ii}^{(k)} = \sum_{i=1}^{s} \sum_{j=1}^{i} |r_{ji}^{(k)}|^2 = \sum_{i=1}^{s} \sum_{j=i}^{s} |r_{ij}^{(k)}|^2 \leq \sum_{i=1}^{s} \sum_{j=i}^{n} |r_{ij}^{(k)}|^2 = \delta_{k+1}(s)$$

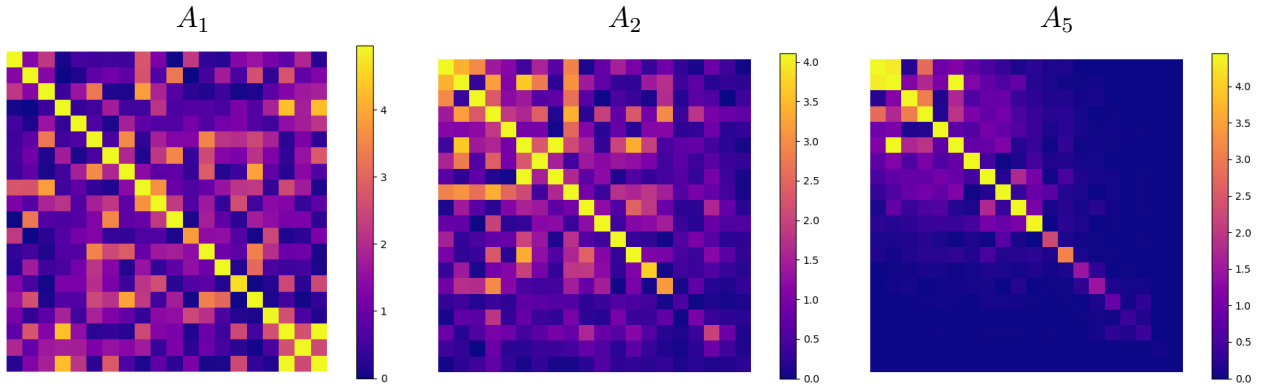
[9] See [Dongarra-Sullivan, *Top Ten Algorithms of the Century*, Comput. Sci. Eng. 2 (2000), 22–23]

$(\delta_k(s))_k$ are monotonically increasing sequences for $s = 0, \dots, n$. Because $\delta_k(s) \leq \delta_k(n) = \operatorname{tr}(A_k) = \operatorname{tr}(A)$, the sequences are bounded. Therefore $\delta(s) := \lim_{k\to\infty} \delta_k(s)$ exists for $s = 0, \dots, n$. It follows that $\lim_{k\to\infty} a_{ss}^{(k)} = \delta(s) - \delta(s-1)$ for $s = 1, \dots, n$. Since the differences

$$\delta_{k+1}(s) - \delta_k(s) = \sum_{i=1}^{s} \sum_{j=s+1}^{n} |r_{ij}^{(k)}|^2$$

tend to 0 as $k \to \infty$, $\lim_{k\to\infty} r_{ij}^{(k)} = 0$ must hold for all $i \neq j$. Along with R_k, A_k must also converge to a diagonal matrix. □

Example 17.80. Let $A \in \mathbb{R}^{20\times 20}$ be a randomly chosen positive definite matrix (constructed as $A = S^{\mathrm{t}}S$). The following graphic illustrates the progress of the Cholesky method with A. The color scale represents the absolute values of all entries outside the main diagonal.



Apparently, the method works from the bottom right to the top left.

Remark 17.81. The approximate location of eigenvalues of a matrix can be narrowed down without computational effort using the following theorem.

Theorem 17.82 (Gershgorin). *For every eigenvalue $\lambda \in \mathbb{C}$ of $(a_{ij}) \in \mathbb{C}^{n\times n}$, there exists an $i \in \{1, \dots, n\}$ such that $|\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|$.*

Proof. Let $x = (x_1, \dots, x_n) \in \mathbb{C}^n$ be an eigenvector for the eigenvalue λ. Let

$$|x_i| = \max\{|x_j| : j = 1, \dots, n\} > 0.$$

After normalization, we can assume $x_i = 1$ and $|x_j| \leq 1$ for $j \neq i$. According to the triangle inequality, it then holds that

$$|\lambda - a_{ii}| = |(Ax)_i - a_{ii}x_i| = \left|\sum_{j=1}^{n} a_{ij}x_j - a_{ii}x_i\right| = \left|\sum_{j\neq i} a_{ij}x_j\right| \leq \sum_{j\neq i} |a_{ij}|. \qquad \square$$

Corollary 17.83. *Every matrix $A \in \mathbb{C}^{n\times n}$ with $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for $i = 1, \dots, n$ is invertible.*[10]

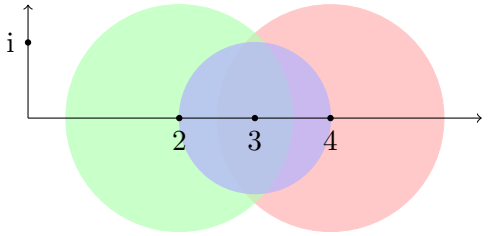
[10] This property is called *diagonal dominance*.

Proof. According to Gershgorin, 0 is not an eigenvalue of A. □

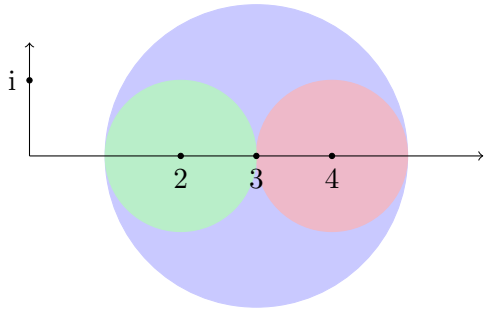
Example 17.84.

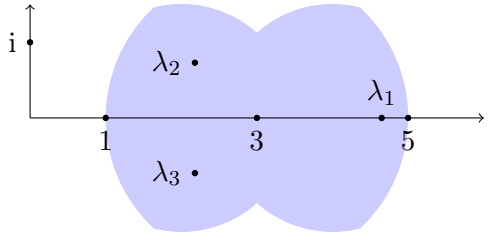
(a) Every eigenvalue lies in one of the depicted circles:

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 0.5 & 4 & 1 \\ 1.5 & 0 & 2 \end{pmatrix}$$



Since A and A^{t} have the same eigenvalues, one can also use the columns and subsequently intersect both sets:





(b) Applying Theorem 17.82 to the companion matrix of $\alpha = X^n + a_1 X^{n-1} + \ldots + a_n \in \mathbb{C}[X]$, one obtains

$$|x| \leq \max\{|a_n|, |a_{n-1}| + 1, \ldots, |a_1| + 1\}$$

for every root $x \in \mathbb{C}$ of α (note $|x + a_1| \leq 1 \Longrightarrow |x| \leq |a_1| + 1$).

17.7 Orthonormalization

Remark 17.85. In our formulation of the Gram-Schmidt process (11.10), we first constructed an orthogonal basis and subsequently obtained an orthonormal basis by normalization. However, this can lead to very large or very small factors, which result in numerical instabilities. It is therefore more favorable to normalize each found vector immediately:

$$b'_s := v_s - \sum_{i=1}^{s-1} [v_s, b_i] b_i, \qquad b_s := \frac{b'_s}{|b'_s|}$$

In the summation, rounding errors can additionally accumulate. An iterated procedure is therefore more stable.

Theorem 17.86 (Modified Gram-Schmidt process)**.** *Let V be a unitary space and $v_1, \ldots, v_k \in V$ linearly independent. For $s = 1, \ldots, k$ we define*

$$b_{s,0} := v_s, \qquad b_{s,i} := b_{s,i-1} - [b_{s,i-1}, b_i] b_i \quad (i = 1, \ldots, s-1), \qquad b_s := \frac{b_{s,s-1}}{|b_{s,s-1}|}.$$

Then $b_1, \ldots, b_k$ is an orthonormal basis of $\langle v_1, \ldots, v_k \rangle$.

Proof. It suffices to show that the vectors $b_1, \dots, b_k$ coincide with the vectors obtained in Theorem 11.10. This is clear for $b_1 = \frac{v_1}{|v_1|}$. Let $s \geq 2$ and assume the claim has already been shown for $b_1, \dots, b_{s-1}$. Then $b_{s,1} = v_s - [v_s, b_1]b_1$. Inductively, let

$$b_{s,i} = v_s - \sum_{j=1}^{i} [v_s, b_j]b_j$$

already be shown for some $i < s-1$. Because $[b_j, b_{i+1}] = 0$ for $j \leq i$, we have

$$b_{s,i+1} = v_s - \sum_{j=1}^{i} [v_s, b_j]b_j - [v_s, b_{i+1}]b_{i+1} = v_s - \sum_{j=1}^{i+1} [v_s, b_j]b_j.$$

This shows that $b_{s,s-1}$ is the vector calculated in Theorem 11.10 (before normalization). □

Remark 17.87. When calculating the QR decomposition of an invertible matrix $A \in \mathrm{GL}(n, \mathbb{C})$, one requires not only the orthonormalized columns of A (as columns of Q) but also the coefficient matrix R. There are two common methods for this, which we present for real A (see Exercise III.15 for the general case):

(a) We perform the orthonormalization as a composition of reflections. For this, let a_1 be the first column of A and $b := a_1 - |a_1|e_1$. The reflection $Q_1 \in \mathrm{O}(n, \mathbb{R})$ across the hyperplane $\langle b \rangle^\perp$ can be calculated according to Exercise II.11 as a HOUSEHOLDER *transformation*:

$$Q_1 = 1_n - \frac{2}{|b|^2} bb^{\mathrm{t}}$$

(for $b = 0$ let $Q_1 = 1_n$). Because of

$$[a_1 + |a_1|e_1, b] = |a_1|^2 - |a_1|[a_1, e_1] + |a_1|[e_1, a_1] - |a_1|^2 = 0$$

it follows that $a_1 + |a_1|e_1 \in \langle b \rangle^\perp$ and

$$Q_1 a_1 = \frac{1}{2}\big(Q_1 b + Q_1(a_1 + |a_1|e_1)\big) = \frac{1}{2}\big(-b + a_1 + |a_1|e_1\big) = |a_1|e_1.$$

It follows that $Q_1 A = \left(\begin{smallmatrix} |a_1| & * \\ 0 & * \end{smallmatrix}\right)$. One proceeds analogously with columns $2, \dots, n$ and obtains corresponding matrices $Q_2, \dots, Q_n \in \mathrm{O}(n, \mathbb{R})$ (in fact, it suffices to consider the shortened columns $(a_{ii}, \dots, a_{ni})^{\mathrm{t}}$). Now $R := Q_n \dots Q_1 A$ is an upper triangular matrix and $Q := Q_1 \dots Q_n \in \mathrm{O}(n, \mathbb{R})$ with $A = QR$, since $Q_i^{-1} = Q_i$.

(b) Instead of reflections, one can use GIVENS *rotations* (cf. Remark 11.43):

$$D_{st}(\varphi) := \begin{pmatrix} 1_{s-1} & & & & \\ & \cos\varphi & & -\sin\varphi & \\ & & 1_{t-s-1} & & \\ & \sin\varphi & & \cos\varphi & \\ & & & & 1_{n-t} \end{pmatrix} \in \mathrm{O}(n, \mathbb{R}).$$

We proceed as in the Gaussian algorithm. Suppose the first $k-1$ columns of A are already in upper triangular form. After row swapping, one can assume $a_{kk} \neq 0$ (A invertible). To eliminate the entry a_{ik} for $i > k$, we set $c := |(a_{kk}, a_{ik})| = \sqrt{|a_{kk}|^2 + |a_{ik}|^2}$ and

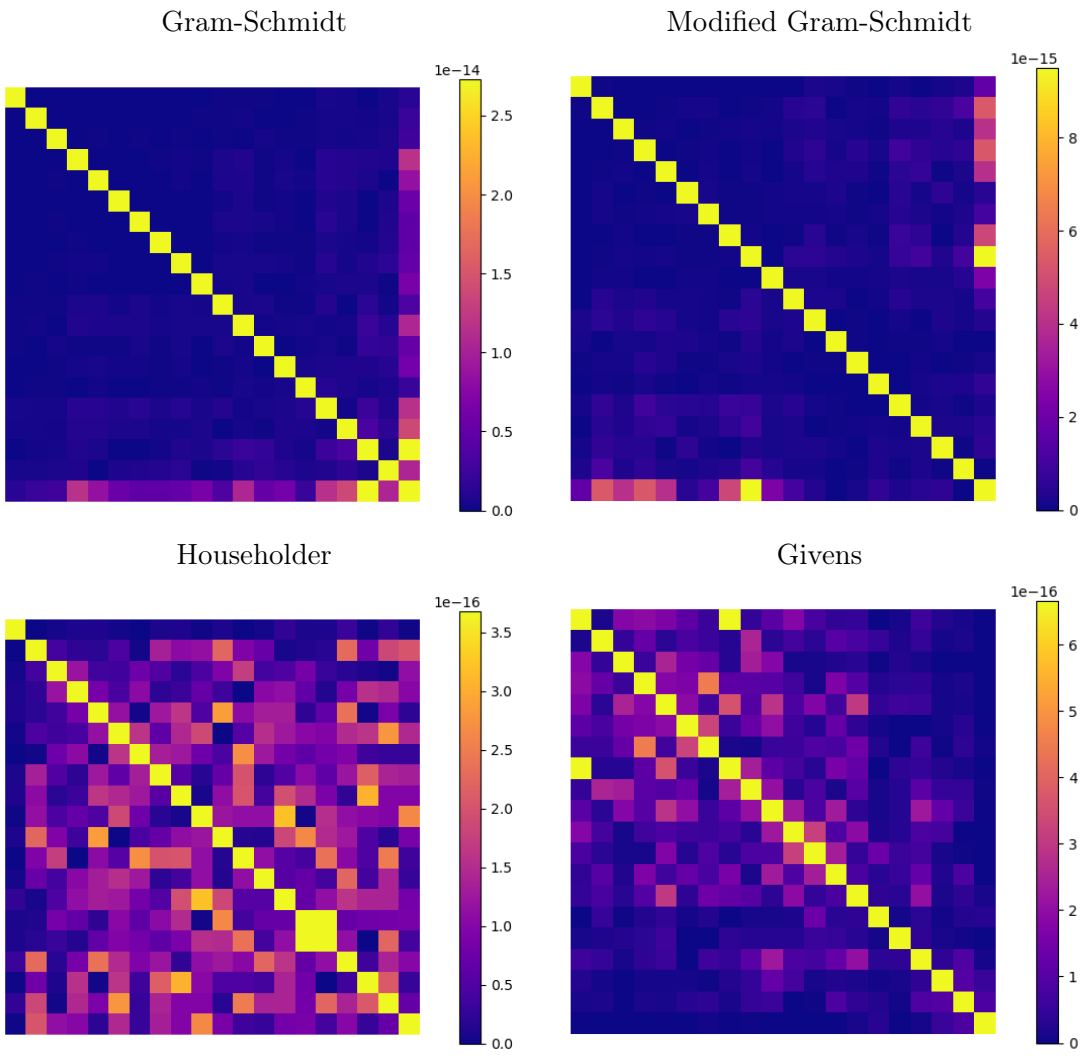
$$\cos\varphi := \frac{[(a_{kk}, a_{ik}), (1, 0)]}{c} = \frac{a_{kk}}{c}$$

Because of $\sin^2\varphi = 1 - \cos^2\varphi = a_{ij}^2/c^2$, we can assume $\sin\varphi = -a_{ik}/c$ (if necessary, replace φ with $-\varphi$). Now

$$D_{ki}(\varphi) = \begin{pmatrix} 1_{k-1} & & & & \\ & a_{kk}/c & & a_{ik}/c & \\ & & 1_{i-k-1} & & \\ & -a_{ik}/c & & a_{kk}/c & \\ & & & & 1_{n-i} \end{pmatrix}.$$

By replacing A with $D_{ki}(\varphi)A$, a_{ik} becomes 0. In this way, one obtains a sequence of Givens rotations $Q_1, \ldots, Q_l \in \mathrm{O}(n, \mathbb{R})$ such that $R := Q_l \ldots Q_1 A$ is an upper triangular matrix. For $Q := Q_1^{\mathrm{t}} \ldots Q_l^{\mathrm{t}} \in \mathrm{O}(n, \mathbb{R})$, it holds that $A = QR$. Although this requires significantly more matrices than (a), the Q_i are sparse, so the multiplications are inexpensive.

Example 17.88. We calculate the QR decomposition $A = QR$ of a random (invertible) matrix $A \in \mathbb{R}^{20\times 20}$ using the four presented methods. The quality of the results is shown by a comparison of $Q^{\mathrm{t}}Q$ with 1_{20}. The color scale refers to the absolute values of the entries outside the main diagonal (the smaller, the more accurate):



18 Analytical Aspects

18.1 Eigenvalue Estimates

Remark 18.1. A Hermitian matrix $A \in \mathbb{C}^{n\times n}$ has, according to Corollary 13.20, real eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. We write $\lambda_k(A) := \lambda_k$ for the k-th largest eigenvalue. The following *Min-Max Theorem* relates these numbers to the so-called RAYLEIGH *quotients* $\frac{vAv^*}{vv^*}$ and generalizes Lemma 17.55.

Theorem 18.2 (COURANT-FISCHER)**.** *For every Hermitian matrix $A \in \mathbb{C}^{n\times n}$, it holds that*

$$\lambda_k(A) = \max_{\substack{V\leq\mathbb{C}^n \\ \dim V=k}} \min_{0\neq v\in V} \frac{vAv^*}{vv^*} = \min_{\substack{V\leq\mathbb{C}^n \\ \dim V=n-k+1}} \max_{0\neq v\in V} \frac{vAv^*}{vv^*}.$$

Proof. Let $\mu_k(A)$ be the middle part of the formula. According to the spectral theorem, there exists an $S \in \mathrm{U}(n,\mathbb{C})$ with $SAS^* = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $\lambda_1 \geq \ldots \geq \lambda_n$. By replacing $V \leq \mathbb{C}^n$ with $\{vS : v \in V\}$, one can assume $A = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Furthermore, one can restrict oneself to normalized $v \in V$ by replacing v with $v/|v|$. For $V := \langle e_1,\ldots,e_k\rangle$, it holds that

$$\mu_k(A) \geq \min_{\substack{v\in V \\ |v|=1}} vAv^* = \min_{\substack{v\in V \\ |v|=1}} \sum_{i=1}^{k} \lambda_k |v_k|^2 \geq \min_{|v|=1} \lambda_k |v|^2 = \lambda_k.$$

Conversely, let $V \leq \mathbb{C}^n$ with $\dim V = k$ be arbitrary. According to the dimension formula, there exists a normalized vector $w \in V \cap \langle e_k,\ldots,e_n\rangle$. It follows that

$$\min_{\substack{v\in V \\ |v|=1}} vAv^* \leq wAw^* = \sum_{i=k}^{n} \lambda_i |w_i|^2 \leq \lambda_k.$$

Since V is arbitrary, $\mu_k(A) \leq \lambda_k = \lambda_k(A)$ holds. For the second equality, one considers analogously $V := \langle e_k,\ldots,e_n\rangle$ and $V \cap \langle e_1,\ldots,e_k\rangle$. □

Theorem 18.3 (CAUCHY's Interlace Theorem)**.** *Let $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ be Hermitian and $P \in \mathbb{C}^{n\times k}$ with orthonormal columns ($k \leq n$). Then $\lambda_i(A) \geq \lambda_i(P^*AP) \geq \lambda_{n-k+i}(A)$ for $i = 1,\ldots,k$.*

Proof. Since A is hermitian, $B := P^*AP \in \mathbb{C}^{k\times k}$ is also hermitian. For $U \leq \mathbb{C}^k$ let $P(U) := \{uP^* : u \in U\} \leq \mathbb{C}^n$. Since P has full rank, $\dim U = \dim P(U)$ holds. Furthermore, $|uP^*| = |u|$ for all $u \in U$. By Courant-Fischer it follows that

$$\lambda_i(B) = \max_{\substack{U\leq\mathbb{C}^k \\ \dim U=i}} \min_{\substack{v\in P(U) \\ |v|=1}} vAv^* \leq \max_{\substack{V\leq\mathbb{C}^n \\ \dim V=i}} \min_{\substack{v\in V \\ |v|=1}} vAv^* = \lambda_i(A).$$

Let b_j be an eigenvector of B for the eigenvalue $\lambda_j(B)$ and $U := \langle b_i, \ldots, b_k \rangle$. Then

$$\lambda_{n-k+i}(A) = \min_{\substack{V \leq \mathbb{C}^n \\ \dim V = k-i+1}} \max_{\substack{v \in V \\ |v|=1}} vAv^* \leq \max_{\substack{v \in P(U) \\ |v|=1}} vAv^* = \max_{\substack{u \in U \\ |u|=1}} uBu^* = \lambda_i(B)$$

as in the proof of Theorem 18.2. □

Example 18.4. For $I \subseteq \{1, \ldots, n\}$, $P = (e_i : i \in I)$ has orthonormal columns. Cauchy's interlacing theorem then applies to the submatrix $B := P^*AP = (a_{ij} : i, j \in I)$. In the case $|I| = n - 1$ one obtains

$$\lambda_1(A) \geq \lambda_1(B) \geq \lambda_2(A) \geq \lambda_2(B) \geq \ldots \geq \lambda_{n-1}(B) \geq \lambda_n(A).$$

18.2 The Spectral Radius

Definition 18.5. For $A \in \mathbb{C}^{n \times n}$,

$$\rho(A) := \max\{|\lambda| : \lambda \in \mathbb{C} \text{ eigenvalue of } A\}$$

is called the *spectral radius* of A.[1]

Example 18.6.

(a) For unitary matrices A, $\rho(A) = 1$ holds.

(b) $A \in \mathbb{C}^{n \times n}$ is nilpotent if and only if $\rho(A) = 0$.

(c) From the Jordan normal form it follows that $\rho(A^{\mathrm{t}}) = \rho(A)$ and $\rho(A^k) = \rho(A)^k$ for $k \in \mathbb{N}_0$.

Theorem 18.7. *For all $A \in \mathbb{C}^{n \times n}$:*

(a) $\|A\| \geq \rho(A)$ for every submultiplicative matrix norm.

(b) For all $\epsilon > 0$ there exists a submultiplicative matrix norm with $\|A\| \leq \rho(A) + \epsilon$.

Proof.

(a) Let $\lambda \in \mathbb{C}$ be an eigenvalue of A with absolute value $\rho(A)$ and v a corresponding eigenvector. Let $B \in \mathbb{C}^{n \times n}$ with first column v and otherwise only zeros. Then

$$\rho(A)\|B\| = \|\lambda B\| = \|AB\| \leq \|A\|\|B\|$$

and $\rho(A) \leq \|A\|$.

(b) Let $J = S^{-1}AS$ be the Jordan normal form of A with $S \in \mathrm{GL}(n, \mathbb{C})$. Let $D := \mathrm{diag}(\epsilon, \epsilon^2, \ldots, \epsilon^n)$. The main diagonals of $N := DJD^{-1}$ and J are identical, while the entries below the main diagonal are ϵ or 0. For the matrix norm $\|B\| := \|DS^{-1}BSD^{-1}\|_1$, it now holds that $\|A\| := \|N\|_1 \leq \rho(A) + \epsilon$ according to Lemma 17.62. □

Theorem 18.8 (Geometric Series[2])**.** *For all $A \in \mathbb{C}^{n \times n}$, the following statements are equivalent:*

(1) $\rho(A) < 1$.

[1] All eigenvalues lie in the complex plane within the circle with center 0 and *radius* $\rho(A)$.

[2] In this context also called Neumann *series*.

(2) $\lim_{k\to\infty} A^k = 0_n$.

(3) $\sum_{k=0}^{\infty} A^k = (1_n - A)^{-1}$.

Proof.

(1) ⇒ (2): Let $\epsilon > 0$ with $q := \rho(A)+\epsilon < 1$. According to Theorem 18.7, there exists a submultiplicative matrix norm with $\|A\| \leq q$. It follows that $\|A^k\| \leq \|A\|^k \leq q^k \to 0$ as $k \to \infty$.

(2) ⇒ (1): Let λ be an eigenvalue of A with magnitude $\rho(A)$ and eigenvector v. Because $\lambda^k v = A^k v \to 0$, it follows that $\rho(A) = |\lambda| < 1$.

(1) ⇒ (3): As before, there exists a submultiplicative matrix norm with $q := \|A\| < 1$. Let $B_m := \sum_{k=0}^{m} A^k$. By the triangle inequality, we have

$$\|B_m\| \leq \sum_{k=0}^{m} q^k \leq \sum_{k=0}^{\infty} q^k = \frac{1}{1-q}.$$

By Bolzano-Weierstraß, the bounded sequence $(B_m)_m$ has a convergent subsequence $(B_{m_k})_k$. For $B := \lim_{k\to\infty} B_{m_k}$, it holds that

$$(1_n - A)B = \lim_{k\to\infty} (1_n - A) \sum_{i=0}^{m_k} A^i = \lim_{k\to\infty} (1_n - A^{m_k+1}) \overset{(2)}{=} 1_n,$$

i. e. $B = (1_n - A)^{-1}$. Since all convergent subsequences of (B_m) have the same limit, the entire sequence must also converge to B.

(3) ⇒ (2): Since the partial sums B_m converge, their differences $B_m - B_{m-1} = A^m$ must form a null sequence. □

Corollary 18.9. *Let $A \in \mathbb{C}^{n\times n}$ with $\|A\| < 1$ for a submultiplicative matrix norm. Then $\sum_{k=0}^{n} A^k = (1_n - A)^{-1}$ holds.*

Proof. From Theorem 18.7 it follows that $\rho(A) < 1$. □

Example 18.10. The submultiplicativity in Corollary 18.9 is indispensable: For $A = (3/4)_{i,j=1}^{2}$, we have $\|A\|_{\max} = 3/4 < 1$ with the matrix norm from Example 17.61, but $A^k = \left(\frac{3}{2}\right)^{k-1} A$.

Theorem 18.11. *For every matrix norm $\|.\|$ and all $A \in \mathbb{C}^{n\times n}$, it holds that*

$$\boxed{\rho(A) = \lim_{k\to\infty} \sqrt[k]{\|A^k\|}.}$$

Proof. For an arbitrary (not necessarily submultiplicative) matrix norm $\|.\|'$, there exist constants $\lambda, \mu > 0$ according to Lemma 17.51 such that $\lambda\|A^k\| \leq \|A^k\|' \leq \mu\|A^k\|$ for all k. It follows that

$$\sqrt[k]{\lambda}\sqrt[k]{\|A^k\|} \leq \sqrt[k]{\|A^k\|'} \leq \sqrt[k]{\mu}\sqrt[k]{\|A^k\|}.$$

Due to $\lim_{k\to\infty} \sqrt[k]{\lambda} = 1 = \lim_{k\to\infty} \sqrt[k]{\mu}$, it suffices to prove the statement for a specific matrix norm. Let

$$S^{-1}AS = \operatorname{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$$

be the Jordan normal form of A with $S \in \mathrm{GL}(n, \mathbb{C})$. We define $\|B\| := \|S^{-1}BS\|_{\max}$ for $B \in \mathbb{C}^{n\times n}$ with the norm from Example 17.61. Every non-zero entry of A^k has, according to Lemma 14.41, the form $\binom{k}{l}\lambda_i^{k-l}$ with $1 \le i \le s$ and $0 \le l \le n-1$. For every $\epsilon > 0$, $k < (1+\epsilon)^k$ holds if k is large enough. It follows that

$$\lim_{k\to\infty} \sqrt[k]{\binom{k}{l}|\lambda_i|^{k-l}} = |\lambda_i| \lim_{k\to\infty} \sqrt[k]{k(k-1)\dots(k-l+1)} \lim_{k\to\infty} |\lambda_i|^{-l/k} = |\lambda_i|.$$

This shows $\lim_{k\to\infty} \sqrt[k]{\|A^k\|} = \rho(A)$. □

18.3 The Exponential Function of a Matrix

Definition 18.12. For $A \in \mathbb{C}^{n\times n}$, let

$$\exp(A) := \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

be the *exponential function* of A.

Remark 18.13.

(a) Due to $\left|\frac{A^k}{k!}\right| \le \frac{|A|^k}{k!} \to 0$, $\exp(A)$ is well-defined according to Theorem 18.8.

(b) Through $\exp(\lambda 1_n) = e^\lambda 1_n$ for $\lambda \in \mathbb{C}$, exp extends the ordinary exponential function on $\mathbb{C}$.

(c) Let

$$J := S^{-1}AS = \mathrm{diag}(J_{n_1}(\lambda_1), \dots, J_{n_s}(\lambda_s))$$

be the Jordan normal form of A with $S \in \mathrm{GL}(n, \mathbb{C})$. Due to the continuity of matrix multiplication, it holds that

$$S^{-1}\exp(A)S = \exp(J) = \mathrm{diag}\big(\exp(J_{n_1}(\lambda_1)), \dots, \exp(J_{n_s}(\lambda_s))\big).$$

According to Lemma 14.41, the non-zero entries of $\exp(J_{n_i}(\lambda_i))$ have the form

$$\sum_{k=l}^{\infty} \frac{\binom{k}{l}\lambda_i^{k-l}}{k!} = \sum_{k=l}^{\infty} \frac{\lambda_i^{k-l}}{l!(k-l)!} = \frac{e^{\lambda_i}}{l!}.$$

Thus, it holds that

$$L_m(\lambda) := \exp(J_m(\lambda)) = e^\lambda \begin{pmatrix} 1 & & & \\ \frac{1}{1!} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \frac{1}{(m-1)!} & \cdots & \frac{1}{1!} & 1 \end{pmatrix} \in \mathbb{C}^{m\times m}$$

and

$$\exp(A) = S\,\mathrm{diag}\big(L_{n_1}(\lambda_1), \dots, L_{n_s}(\lambda_s)\big)S^{-1}.$$

Example 18.14.

(a) For the permutation matrix $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ we have

$$\exp(P) = \sum_{k=0}^{\infty} \frac{1}{(2k)!} 1_n + \sum_{k=0}^{\infty} \frac{1}{(2k+1)!} P = \begin{pmatrix} \cosh(1) & \sinh(1) \\ \sinh(1) & \cosh(1) \end{pmatrix} = \begin{pmatrix} 1.54\ldots & 1.17\ldots \\ 1.17\ldots & 1.54\ldots \end{pmatrix}$$

with the hyperbolic trigonometric functions.

(b) For the matrix

$$A = \begin{pmatrix} 5 & 0 & 1 \\ -5-\mathrm{i} & -\mathrm{i} & -1 \\ -9 & 0 & -1 \end{pmatrix}$$

from Example 14.34 we have

$$\exp(A) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} e^2 & 0 & 0 \\ e^2 & e^2 & 0 \\ 0 & 0 & e^{-\mathrm{i}} \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & -3 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} 4e^2 & 0 & e^2 \\ -4e^2 + e^{-\mathrm{i}} & e^{-\mathrm{i}} & -e^2 \\ -9e^2 & 0 & -2e^2 \end{pmatrix}.$$

(c) According to Remark 18.13, $\exp(A) = 1_n$ holds if and only if A is diagonalizable and the eigenvalues are integer multiples of $2\pi\mathrm{i}$.

Theorem 18.15 (Jacobi)**.** *For $A \in \mathbb{C}^{n\times n}$ we have* $\boxed{\det(\exp(A)) = e^{\operatorname{tr}(A)}.}$

Proof. Wlog. let A be in Jordan normal form with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Then $\exp(A)$ is a lower triangular matrix with diagonal $(e^{\lambda_1}, \ldots, e^{\lambda_n})$. This shows $\det(\exp(A)) = e^{\lambda_1 + \ldots + \lambda_n} = e^{\operatorname{tr}(A)}$. □

Example 18.16. If A is real, then $\det(\exp(A)) > 0$.

Theorem 18.17 (Functional equation)**.** *For commuting matrices $A, B \in \mathbb{C}^{n\times n}$ we have*

$$\boxed{\exp(A+B) = \exp(A)\exp(B).}$$

In particular, $\exp(A)$ *is invertible with* $\exp(A)^{-1} = \exp(-A)$.

Proof. We use the Jordan-Chevalley decomposition $A = D_A + N_A$ and $B = D_B + N_B$ from Corollary 16.20.[3] Since all involved matrices are polynomials in A or B, they are pairwise commuting. According to Lemma 14.11, D_A and D_B can be simultaneously diagonalized. Wlog. let $D_A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $D_B = \operatorname{diag}(\mu_1, \ldots, \mu_n)$. Let $D := D_A + D_B$ and $N := N_A + N_B$. From the functional equation of the ordinary exponential function, we obtain

$$\exp(D) = \operatorname{diag}(e^{\lambda_1+\mu_1}, \ldots, e^{\lambda_n+\mu_n}) = \operatorname{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n}) \operatorname{diag}(e^{\mu_1}, \ldots, e^{\mu_n}) = \exp(D_A)\exp(D_B).$$

Since N_A and N_B are nilpotent, it holds that

$$\exp(N) = \sum_{k=0}^{\infty} \frac{1}{k!}(N_A + N_B)^k = \sum_{k=0}^{2n} \sum_{l=0}^{k} \frac{1}{k!} \binom{k}{l} N_A^l N_B^{k-l} = \sum_{k=0}^{2n} \sum_{l=0}^{k} \frac{N_A^l}{l!} \frac{N_B^{k-l}}{(k-l)!} = \exp(N_A)\exp(N_B).$$

[3] This can be avoided by using the Cauchy product for absolutely convergent matrix series instead.

Analogously,

$$\exp(D+N) = \sum_{k=0}^{\infty} \sum_{l=0}^{k} \frac{D^{k-l}}{(k-l)!} \frac{N^l}{l!} = \sum_{l=0}^{n} \frac{N^l}{l!} \sum_{k=l}^{\infty} \frac{D^{k-l}}{(k-l)!} = \exp(N)\exp(D).$$

The corresponding equations also hold for $D_A + N_A$ and $D_B + N_B$. Overall, it follows that

$$\begin{aligned}\exp(A+B) &= \exp(D+N) = \exp(D)\exp(N) = \exp(D_A)\exp(D_B)\exp(N_A)\exp(N_B) \\ &= \exp(D_A)\exp(N_A)\exp(D_B)\exp(N_B) = \exp(A)\exp(B).\end{aligned}$$

The second assertion follows from $\exp(A)\exp(-A) = \exp(A-A) = \exp(0_n) = 1_n$. □

Remark 18.18.

(a) The commutativity of A and B in the functional equation cannot be omitted, because otherwise $\exp(A)\exp(B) = \exp(A+B) = \exp(B+A) = \exp(B)\exp(A)$ would hold for all A, B in contradiction to Theorem 18.19. In fact, the GOLDEN-THOMPSON *inequality*

$$|\exp(A+B)| \leq |\exp(A)\exp(B)|$$

holds for all Hermitian matrices A, B with equality if and only if $AB = BA$ (without proof).

(b) As is well known, every $z \in \mathbb{C}^\times$ can be uniquely written in the form $|z|e^{\varphi \mathrm{i}}$ with $-\pi < \varphi \leq \pi$. The map $\log\colon \mathbb{C}^\times \to \mathbb{C}$, $z \mapsto \ln(|z|) + \varphi\mathrm{i}$ is called the *principal branch of the natural logarithm.*

Theorem 18.19. *For every $A \in \mathrm{GL}(n,\mathbb{C})$ there exists exactly one $B \in \mathbb{C}^{n\times n}$ with $\exp(B) = A$ and $-\pi < \mathrm{Im}(\lambda) \leq \pi$ for every eigenvalue λ of B.*

Proof. For existence, we can assume $A = J_n(\mu)$ for a $\mu \neq 0$ according to Remark 18.13. Let $\lambda = \log(\mu)$ and $B := J_n(\lambda)$. According to Remark 18.13, μ is an eigenvalue of $\exp(B) = L_n(\lambda)$ with algebraic multiplicity n and geometric multiplicity 1. From Theorem 14.28 it follows that $\exp(B) \approx A$. Therefore, there exists an $S \in \mathrm{GL}(n,\mathbb{C})$ with $\exp(S^{-1}BS) = S^{-1}\exp(B)S = A$.

Now let $A \in \mathrm{GL}(n,\mathbb{C})$ be arbitrary and B, C with the specified properties. After a change of basis, we can assume $B = \mathrm{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$. According to Remark 18.13

$$A = \exp(B) = \mathrm{diag}(L_{n_1}(\lambda_1), \ldots, L_{n_s}(\lambda_s)).$$

Due to the special choice of eigenvalues with the principal branch of the logarithm, C must have the same Jordan normal form as B. So let $S \in \mathrm{GL}(n,\mathbb{C})$ with $S^{-1}BS = C$. From

$$A = \exp(C) = S^{-1}\exp(B)S = S^{-1}AS$$

it follows that $S \in \mathrm{C}(A)$. We must show $S \in \mathrm{C}(B)$. Obviously $\chi_A = (X - e^{\lambda_1})^{n_1} \ldots (X - e^{\lambda_s})^{n_s}$. Wlog. let $\lambda_1, \ldots, \lambda_s$ be sorted such that identical λ_i are adjacent. According to Exercise II.30, $C = \mathrm{diag}(C_1, \ldots, C_t)$, where each C_k belongs to a block $\mathrm{diag}(L_{n_i}(\lambda_i), \ldots, L_{n_j}(\lambda_j))$ with $\lambda_i = \ldots = \lambda_j$. We can therefore assume $\lambda_1 = \ldots = \lambda_s$ with $n_1 \geq \ldots \geq n_s$. Now A is a linear combination of the powers of $J := \mathrm{diag}(J_{n_1}(0), \ldots, J_{n_s}(0))$. In particular, $\mathrm{C}(J) \subseteq \mathrm{C}(A)$. According to Frobenius (Theorem 15.32)

$$\dim \mathrm{C}(A) = \sum_{i=1}^{s} (2i-1)n_i = \dim \mathrm{C}(J).$$

It follows that $\mathrm{C}(J) = \mathrm{C}(A)$. On the other hand, B is also a linear combination of J. This shows $S \in \mathrm{C}(A) = \mathrm{C}(J) \subseteq \mathrm{C}(B)$, as desired. □

Example 18.20. The eigenvalues of a unitary matrix U can be written in the form $e^{\varphi_1 \mathrm{i}}, \ldots, e^{\varphi_n \mathrm{i}}$ with $\varphi_1, \ldots, \varphi_n \in \mathbb{R}$ according to Corollary 13.20. According to the spectral theorem, there exists an $S \in \mathrm{U}(n, \mathbb{C})$ with $U = S \operatorname{diag}(e^{\varphi_1 \mathrm{i}}, \ldots, e^{\varphi_n \mathrm{i}}) S^*$. For the Hermitian matrix $H := S \operatorname{diag}(\varphi_1, \ldots, \varphi_n) S^*$, it holds that

$$U = \exp(H\mathrm{i}).$$

This connection plays a role in quantum mechanics. If one dispenses with the special choice of eigenvalues in Theorem 18.19, one can choose $\varphi_1, \ldots, \varphi_n > 0$. Then H is even positive definite.

Remark 18.21.

(a) In analysis, one can calculate $\log(1 - x)$ for $x \in \mathbb{R}$ with $|x| < 1$ using the *Mercator series*

$$\log(1 - x) = -\sum_{k=1}^{\infty} \frac{1}{k} x^k$$

For $\rho(A) < 1$, according to Theorem 18.11, the corresponding series with A instead of x also converges. If A is nilpotent (i. e. $\rho(A) = 0$), the series terminates after $n-1$ summands. If applicable, $B := \log(1 - A)$ is the uniquely determined matrix from Theorem 18.19 (Exercise III.18).

(b) A system of ordinary homogeneous *differential equations* of first order has the form $f'(t) = Af(t)$, where $f \colon \mathbb{R} \to \mathbb{R}^n$ is a differentiable function with derivative f' and $A \in \mathbb{R}^{n \times n}$. One can show that for a given A, all solutions f have the form $f(t) = \exp(At)c$ with $c \in \mathbb{R}$ (note: $t \in \mathbb{R}$). This is a generalization of the well-known derivative rule $(e^{at})' = ae^{at}$.

18.4 Non-negative Matrices

Remark 18.22. In probability theory and other practical fields, real matrices with only non-negative entries occur. We show that the eigenvalues and eigenvectors of such matrices exhibit a special structure. This is the theoretical basis for important applications such as the Google search algorithm.

Definition 18.23. One calls $A = (a_{ij}) \in \mathbb{R}^{n \times m}$

- *positive* (resp. *non-negative*), if $a_{ij} > 0$ (resp. $a_{ij} \geq 0$) holds for all i, j.
- *decomposable*, if $n = m$ and there exists $\varnothing \neq I \subsetneq \{1, \ldots, n\}$ with $a_{ij} = 0$ for all $i \in I$ and $j \notin I$.
- *indecomposable*, if $n = m$ and A is not decomposable.

We write $A < B$ (resp. $A \leq B$), if $B - A$ is positive (resp. non-negative). Furthermore, let $A_+ := (|a_{ij}|)_{ij}$.

Remark 18.24. One easily sees that $\leq$ defines an order relation on $\mathbb{R}^{n \times m}$. We use the notation $v \geq w$ and v_+ also for vectors v, w (viewed as $n \times 1$ or $1 \times n$ matrices). In contrast to $\mathbb{R}$, $\leq$ is not total on $\mathbb{R}^{n \times m}$. For example, neither $(1, -1) \leq 0$ nor $(1, -1) \geq 0$ holds.

Example 18.25.

(a) In stochastics, the long-term development of random events is studied using *Markov chains*. In the simplest case, a Markov chain consists of states $Z_1, \ldots, Z_n$. Let the probability that a transition from Z_i to Z_j occurs be w_{ij}. $W = (w_{ij})$ is called the *transition matrix* of the Markov chain. If the process starts in Z_i, then the vector $e_i W^k$ describes the state probabilities after k time

units. One is therefore interested in $\lim_{k\to\infty} W^k$. Obviously, W is non-negative and every row sum is 1. Matrices with this property are called (row-)*stochastic*. Obviously, $v = (1, \ldots, 1)^{\mathrm{t}}$ is an eigenvector of W for the eigenvalue 1. According to Exercise III.20, W^k is also stochastic for $k \geq 0$. From Lemma 17.62 and Theorem 18.11 it follows that

$$\rho(W) = \lim_{k\to\infty} \sqrt[k]{\|W^k\|_\infty} = 1.$$

(b) Obviously, every positive matrix is irreducible.

(c) A permutation matrix $P_\sigma = (\delta_{i\sigma(j)})$ is irreducible if and only if $\sigma \in S_n$ is an n-cycle. For if $I \subseteq \{1, \ldots, n\}$ is the set of digits of a cycle of σ, then $\delta_{i\sigma(j)} = 0$ for all $i \in I$ and $j \notin I$.

(d) For every reducible matrix $A \in \mathbb{R}^{n\times m}$ there exists a permutation matrix P and $1 \leq k < n$ with $PAP^{\mathrm{t}} = \begin{pmatrix} A_1 & A_2 \\ 0_{k\times(n-k)} & A_3 \end{pmatrix}$.

(e) If A is irreducible, then A^{t} is also irreducible.

Theorem 18.26 (Perron)**.** *For every non-negative matrix $A \in \mathbb{R}^{n\times n}$, $\rho(A)$ is an eigenvalue with a non-negative eigenvector.*

Proof.
Step 1: $\rho(A)$ is an eigenvalue of A.
In the case $\rho(A) = 0$, $\rho(A)$ is an eigenvalue. So let $\rho(A) > 0$. By replacing A with $\rho(A)^{-1}A \geq 0$, one can assume $\rho(A) = 1$. For $0 < t < 1$ and $m \in \mathbb{N}$, we have $\rho(tA) = t < 1$ and

$$(1 - tA)^{-1} \overset{18.8}{=} \sum_{k=0}^{\infty} (tA)^k \geq 1_n + tA + \ldots + (tA)^m.$$

If 1 is not an eigenvalue of A, then $1 - A$ is invertible and it holds that

$$(1 - A)^{-1} = (\lim_{t\to 1^-} (1 - tA))^{-1} = \lim_{t\to 1^-} (1 - tA)^{-1} \geq \lim_{t\to 1^-} (1_n + tA + \ldots + (tA)^m) = 1_n + A + \ldots + A^m$$

for all $m \in \mathbb{N}$. In particular, $\lim_{m\to\infty} A^m = 0$ in contradiction to Theorem 18.8.

Step 2: Every positive matrix A possesses a positive eigenvector for the eigenvalue $\rho(A)$.[4]
As above, we can assume $\rho(A) = 1$. Let $v = (v_1, \ldots, v_n)$ be an eigenvector for the eigenvalue 1 of A. Then

$$v_+ = (Av)_+ \leq A_+ v_+ = Av_+,$$

thus $w := (A - 1_n)v_+ \geq 0$. In the case $w = 0$, $v_+ = Av_+ > 0$ is a positive eigenvector of A. Now let $w \neq 0$. Then $Aw > 0$ because of $A > 0$. Therefore there exists an $\epsilon > 0$ with $Aw \geq \epsilon Av_+$. For $z := Av_+ > 0$ it holds that

$$(A - 1_n)z = A(A - 1_n)v_+ = Aw \geq \epsilon z,$$

thus $Az \geq (1 + \epsilon)z$. For $B := (1 + \epsilon)^{-1}A$ it follows that $Bz \geq z$ and $B^k z \geq z$ for all $k \in \mathbb{N}$. On the other hand, $\rho(B) = (1 + \epsilon)^{-1}\rho(A) < 1$ and $\lim_{k\to\infty} B^k = 0$ according to Theorem 18.8. Contradiction.

Step 3: Every non-negative matrix A possesses a non-negative eigenvector for the eigenvalue $\rho(A)$.
For $k \in \mathbb{N}$ let $A_k := A + \frac{1}{k}J > 0$, where $J = (1)_{i,j=1}^n$. Certainly $A_1 > A_2 > \ldots > A$ and $A_1^m > A_2^m >$

[4] This was originally proven by Perron.

$\ldots > A^m$ for all $m \in \mathbb{N}$. From Theorem 18.11 (applied for example with the Euclidean norm) it follows that $\rho(A_1) \geq \rho(A_2) \geq \ldots \geq \rho(A)$. In particular, there exists

$$\mu := \lim_{k\to\infty} \rho(A_k) \geq \rho(A).$$

According to Step 2, there exist positive eigenvectors v_k of A_k for the eigenvalue $\rho(A_k)$. After normalization, $|v_k| = 1$ holds. By Bolzano-Weierstraß, $(v_k)_k$ has a convergent subsequence. Wlog. let $v := \lim_{k\to\infty} v_k \geq 0$. Because of $|v| = 1$, $v \neq 0$. Furthermore,

$$Av = \lim_{k\to\infty} A_k \lim_{k\to\infty} v_k = \lim_{k\to\infty} A_k v_k = \lim_{k\to\infty} \rho(A_k) v_k = \mu v.$$

Because of $\mu \geq \rho(A)$, it follows that $\mu = \rho(A)$. □

Theorem 18.27 (Perron-Frobenius)**.** *For every non-negative irreducible matrix $A \in \mathbb{R}^{n\times n}$ the following holds:*

(a) The algebraic multiplicity of $\rho(A)$ as an eigenvalue is 1*.*

(b) $E_{\rho(A)}(A) = \langle v \rangle$ with $v > 0$.

(c) Up to scalar multiplication, v is the only non-negative eigenvector of A.

Proof.

(a,b) According to Perron, there exists an eigenvector $v \geq 0$ for the eigenvalue $\rho(A)$. Let $I := \{1 \leq i \leq n : v_i = 0\}$. For $i \in I$ it holds that

$$0 = \rho(A) v_i = (Av)_i = \sum_{j=1}^{n} a_{ij} v_j = \sum_{j\notin I} a_{ij} v_j.$$

This only shows if $a_{ij} = 0$ for all $i \in I$ and $j \notin I$. Since A is irreducible, $I = \varnothing$ must hold, i. e. $v > 0$.

Now let also $Aw = \rho(A)w$ with $w \in \mathbb{R}^{n\times 1}$. Then there exists a $\lambda \in \mathbb{R}$ such that $w - \lambda v \geq 0$ vanishes at at least one coordinate. Since we have already seen that every non-negative eigenvector for the eigenvalue $\rho(A)$ is positive, it follows that $w = \lambda v$. Therefore, at least the geometric multiplicity of $\rho(A)$ is equal to 1. The Jordan normal form of A thus possesses only one block for $\rho(A)$. It is therefore sufficient to show

$$\operatorname{Ker}\big((A - \rho(A)1_n)^2\big) = \langle v \rangle.$$

Let $w \in \operatorname{Ker}\big((A - \rho(A)1_n)^2\big)$. Because $(A - \rho(A)1_n)w \in \langle v \rangle$, there exists a $\lambda \in \mathbb{C}$ with $Aw - \rho(A)w = \lambda v$. Since A^{t} is also irreducible with $\rho(A^{\mathrm{t}}) = \rho(A)$, there exists $u > 0$ with $A^{\mathrm{t}}u = \rho(A)u$. It follows that

$$\lambda[v,u] = [Aw,u] - \rho(A)[w,u] = w^{\mathrm{t}}A^{\mathrm{t}}u - \rho(A)[w,u] = \rho(A)\big([w,u] - [w,u]\big) = 0.$$

Because $[v,u] > 0$, we have $\lambda = 0$ and $w \in \langle v \rangle$ as desired.

(c) Let $w \geq 0$ be an arbitrary eigenvector of A for the eigenvalue $\lambda \in \mathbb{C}$. As above, $v > 0$ follows from the irreducibility of A. As before, let $u > 0$ with $A^{\mathrm{t}}u = \rho(A)u$. Because $[w,u] > 0$ and

$$\lambda[w,u] = [Aw,u] = w^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[w,u]$$

it holds that $\lambda = \rho(A)$. Thus v is, up to scaling, the only non-negative eigenvector of A. □

Remark 18.28. The next theorem generalizes the estimate $\rho(A) \leq \|A\|_\infty$ from Theorem 18.7 (set $x = (1, \ldots, 1)$ and use Lemma 17.62).

Theorem 18.29 (COLLATZ-WIELANDT)**.** *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}_{\geq 0}$ be irreducible and $x > 0$. Then*

$$\min\Big\{\sum_{j=1}^{n} a_{ij}x_j/x_i : i = 1, \ldots, n\Big\} \leq \rho(A) \leq \max\Big\{\sum_{j=1}^{n} a_{ij}x_j/x_i : i = 1, \ldots, n\Big\}.$$

Proof. Let $u > 0$ with $A^{\mathrm{t}}u = \rho(A)u$. Let $y_i := \sum_{j=1}^{n} a_{ij}x_j$ and $z_i := y_i/x_i$. Then

$$\sum_{j=1}^{n}(z_j - \rho(A))x_ju_j = \sum_{j=1}^{n} y_ju_j - \sum_{j=1}^{n} x_j \sum_{k=1}^{n} a_{kj}u_k = \sum_{j=1}^{n} y_ju_j - \sum_{k=1}^{n} u_k \sum_{j=1}^{n} a_{kj}x_j = 0.$$

Because $x, u > 0$, there exist $1 \leq s, t \leq n$ with $z_s \leq \rho(A)$ and $z_t \geq \rho(A)$. Thus $\min_i z_i \leq \rho(A) \leq \max_i z_i$. □

Theorem 18.30 (VON MISES)**.** *Let $A \in \mathbb{R}^{n\times n}$ be non-negative and irreducible. Then the sequence*

$$x_{k+1} := \frac{Ax_k}{|Ax_k|} \qquad (k = 1, 2, \ldots)$$

converges for every positive starting vector $x_1 \in \mathbb{R}^{n\times 1}$ to a positive eigenvector for $\rho(A)$. In particular, $\rho(A) = \lim_{k\to\infty} |Ax_k|$.

Proof. According to Perron-Frobenius, the requirements of the power method (Theorem 17.70) are satisfied. However, we must verify that the iteration converges for *every* positive starting vector x_1. Let $v, u > 0$ with $Av = \rho(A)v$ and $A^{\mathrm{t}}u = \rho(A)u$. Because $[u, v] > 0$, we have $\mathbb{R}^n = \langle v\rangle \oplus u^\perp$. For $x \in u^\perp$, it holds that

$$[Ax, u] = x^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[x, u] = 0.$$

This shows that $u^\perp$ is A-invariant. If one extends v with vectors from $u^\perp$ to a basis of $\mathbb{R}^n$, then A is transformed into the form $\left(\begin{smallmatrix}\rho(A) & 0\\ 0 & *\end{smallmatrix}\right)$. Because $[x_1, u] > 0$, it follows that $x_1 \notin u^\perp$. The proof of Theorem 17.70 now shows that the power method converges for x_1. □

Lemma 18.31. *Let $A \geq 0$ be irreducible and λ be an eigenvalue of A with absolute value $\rho(A)$. Then there exists a diagonal matrix $U \in \mathrm{U}(n, \mathbb{C})$ with $\rho(A)A = \lambda UAU^*$.*

Proof. Wlog. let $\rho(A) > 0$. Let $w \in \mathbb{C}^{n\times 1}$ be an eigenvector of A for the eigenvalue λ. For $z := w_+ \geq 0$, it holds that

$$\rho(A)z = (\lambda w)_+ = (Aw)_+ \leq Az.$$

Let $u > 0$ with $A^{\mathrm{t}}u = \rho(A)u$. From

$$0 < \rho(A)[z, u] \leq [Az, u] = z^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[z, u]$$

it follows that $Az = \rho(A)z$ and $z > 0$. Let $U := \mathrm{diag}(w_1/z_1, \ldots, w_n/z_n) \in \mathrm{U}(n, \mathbb{C})$. For $B := (b_{ij}) = \rho(A)\lambda^{-1}U^*AU \in \mathbb{C}^{n\times n}$, it holds that

$$Bz = \frac{\rho(A)}{\lambda}U^*Aw = \rho(A)U^*w = \rho(A)z = Az.$$

Since the entries of B and A differ only by factors of absolute value 1, we have $B_+ = A$. For $i = 1, \ldots, n$, we have

$$(Az)_i = (Bz)_i = \sum_{j=1}^{n} b_{ij} z_j \overset{\in\mathbb{R}}{=} \sum_{j=1}^{n} \mathrm{Re}(b_{ij}) z_j \leq \sum_{j=1}^{n} |\mathrm{Re}(b_{ij})| z_j \leq \sum_{j=1}^{n} |b_{ij}| z_j = (Az)_i.$$

This is only possible if $B = B_+ = A$. □

Remark 18.32. Note that A and UAU^* in Lemma 18.31 have the same main diagonal. Thus, if $a_{ii} \neq 0$ for some $1 \leq i \leq n$, then $\rho(A)$ is the only eigenvalue of absolute value $\rho(A)$.

Corollary 18.33. *For every positive matrix A, $\rho(A)$ is the only eigenvalue of absolute value $\rho(A)$.*

Proof. Follows from Remark 18.32. □

Example 18.34. We consider a Markov chain with transition matrix

$$W = \frac{1}{10} \begin{pmatrix} 5 & 1 & 4 \\ 2 & 4 & 4 \\ 0 & 9 & 1 \end{pmatrix}.$$

Obviously, W is irreducible. According to Remark 18.32, $\rho(A) = 1$ is the only eigenvalue of magnitude 1. According to Theorem 17.66, $W_\infty := \lim_{k\to\infty} W^k$ exists. To calculate W_∞, we must determine eigenvectors of W and W^{t} according to Corollary 17.67. Since W is stochastic, $v := (1, 1, 1)^{\mathrm{t}}$ is an eigenvector of W for the eigenvalue 1. One calculates that $w := \frac{1}{91}(18, 45, 28)^{\mathrm{t}}$ is an eigenvector of W^{t} with $[v, w] = 1$. Now it holds that

$$W_\infty = vw^{\mathrm{t}} = \frac{1}{91} \begin{pmatrix} 18 & 45 & 28 \\ 18 & 45 & 28 \\ 18 & 45 & 28 \end{pmatrix}.$$

In the long run, one is therefore in state Z_1, Z_2, or Z_3 with probability $\frac{18}{91} \approx 20\%$, 49%, or 31%, respectively.

Theorem 18.35 (Wielandt)**.** *Let $A \in \mathbb{R}^{n\times n}$ be non-negative and irreducible with exactly k eigenvalues of magnitude $\rho(A)$. Then:*

(a) The eigenvalues of magnitude $\rho(A)$ are $e^{2\pi \mathrm{i} j/k}\rho(A)$ for $j = 1, \ldots, k$, i.e., they are distributed uniformly on a circle in the complex plane.

(b) If μ is any eigenvalue of A, then $e^{2\pi \mathrm{i}/k}\mu$ is also an eigenvalue of A.

Proof.

(a) Let $\lambda_1\rho(A), \ldots, \lambda_k\rho(A) \in \mathbb{C}$ be the eigenvalues of A of magnitude $\rho(A)$, i.e., $|\lambda_1| = \ldots = |\lambda_k| = 1$. According to Lemma 18.31, for each i there exists a $U_i \in \mathrm{GL}(n, \mathbb{C})$ with $\lambda_i A = U_i^{-1} A U_i \approx A$ (unitarity is not needed). A comparison of the eigenvalues shows $\{\lambda_i\lambda_j : j = 1, \ldots, k\} = \{\lambda_1, \ldots, \lambda_k\}$ for $i = 1, \ldots, k$. It follows that

$$\lambda_i^k \prod_{j=1}^{k} \lambda_j = \prod_{j=1}^{k} (\lambda_i\lambda_j) = \prod_{j=1}^{k} \lambda_j$$

and $\lambda_i^k = 1$ for $i = 1, \dots, k$. Therefore, $\lambda_1, \dots, \lambda_k$ are exactly the k-th roots of unity defined in Example 11.28 and Definition 17.6.

(b) Wlog. let $\lambda_1 = e^{2\pi \mathrm{i}/k}$. The similarity $\lambda_1 A \approx A$ shows that $\lambda_1 \mu$ is an eigenvalue of A. □

Theorem 18.36. *Let $A \in \mathbb{R}^{n \times n}$ be non-negative. Then for every eigenvalue $\lambda \in \mathbb{C}$ of A of magnitude $\rho(A)$, there exists an $m \leq n$ with $\lambda^m = \rho(A)^m$.*

Proof. According to Example 18.25, there exists a permutation matrix P such that

$$P^{\mathrm{t}} A P = \begin{pmatrix} A_1 & & * \\ & \ddots & \\ 0 & & A_k \end{pmatrix}$$

with square non-negative irreducible matrices $A_1, \dots, A_k$. Every eigenvalue of A_i is also an eigenvalue of A (one extends a corresponding eigenvector with zeros). Since A has exactly n complex eigenvalues, for every eigenvalue λ of A there exists an i such that λ is an eigenvalue of A_i. The claim follows from Theorem 18.35. □

Example 18.37. Let P_σ be a permutation matrix and $\sigma = \sigma_1 \dots \sigma_k$ the decomposition into pairwise disjoint cycles (including 1-cycles). Let l_i be the length of σ_i. According to Example 15.19, P has the eigenvalues

$$\zeta_{l_1}, \zeta_{l_1}^2, \dots, \zeta_{l_1}^{l_1}, \zeta_{l_2}, \dots, \zeta_{l_2}^{l_2}, \dots, \zeta_{l_k}^{l_k}$$

with $\zeta_m = e^{2\pi \mathrm{i}/m}$ for $m \in \mathbb{N}$.

18.5 The PageRank

Remark 18.38. To sort results, the search engine Google assigns a weight to each website w, which measures how many other websites link to w. We write $w_i \to w_j$ if the website w_i contains a hyperlink to w_j. Let l_i be the number of (indexed) hyperlinks on w_i. One "defines" the PAGE *Rank*[5] of w_i by

$$p_i := \frac{1-d}{n} + d \sum_{\substack{1 \leq j \leq n \\ w_j \to w_i}} \frac{p_j}{l_j}, \tag{18.1}$$

where n is the number of all indexed pages and $0 < d < 1$ is a constant damping factor.[6] The PageRank of w_i is thus "large" if many pages w_j (with high p_j and low l_j) link to w_i. The condition $d < 1$ prevents isolated websites (which are not linked to) from receiving PageRank 0. The p_i cannot be calculated directly through (18.1). It is not even clear whether they are uniquely determined by (18.1).

Definition 18.39. With the above notation, $G = (g_{ij})$ with

$$g_{ij} = \begin{cases} \frac{d}{l_i} + \frac{1-d}{n} & \text{if } w_i \to w_j, \\ \frac{1-d}{n} & \text{otherwise} \end{cases}$$

is called the *Google matrix*.

[5] Named after one of the Google founders Larry Page (not after "web page").
[6] Google uses $d = 0.85$. The exact value of n is unknown, but is likely to be greater than 10^{10}.

Theorem 18.40. *The Google matrix is stochastic. Furthermore, $(p_1, \ldots, p_n)$ is the unique eigenvector of G^{t} for the eigenvalue 1 with $p_1 + \ldots + p_n = 1$. In particular, $p_1, \ldots, p_n$ are uniquely determined.*

Proof. Because $0 < d < 1$, we have $G > 0$. The i-th row sum of G is

$$l_i\Big(\frac{d}{l_i} + \frac{1-d}{n}\Big) + (n - l_i)\frac{1-d}{n} = 1,$$

i. e. G is stochastic. By the definition of p_i, it holds that

$$\sum_{i=1}^{n} p_i = 1 - d + d\sum_{i=1}^{n} \sum_{\substack{1 \le j \le n \\ w_j \to w_i}} \frac{p_j}{l_j} = 1 - d + d\sum_{j=1}^{n} p_j \sum_{\substack{1 \le i \le n \\ w_j \to w_i}} \frac{1}{l_j} = 1 - d + d\sum_{j=1}^{n} p_j.$$

It follows that $\sum_{i=1}^{n} p_i = 1$. For $v := (p_1, \ldots, p_n)$, we have

$$(G^{\mathrm{t}}v)_i = \sum_{j=1}^{n} g_{ji} p_j = \frac{1-d}{n}\sum_{i=j}^{n} p_j + \sum_{\substack{1 \le j \le n \\ w_j \to w_i}} \frac{dp_j}{l_j} = p_i,$$

i. e. v is an eigenvector of G^{t} for the eigenvalue $\rho(G^{\mathrm{t}}) = \rho(G) = 1$. According to Perron-Frobenius, v is uniquely determined by $\sum p_i = 1$. □

Remark 18.41.

(a) In practice, p_i is determined (approximately) using the von Mises theorem. One can show that $|\lambda| \le d$ holds for all eigenvalues $\lambda \neq 1$ of G. Thus, the power method converges at least by the factor $\frac{1}{d}$ (see proof of Theorem 17.70).

(b) One can also view G as the transition matrix of a Markov chain, in which a surfer on w_i clicks on one of the hyperlinks on w_i with probability d and accesses a random other website with probability $1 - d$. Once the p_i have been determined, one can use Corollary 17.67 to determine the long-term residence probabilities of the surfer (see Example 18.34).

19 Linear Optimization

19.1 Linear Programs

Remark 19.1.

(a) In linear optimization, one maximizes (or minimizes) an *objective function* (for example, profit or costs of a company) under constraints (for example, demand or capacity).

(b) For the sake of consistency, we will only use column vectors. Deviating from previous chapters, let $\mathbb{R}^n$ be the space of real *column* vectors.

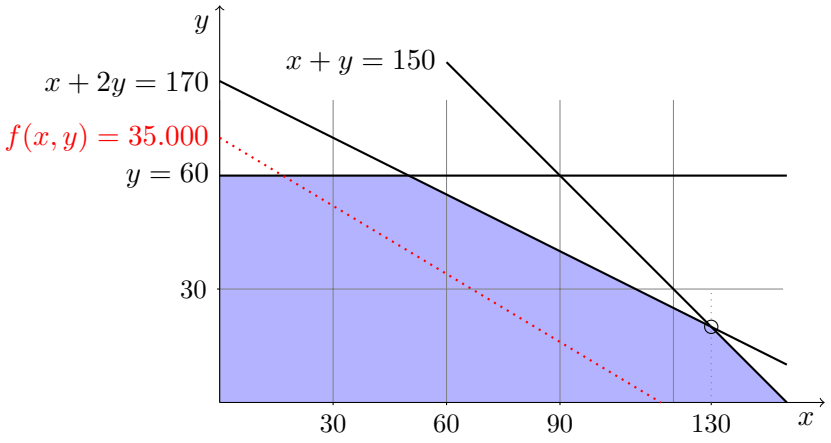
Example 19.2. A company produces two products P and Q using three machines A, B, C. The machines have a maximum runtime of 170 hours (A), 150 hours (B), and 180 hours (C) during the production period. The production of P and Q requires the following resources:

Product	Runtime A	Runtime B	Runtime C	Revenue
P	$1h$	$1h$	0	300 €
Q	$2h$	$1h$	$3h$	500 €

In what ratio should units of P and Q be produced to maximize the total revenue? Let x and y be the number of products P and Q to be produced. The objective function $f\colon \mathbb{N}_0^2 \to \mathbb{R}$, $(x, y) \mapsto 300x + 500y$ is to be maximized under the constraints

$$x, y \geq 0, \qquad x + 2y \leq 170, \qquad x + y \leq 150, \qquad 3y \leq 180$$

These inequalities describe the area marked in blue in $\mathbb{R}^2$:



The parameters for a revenue of $f(x, y) = 35.000$ are marked with the red dotted line. Since the slope $-3/5$ of this line does not depend on the revenue, it is easy to see that f is maximal for $(x, y) = (130, 20)$. The revenue then amounts to 49.000 €. We are lucky that the solution (x, y) here is

unique and integer-valued. For problems with significantly more parameters, this graphical approach is unsuitable.

Remark 19.3.

(i) A "linear" objective function has the form $f\colon \mathbb{R}^m \to \mathbb{R}$, $x \mapsto c^{\mathrm{t}}x + a$ for some $c \in \mathbb{R}^m$ and $a \in \mathbb{R}$. Since constants have no influence on the position of extrema, one can assume $a = 0$. Because of $\max f = -\min(-f)$, one can restrict oneself to the search for maxima.

(ii) Linear constraints on $x \in \mathbb{R}^m$ can occur in the form $c^{\mathrm{t}}x \leq b$, $c^{\mathrm{t}}x = b$ or $c^{\mathrm{t}}x \geq b$ with $c \in \mathbb{R}^m$ and $b \in \mathbb{R}$. Because of $c^{\mathrm{t}}x \geq b \iff (-c)^{\mathrm{t}}x \leq -b$, one can do without the third variant. By adding a so-called *slack variable* y, one can replace $c^{\mathrm{t}}x \leq b$ by $c^{\mathrm{t}}x + y = b$ and $y \geq 0$. Slack variables do not appear in the objective function. In general, one can assume $x_i \geq 0$ for $i = 1, \ldots, m$ by replacing x_i with two variables $x_i^+, x_i^- \geq 0$ with $x = x_+ - x_-$. All constraints can now be summarized in matrix form $Ax = b$ with $b \in \mathbb{R}^n$ and $x \geq 0$. Using the Gaussian algorithm, one can transform A into row echelon form and delete zero rows. In particular, one can assume that A has full rank. In the case $n \geq m$, $Ax = b$ can have at most one solution according to Remark 6.7. The search for a maximum of the objective function is uninteresting in this case. We will therefore assume $n < m$. By multiplying rows by -1, one achieves $b \geq 0$.

(iii) Alternatively, constraints of the form $c^{\mathrm{t}}x = b$ can be split into $c^{\mathrm{t}}x \leq b$ and $-c^{\mathrm{t}}x \leq -b$. One thus achieves $Ax \leq b$ without adding slack variables and without restrictions on x. We will return to this in Definition 19.14.

Definition 19.4. A *linear program* (in standard form) $L = (A, b, c)$ consists of $A \in \mathbb{R}^{n\times m}$ with rank $n < m$, $b \in \mathbb{R}^n$ with $b \geq 0$ and $c \in \mathbb{R}^m$. A non-negative vector $x \in \mathbb{R}^m$ is called *feasible*, if x satisfies the constraint $Ax = b$. We are looking for feasible $x_{\max}$ that maximize the objective function $f\colon \mathbb{R}^m \to \mathbb{R}$, $x \mapsto c^{\mathrm{t}}x$, i.e., $f(x_{\max}) \geq f(x)$ for all feasible $x \in \mathbb{R}^m$. Furthermore, L is called

- *solvable*, if at least one $x_{\max}$ exists.
- *infeasible*, if there are no feasible x.
- *unbounded*, if f becomes arbitrarily large on the set of feasible x.

Remark 19.5. Let L be a linear program with objective function f. Suppose f is bounded on the set M of feasible x. By Bolzano-Weierstraß, there exists a sequence of feasible points $(x_i)_i$ with $s := \lim_{i\to\infty} f(x_i) = \sup_{x\in M} f(x)$. Since M is closed, $x_{\max} := \lim_{i\to\infty} x_i$ exists. Since f is continuous as a linear function, $f(x_{\max}) = s = \max_{x\in M} f(x)$, i. e. L is solvable. This shows that every linear program is either solvable, infeasible, or unbounded.

Example 19.6. The problem from Example 19.2 can be written as a linear program with three slack variables:

$$A = \begin{pmatrix} 1 & 2 & 1 & . & . \\ 1 & 1 & . & 1 & . \\ . & 3 & . & . & 1 \end{pmatrix}, \qquad b = (170, 150, 180)^{\mathrm{t}}, \qquad c = (300, 500, 0, 0, 0)^{\mathrm{t}}.$$

19.2 Convex Sets

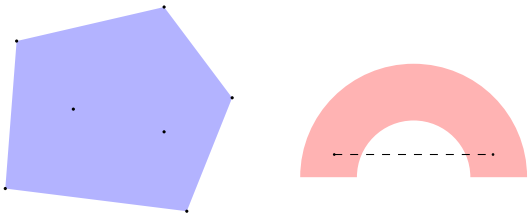
Definition 19.7.

- A subset $\Delta \subseteq \mathbb{R}^n$ is called *convex*, if $\lambda x + (1-\lambda)y \in \Delta$ holds for all $x, y \in \Delta$ and $0 \leq \lambda \leq 1$. Intuitively, this means that for any two points in Δ, their connecting line segment also lies in Δ.
- For $v_1, \ldots, v_k \in \mathbb{R}^n$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ with $\lambda_1 + \ldots + \lambda_k = 1$, one calls $\lambda_1 v_1 + \ldots + \lambda_k v_k$ a *convex combination* of $v_1, \ldots, v_k$.
- The set of all convex combinations of elements from Δ is called the *convex hull* of Δ and is denoted by $\operatorname{con}(\Delta)$ (or $\operatorname{con}(x_1, \ldots, x_k)$ if $\Delta = \{x_1, \ldots, x_k\}$). Obviously, $\operatorname{con}(\Delta)$ is the "smallest" convex subset containing Δ, i. e.

$$\operatorname{con}(\Delta) = \bigcap_{\substack{\Delta \subseteq \Gamma \subseteq \mathbb{R}^n \\ \Gamma \text{ convex}}} \Gamma.$$

Example 19.8.

(a) Convex sets in $\mathbb{R}$ are (open or (half-)closed) intervals. In $\mathbb{R}^2$, convex sets have neither holes nor inward-facing "kinks". Here is the convex hull of seven points (blue) and a non-convex set (red):



(b) For every norm, $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$ is convex by the triangle inequality. For $\|.\|_2$ one obtains the unit ball, for $\|.\|_\infty$ the cube with edge length 2, and for $\|.\|_1$ the so-called *standard simplex* $\operatorname{con}(\pm e_1, \ldots, \pm e_n)$ (for $n = 3$ this is an octahedron).

Theorem 19.9 (Carathéodory). *Let $\Delta \subseteq \mathbb{R}^n$ and $x \in \operatorname{con}(\Delta)$. Then x is a convex combination of at most $n+1$ elements in Δ.*

Proof. Let $x = \lambda_1 v_1 + \ldots + \lambda_k v_k$ be a convex combination with $v_1, \ldots, v_k \in \Delta$ and k minimal. Suppose $k > n+1$. We form the matrix $A \in \mathbb{R}^{(n+1)\times k}$ with columns $v_1, \ldots, v_k$ and the additional row $(1, \ldots, 1)$. Because $\operatorname{rk}(A) \leq n+1 < k$, there exists a $y \in \mathbb{R}^k \setminus \{0\}$ with $Ay = 0$. Thus $y_1 v_1 + \ldots + y_k v_k = 0$ and $y_1 + \ldots + y_k = 0$. Let $1 \leq s \leq k$ with $\frac{y_s}{\lambda_s} \geq \frac{y_i}{\lambda_i}$ for $i = 1, \ldots, k$. For $\lambda_i' := \lambda_i\big(1 - \frac{y_i \lambda_s}{\lambda_i y_s}\big) \geq 0$ it holds that

$$\sum_{i \neq s} \lambda_i' v_i = \sum_{i=1}^{k} \lambda_i v_i - \frac{\lambda_s}{y_s} \sum_{i=1}^{k} y_i v_i = x,$$

$$\sum_{i \neq s} \lambda_i' = \sum_{i=1}^{k} \lambda_i - \frac{\lambda_s}{y_s} \sum_{i=1}^{k} y_i = 1$$

in contradiction to the choice of k. □

Example 19.10. In contrast to subspaces, $\operatorname{con}(\Delta) \subseteq \mathbb{R}^n$ cannot always be "spanned" (as a convex combination) by finitely (or countably) many elements in Δ. To see this, consider the convex unit disk $\Delta := \{x \in \mathbb{R}^2 : |x| \leq 1\}$. Suppose there exist countably many $v_1, v_2, \ldots \in \Delta$ such that every element in Δ is a convex combination of the v_i. As is well known, there are uncountably many $x \in \Delta$ with $|x| = 1$ (parametrized by $(\cos(\varphi), \sin(\varphi))$ with $\varphi \in \mathbb{R}$). We can assume $x \neq \pm v_i$ for $i \in \mathbb{N}$. Then wlog. $x = \lambda_1 v_1 + \ldots + \lambda_k v_k$ with $k \geq 2$, $\lambda_i > 0$ and $\lambda_1 + \ldots + \lambda_k = 1$. According to the triangle inequality and the Cauchy-Schwarz inequality, it holds that

$$1 = [x, x] \leq \sum_{i=1}^{k} \lambda_i |[x, v_i]| \leq \sum_{i=1}^{k} \lambda_i |x||v_i| \leq \sum_{i=1}^{k} \lambda_i = 1$$

only if x is linearly dependent on each v_i and $|v_1| = \ldots = |v_k| = 1$. But then $x = \pm v_1$. Thus, one needs uncountably many elements from Δ to represent every element as a convex combination.

Lemma 19.11 (Farkas)**.** *For all $U \leq \mathbb{R}^n$, exactly one of the following statements holds:*

(1) There exists a $u \in U$ with $u \geq 0$ and $u_1 > 0$.

(2) There exists a $v \in U^\perp$ with $v \geq 0$ and $v_1 > 0$.

Proof. If (1) and (2) were to hold simultaneously with u and v respectively, then $0 = [u, v] = u_1 v_1 + \ldots + u_n v_n \geq u_1 v_1 > 0$. We argue by induction on n. For $n = 1$, $U = \mathbb{R}$ or $U^\perp = \mathbb{R}$. So let $n \geq 2$ and

$$\widetilde{U} := \left\{x \in \mathbb{R}^{n-1} : (x, 0) \in U\right\} \leq \mathbb{R}^{n-1},$$
$$\widehat{U} := \left\{x \in \mathbb{R}^{n-1} : \exists y \in \mathbb{R} : (x, y) \in U\right\} \leq \mathbb{R}^{n-1}.$$

For dimensional reasons, there exists a $u_0 \in U$ such that $U = (\widetilde{U} \times \{0\}) + \langle u_0 \rangle$. From Exercise II.6 it follows that

$$U^\perp = (\widetilde{U}^\perp \times \mathbb{R}) \cap u_0^\perp. \tag{19.1}$$

By induction, there are three alternatives:

(a) There exists a $\tilde{u} \in \widetilde{U}$ with $\tilde{u} \geq 0$ and $\tilde{u}_1 > 0$. Then (1) holds with $u := (\tilde{u}, 0) \in U$.

(b) There exists a $\hat{v} \in \widehat{U}^\perp$ with $\hat{v} \geq 0$ and $\hat{v}_1 > 0$. For $v := (\hat{v}, 0)$ and all $u = (\hat{u}, u_n) \in U$ it holds that $[u, v] = [\hat{u}, \hat{v}] = 0$, i. e. (2) holds for v.

(c) There exist $\tilde{v} \in \widetilde{U}^\perp$ and $\hat{u} \in \widehat{U}$ with $\tilde{v}, \hat{u} \geq 0$ and $\tilde{v}_1, \hat{u}_1 > 0$. Let $y \in \mathbb{R}$ with $u = (\hat{u}, y) \in U$. In the case $y \geq 0$, (1) holds for u. Therefore, let $y < 0$. According to (19.1), there exists a $z \in \mathbb{R}$ with $v = (\tilde{v}, z) \in u_0^\perp \subseteq U^\perp$. Because $\hat{u}, \tilde{v} \geq 0$, we have

$$0 = [u, v] = [\hat{u}, \tilde{v}] + yz \geq yz$$

and $z \geq 0$. Thus (2) holds for v. □

Example 19.12. In $\mathbb{R}^2$, Farkas' Lemma states that for every line g through the origin, either g or the line $g^\perp$ perpendicular to it passes through the first quadrant ($x, y \geq 0$).

Corollary 19.13. *For $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$, exactly one of the following statements holds:*

(1) There exists an $x \in \mathbb{R}^m$ with $x \geq 0$ and $Ax = b$.

(2) There exists a $y \in \mathbb{R}^n$ with $b^t y < 0$ and $A^t y \geq 0$.

Proof. If x and y are solutions of (1) and (2) respectively, then $0 \leq (A^t y)^t x = y^t(Ax) = y^t b < 0$. Thus, at most one of the statements can hold. We append $-b$ on the left as a new column to A and obtain $S := (-b|A) \in \mathbb{R}^{n\times(m+1)}$. Let $U := \mathrm{Ker}(S) \leq \mathbb{R}^{m+1}$. Suppose there exists an $\tilde{x} = (x_1, x)^t \in U$ with $x \geq 0$ and $x_1 > 0$. After scaling, we can assume $x_1 = 1$. Then x^t is a solution of (1).

By Farkas' Lemma, we can now assume that there exists a $\tilde{z} = (z_1, z)^t \in U^\perp$ with $z \geq 0$ and $z_1 > 0$. According to Example 11.16, $U^\perp = \{S^t y : y \in \mathbb{R}^n\}$. So let $y \in \mathbb{R}^n$ with $S^t y = \tilde{z}$. Then $b^t y = -z_1 < 0$ and $A^t y = z \geq 0$. Thus (2) holds for y. □

Definition 19.14. Let $L = (A, b, c)$ be a linear program. The *dual* linear program L^* to L minimizes the objective function $f^*\colon \mathbb{R}^n \to \mathbb{R}$, $y \mapsto b^t y$ under the constraint $A^t y \geq c$. Analogous to Definition 19.4, one defines solvability, feasibility, and boundedness of L^*.

Example 19.15. The dual program to $L = (A, b, c)$ from Example 19.6 is

$$\min_{y\in\mathbb{R}^3}(170, 150, 180)y \qquad \begin{pmatrix} 1 & 1 & . \\ 2 & 1 & 3 \\ 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{pmatrix} y \geq \begin{pmatrix} 300 \\ 500 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Because of $y, c \geq 0$, L^* is bounded. Obviously, y is a minimal solution only if equality holds in the first two rows, i.e., $y_1 + y_2 = 300$ and $2y_1 + y_2 + 3y_3 = 500$. Then $y = (y_1, 300 - y_1, (200 - y_1)/3)$ and

$$f^*(y) = c^t y = 170y_1 + 150(300 - y_1) + 60(200 - y_1) = 57.000 - 40y_1.$$

The minimum $f^*(y) = 49.000$ is obtained for $y_1 = 200$ because of $y_3 \geq 0$. This is exactly the maximum value of L. The next theorem shows that this is no coincidence.

Theorem 19.16 (Duality Theorem)**.** *For every linear program $L = (c, A, b)$, the following holds:*

(a) L is solvable if and only if L^* is solvable. For solutions $x_{\max}$ and $y_{\min}$, it holds that $f(x_{\max}) = f^*(y_{\min})$.

(b) If L (resp. L^*) is unbounded, then L^* (resp. L) is infeasible.

Proof.

(b) If $x \in \mathbb{R}^m$ is feasible for L and $y \in \mathbb{R}^n$ is feasible for L^*, then

$$f(x) = c^t x \leq y^t A x = y^t b = f^*(y). \tag{19.2}$$

If L (resp. L^*) is unbounded, then L^* (resp. L) must be infeasible.

(a) Let $x_{\max} \in \mathbb{R}^m$ be a solution of L. Let

$$A_1 := \begin{pmatrix} 0 & A \\ -1 & c^t \end{pmatrix} \in \mathbb{R}^{(n+1)\times(m+1)}$$

and $U := \mathrm{Ker}(A_1)$. If there existed a $\binom{t}{w} \in \mathrm{Ker}(A_1)$ with $t > 0$ and $w \in \mathbb{R}^m_{\geq 0}$, then $x_{\max} + w \geq 0$, $A(x_{\max} + w) = b$ and

$$f(x_{\max} + w) = f(x_{\max}) + t > f(x_{\max}).$$

By Farkas' Lemma and Example 11.16, there exists a $y = A_1^{\mathrm{t}}\binom{u}{s} \geq 0$ with $u \in \mathbb{R}^n$ and $y_1 > 0$. Now $s < 0$ and $A^{\mathrm{t}}u \geq -sc$. For $v := -\frac{1}{s}u$, it finally holds that $A^{\mathrm{t}}v \geq c$, i. e. v is feasible for L^*.

Now we consider

$$A_2 = \begin{pmatrix} c & -A^{\mathrm{t}} & A^{\mathrm{t}} & 1_m & 0 & 0 \\ -b & 0 & 0 & 0 & A & 0 \\ 0 & b^{\mathrm{t}} & -b^{\mathrm{t}} & 0 & -c^{\mathrm{t}} & 1 \end{pmatrix} \in \mathbb{R}^{(n+m+1)\times 2(n+m+1)}.$$

Suppose there exist $x \in \mathbb{R}^m$, $u \in \mathbb{R}^n$ and $s \in \mathbb{R}$ with $(x^{\mathrm{t}}, u^{\mathrm{t}}, s)A_2 \geq 0$ and a positive first coordinate. Then

$$c^{\mathrm{t}}x > u^{\mathrm{t}}b, \qquad Ax \leq sb, \qquad Ax \geq sb, \qquad x, s \geq 0, \qquad A^{\mathrm{t}}u \geq sc.$$

Suppose $s > 0$. By passing to $\frac{1}{s}(x^{\mathrm{t}}, u^{\mathrm{t}}, s)$, one can assume $s = 1$. Then x is feasible for L and u for L^*, but (19.2) is violated. Thus $s = 0$ and $Ax = 0$. For all $\lambda, \mu > 0$, $x_{\max} + \lambda x$ is feasible for L and $v + \mu u$ is feasible for L^*. By choosing λ or μ (if $c^{\mathrm{t}}x = 0$) sufficiently large, (19.2) becomes invalid again. Therefore, $(x^{\mathrm{t}}, u^{\mathrm{t}}, s)$ cannot exist.

Using Farkas' Lemma, one finds a non-negative vector $(1, u_+, u_-, z, x, s) \in \mathrm{Ker}(A_2)$. This means

$$Ax = b, \qquad A^{\mathrm{t}}(u_+ - u_-) = c + z \geq c, \qquad b^{\mathrm{t}}(u_+ - u_-) = c^{\mathrm{t}}x - s \leq c^{\mathrm{t}}x.$$

Thus x is feasible for L and $u := u_+ - u_-$ is feasible for L^*. Because of $f(x) = c^{\mathrm{t}}x \geq b^{\mathrm{t}}u = f^*(u)$, it follows that $s = 0$ from (19.2). Furthermore, u must be a solution for L^* and $f(x_{\max}) = c^{\mathrm{t}}x = b^{\mathrm{t}}u = f^*(u)$.

Finally, let $y_{\min}$ be a solution for L^*. If (2) in Corollary 19.13 holds for y, then $A^{\mathrm{t}}(y_{\min} + y) \geq c$ and $f^*(y_{\min} + y) = b^{\mathrm{t}}(y_{\min} + y) < b^{\mathrm{t}}y_{\min} = f^*(y_{\min})$. Thus (1) must hold, i. e. L is solvable. □

Example 19.17. It can happen that both L and L^* are infeasible. For example, for $A = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, $b = c = (1, 1)^{\mathrm{t}}$.

19.3 The Simplex Algorithm

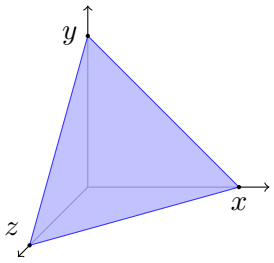
Definition 19.18.

- The feasible set $P := \{x \in \mathbb{R}^m_{\geq 0} : Ax = b\}$ of a linear program $L = (A, b, c)$ describes a *polyhedron*. Obviously, P is convex. One calls $x \in P$ a *vertex* of P or L if $P \setminus \{x\}$ is also convex.
- For $x \in P$, one calls $\mathrm{supp}(x) := \{i : x_i > 0\} \subseteq \{1, \ldots, m\}$ the *support* of x.
- For $I \subseteq \{1, \ldots, m\}$ and $x \in \mathbb{R}^m$, let $A_I := (a_{ij} : i = 1, \ldots, n,\ j \in I)$, $x_I := (x_i : i \in I)$ and $I' := \{1, \ldots, m\} \setminus I$. One calls I a *basis set* of L if A_I is invertible. If additionally $A_I^{-1}b \geq 0$, then I is called *feasible*.

Remark 19.19.

(a) In geometry, polyhedra are defined as sets of the form $P = \{x \in \mathbb{R}^n : Ax \leq b\}$. Bounded polyhedra are called *polytopes*. In $\mathbb{R}^2$, polytopes are convex polygons like the feasible region in Example 19.2. The vertices are exactly those points that one would colloquially call corners. The condition that $P \setminus \{x\}$ remains convex intuitively means that one can "round off" vertices.

(b) We saw in Remark 19.3 how to transform polyhedra of the form $\{x : Ax \leq b\}$ into polyhedra of the form $\{x \geq 0 : Ax = b\}$ by introducing slack variables. The geometric visualization of the polyhedra defined in Definition 19.18 thus refers to a proper subspace. For example, $P = \{(x, y, z) \geq 0 : x + y + z = 1\}$ describes an equilateral triangle on a plane in $\mathbb{R}^3$ with the vertices e_1, e_2 and e_3.



Lemma 19.20. *For every feasible point x of a linear program $L = (A, b, c)$, the following are equivalent:*

(1) x is a vertex.

(2) For $I := \operatorname{supp}(x)$, it holds that $\operatorname{rk}(A_I) = |I|$.

Proof.

(1) $\Rightarrow$ (2): Assume the columns of A_I are linearly dependent. Then there exists a $y \in \mathbb{R}^m$ with $\operatorname{supp}(y) \subseteq I$ and $Ay = 0$. Furthermore, there exists an $\epsilon > 0$ with $x \pm \epsilon y \geq 0$. Now $x \pm \epsilon y$ lie in the polyhedron P of L and $x = \frac{1}{2}(x - \epsilon y) + \frac{1}{2}(x + \epsilon y)$. Then $P \setminus \{x\}$ would be not convex. Contradiction.

(2) $\Rightarrow$ (1): Assume $P \setminus \{x\}$ is not convex. Then there exist $y, z \in P \setminus \{x\}$ and $0 < \lambda < 1$ with $x = \lambda y + (1 - \lambda)z$. Because of $y, z \geq 0$, it holds that $\operatorname{supp}(y), \operatorname{supp}(z) \subseteq I$. This shows $A_I y_I = b = A_I z_I$ and $y_I - z_I \in \operatorname{Ker}(A_I)$. Since A_I has full rank, it would follow that $y = z = x$. □

Example 19.21.

(a) In the case $b = 0$, $x = 0$ is a vertex with $I = \varnothing$.

(b) If I is a feasible basis set for L, then there exists exactly one feasible vertex x with $\operatorname{supp}(x) \subseteq I$, because $x_I = A_I^{-1} b \geq 0$. In the case $\operatorname{supp}(x) \subsetneq I$, x can belong to several basis sets.

Corollary 19.22. *Every linear program in m variables has at most $\sum_{k=0}^{n} \binom{m}{k} \leq 2^m$ vertices.*

Proof. Each vertex x is uniquely determined by $I := \operatorname{supp}(x)$, because $A_I x_I = b$ has (at most) one solution according to Lemma 19.20. The number of subsets $I \subseteq \{1, \ldots, m\}$ with $|I| \leq n$ is known to be

$$\sum_{k=0}^{n} \binom{m}{k} \leq \sum_{k=0}^{m} \binom{m}{k} = (1+1)^m = 2^m. \qquad \square$$

Remark 19.23. The estimate in Corollary 19.22 is not optimal for $n \geq 2$. Indeed, if x is a vertex, then every index set I with $\operatorname{supp}(x) \subseteq I$ and $\operatorname{rk}(A_I) = |I|$ leads to the same vertex. Thus, one only needs to consider so-called "antichains" $\{I_1, \ldots, I_k\}$ of $\{1, \ldots, m\}$, i.e., $I_s \not\subseteq I_t$ for all $s \neq t$. A theorem by Sperner provides the sharper estimate $k \leq \binom{m}{\lfloor m/2 \rfloor}$.[1]

[1] see notes on Logic and Set Theory

Nevertheless, the number of vertices can grow exponentially with m. For example, the linear program with $A = (1_n, 1_n) \in \mathbb{R}^{n \times 2n}$ and $b = (1, \ldots, 1)^{\mathrm{t}}$ has exactly $2^n = \sqrt{2}^m$ vertices (each I contains exactly one element from $\{i, i+n\}$ for $i = 1, \ldots, n$).

Theorem 19.24 (Fundamental Theorem of Linear Programming)**.** *If the linear program $L = (A, b, c)$ is solvable, then the maximum is attained at a vertex.*

Proof. Let $x_{\max}$ be a solution of L and $I := \operatorname{supp}(x_{\max})$. We can assume that $x_{\max}$ is not a vertex. As in the proof of Lemma 19.20, there exist $y \in \mathbb{R}^m \setminus \{0\}$ and $\epsilon > 0$ with $\operatorname{supp}(y) \subseteq I$, $Ay = 0$ and $x_\pm := x_{\max} \pm \epsilon y \geq 0$. Because of

$$f(x_{\max}) \pm \epsilon f(y) = f(x_\pm) \leq f(x_{\max})$$

it follows that $f(y) = 0$. We can choose ϵ such that $\operatorname{supp}(x_+) \subsetneq I$ or $\operatorname{supp}(x_-) \subsetneq I$ holds. If necessary, x_+ or x_- is a solution of L with a smaller support. If we continue in this way, we arrive after finitely many steps at a vertex x with $f(x) = f(x_{\max})$. □

Remark 19.25. According to Theorem 19.24 and Corollary 19.22, one only needs to consider finitely many (but possibly exponentially many) feasible vertices to solve a linear program. Nevertheless, it is possible that the maximum is attained at infinitely many points, for example in the trivial case $c = 0$. With the following theorem, one can check whether a found vertex is a solution.

Theorem 19.26 (Simplex Criterion)**.** *Let $L = (A, b, c)$ be a linear program with a feasible basis set I. If*

$$\gamma(I) := c_{I'}^{\mathrm{t}} - c_I^{\mathrm{t}} A_I^{-1} A_{I'} \leq 0,$$

then the vertex corresponding to I is a solution of L.

Proof. Wlog. let $I = \{1, \ldots, n\}$ and $A = (A_I, A_{I'})$. For every feasible vector x, $A_I x_I + A_{I'} x_{I'} = b$ holds. It follows that $x_I + A_I^{-1} A_{I'} x_{I'} = A_I^{-1} b$ and

$$f(x) = c^{\mathrm{t}} \begin{pmatrix} x_I \\ x_{I'} \end{pmatrix} = c_I^{\mathrm{t}}(A_I^{-1} b - A_I^{-1} A_{I'} x_{I'}) + c_{I'}^{\mathrm{t}} x_{I'} = c_I^{\mathrm{t}} A_I^{-1} b + \gamma(I) x_{I'} \overset{x \geq 0}{\leq} c_I^{\mathrm{t}} A_I^{-1} b$$

with equality if $x_{I'} = 0$, i. e. if x is the vertex with $\operatorname{supp}(x) \subseteq I$. □

Remark 19.27. In practice, it is costly to enumerate all vertices. The idea of the simplex method is to move from a vertex x to a "neighboring" vertex y with $f(y) > f(x)$. Intuitively, neighboring means that the connecting segment $S := \{\lambda x + (1-\lambda) y : 0 \leq \lambda \leq 1\}$ forms an edge of the polyhedron P, i. e. $P \setminus S$ is convex.

Lemma 19.28. *Let $L = (A, b, c)$ be a linear program with a feasible basis set I. Let x be the corresponding vertex (Example 19.21), $y := x_I$ and $M := (m_{ij}) = A_I^{-1} A_{I'}$. Let $j \in I'$ with $\gamma(I)_j > 0$.[2] Then the following holds:*

(a) If the j-th column of M is non-positive (i. e. ≤ 0), then L is unbounded.

[2] We index the components of $\gamma(I)$ and the columns of M with I' instead of $1, \ldots, n-m$.

(b) Otherwise, let $i \in I$ with

$$\frac{y_i}{m_{ij}} = \min\Big\{\frac{y_k}{m_{kj}} : k \in I,\ m_{kj} > 0\Big\}.$$

Then $J := I \setminus \{i\} \cup \{j\}$ is also a feasible basis set and for the corresponding vertex x' it holds that $f(x') \geq f(x)$.

(c) If $y > 0$, then $f(x') > f(x)$ in (b).

Proof.

(a) Let $\mu > 0$ and $z \in \mathbb{R}^n$ with $z_{I'} = \mu e_j \geq 0$ and $z_I = y - \mu M_j \geq y = x_I \geq 0$, where $M_j \leq 0$ denotes the j-th column of M. Because of

$$Az = A_I z_I + A_{I'} z_{I'} = Ax - \mu A_I M_j + \mu A_{I'} e_j = b$$

z is feasible and

$$f(z) = c_I^{\mathrm{t}} z_I + \mu c_{I'}^{\mathrm{t}} e_j = c^{\mathrm{t}} x - \mu c_I^{\mathrm{t}} M_j + \mu c_{I'}^{\mathrm{t}} e_j = c^{\mathrm{t}} x + \mu \gamma(I)_j \to \infty \qquad (\mu \to \infty).$$

Thus L is unbounded.

(b,c) Apparently $S := A_I^{-1} A_J = (e_1, \dots, e_{i-1}, M_j, e_{i+1}, \dots, e_n)$. By expansion along the i-th row, one obtains $\det(S) = m_{ij} > 0$. In particular, S is invertible. Thus A_J must also be invertible, i. e. J is a basis set. We define $x' \in \mathbb{R}^m$ by

$$x'_k = \begin{cases} y_k - \frac{y_i}{m_{ij}} m_{kj} & \text{if } k \in I \setminus \{i\}, \\ \frac{1}{m_{ij}} y_i & \text{if } k = j, \\ 0 & \text{otherwise.} \end{cases}$$

By the choice of i, it holds that $y_k \geq \frac{y_i}{m_{ij}} m_{kj}$ for $k \in I \setminus \{i\}$. This shows $x' \geq 0$. Furthermore,

$$\begin{aligned} A_I^{-1} A x' = A_I^{-1} A_J x'_J &= \sum_{k \neq i} \Big(y_k - \frac{y_i}{m_{ij}} m_{kj}\Big) e_k + \frac{y_i}{m_{ij}} M_j \\ &= \sum_{k \in I} \Big(y_k - \frac{y_i}{m_{ij}} m_{kj}\Big) e_k + \frac{y_i}{m_{ij}} M_j = y = x_I \end{aligned}$$

and $Ax' = A_I x_I = Ax = b$. Thus x' is a feasible vertex for J. Finally,

$$f(x') = \sum_{k \in I} c_k \Big(y_k - \frac{y_i}{m_{ij}} m_{kj}\Big) + \frac{y_i}{m_{ij}} c_j = f(x) - \frac{y_i}{m_{ij}} c_I^{\mathrm{t}} M_j + \frac{y_i}{m_{ij}} c_j = f(x) + \frac{y_i}{m_{ij}} \gamma(I)_j \geq f(x)$$

with equality if and only if $y_i = 0$. □

Remark 19.29 (Simplex Algorithm)**.** Let a linear program $L = (A, b, c)$ be given.

(1) One determines a feasible basis set I_1 and the corresponding vertex x_1. This can be achieved if necessary by appending b as a column of A and completing $\{m+1\}$ to a basis set.

(2) For $k = 1, 2, \dots$ repeat:

  (a) If the simplex criterion $\gamma(I_k) \leq 0$ is satisfied, then x_k is a solution for L.

  (b) Otherwise, one can apply Lemma 19.28. If Lemma 19.28(a) is satisfied, then L is unbounded and one can terminate.

(c) Otherwise, one finds with Lemma 19.28(a) a vertex x_{k+1} with $f(x_{k+1}) \geq f(x_k)$.

(3) In practice, it can happen that the loop enters an infinite cycle of vertices $x_k, x_{k+1}, \ldots, x_l, x_k, \ldots$. In this case, one can start with a new starting vertex x_1 or incorporate artificial perturbations.

One can construct examples in which the simplex algorithm traverses all (exponentially many) vertices, but in practice this occurs only very rarely.

Example 19.30. We consider again the linear program from Example 19.6:

$$A = \begin{pmatrix} 1 & 2 & 1 & . & . \\ 1 & 1 & . & 1 & . \\ . & 3 & . & . & 1 \end{pmatrix}, \qquad b = (170, 150, 180), \qquad c = (300, 500, 0, 0, 0).$$

Obviously $I_1 = \{3, 4, 5\}$ is a feasible basis set with vertex $x_1 = (0, 0, 170, 150, 180)^{\mathrm{t}}$, $f(x_1) = 0$ and $A_{I_1} = 1_3$. It holds that $\gamma(I) = (300, 500)$ and

$$M_1 = A_{I_1'} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 3 \end{pmatrix}.$$

We can choose $j = 1$ in Lemma 19.28. Then $i = 4$ because of $170 > 150$. Now $I_2 = \{1, 3, 5\}$ and $x_2 = (150, 0, 20, 0, 180)^{\mathrm{t}}$ with $f(x_2) = 45.000$. One calculates

$$M_2 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 3 & 0 \end{pmatrix}$$

and $\gamma(I_2) = (500, 0) - (300, 0, 0)M = (200, -300)$. Here one must choose $j = 2$ and $i = 3$. Thus $I_3 = \{1, 2, 5\}$ and $x_3 = (130, 20, 0, 0, 120)^{\mathrm{t}}$ with $f(x_3) = 49.000$. Furthermore

$$M_3 = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 & 0 \\ 1 & -1 & 0 \\ -3 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \\ -3 & 3 \end{pmatrix}$$

and $\gamma(I_3) = (0, 0) - (300, 500, 0)M = (-200, -100)$. According to the simplex criterion, x_3 is a solution. Except for the slack variable x_5, this is the solution found in Example 19.2.

20 Lattices and Quadratic Forms

20.1 Lattices

Remark 20.1. In the vector space axioms, it is not used that the field K possesses inverse elements wrt. multiplication. One can therefore replace K with an arbitrary ring R and then speaks of *R-modules* (more precisely: *left modules*). In general, the theory of modules is arbitrarily complicated and does not have much to do with linear algebra anymore. Even $\mathbb{Z}$-modules usually do not possess a basis (for example, $\mathbb{F}_2$ is in a natural way a $\mathbb{Z}$-module in which every element is linearly dependent because of $2 \cdot x = 0$ for $x \in \mathbb{F}_2$). In this section, we investigate a family of $\mathbb{Z}$-modules within $\mathbb{R}^n$ which, by definition, possess a basis.

Definition 20.2. A *lattice* is a subgroup $\Delta \leq (\mathbb{R}^n, +)$ of the form

$$\Delta = \Big\{ \sum_{i=1}^{m} \lambda_i b_i : \lambda_1, \ldots, \lambda_m \in \mathbb{Z} \Big\} = \mathbb{Z}b_1 + \ldots + \mathbb{Z}b_m,$$

where $B := \{b_1, \ldots, b_m\} \subseteq \mathbb{R}^n$ is linearly independent. If applicable, one calls B a *basis* of Δ and $\mathrm{rk}(\Delta) := |B|$ the *rank* of Δ. In the case $m = n$, one says: Δ has *full* rank. The matrix $A \in \mathbb{R}^{m \times n}$ with rows $b_1, \ldots, b_m$ is called a *generator matrix* of Δ. Furthermore, one calls

$${}_B[\Delta]_B := AA^{\mathrm{t}} = ([b_i, b_j])_{ij} \in \mathbb{R}^{m \times m}$$

the *Gram matrix* of Δ wrt. B.

Remark 20.3.

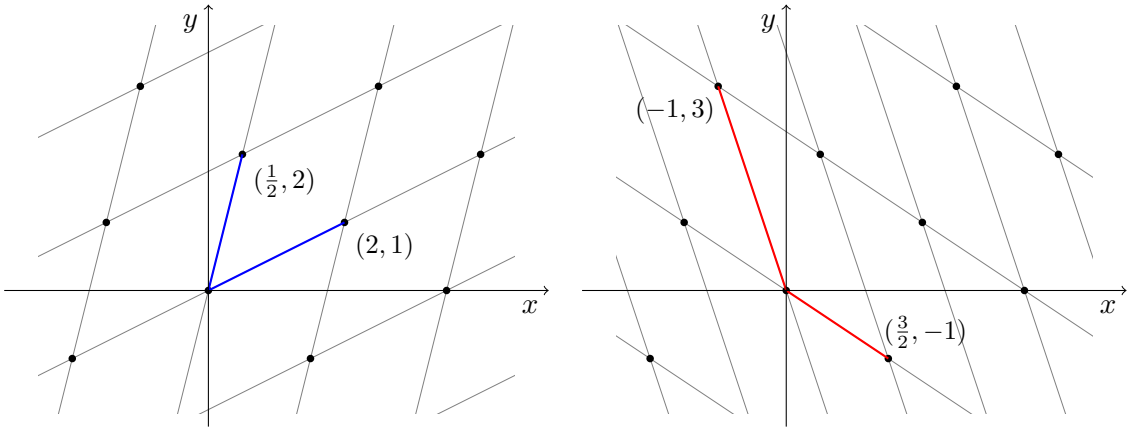
(a) For every basis B of a lattice Δ, it holds that $|B| = \dim\langle B \rangle = \dim\langle \Delta \rangle$. Therefore, $\mathrm{rk}(\Delta)$ does not depend on the choice of the basis. By replacing $\mathbb{R}^n$ with $\langle \Delta \rangle$, one can often assume that Δ has full rank.

(b) Since the generator matrix A has full rank, the Gram matrix AA^{t} is positive definite (Lemma 12.40).

Example 20.4.

(a) The *trivial* lattice with the standard basis $B := \{e_1, \ldots, e_n\}$ consists of all points in $\mathbb{R}^n$ with integer coordinates. The generator matrix and the Gram matrix wrt. B are 1_n.

(b) A section of a lattice $\Delta \subseteq \mathbb{R}^2$ with two different bases:



(c) Let $A \in \mathbb{R}^{n\times n}$ be positive definite with Cholesky decomposition $A = R^{\mathrm{t}}R$. Then there exists a lattice with generator matrix R^{t} and Gram matrix A. In this way, one can assign a "canonical" lattice to A.

(d) For a lattice $\Delta \subseteq \mathbb{R}^n$ with generator matrix $A \in \mathbb{R}^{m\times n}$ let

$$\Delta^* := \{x \in \langle\Delta\rangle : \forall d \in \Delta : [x, d] \in \mathbb{Z}\} \subseteq \mathbb{R}^n.$$

For every $x \in \langle\Delta\rangle$ there exists a $y \in \mathbb{R}^m$ with $x = yA$. It holds that

$$x \in \Delta^* \iff \forall b \in B : [x, b] \in \mathbb{Z} \iff xA^{\mathrm{t}} \in \mathbb{Z}^m \iff z := yAA^{\mathrm{t}} \in \mathbb{Z}^m.$$

For $S := (AA^{\mathrm{t}})^{-1}A = {}_B[\Delta]_B^{-1}A \in \mathbb{R}^{m\times n}$ it therefore holds that $\Delta^* = \{zS : z \in \mathbb{Z}^m\}$.[1] This shows that Δ^* is a lattice with generator matrix S. One calls Δ^* the *dual* lattice to Δ. The Gram matrix of Δ^* wrt. S is

$$SS^{\mathrm{t}} = {}_B[\Delta]_B^{-1}AA^{\mathrm{t}}{}_B[\Delta]_B^{-1} = {}_B[\Delta]_B^{-1}.$$

The trivial lattice Δ is obviously *self-dual*, i.e., $\Delta^* = \Delta$. In general, $(\Delta^*)^* = \Delta$, because

$$(SS^{\mathrm{t}})^{-1}S = {}_B[\Delta]_B S = A.$$

Definition 20.5. For $n \geq 1$ let $\mathrm{GL}(n, \mathbb{Z}) := \{A \in \mathrm{GL}(n, \mathbb{Q}) \cap \mathbb{Z}^{n\times n} : A^{-1} \in \mathbb{Z}^{n\times n}\}$.

Lemma 20.6. *For $A \in \mathbb{Z}^{n\times n}$ it holds that $A \in \mathrm{GL}(n, \mathbb{Z})$ if and only if $\det A = \pm 1$.*

Proof. For $A \in \mathrm{GL}(n, \mathbb{Z})$ it holds that $\det A \in \mathbb{Z}$ and $\det(A)^{-1} = \det(A^{-1}) \in \mathbb{Z}$, thus $\det A = \pm 1$. The reverse inclusion follows from Remark 9.24. □

Lemma 20.7. *Let B and C be bases of a lattice $\Delta \subseteq \mathbb{R}^n$. Then there exists an $S \in \mathrm{GL}(m, \mathbb{Z})$ with $S^{\mathrm{t}}{}_B[\Delta]_B S = {}_C[\Delta]_C$.*

Proof. Let $G_B, G_C \in \mathbb{R}^{m\times n}$ be the generator matrices of Δ wrt. B and C, respectively. Then there exists $S \in \mathbb{Z}^{m\times m}$ with $SG_B = G_C$. It follows that

$${}_C[\Delta]_C = G_C G_C^{\mathrm{t}} = SG_B G_B^{\mathrm{t}} S^{\mathrm{t}} = S\,{}_B[\Delta]_B S^{\mathrm{t}}$$

and $\det({}_C[\Delta]_C) = \det(S)^2 \det({}_B[\Delta]_B) \geq \det({}_B[\Delta]_B)$. For reasons of symmetry, $\det({}_B[\Delta]_B) \geq \det({}_C[\Delta]_C)$ also holds. This shows $\det(S) = \pm 1$ and $S \in \mathrm{GL}(m, \mathbb{Z})$ according to Lemma 20.6. □

[1] By the way, $S^{\mathrm{t}} = A^+$ is the pseudoinverse of A.

Definition 20.8.

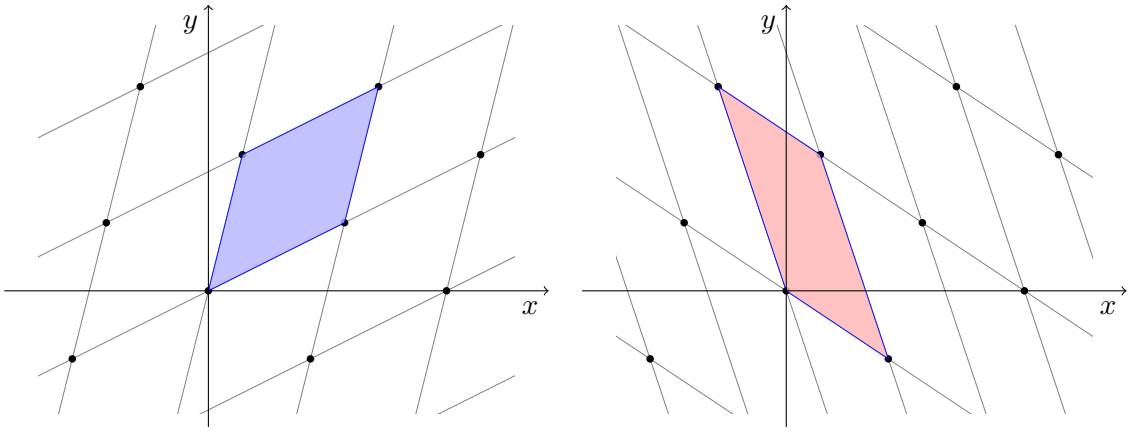
- A lattice Δ is called *integral*, if ${}_B[\Delta]_B$ is integral for a basis B. According to Lemma 20.7, this property does not depend on the choice of B.
- One calls $\operatorname{disc}(\Delta) := \sqrt{\det({}_B[\Delta]_B)}$ the *discriminant* of Δ. This is also independent of B.

Example 20.9.

(a) If Δ has full rank with generator matrix A, then $\operatorname{disc}(\Delta) = |\det(A)|$. Geometrically, $\operatorname{disc}(\Delta)$ is in this case the volume of the so-called *fundamental mesh*

$$\{\lambda_1 b_1 + \ldots + \lambda_n b_n : 0 \leq \lambda_1, \ldots, \lambda_n \leq 1\}$$

for a basis $b_1, \ldots, b_n$ of Δ (cf. Example 9.2). In particular, $\operatorname{disc}(\Delta) = 7/2$ holds for the lattice from Example 20.4(b).



(b) A lattice Δ is integral if and only if $\Delta \subseteq \Delta^*$. In general, $\operatorname{disc}(\Delta^*) = \operatorname{disc}(\Delta)^{-1}$ holds.

20.2 The Minimal Norm

Definition 20.10. For a lattice $\Delta \subseteq \mathbb{R}^n$, let

$$\min \Delta := \min\big\{|x| : x \in \Delta \setminus \{0\}\big\}$$

be the *minimal norm* of Δ.

Remark 20.11.

(a) Since a lattice Δ is a group, $\min \Delta = \min\big\{|x - y| : x, y \in \Delta, x \neq y\big\}$ holds.

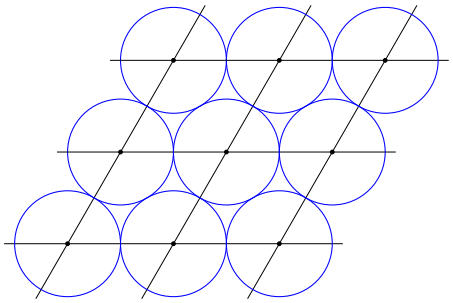
(b) Let A be a generator matrix of Δ and $G := AA^{\mathrm{t}}$ the Gram matrix. For $x \in \Delta$ there exists an $s \in \mathbb{Z}^m$ with $x = sA$. It follows that $|x| = |sA| = \sqrt{sGs^{\mathrm{t}}}$. Let λ (> 0) be the smallest eigenvalue of G. According to Lemma 17.55, $|x| \geq \sqrt{\lambda}|s|$ holds. In particular, $\min \Delta \geq \sqrt{\lambda}$. Equality will generally not occur, as there is not necessarily a normalized integer eigenvector of G. However, there exist only finitely many vectors $x \in \Delta$ with $|x| = \min \Delta$. They are called *shortest* vectors. Both the determination of the shortest vectors and the determination of $\min \Delta$ are difficult algorithmic problems for $n > 10$. Modern encryption methods such as LWE[2] are based on this, which — in contrast to classical methods like RSA — cannot yet be broken by quantum computers.

[2]Learning with errors

(c) HARRIOT investigated in 1587 how (cannon) balls can be arranged as densely as possible in $\mathbb{R}^n$. If the unit balls are arranged regularly, the centers form a lattice Δ with $\min \Delta \geq 2$ (otherwise balls would overlap). The density of the arrangement can be measured by dividing the volume of the unit ball τ_n by the volume of the fundamental mesh $\operatorname{disc}(\Delta)$: $\rho(\Delta) = \frac{\tau_n}{\operatorname{disc}(\Delta)}$.

(1) For $n = 1$, $\Delta = \mathbb{Z}2$ is optimal with $\rho(\Delta) = 1$.

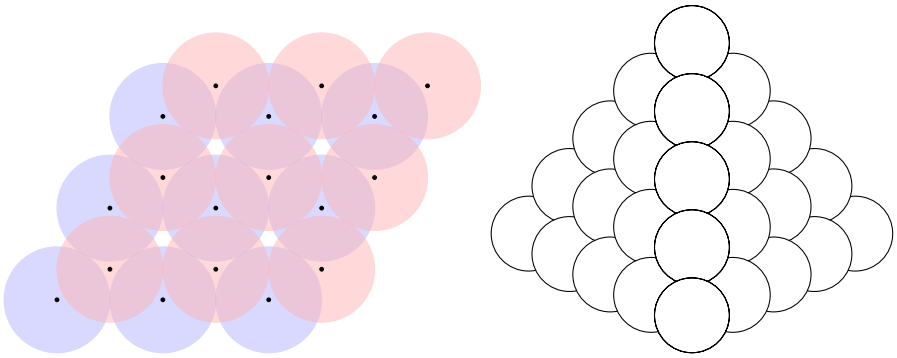
(2) For $n = 2$, an optimal arrangement is obtained by the hexagonal lattice $\Delta = \mathbb{Z}(2,0)+\mathbb{Z}(1,\sqrt{3})$



with $\min \Delta = 2$ and $\rho(\Delta) = \frac{\pi}{2\sqrt{3}} \approx 0.91$. The square lattice $\mathbb{Z}(2,0) + \mathbb{Z}(0,2)$ has only density $\pi/4 \approx 0.79$.

(3) For $n = 3$, one takes the optimal arrangement in the plane and stacks these offset on top of each other:

$$\Delta = \mathbb{Z}(2,0,0) + \mathbb{Z}(1,\sqrt{3},0) + \mathbb{Z}(1,1/\sqrt{3},\sqrt{8/3})$$



It holds that $\min \Delta = 2$ and $\rho(\Delta) = \frac{\pi}{3\sqrt{2}} \approx 0.74$. Gauss proved that there is no lattice with higher density (Theorem A.81, cf. Exercise III.27). KEPLER conjectured in 1611 that it is generally impossible to arrange balls even more densely (even irregularly).[3] This was only proven in 1998 by HALES using a computer.[4]

(4) For $4 \leq n \leq 7$, the densest sphere packing is not known. For $n = 8$, VIAZOVSKA proved in

[3]There are infinitely many ways to stack the planes offset with the same density, but these are formally not lattices.
[4]Since 2017, there has been a "formal" proof verified by computer.

2016 that the so-called E_8-lattice with generator and Gram matrix

$$\begin{pmatrix} 2 & . & . & . & . & . & . & . \\ -1 & 1 & . & . & . & . & . & . \\ . & -1 & 1 & . & . & . & . & . \\ . & . & -1 & 1 & . & . & . & . \\ . & . & . & -1 & 1 & . & . & . \\ . & . & . & . & -1 & 1 & . & . \\ . & . & . & . & . & -1 & 1 & . \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix} \quad \begin{pmatrix} 4 & -2 & . & . & . & . & . & 1 \\ -2 & 2 & -1 & . & . & . & . & . \\ . & -1 & 2 & -1 & . & . & . & . \\ . & . & -1 & 2 & -1 & . & . & . \\ . & . & . & -1 & 2 & -1 & . & . \\ . & . & . & . & -1 & 2 & -1 & . \\ . & . & . & . & . & -1 & 2 & . \\ 1 & . & . & . & . & . & . & 2 \end{pmatrix}$$

is the densest arrangement. It holds that $\rho(E_8) = \frac{\pi^4}{384} \approx 0.25$. One sees that E_8 consists of the vectors x and $x + \frac{1}{2}(1, \ldots, 1)$ for $x \in \mathbb{Z}^8$ with $2 \mid x_1 + \ldots + x_8$. In particular, $\min E_8 = 2$. One year later, the case $n = 24$ was also solved. Viazovska received the Fields Medal in 2022 for these achievements.

Example 20.12. For the lattice $\Delta = \mathbb{Z}(2,1) + \mathbb{Z}(1/2,2)$ from Example 20.4(b), $\pm(3/2,-1)$ are the shortest vectors with $\min \Delta = \sqrt{13}/2$.

Theorem 20.13. *A subgroup $U \leq (\mathbb{R}^n, +)$ is a lattice if and only if for all $d \in \mathbb{N}$ there exist only finitely many $x \in U$ with $|x| \leq d$.*

Proof. For every lattice, the specified restriction holds according to Remark 20.11. Conversely, let $U \leq (\mathbb{R}^n, +)$ such that the condition holds. Let $u_1, \ldots, u_k \in U$ be a maximal set of linearly independent vectors. Then $U \subseteq \langle u_1, \ldots, u_k \rangle =: V$. In the case $k = 0$, $U = \{0\} = \langle \varnothing \rangle$ is a lattice of rank 0. Now let the claim already be proven for $k-1$. Since $W := U \cap \langle u_1, \ldots, u_{k-1} \rangle$ also satisfies the assumption of the theorem, W is a lattice with basis $b_1, \ldots, b_{k-1}$. We consider

$$S := \{\lambda_1 b_1 + \ldots + \lambda_{k-1} b_{k-1} + \lambda u_k : 0 \leq \lambda_1, \ldots, \lambda_{k-1} < 1,\ 0 \leq \lambda \leq 1\} \cap U.$$

By the triangle inequality, S is bounded and therefore finite by assumption. Because $u_k \in S$, there exists a $b_n = \lambda_1 b_1 + \ldots + \lambda_{k-1} b_{k-1} + \lambda u_k \in S$ with $\lambda > 0$ minimal. Obviously $b_1, \ldots, b_n \in U$ are linearly independent. Every element $u \in U$ has the form $u = \mu_1 b_1 + \ldots + \mu_k b_k$ with $\mu_1, \ldots, \mu_k \in \mathbb{R}$. Let $x_k \in \mathbb{Z}$ with $0 \leq \mu_k - x_k \lambda < \lambda$. Then there exist $x_1, \ldots, x_{k-1} \in \mathbb{Z}$ such that

$$u - x_1 b_1 - \ldots - x_k b_k = \lambda'_1 b_1 + \ldots + \lambda'_{k-1} b_{k-1} + \lambda'_k u_k \in S$$

with $0 \leq \lambda'_1, \ldots, \lambda'_{k-1} < 1$ and $0 \leq \lambda'_k < \lambda$ holds. By the choice of b_n, it follows that $\lambda'_k = 0$ and $\lambda'_1 b_1 + \ldots + \lambda'_{k-1} b_{k-1} \in W$. Since $b_1, \ldots, b_{k-1}$ is a basis of W, it must hold that $\lambda'_1 = \ldots = \lambda'_{k-1} = 0$. This shows $u = x_1 b_1 + \ldots + x_k b_k$. Thus U is a lattice with basis $b_1, \ldots, b_k$. □

Corollary 20.14. *If $\Delta, \Lambda \subseteq \mathbb{R}^n$ are lattices, then so is $\Delta \cap \Lambda$.*

Proof. As is well known, $\Delta \cap \Lambda$ is a subgroup of $(\mathbb{R}^n, +)$. Like Δ, $\Delta \cap \Lambda$ can also contain only finitely many elements with bounded norm. □

20.3 Integer Matrices

Remark 20.15. The study of integral lattices leads to matrices with integer entries. In section 10.3, we had already constructed matrices with polynomial entries and noted that the Gaussian algorithm is not applicable because it would require division. With some number theory[5], we develop a substitute for the Gaussian algorithm in $\mathbb{Z}^{n\times m}$, with which systems of equations over $\mathbb{Z}$ can be solved.

Definition 20.16. For $a, b \in \mathbb{Z}$ we write $a \mid b$ (a *divides* b), if a $c \in \mathbb{Z}$ with $b = ac$ exists. A $d \in \mathbb{Z}$ is called a *common divisor* of $a_1, \ldots, a_n \in \mathbb{Z}$, if $d \mid a_1, \ldots, d \mid a_n$. Let $\mathrm{cd}(a_1, \ldots, a_n)$ be the set of common divisors of $a_1, \ldots, a_n$. One calls $d \in \mathrm{cd}(a_1, \ldots, a_n)$ the *greatest common divisor* and writes $\gcd(a_1, \ldots, a_n) := d$, if $d \geq 0$ and $e \mid d$ holds for all $e \in \mathrm{cd}(a_1, \ldots, a_n)$. In the case $d = 1$ we call $a_1, \ldots, a_n$ *coprime* (as with polynomials).

Remark 20.17.

(a) If g and g' are greatest common divisors of $a_1, \ldots, a_n$, then $g \mid g' \mid g$ and $g = \pm g'$ holds. Because of $g, g' \geq 0$ it follows that $g = g'$, i. e. there exists at most one greatest common divisor of $a_1, \ldots, a_n$ (this justifies the notation gcd).

(b) Contrary to its name, the ggT is not necessarily the *largest* common divisor. For example, $\mathrm{cd}(0, 0) = \mathbb{Z}$, but $\gcd(0, 0) = 0$ (note $0 \mid 0$). More generally, $\gcd(a_1, \ldots, a_n) = 0$ if and only if $a_1 = \ldots = a_n = 0$.

(c) One should compare the following lemma with Lemma 15.6.

Lemma 20.18 (Bézout)**.** *For $a_1, \ldots, a_n \in \mathbb{Z}$ there exist $b_1, \ldots, b_n \in \mathbb{Z}$ with $a_1b_1 + \ldots + a_nb_n = \gcd(a_1, \ldots, a_n)$ and $\gcd(b_1, \ldots, b_n) = 1$.*

Proof. Induction on n: In the case $n = 1$ one sets $b_1 = 1$. Let $n \geq 2$ and the claim be already proven for $n - 1$. Then there exist $b_1, \ldots, b_{n-1} \in \mathbb{Z}$ with $\gcd(a_1, \ldots, a_{n-1}) = a_1b_1 + \ldots + a_{n-1}b_{n-1}$. Because of $\mathrm{cd}(a_1, \ldots, a_n) = \mathrm{cd}(\gcd(a_1, \ldots, a_{n-1}), a_n)$ we have $\gcd(a_1, \ldots, a_n) = \gcd(a_1b_1 + \ldots + a_{n-1}b_{n-1}, a_n)$. Therefore we can assume $n = 2$. Let

$$d := \min\{e \in \mathbb{N} : \exists b_1, b_2 \in \mathbb{Z} : e = a_1b_1 + a_2b_2\}.$$

For $e \in \mathrm{cd}(a_1, a_2)$ it then holds that $e \mid d$, thus also $\gcd(a_1, a_2) \mid d$. Euclidean division yields $a_1 = qd + r$ with $q \in \mathbb{Z}$ and $0 \leq r < d$. Because of $r = a_1 - qd \in \mathbb{Z}a_1 + \mathbb{Z}a_2$, it follows that $r = 0$ from the minimality of d. Thus $d \mid a_1$ and analogously $d \mid a_2$. This shows $d \mid \gcd(a_1, a_2)$ and $\gcd(a_1, a_2) = d = a_1b_1 + a_2b_2$. For $g := \gcd(b_1, b_2)$ we have

$$d \mid a_1\frac{b_1}{g} + a_2\frac{b_2}{g} = \frac{d}{g},$$

hence $g = 1$. □

Lemma 20.19. *For $a_1, \ldots, a_n \in \mathbb{Z}$ there exists a matrix in $\mathbb{Z}^{n\times n}$ with first row (or column) $(a_1, \ldots, a_n)$ and determinant $\gcd(a_1, \ldots, a_n)$.*

[5] More details can be found in my Number Theory notes

Proof. In the case $a_1 = \ldots = a_n = 0$, 0_n satisfies the claim. So let wlog. $a_1 \neq 0$ (swap columns/rows if necessary). Since the determinant is linear in each row (and column), we can assume $\gcd(a_1, \ldots, a_n) = 1$. We argue by induction on n. For $n = 1$, choose $A = (a_1)$. Now let $n \geq 2$. By the induction hypothesis, there exists $A_1 \in \mathbb{Z}^{(n-1)\times(n-1)}$ with first row $(a_1, \ldots, a_{n-1})$ and

$$g := \det(A_1) = \gcd(a_1, \ldots, a_{n-1}) > 0.$$

By Bézout, there exist $s, t \in \mathbb{Z}$ with $gs + a_n t = 1$. Let $b_i := a_i t/g \in \mathbb{Z}$ for $i = 1, \ldots, n-1$. We construct $A_2 \in \mathbb{Z}^{(n-1)\times(n-1)}$ by deleting the first row of A_1 and instead adding $(b_1, \ldots, b_{n-1}) = \frac{t}{g}(a_1, \ldots, a_{n-1})$ as the last column. Then $\det(A_2) = (-1)^n t$ holds. Let

$$A := \begin{pmatrix} & A_1 & & a_n \\ & & & 0 \\ b_1 & \cdots & b_{n-1} & s \end{pmatrix}.$$

By expansion along the last column, one sees

$$\det(A) = (-1)^n a_n \det(A_2) + s \det(A_1) = a_n t + gs = 1.$$

The analogous statement for columns is obtained by transposition. □

Remark 20.20. The next theorem provides an integer version of the row echelon form (cf. Theorem 6.10).

Theorem 20.21 (Hermite Normal Form)**.** *For every matrix $A \in \mathbb{Z}^{n\times m}$ there exists exactly one non-negative matrix $H \in \mathbb{Z}^{n\times m}$ with the following properties:*

(i) $H = SA$ *for some* $S \in \mathrm{GL}(n, \mathbb{Z})$.

(ii) Zero rows are located at the bottom of H.

(iii) For $\tau_i := \min\{1 \leq j \leq m : a_{ij} \neq 0\}$, *it holds that* $\tau_1 < \tau_2 < \ldots$ *(if defined) and* $a_{j,\tau_i} < a_{i,\tau_i}$ *for* $j = 1, \ldots, i-1$.

In particular, H is an upper triangular matrix.

Proof. Existence: The following algorithm transforms A into a matrix H with the desired properties. We traverse the columns of A from left to right. Let $(a_1, \ldots, a_n)$ be the first column of A. By Bézout, there exist $b_1, \ldots, b_n \in \mathbb{Z}$ with $d := a_1 b_1 + \ldots + a_n b_n$ and $\gcd(b_1, \ldots, b_n) = 1$. By Lemma 20.19, there exists an $S \in \mathrm{GL}(n, \mathbb{Z})$ with first row $(b_1, \ldots, b_n)$. By replacing A with SA, we achieve $a_1 = d \geq 0$. However, this also changes $a_2, \ldots, a_n$. If we repeat this step, d can at most become smaller. After finitely many repetitions, $a_1 \in \mathrm{cd}(a_2, \ldots, a_n)$. For $i \geq 2$, there exist coprime $b_1, b_i \in \mathbb{Z}$ with $a_1 b_1 = a_i b_i$. By Lemma 20.19, there exists an $S \in \mathrm{GL}(n, \mathbb{Z})$ with i-th row $(b_1, 0, \ldots, 0, -b_i, 0, \ldots, 0)$. After replacing A with SA, $a_i = 0$ holds. In this way, we achieve $a_2 = \ldots = a_n = 0$.

In the second column of A, we can analogously achieve $a_{22} \geq 0 = a_{32} = \ldots = a_{n2}$. In the case $a_{22} = 0$, we proceed to the third column. In the case $a_{12} < 0$ or $a_{12} > a_{22}$, we add or subtract a_{22} from a_{12} by multiplication with an elementary matrix from the left (this does not change the first column of A). By repeating this step sufficiently many times, one obtains $0 \leq a_{12} < a_{22}$. Once all columns have been traversed in this manner, A (or H) has the desired properties.

Uniqueness: For the uniqueness of H, we argue as in the proof of Theorem 6.10. We can assume wlog. that A itself satisfies the properties of H. Let a_i (or s_i) be the i-th column of A (or S). Then

$h_i = Sa_i$ is the i-th column of H. Let $a_i \in \langle e_1^{\mathrm{t}}, \ldots, e_k^{\mathrm{t}} \rangle$. We show $a_i = h_i$ and $s_k = e_k^{\mathrm{t}}$ by induction on k. In the case $k = 0$, $a_i = 0$ and $h_i = Sa_i = 0$. Now let the claim be already proven up to $k-1$. The first column (from the left) of A that does not lie in $\langle e_1^{\mathrm{t}}, \ldots, e_{k-1}^{\mathrm{t}} \rangle$ has the form $a_i = (a_{1i}, \ldots, a_{ki}, 0, \ldots, 0)^{\mathrm{t}}$ with $a_{ki} > 0$. By the induction hypothesis, $h_{ji} = a_{ji} + s_{jk}a_{ki}$ for $j < k$ and $h_{ki} = s_{kk}a_{ki}$. Since S is invertible, $s_{kk} \neq 0$ and $h_{ki} \neq 0$. Therefore, $i = \tau_k$ and $h_{ki} > 0$ must hold. It follows that $s_{jk}a_{ki} = h_{ji} = 0$ and $s_{ji} = 0$ for $j > k$. From $\det(S) = 1$, it follows that $s_{kk} = 1$ and $h_{ki} = a_{ki}$. For $j < k$, it holds that

$$0 \leq h_{ji} = a_{ji} + s_{jk}a_{ki} < h_{ki} = a_{ki}.$$

From $0 \leq a_{ji} < a_{ki}$, it follows that $s_{jk} = 0$ and $s_k = e_k$. Furthermore, $a_i = h_i$ as claimed. □

Example 20.22. If one has a lattice Δ with an integer generator matrix A, one can determine a canonical basis using the Hermite normal form:

$$A := \begin{pmatrix} 0 & 2 & 6 \\ 3 & 3 & 0 \\ 2 & 2 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & -1 \\ 0 & 2 & 6 \\ 2 & 2 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & -1 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Remark 20.23.

(a) The properties of the Hermite normal form do not make use of integrality. One can therefore also transform rational matrices into Hermite normal form by temporarily multiplying by the least common denominator of all entries and applying the algorithm to the resulting integer matrix.

(b) In the Hermite normal form, one only operates on the rows (i.e., from the left) of a given matrix. If one also allows column operations, one arrives at a uniquely determined diagonal matrix.

Theorem 20.24. *For $A \in \mathbb{Z}^{n \times m}$ there exist uniquely determined non-negative numbers $d_1 \mid d_2 \mid \ldots$ with*

$$SAT = \begin{pmatrix} d_1 & & 0 \\ & d_2 & \\ 0 & & \ddots \end{pmatrix} \qquad \text{(Smith normal form)}$$

for certain $S \in \mathrm{GL}(n, \mathbb{Z})$ and $T \in \mathrm{GL}(m, \mathbb{Z})$. One calls $d_1, \ldots, d_{\min\{n,m\}}$ elementary divisors *of A.*

Proof. Existence: For the zero matrix, the claim is clear (note $0 \mid 0$). As in the proof of Theorem 20.21, one can assume $a_{11} > 0 = a_{21} = \ldots = a_{n1}$. The same procedure with the columns (i.e., by multiplication with $T \in \mathrm{GL}(m, \mathbb{Z})$ from the right) leads to $A = \mathrm{diag}(d, A_1)$. If not every entry of A_1 is divisible by d, one adds a corresponding column of A_1 to the first column of A (the corresponding elementary matrix has determinant 1). Then one starts from the beginning and reduces d further in this way. After finitely many steps, all entries of A_1 are divisible by d. Now one applies the procedure to A_1. In doing so, the divisibility by d is preserved in each step. In the end, one obtains the desired diagonal matrix.

Uniqueness: Let $D_k(A)$ be the gcd of all determinants of $k \times k$ submatrices of A. The columns of AT are integer linear combinations of the columns of A. This also holds for $k \times k$ submatrices. Thus $D_k(A) \mid D_k(AT) \mid D_k(ATT^{-1}) = D_k(A)$. Analogously, $D_k(A) = D_k(SA)$ and therefore $D_k(A) = D_k((d_i\delta_{ij})) = d_1 \ldots d_k$. Thus the elementary divisors are uniquely determined by A. □

Remark 20.25.

(i) If A is square, then $|\det A|$ is the product of the elementary divisors of A. The above proof shows how the elementary divisors can be characterized by subdeterminants.

(ii) In contrast to the Gaussian elimination, the procedure for the Smith normal form is significantly more complex.

(iii) The Smith normal form allows for solving integer systems of linear equations $Ax = b$. For this, let $SAT = D = (d_i\delta_{ij})$ as in Theorem 20.24 with $d_1, \ldots, d_k > 0 = d_{k+1} = \ldots = d_{\min\{n,m\}}$. The simpler system $Dy = Sb = (b_1, \ldots, b_n)^{\mathrm{t}}$ is solvable if and only if $d_i \mid b_i$ for $i = 1, \ldots, n$. The solutions are then $y = (b_1/d_1, \ldots, b_k/d_k, *, \ldots, *)$, where $*$ stands for an arbitrary integer. The solutions of $Ax = b$ are obtained by $x = Ty$.

Example 20.26. We are looking for all $x \in \mathbb{Z}^3$ with

$$Ax := \begin{pmatrix} 21 & -5 & 26 \\ 3 & -1 & 4 \\ -8 & 2 & -10 \end{pmatrix} x = \begin{pmatrix} 47 \\ 7 \\ -18 \end{pmatrix} =: b. \tag{20.1}$$

The algorithm in the proof of Theorem 20.24 (modulo obvious simplifications) yields:

$$\begin{pmatrix} 21 & -5 & 26 \\ 3 & -1 & 4 \\ -8 & 2 & -10 \end{pmatrix} \sim \begin{pmatrix} -1 & 3 & 4 \\ -5 & 21 & 26 \\ 2 & -8 & -10 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 5 & 6 & 6 \\ -2 & -2 & -2 \end{pmatrix}$$
$$\sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & 6 \\ 0 & -2 & -2 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 6 & 6 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} =: D.$$

For

$$S := \begin{pmatrix} -5 & -8 & -2 \\ -1 & -1 & 0 \\ 2 & 3 & 1 \end{pmatrix} \in \mathrm{GL}(3, \mathbb{Z}), \qquad T := \begin{pmatrix} -1 & 1 & -2 \\ -1 & 0 & -1 \\ 2 & -2 & 3 \end{pmatrix} \in \mathrm{GL}(3, \mathbb{Z})$$

it holds that $SDT = A$. Therefore, (20.1) is equivalent to

$$Dy = S^{-1}b = \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix}$$

with $y := Tx$. This shows $y = (-3, -2, a)$ with $a \in \mathbb{Z}$. Thus $x = T^{-1}y = (a - 4, 5 - a, 6 - a)$ for $a \in \mathbb{Z}$.

Theorem 20.27. *Let $\Lambda \subseteq \Delta \subseteq \mathbb{R}^n$ be lattices. Then there exists a basis $b_1, \ldots, b_k$ of Δ and uniquely determined natural numbers $d_1 \mid d_2 \mid \ldots \mid d_l$, such that $d_1b_1, \ldots, d_lb_l$ is a basis of Λ.*

Proof. Existence: Let $a_1, \ldots, a_k$ and $c_1, \ldots, c_l$ initially be arbitrary bases of Δ and Λ, respectively. Here $l = \mathrm{rk}(\Lambda) \leq \mathrm{rk}(\Delta) = k$ holds. We write $c_i = \sum_{j=1}^{k} x_{ij}a_j$ with $x_{ij} \in \mathbb{Z}$ for $i = 1, \ldots, l$ and set $A := (x_{ij}) \in \mathbb{Z}^{l \times k}$. According to Theorem 20.24, there exist $S = (s_{ij}) \in \mathrm{GL}(l, \mathbb{Z})$ and $T = (t_{ij}) \in \mathrm{GL}(k, \mathbb{Z})$ with $A = S(\delta_{ij}d_i)T$ and $d_1 \mid \ldots \mid d_l$. The elements $b_i := \sum_{j=1}^{k} t_{ij}a_j \in \Delta$ for $i = 1, \ldots, k$ form a basis of Δ. Because of

$$c_i = \sum_{j=1}^{k} x_{ij}a_j = \sum_{j=1}^{k} \sum_{m=1}^{l} s_{im}d_m t_{mj}a_j = \sum_{m=1}^{l} s_{im}d_m b_m$$

$\{d_1b_1, \ldots, d_lb_l\}$ is a basis of Λ.

Uniqueness: Let $b'_1, \ldots, b'_k$ be another basis of Δ and $d'_1 \mid \ldots \mid d'_l$, such that $d'_1 b'_1, \ldots, d'_l b'_l$ is a basis of Λ. Then there exist $S = (s_{ij}) \in \mathrm{GL}(k, \mathbb{Z})$ and $T = (t_{ij}) \in \mathrm{GL}(l, \mathbb{Z})$ with $b'_i = \sum_{j=1}^{k} s_{ij} b_j$ and $d'_i b'_i = \sum_{j=1}^{l} t_{ij} d_j b_j$ for all i. A comparison of coefficients shows $d'_i s_{ij} = t_{ij} d_j$ and

$$(\delta_{ij} d'_i) S = T(\delta_{ij} d_i).$$

From the uniqueness of the Smith normal form, it follows that $d'_i = d_i$ for $i = 1, \ldots, l$. □

20.4 Theorems of Hermite and Minkowski

Lemma 20.28. *Every lattice $\Delta \subseteq \mathbb{R}^n$ possesses a basis $b_1, \ldots, b_k$ with $|b_1| = \min \Delta$.*

Proof. Let $b_1, \ldots, b_k \in \Delta$ be an arbitrary basis. A shortest vector $v \in \Delta$ has the form $v = \lambda_1 b_1 + \ldots + \lambda_k b_k$ with coprime $\lambda_1, \ldots, \lambda_k \in \mathbb{Z}$. According to Lemma 20.19, there exists an $S \in \mathrm{GL}(k, \mathbb{Z})$ with first column $\lambda_1, \ldots, \lambda_k$. By change of basis (Lemma 20.7), one can therefore assume $|b_1| = |v| = \min \Delta$. □

Theorem 20.29 (Hermite)**.** *For every lattice $\Delta \subseteq \mathbb{R}^n$ of rank k, it holds that*

$$\boxed{\min \Delta \leq \left(\frac{4}{3}\right)^{(k-1)/4} \sqrt[k]{\mathrm{disc}(\Delta)}.}$$

Proof. Let $b_1, \ldots, b_k$ be a basis of Δ and $G = (g_{ij}) \in \mathbb{R}^{k \times k}$ the corresponding Gram matrix. According to Lemma 20.28, we can assume $\min \Delta = |b_1| = \sqrt{g_{11}}$. For $k = 1$, $\sqrt{g_{11}} = \mathrm{disc}(\Delta)$ holds as claimed. Let $k \geq 2$ and the claim be already proven for $k - 1$. The vectors

$$c_i := b_i - \frac{[b_i, b_1]}{|b_1|^2} b_1 = b_i - \frac{g_{1i}}{g_{11}} b_1 \in \mathbb{R}^n$$

for $2 \leq i \leq k$ are orthogonal to b_1. We consider the lattice Λ with basis $c_2, \ldots, c_k$ and Λ_1 with basis $b_1, c_2, \ldots, c_k$. For the Gram matrices, $G_{\Lambda_1} = S^{\mathrm{t}} G S = \mathrm{diag}(g_{11}, G_\Lambda)$ holds, where

$$S := \begin{pmatrix} 1 & -g_{12}/g_{11} & \cdots & -g_{1k}/g_{11} \\ & \ddots & 0 & 0 \\ & & \ddots & 0 \\ & & & 1 \end{pmatrix} \in \mathrm{GL}(k, \mathbb{Q})$$

has determinant 1. This shows $\mathrm{disc}(\Delta) = \mathrm{disc}(\Lambda_1) = \sqrt{g_{11}}\,\mathrm{disc}(\Lambda)$. By induction, there exists a $y = x_2 c_2 + \ldots + x_k c_k \in \Lambda$ with $x_2, \ldots, x_k \in \mathbb{Z}$ and

$$|y| = \min \Lambda \leq \left(\frac{4}{3}\right)^{(k-2)/4} \left(\frac{\mathrm{disc}(\Lambda)}{\sqrt{g_{11}}}\right)^{\frac{1}{k-1}}.$$

Furthermore, there exists an $x_1 \in \mathbb{Z}$ with

$$\mu := \left| x_1 + \frac{g_{12}}{g_{11}} x_2 + \ldots + \frac{g_{1k}}{g_{11}} x_k \right| \leq \frac{1}{2}.$$

One easily verifies $S^{-1} = 2 \cdot 1_k - S$ (the entries outside the main diagonal change sign). For $v := x_1 b_1 + \ldots + x_k b_k \in \Delta$, it therefore holds that

$$g_{11} = (\min \Delta)^2 \leq |v|^2 = x G x^{\mathrm{t}} = x S^{-\mathrm{t}} G_{\Lambda_1} S^{-1} x^{\mathrm{t}} = \mu^2 g_{11} + (x_2, \ldots, x_k) G_\Lambda (x_2, \ldots, x_k)^{\mathrm{t}}$$

$$\leq \frac{g_{11}}{4} + |y|^2 \leq \frac{g_{11}}{4} + \Big(\frac{4}{3}\Big)^{(k-2)/2} \Big(\frac{\operatorname{disc}(\Delta)^2}{g_{11}}\Big)^{\frac{1}{k-1}}.$$

Rearranging yields

$$g_{11} \leq \Big(\frac{4}{3}\Big)^{k/2} \Big(\frac{\operatorname{disc}(\Delta)^2}{g_{11}}\Big)^{\frac{1}{k-1}}, \qquad g_{11}^k \leq \Big(\frac{4}{3}\Big)^{k(k-1)/2} \operatorname{disc}(\Delta)^2. \qquad \square$$

Example 20.30. A lattice with Gram matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ yields equality in Hermite's bound. We will replace the exponential factor $(4/3)^{(k-1)/4}$ by the polynomial factor $\sqrt{k}$.

Theorem 20.31 (Minkowski's theorem on linear forms)**.** *Let $\Delta \subseteq \mathbb{R}^n$ be a lattice with full rank. For all $d_1, \ldots, d_n \in \mathbb{R}_+$ with $d_1 \ldots d_n \geq \operatorname{disc}(\Delta)$ there exists an $x \in \Delta \setminus \{0\}$ with $|x_i| \leq d_i$ for $i = 1, \ldots, n$.*

Proof (Rado). Let $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ be a generator matrix of Δ.

Case 1: $\Delta \subseteq \mathbb{Z}^n$.
According to the Hermite normal form, we can assume that A is an upper triangular matrix. Then

$$d_1 \ldots d_n \geq \operatorname{disc}(\Delta) = |\det(A)| = |a_{11} \ldots a_{nn}|.$$

For integers $0 \leq \alpha_i < |a_{ii}|$ and $0 \leq \delta_i \leq d_i$, we consider the system of equations $xA = \alpha + \delta$. We show by induction on n that for a given δ, there exists exactly one α such that the system has an integer solution x. In the case $n = 1$, α_1 and x_1 are uniquely determined by

$$\alpha_1 = x_1 a_{11} - \delta_1 \equiv -\delta_1 \pmod{|a_{11}|}.$$

Let $n \geq 2$ and $B := A_{nn}$. Inductively, there exist unique $\alpha_1, \ldots, a_{n-1}$ such that $x_1, \ldots, x_{n-1}$ is an integer solution of $xB = \alpha + \delta$. Now α_n and x_n are uniquely determined by

$$\alpha_n = x_1 a_{1n} + \ldots + x_n a_{nn} - \delta_n \equiv a_{1n} x_1 + \ldots + x_{n-1} a_{n-1,n} - \delta_n \pmod{|a_{nn}|}$$

(because $a_{ni} = 0$ for $i < n$, x_n only appears in this equation).

Obviously, there are $(\lfloor d_1 \rfloor + 1) \ldots (\lfloor d_n \rfloor + 1) > d_1 \ldots d_n$ possible vectors δ, but only $|a_{11} \ldots a_{nn}| \leq d_1 \ldots d_n$ possible α. Thus, for at least one α, there must exist $\delta \neq \delta'$ such that $yA = \alpha + \delta$ and $zA = \alpha + \delta'$ have integer solutions $y \neq z$. For $x := (y - z)A \in \Delta \setminus \{0\}$, it holds that $|x_i| = |\delta_i - \delta_i'| \leq d_i$ for $i = 1, \ldots, n$.

Case 2: $\Delta \subseteq \mathbb{Q}^n$.
Let m be a common denominator of all entries in $A \in \mathbb{Q}^{n \times n}$. Then one can apply Case 1 to $m\Delta \subseteq \mathbb{Z}^n$ and $md_1, \ldots, md_n$. This yields an $x \in m\Delta \setminus \{0\}$ with $|x_i| \leq md_i$ for $i = 1, \ldots, n$. Obviously, $\frac{1}{m} x \in \Delta$ satisfies the claim.

Case 3: $\Delta \subseteq \mathbb{R}^n$.
Let $A_1, A_2, \ldots \in \mathbb{Q}^{n \times n}$ with $\lim_{k \to \infty} A_k = A$. According to Case 2, there exist $x_1, x_2, \ldots \in \mathbb{Z}^n \setminus \{0\}$ with $|(x_k A_k)_i| \leq d_i$ for $k \in \mathbb{N}$ and $i = 1, \ldots, n$. Because of $|x_k| = |(x_k A_k) A_k^{-1}| \leq |d| |A_k^{-1}|$, the sequence $(x_k)_k$ is bounded. It must therefore contain a constant subsequence $x := x_{k_1} = x_{k_2} = \ldots$. Now it holds that

$$|(xA)_i| = \lim_{j \to \infty} |(x A_{k_j})_i| \leq d_i$$

for $i = 1, \ldots, n$. $\square$

Remark 20.32. The linear forms theorem is a special case of *Minkowski's lattice point theorem*, which guarantees the existence of lattice vectors in certain convex sets. The corresponding geometric argument is sketched in Exercise III.28.

Corollary 20.33. *For every lattice Δ of rank k, it holds that* $\boxed{\min \Delta \leq \sqrt{k}\, \sqrt[k]{\operatorname{disc}(\Delta)}.}$

Proof. Since the statement only depends on the Gram matrix of Δ, we can assume $\Delta \subseteq \mathbb{R}^k$ according to Example 20.4. By the linear forms theorem, there exists an $x \in \Delta$ with $|x_i| \leq \sqrt[k]{\operatorname{disc}(\Delta)}$ for $i = 1, \ldots, k$. It follows

$$\min \Delta \leq |x| \leq \sqrt{k \operatorname{disc}(\Delta)^{2/k}} = \sqrt{k}\, \sqrt[k]{\operatorname{disc}(\Delta)}. \qquad \square$$

Remark 20.34. A calculation shows that the estimate from Corollary 20.33 is sharper than Hermite's bound for $k \geq 23$. However, both bounds are not optimal for $k \geq 3$. The smallest number γ_k with $\min \Delta \leq \sqrt{\gamma_k}\, \sqrt[k]{\operatorname{disc}(\Delta)}$ for all lattices Δ of rank k is called the k-th *Hermite constant* (cf. Remark 20.50). Using measure theory, Blichfeldt proved

$$\gamma_k \leq \frac{2}{\pi} \Gamma\Big(\frac{k}{2} + 2\Big)^{2/k} \leq \frac{2}{\pi} \Big(\Big\lfloor \frac{k+3}{2} \Big\rfloor !\Big)^{2/k}$$

where Γ is the Euler gamma function. So far, only the following values are known:

n	1	2	3	4	5	6	7	8	24
γ_n^n	1	$\frac{4}{3}$	2	4	8	$\frac{64}{3}$	64	2^8	2^{48}

The values $\gamma_2^2 = 4/3$ and $\gamma_3^3 = 2$ are derived in Theorem 20.59 and Theorem A.80. For $n = 8$, E_8 from Remark 20.11 yields the optimal value $\gamma_8 = 2$.

20.5 The LLL Algorithm

Remark 20.35. Many problems for lattices Δ can be solved algorithmically by constructing a basis of "short", "nearly orthogonal" vectors. Over $K \in \{\mathbb{R}, \mathbb{C}\}$, one can always construct an orthonormal basis using Gram-Schmidt. Over $\mathbb{Z}$, the situation is more complicated. In the first step, we bring the Gram matrix into a block diagonal form with blocks as small as possible.

Definition 20.36. A lattice $\Delta \subseteq \mathbb{R}^n$ is called *decomposable*, if there exist lattices $\Delta_1, \Delta_2 \subsetneq \Delta$ with $\Delta = \Delta_1 \oplus \Delta_2$ and $\Delta_1 \subseteq \Delta_2^{\perp}$. If applicable, $\Delta = \Delta_1 \perp \Delta_2$ is called an *orthogonal decomposition*. Otherwise, Δ is called *indecomposable*.

Remark 20.37. A lattice Δ is decomposable if and only if there exists a basis B with ${}_B[\Delta]_B = \operatorname{diag}(A_1, A_2)$.

Theorem 20.38 (Eichler)**.** *Every lattice Δ possesses an orthogonal decomposition $\Delta = \Delta_1 \perp \ldots \perp \Delta_k$ into indecomposable lattices $\Delta_1, \ldots, \Delta_k$ that are uniquely determined up to their order.*

Proof. Since the rank can only decrease finitely many times, Δ can always be orthogonally decomposed into indecomposable lattices $\Delta_1, \ldots, \Delta_k$. We call $x \in \Delta$ *minimal* if x is not the sum of shorter vectors, i. e. for $y, z \in \Delta$ with $x = y + z$ it holds that $|y| \geq |x|$ or $|z| \geq |x|$. Since only finitely many vectors in Δ have a given norm (Theorem 20.13), every $x \in \Delta$ can be written as a sum of minimal elements. Because $|x_1 + \ldots + x_k| = |x_1| + \ldots + |x_k|$ for $x_i \in \Delta_i$, every minimal element lies in one Δ_i. Let $x \in \Delta_i$ be minimal as an element of Δ_i. Let $y, z \in \Delta$ with $x = y + z$. Let $y = y_1 + \ldots + y_k$ and $z = z_1 + \ldots + z_k$ be the unique decompositions with $y_j, z_j \in \Delta_i$ for $j = 1, \ldots, k$. Then $x = y_i + z_i$ and $|y| = |y_1| + \ldots + |y_k| \geq |y_i| \geq |x|$ or $|z| \geq |x|$. Thus x is also minimal in Δ.

Let $M \subseteq \Delta$ be the set of all minimal elements. We define an equivalence relation on M by

$$x \sim y \ :\Longleftrightarrow\ \exists z_1, \ldots, z_s \in M : x = z_1, y = z_s, [z_i, z_{i+1}] \neq 0 \text{ for } i = 1, \ldots, s-1.$$

Elements from different equivalence classes are orthogonal and thus linearly independent. Therefore, there can only be finitely many equivalence classes $M_1, \ldots, M_l \subseteq M$. Each M_i lies in one Δ_j and generates as a group a lattice $\Lambda_i \subseteq \Delta_j$ (the set of all integer linear combinations of elements from M_i). By definition,

$$\Lambda_1 + \ldots + \Lambda_l = \Lambda_1 \perp \ldots \perp \Lambda_l.$$

If there exists an $x \in \Delta_j \setminus \Lambda_i$, then x can be decomposed into minimal elements (within Δ_j), of which at least one does not lie in M_i. Then Δ_j would have to contain further equivalence classes of M and one would have an orthogonal decomposition $\Delta_j = \Lambda_i \perp \Gamma$. This contradicts the indecomposability of Δ_j. Thus $\Delta_j = \Lambda_i$. Analogously, it follows that $k = l$ and $\Delta_i = \Lambda_i$ for $i = 1, \ldots, k$ after renumbering. □

Theorem 20.39. *For every indecomposable integral lattice Δ of rank $m \geq 2$, it holds that* $\min \Delta \geq 2$.

Proof. Let $A = (a_{ij}) \in \mathbb{Z}^{m \times m}$ be the Gram matrix of Δ with respect to a basis. Suppose there exists an $x \in \mathbb{Z}^m$ with $xAx^{\mathrm{t}} = 1$. Then $\gcd(x_1, \ldots, x_m) = 1$. According to Lemma 20.19, there exists an $S \in \mathrm{GL}(m, \mathbb{Z})$ with first column x. By replacing A with $S^{\mathrm{t}}AS$, we achieve $a_{11} = 1$. Now there exists an upper triangular matrix $T \in \mathrm{GL}(m, \mathbb{Z})$ with ones on the main diagonal such that $T^{\mathrm{t}}AT = \mathrm{diag}(1, A_1)$ with $A_1 \in \mathbb{Z}^{(m-1)\times(m-1)}$ holds. This contradicts Remark 20.37. □

Lemma 20.40. *Let $\Delta \subseteq \mathbb{R}^k$ be a lattice with basis $v_1, \ldots, v_k$. Let $b_1, \ldots, b_k$ be the orthogonal basis from the Gram-Schmidt process applied to $v_1, \ldots, v_k$ (without normalization). Then* $\mathrm{disc}(\Delta) = |b_1| \ldots |b_k|$ *holds.*

Proof. Let $A \in \mathbb{R}^{k \times n}$ be the generator matrix with rows $v_1, \ldots, v_k$. Let B be the matrix with rows $b_1, \ldots, b_k$. According to the Gram-Schmidt process, there exists a lower triangular matrix $R \in \mathbb{R}^{k \times k}$ with ones on the main diagonal and $RA = B$. Because of $\det(R) = 1$, it holds that

$$\mathrm{disc}(\Delta) = \sqrt{\det(AA^{\mathrm{t}})} = \sqrt{\det(BB^{\mathrm{t}})} = \sqrt{\det(\mathrm{diag}(|b_1|^2, \ldots, |b_k|^2))} = |b_1| \ldots |b_k|. \qquad \square$$

Remark 20.41.

(a) In the following, let $\lfloor x \rceil \in \mathbb{Z}$ be the rounded value of $x \in \mathbb{R}$, i. e. $|x - \lfloor x \rceil| \leq \frac{1}{2}$ (it does not matter whether one rounds $*.5$ up or down).

(b) Let $v_1, \ldots, v_k \in \mathbb{R}^n$ be linearly independent and $b_1, \ldots, b_k$ the corresponding Gram-Schmidt orthogonal basis. Then there exists a $w_k \in \langle b_1, \ldots, b_{k-1} \rangle = \langle v_1, \ldots, v_{k-1} \rangle$ with $v_k = b_k + w_k$. Therefore, b_k is the image of the projection of v_k onto $\langle b_1, \ldots, b_{k-1} \rangle^{\perp}$. In particular, b_k does not change if one replaces v_k with an element from $v_k + \langle v_1, \ldots, v_{k-1} \rangle$.

Definition 20.42. Let $\frac{1}{4} < \delta < 1$. Let $v_1, \dots, v_k$ be a basis of a lattice $\Delta \subseteq \mathbb{R}^n$ and $b_1, \dots, b_k$ the corresponding Gram-Schmidt orthogonal basis. Let $\mu_{ij} := \frac{[v_i, b_j]}{|b_j|^2}$ for $j < i$. One calls $v_1, \dots, v_k$ *δ-reduced*, if the following conditions hold:

- (Length condition) $|\mu_{ij}| \leq \frac{1}{2}$ for all $1 \leq j < i \leq k$,
- (Lovász condition) $|b_i|^2 \geq (\delta - \mu_{i,i-1}^2)|b_{i-1}|^2$ for all $2 \leq i \leq k$.

Theorem 20.43 (LLL algorithm[6])**.** *The following algorithm transforms a basis $v_1, \dots, v_k$ of a lattice $\Delta \subseteq \mathbb{R}^n$ into a δ-reduced basis.*

(1) Compute the Gram-Schmidt orthogonal basis $b_1, \dots, b_k$ from $v_1, \dots, v_k$.

(2) Compute μ_{ij} for $1 \leq j < i \leq k$.

(3) Set $i = 2$.

(4) As long as $i \leq k$ repeat:

(a) For $j = i-1, i-2, \dots, 1$:

- *Replace v_i with $v_i - \lfloor \mu_{ij} \rceil v_j$.*
- *Update μ_{il} for $l = 1, \dots, j$.*

(b) If the Lovász condition holds for v_i: increase i.

(c) Otherwise: swap v_i with v_{i+1}?! and go to step (1).

Proof. We first show that the algorithm terminates. The numbers

$$d_s := |b_1| \dots |b_s| \qquad\qquad D := d_1 \dots d_k$$

change only in step (1). Suppose v_i violates the Lovász condition, i. e. $|b_i|^2 < (\delta - \mu_{i,i-1}^2)|b_{i-1}|^2$. By the swap $v_i \leftrightarrow v_{i-1}$, $d_1, \dots, d_{i-2}$ remain untouched, while b_{i-1} is replaced by

$$b'_{i-1} = v_i - \sum_{j<i-1} \mu_{ij} b_j = b_i + \mu_{i,i-1} b_{i-1}$$

replaced. Because of $[b_{i-1}, b_i] = 0$, we have

$$|b'_{i-1}|^2 = |b_i|^2 + \mu_{i,i-1}^2 |b_{i-1}| < \delta |b_{i-1}|^2.$$

Therefore, d_{i-1} is replaced by

$$|b_1| \dots |b_{i-2}||b'_{i-1}| < \sqrt{\delta} d_{i-1}$$

replaced. For $j \geq i$, $v_1, \dots, v_j$ and $v_1, \dots, v_i, v_{i-1}, \dots, v_j$ obviously generate the same lattice Δ_j. Therefore, $d_i, \dots, d_k$ do not change according to Lemma 20.40. Overall, D is thus reduced by at least the factor $\sqrt{\delta} < 1$. According to Lemma 20.40 and Hermite, we have

$$d_j = \operatorname{disc}(\Delta_j) \geq \left(\frac{3}{4}\right)^{j(j-1)/4} \left(\min \Delta_j\right)^j \geq \left(\frac{3}{4}\right)^{j(j-1)/4} \left(\min \Delta\right)^j.$$

Therefore, D is bounded from below by a positive constant that depends only on Δ. This shows that the Lovász condition can only be violated finitely many times. Thus, after finitely many steps, $i = k+1$ is reached and the algorithm terminates.

[6]Pronounced: Triple-L algorithm. Named after A. Lenstra, H. Lenstra and L. Lovász.

At the end of the algorithm, the Lovász condition for $v_1, \dots, v_k$ is satisfied. In step (4), v_i is modified. According to Remark 20.41, $b_1, \dots, b_k$ remain untouched by this. For $j = i-1$, one replaces μ_{ij} by

$$\frac{[v_i - \lfloor \mu_{ij} \rceil v_j, b_j]}{|b_j|^2} = \frac{[v_i, b_j]}{|b_j|^2} - \lfloor \mu_{ij} \rceil = \mu_{ij} - \lfloor \mu_{ij} \rceil$$

(note $v_j \in b_j + \langle b_1, \dots, b_{j-1} \rangle$). Subsequently, $|\mu_{i,i-1}| \leq 1/2$ holds. For $j = i-2$, one analogously obtains $|\mu_{i,i-2}| \leq 1/2$, while $\mu_{i,i-1}$ is no longer changed because j is traversed in descending order. At the end, $|\mu_{ij}| \leq 1/2$ holds for all $j < i$, i. e. the length condition is satisfied. Overall, $v_1, \dots, v_k$ is δ-reduced. □

Remark 20.44. One can show that the LLL algorithm has polynomial runtime in k. Larger values of δ (i. e. close to 1) yield shorter basis vectors with longer runtime. In many implementations, $\delta = 3/4$ is chosen as a compromise. In fact, the algorithm also terminates for $\delta = 1$, but not necessarily in polynomial runtime (without proof).

Theorem 20.45. *Let $v_1, \dots, v_k$ be a δ-reduced basis of a lattice $\Delta \subseteq \mathbb{R}^n$. Let $\rho := \frac{2}{\sqrt{4\delta - 1}}$. Then*

$$|v_1| \leq \rho^{k-1} \min \Delta, \qquad |v_1| \dots |v_k| \leq \rho^{k(k-1)/2} \operatorname{disc}(\Delta).$$

Proof. Let $b_1, \dots, b_k$ be the Gram-Schmidt orthogonal basis for $v_1, \dots, v_k$. It holds that $\sigma := \rho^2 = \frac{1}{\delta - 1/4} > \frac{4}{3}$. From the δ-reduction it follows that

$$|b_i|^2 \geq (\delta - \mu_{i,i-1}^2)^2 |b_{i-1}|^2 \geq \sigma^{-1} |b_{i-1}|^2 \geq \dots \geq \sigma^{j-i} |b_j|^2$$

for $j < i$. Let $v = \lambda_1 v_1 + \dots + \lambda_i v_i = \eta_1 b_1 + \dots + \eta_i b_i \in \Delta$ be a shortest vector with $\lambda_1, \dots, \lambda_i \in \mathbb{Z}$, $\eta_1, \dots, \eta_i \in \mathbb{R}$ and $\lambda_i \neq 0$. By construction of $b_1, \dots, b_k$, it holds that $\eta_i = \lambda_i$. In particular, $|\eta_i| \geq 1$. Since $b_1, \dots, b_k$ are pairwise orthogonal, it follows that

$$(\min \Delta)^2 = |v|^2 = \eta_1^2 |b_1|^2 + \dots + \eta_i^2 |b_i|^2 \geq |b_i|^2 \geq \sigma^{1-i} |b_1|^2 = \sigma^{1-i} |v_1|^2 \geq \sigma^{1-k} |v_1|^2$$

and $|v_1| \leq \rho^{k-1} \min \Delta$. We show

$$1 + \frac{1}{4} \sum_{j=1}^{i-1} \sigma^j \leq \sigma^{i-1}$$

by induction on i. For $i = 1$ it is $1 = \sigma^0$. Let $i \geq 2$ and the assertion for $i-1$ be already proven. Then

$$1 + \frac{1}{4} \sum_{j=1}^{i-1} \sigma^j \leq \sigma^{i-2} + \frac{1}{4} \sigma^{i-1} = \sigma^{i-2} \Big(1 + \frac{1}{4} \sigma \Big) \leq \sigma^{i-1}.$$

As above, it is therefore

$$|v_i|^2 = \Big| b_i + \sum_{j=1}^{i-1} \mu_{ij} b_j \Big|^2 = |b_i|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 |b_j|^2 \leq |b_i|^2 + \frac{1}{4} \sum_{j=1}^{i-1} \sigma^{i-j} |b_i|^2 \leq \sigma^{i-1} |b_i|^2.$$

It follows that

$$|v_1| \dots |v_k| \leq \rho^{k(k-1)/2} |b_1| \dots |b_k| \overset{20.40}{=} \rho^{k(k-1)/2} \operatorname{disc}(\Delta).$$ □

Remark 20.46. For $\delta = 3/4$ one obtains $\rho = \sqrt{2} \approx 1.41$. With $\delta \to 1$, ρ approaches the factor $2/\sqrt{3} \approx 1.15$ from the Hermite bound. We will return to this in Remark 20.67. The estimate for $|v_1|$ suggests that v_1 is the shortest among the basis vectors. However, this is not always the case as the following example shows.

Example 20.47. We consider a lattice Δ with Gram matrix

$$\begin{pmatrix} 14 & 7 & 5 & 3 & -2 \\ 7 & 11 & 3 & 6 & 3 \\ 5 & 3 & 17 & -5 & 4 \\ 3 & 6 & -5 & 11 & 3 \\ -2 & 3 & 4 & 3 & 9 \end{pmatrix}$$

and $\operatorname{disc}(\Delta) = 169$. Hermite yields $\min \Delta \leq \frac{4}{3}\sqrt[5]{169} \approx 3.72$. The LLL algorithm with $\delta = 3/4$ or $\delta = 9/10$ yields reduced bases with Gram matrices

$$\begin{pmatrix} 14 & 7 & 4 & -4 & 5 \\ 7 & 11 & 1 & -5 & -1 \\ 4 & 1 & 7 & -1 & 3 \\ -4 & -5 & -1 & 10 & 1 \\ 5 & -1 & 3 & 1 & 10 \end{pmatrix}, \qquad \begin{pmatrix} 7 & 1 & -1 & 3 & 0 \\ 1 & 11 & -5 & -4 & 3 \\ -1 & -5 & 10 & 1 & 0 \\ 3 & -4 & 1 & 11 & -5 \\ 0 & 3 & 0 & -5 & 9 \end{pmatrix}.$$

Up to sign, there is only one shortest vector and $\min \Delta = \sqrt{7} \approx 2.65$.

20.6 Quadratic Forms

Remark 20.48. We change our perspective on lattices by neglecting the generator matrix and instead only considering Gram matrices. We then know the rank, but no longer the overarching Euclidean space. The situation here is similar to bilinear forms and their Gram matrices. In Remark 12.4, we introduced quadratic forms via bilinear forms. We repeat the definition with $\mathbb{Z}$ instead of a field.

Definition 20.49.

- For a symmetric matrix $C \in \mathbb{R}^{n\times n}$, the map

$$q = q_C \colon \mathbb{Z}^n \to \mathbb{R}, \qquad x \mapsto xCx^{\mathrm{t}} = \sum_{i,j=1}^{n} c_{ij} x_i x_j$$

  is called a *quadratic form* of rank n with *Gram matrix* C.

- One calls q_C
  - *non-degenerate*, if $C \in \mathrm{GL}(n, \mathbb{R})$.
  - *positive* (resp. *negative*), if C is positive (resp. negative) definite, i.e., $q(x) > 0$ (resp. $q(x) < 0$) for all $x \neq 0$. If q is positive, then

$$\min q := \min\bigl\{|q(x)| : x \in \mathbb{Z}^n \setminus \{0\}\bigr\}$$

    is called the *minimum* of q.
  - *integral*, if $C \in \mathbb{Z}^{n\times n}$.

- One calls $\det(q) := \det(C)$ the *determinant* of q.

- Two quadratic forms q_A and q_B are called *equivalent*, if there exists an $S \in \mathrm{GL}(n, \mathbb{Z})$ with $S^{\mathrm{t}}AS = B$.

Remark 20.50.

(a) The symmetry of C is non-essential, because for an arbitrary matrix $A \in \mathbb{R}^{n\times n}$, $C := \frac{1}{2}(C + C^{\mathrm{t}})$ is symmetric with $q_A = q_C$. If A is integral, then C does not have to be integral. Some authors therefore define integral quadratic forms by the condition $q_C(x) \in \mathbb{Z}$ for all $x \in \mathbb{Z}^n$. This means $c_{ii} \in \mathbb{Z}$ and $2c_{ij} \in \mathbb{Z}$ for all $i \neq j$.

(b) If q is degenerate, one can eliminate a variable and reduce the rank by transitioning to an equivalent form. We are mainly interested in positive quadratic forms (cf. Exercise III.30). If q is negative, one can switch to the positive form $-q$ and transfer results.

(c) Every lattice Δ with Gram matrix G defines a positive quadratic form q_G with $|x|^2 = q_G(x_1, \ldots, x_k)$ for all $x \in \Delta$, where $x_1, \ldots, x_k$ are the coefficients wrt. the chosen basis. A change of basis of Δ corresponds, according to Lemma 20.7, to the transition to an equivalent quadratic form. Conversely, we have seen in Example 20.4 how to construct a lattice from a positive definite matrix. In this way, one can switch back and forth between lattices and positive quadratic forms. However, there are many lattices with the same Gram matrix and arbitrary rank. Note that $\min q_G = (\min \Delta)^2$ and $\det(q) = \operatorname{disc}(\Delta)^2$ holds. The Hermite bound therefore has the form

$$\min q \leq \left(\frac{4}{3}\right)^{(k-1)/2} \sqrt[k]{\det(q)}$$

(d) Equivalent quadratic forms q_A and q_B have the same determinant and take the same values. In particular, $\min q_A = \min q_B$ if the forms are positive. Just as we reduced a lattice basis with the LLL algorithm, we will replace a quadratic form with a "simplest" possible equivalent form.

(e) As with lattices, for a given $d > 0$ there are only finitely many $x \in \mathbb{Z}^n$ with $q(x) \leq d$ (Remark 20.11). Furthermore, according to Lemma 20.28, there exists an equivalent form q_C with $c_{11} = \min q$.

Theorem 20.51. *Every integral quadratic form q is equivalent to a form q_C with*

$$C = \begin{pmatrix} * & * & & 0 \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix} \geq 0.$$

Proof. The proof proceeds similarly to Theorem 20.24. Let $q = q_A$. Let $(a_1, \ldots, a_n)$ be the first row/column of A and $b_2, \ldots, b_n \in \mathbb{Z}$ with

$$d := \gcd(a_2, \ldots, a_n) = a_2 b_2 + \ldots + a_n b_n.$$

According to Lemma 20.19, there exists a $U \in \mathrm{GL}(n-1, \mathbb{Z})$ with first column $(b_2, \ldots, b_n)$. Let $S := \operatorname{diag}(1, U) \in \mathrm{GL}(n, \mathbb{Z})$. Then AS has the entry d at position $(1, 2)$. Repeating this step, a_2 eventually divides each of the numbers $a_2, \ldots, a_n$. As usual, one achieves $a_i = 0$ for $i = 3, \ldots, n$. By multiplying S^{t} from the left to AS, the first row is not changed. Since $S^{\mathrm{t}}AS$ is symmetric, the first column of A now has the same form. One proceeds analogously with rows $2, \ldots, n-1$. In the end, A has been transformed into the desired matrix C. □

Example 20.52.

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

Remark 20.53.

(a) Theorem 20.51 is not suitable for counting equivalence classes of quadratic forms, because infinitely many such *tridiagonal matrices* are equivalent to each other. For *binary* quadratic forms (i.e., $n = 2$), Theorem 20.51 merely provides a sign reduction. We treat this case in Theorem 20.59.

(b) An integral positive quadratic form q is called *unimodular* if $\det(q) = 1$. Let q_C be unimodular with $n \leq 5$. Because $(4/3)^2 < 2$, $\min q = 1$ according to Hermite. As in Theorem 20.39, one constructs an equivalent form with $C = \operatorname{diag}(1, C_1)$ and $\det(C_1) = 1$. One can now repeat the same argument with q_{C_1}. Finally, q is equivalent to q_{1_n}. In particular, q is, up to equivalence, the only unimodular quadratic form of rank n. According to a theorem by MORDELL, this also holds for $n \in \{6, 7\}$. For $n = 8$, the E_8-lattice introduced in Remark 20.11 determines a unimodular form q with $\min q = 2$. In particular, q is not equivalent to q_{1_8}.

Theorem 20.54. *For all $n \in \mathbb{N}$ and $d \in \mathbb{R}$, there exist, up to equivalence, only finitely many positive quadratic forms q with rank n and $\det(q) \leq d$.*

Proof. Let q be a quadratic form of rank n with $\det(q) \leq d$. Let Δ be a corresponding lattice (Remark 20.50). Let $v_1, \ldots, v_n$ be a δ-reduced basis of Δ (for example for $\delta = 3/4$). By transitioning to an equivalent quadratic form, one can assume q_C with $c_{ii} = |v_i|^2$ for $i = 1, \ldots, n$. According to Theorem 20.45, $c_{11} \ldots c_{nn}$ is bounded by a function in d. Since C is positive definite, it also holds that $c_{ii}c_{jj} - c_{ij}^2 > 0$, i.e., $|c_{ij}| < \sqrt{c_{ii}c_{jj}}$ for $i \neq j$. Therefore, there are only finitely many possibilities for C. □

Definition 20.55. A positive quadratic form q with Gram matrix $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ is called (*Minkowski*)*-reduced*, if for $i = 1, \ldots, n$ the following holds:

- $c_{ii} \leq q(x)$ for all $x \in \mathbb{Z}^n$ with $\gcd(x_i, x_{i+1}, \ldots, x_n) = 1$,
- $c_{i,i+1} \geq 0$ if $i < n$.

Theorem 20.56 (MINKOWSKI)**.** *Every positive quadratic form is equivalent to a reduced form.*

Proof. We reduce a positive quadratic form q with Gram matrix C step by step. Let $1 \leq i \leq n$. Among the vectors $x \in \mathbb{Z}^n$ with $\gcd(x_i, \ldots, x_n) = 1$, we can choose one with minimal $q(x)$ according to Remark 20.50. According to Lemma 20.19, there exists a $T \in \mathrm{GL}(n - i + 1, \mathbb{Z})$ with first column $(x_i, \ldots, x_n)$. Then

$$S := \begin{pmatrix} 1 & & & x_1 & 0 \\ & \ddots & & \vdots & \vdots \\ & & 1 & x_{i-1} & 0 \\ & & & & \\ & & & T & \end{pmatrix} \in \mathrm{GL}(n, \mathbb{Z}).$$

Replacing C by $S^{\mathrm{t}}CS$, $c_{11}, \ldots, c_{i-1,i-1}$ remain unchanged, while c_{ii} now satisfies the first Minkowski condition. In this way, one ensures that the condition holds for $c_{11}, \ldots, c_{nn}$. The signs of $c_{i,i+1}$ can be changed by $S = \operatorname{diag}(\epsilon_1, \ldots, \epsilon_n)$ with $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ ($c_{11}, \ldots, c_{nn}$ remain unchanged). □

Example 20.57. Let q be the quadratic form for the lattice $\Delta \subseteq \mathbb{R}^5$ from Example 20.47. The Gram matrix

$$\begin{pmatrix} 7 & 0 & 3 & -1 & 3 \\ 0 & 9 & 4 & 0 & 2 \\ 3 & 4 & 10 & 1 & -2 \\ -1 & 0 & 1 & 10 & 3 \\ 3 & 2 & -2 & 3 & 13 \end{pmatrix}$$

defines an equivalent reduced form. Compared to the δ-reduction, the diagonal elements have become even smaller.

Remark 20.58.

(a) Let q_C be reduced. Let $x \in \mathbb{Z}^n \setminus \{0\}$, $g := \gcd(x_1, \ldots, x_n)$ and $y_i := x_i/g$. Then $c_{11} \leq q(y_1, \ldots, y_n) = g^2 q(x)$ holds. This shows $c_{11} = \min q$. Furthermore, $c_{11} \leq q(e_2) = c_{22} \leq q(e_3) = c_{33} \leq \ldots \leq c_{nn}$. For $i < j$ it also holds that

$$c_{jj} \leq q(e_i \pm e_j) = c_{ii} \pm 2c_{ij} + c_{jj},$$

i.e. $2|c_{ij}| \leq c_{ii}$.

(b) One can show that q_C is already reduced if finitely many inequalities of the form $c_{ii} \leq q(x)$ hold (together with $c_{i,i+1} \geq 0$). However, the number of these inequalities grows exponentially with n. For $n \geq 10$ it is very costly to reduce a given quadratic form. The reduction according to Korkine-Zolotarev provides a compromise between the efficiency of the LLL algorithm and the quality of the Minkowski reduction. We will not go into this in detail.

(c) The diagonal entries c_{ii} of a reduced quadratic form q_C are uniquely determined by the Minkowski condition. Furthermore, there exist only finitely many vectors $x \in \mathbb{Z}^n$ with $q_C(x) = c_{ii}$. Therefore, there exist only finitely many matrices $S \in \mathrm{GL}(n, \mathbb{Z})$ such that the equivalent form $q_{S^{\mathrm{t}}CS}$ is also reduced. This shows that every positive quadratic form is equivalent to at most finitely many reduced forms. Among these, one can choose a canonical representative by defining an order on the matrix entries (e.g., among all equivalent forms let c_{12} be minimal etc.). For $n \leq 2$ there is generally only one reduced form per equivalence class.

Theorem 20.59 (Gauss)**.** *A positive quadratic form $q = q_C$ is reduced if and only if $0 \leq 2c_{12} \leq c_{11} \leq c_{22}$. In this case $c_{11}c_{22} \leq \frac{4}{3}\det(q)$. If q_D is reduced and equivalent to q, then $C = D$.*

Proof. Let $a := c_{11}$, $b := c_{12}$ and $c := c_{22}$. If q is reduced, then $0 \leq 2b \leq a \leq c$ holds by Remark 20.58. Conversely, let $0 \leq 2b \leq a \leq c$ be given. For $(x, y) \in \mathbb{Z}^2 \setminus \{0\}$ it holds that

$$\begin{aligned} q(x,y) &= ax^2 + 2bxy + cy^2 \geq a(x^2 + y^2) + 2bxy \\ &\geq a(x^2 + y^2) - a|xy| = a(|x| - |y|)^2 + a|xy| \geq a \end{aligned} \tag{20.2}$$

with equality if $(x, y) = (1, 0)$. Therefore $a = \min q \leq q(x_1, x_2)$ for $\gcd(x_1, x_2) = 1$. Furthermore $c = q(0, 1) \leq q(x_1, x_2)$ for $\gcd(x_2) = 1$. Thus q is reduced.

For uniqueness, let $S = \left(\begin{smallmatrix} x & u \\ y & v \end{smallmatrix}\right) \in \mathrm{GL}(2, \mathbb{Z})$ with $S^{\mathrm{t}}CS = \left(\begin{smallmatrix} a & b' \\ b' & c' \end{smallmatrix}\right)$ and $0 \leq 2b' \leq a \leq c'$ (note $a = \min q$). Wlog. let $c' \leq c$. For (x, y), equality holds in (20.2). There are two cases for this:

Case 1: $y \neq 0$.
From (20.2) it follows that $c = a \leq c' \leq c$. Thus $c = c'$. From $a^2 - b^2 = \det(C) = a^2 - (b')^2$ and $b' \geq 0$ one obtains $b = b'$.

Case 2: $(x, y) = (\pm 1, 0)$.
By replacing S with $-S$, one can assume $x = 1$. Then $v = \det(S) = \pm 1$. From

$$ua - \frac{a}{2} \leq ua \pm b = b' \leq \frac{a}{2}$$

it follows that $u \leq 1$. In the case $u < 0$, b' would be negative. For $u = 0$, $S = 1_2$. So let $u = 1$. Then $b = a/2 = b'$. From $ac - b^2 = \det(C) = ac' - b^2$ it follows that $c = c'$. □

Remark 20.60.

(a) The bound $c_{11}c_{22} \leq \frac{4}{3}\det(q)$ corresponds to the estimate from Theorem 20.45 for $\delta \to 1$.

(b) Let Δ be a lattice with δ-reduced basis v_1, v_2 and Gram matrix $C = (c_{ij})$. Then

$$\frac{|c_{12}|}{c_{11}} = \frac{|[v_2, v_1]|}{|v_1|^2} = |\mu_{21}| \leq \frac{1}{2},$$

i. e. $2|c_{12}| \leq c_{11}$. The (only) Lovász condition is

$$(\delta - \mu_{21}^2)|v_1|^2 \leq |b_2|^2 = |v_2 - \mu_{21}v_1|^2 = |v_2|^2 - 2\mu_{21}[v_1, v_2] + \mu_{21}^2|v_1|^2 = |v_2|^2 - \mu_{21}^2|v_1|^2,$$

i. e. $\delta c_{11} \geq c_{22}$. For $\delta = 1$, q_C is reduced if one ignores the sign of c_{12}.

(c) The following algorithm reduces a positive binary quadratic form q_C:

(1) If $c_{11} > c_{22}$, swap c_{11} and c_{22}.

(2) Let $\lambda := \lfloor c_{12}/c_{11} \rceil$. Subtract λ times the first row/column from the second row/column of C. Afterwards, $2|c_{12}| \leq c_{11}$ holds.

(3) If $c_{11} > c_{22}$, then go to (1).

(4) Replace c_{12} by $|c_{12}|$.

(d) The equivalence classes of positive quadratic forms q_C with rank 2 and $\det(q) \leq 10$ are given by the following parameters:

$\det(q)$	1	2	3	3	4	4	5	5	6	6	7	7	8	8	8	9	9	9	10	10
c_{11}	1	1	1	2	1	2	2	1	2	1	2	1	3	2	1	3	2	1	2	1
c_{12}	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0
c_{22}	1	2	3	2	4	2	3	5	3	6	4	7	3	4	8	3	5	9	5	10

There are such tables for $n \leq 5$ on the internet (cf. Exercise III.32).[7]

(e) For $n \geq 3$, there exist equivalent positive reduced quadratic forms with different Gram matrices. To see this, consider $q_k := q_{C_k}$ with

$$C_k := \begin{pmatrix} 2 & 1 & k \\ 1 & 2 & 1 \\ k & 1 & 2 \end{pmatrix}$$

for $k \in \{0, 1\}$. According to Example 20.52, q_0 and q_1 are equivalent. According to Example 12.35, q_0 is positive and $\min q_0 = 2$. Because $c_{ii} = 2 = \min q_k$ for $i = 1, 2, 3$ and $c_{12} = c_{23} = 1$, q_0 and q_1 are reduced.

[7] `https://www.math.rwth-aachen.de/homes/Gabriele.Nebe/LATTICES/`

20.7 Successive Minima

Definition 20.61. Let $q\colon \mathbb{Z}^n \to \mathbb{R}$ be a positive quadratic form. Following Courant-Fischer, we define the *successive minima*

$$\mu_i(q) := \min_{\substack{s_1,\ldots,s_i \in \mathbb{Z} \\ \text{linearly independent}}} \max\{q(s_1), \ldots, q(s_i)\}.$$

Remark 20.62. Obviously $\min q = \mu_1(q) \leq \ldots \leq \mu_n(q)$. If q_C is reduced, then

$$\mu_i(q) \leq \max\{q(e_1), \ldots, q(e_i)\} = c_{ii}$$

for $i = 1, \ldots, n$. In general, however, equality does not necessarily hold.

Example 20.63.

(a) Let

$$C = \begin{pmatrix} 1 & . & . & . & 1/2 \\ . & 1 & . & . & 1/2 \\ . & . & 1 & . & 1/2 \\ . & . & . & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 & 5/4 \end{pmatrix}.$$

Because of

$$q(x) = x_1^2 + \ldots + x_4^2 + \frac{5}{4}x_5^2 + x_1x_5 + \ldots + x_4x_5 = \sum_{i=1}^{4}\Big(x_i + \frac{1}{2}x_5\Big)^2 + \frac{1}{4}x_5^2 \geq 1$$

for $x \in \mathbb{Z}^n \setminus \{0\}$, q is positive and reduced. On the other hand,

$$B := \big\{e_1, \ldots, e_4, (-1,-1,-1,-1,2)\big\}$$

is linearly independent with $q(b) = 1$ for $b \in B$. This shows $\mu_5(q) = 1 < \frac{5}{4} = c_{55}$.

(b) The successive minima of the form q from Example 20.57 are $7, 9, 10, 10, 11$. Here too, $\mu_5(q) = 11 < 13 = c_{55}$ holds.

Lemma 20.64. *For every positive quadratic form q of rank n, there exist linearly independent vectors $s_1, \ldots, s_n \in \mathbb{Z}^n$ with $q(s_i) = \mu_i(q)$ for $i = 1, \ldots, n$.*

Proof. Obviously, there exists $s_1 \in \mathbb{Z}^n$ with $q(s_1) = \mu_1(q)$. Suppose $s_1, \ldots, s_k$ with $q(s_i) = \mu_i(q)$ for $i = 1, \ldots, k$ have already been chosen. By definition, there exist linearly independent $t_1, \ldots, t_{k+1}$ with $q(t_i) \leq \mu_{k+1}(q)$ for $i = 1, \ldots, k+1$. At least one of the t_i, say t_1, must be linearly independent of $s_1, \ldots, s_k$. Let $i \leq k+1$ be minimal with $\mu_i(q) = \mu_{k+1}(q)$. From

$$\mu_i(q) \leq \max\big\{q(s_1), \ldots, q(s_{i-1}), q(t_1)\big\} \leq \mu_{k+1}(q)$$

it follows that $q(t_1) = \mu_{k+1}(q)$. We can therefore set $s_{k+1} := t_1$. The claim now follows inductively. □

Remark 20.65. The following theorem improves Hermite's bound (Remark 20.50).

Theorem 20.66 (MINKOWSKI)**.** *For every positive quadratic form q of rank n, it holds that*

$$\boxed{\mu_1(q)\dots\mu_n(q) \le \left(\frac{4}{3}\right)^{n(n-1)/2} \det(q).}$$

Proof. Let $S \in \mathbb{Z}^{n\times n}$ with columns $s_1, \dots, s_n$ as in Lemma 20.64 (Note: S is not necessarily in $\mathrm{GL}(n,\mathbb{Z})$). Let $q = q_C$ and $S^{\mathrm{t}}CS = R^{\mathrm{t}}R$ be the Cholesky decomposition of $S^{\mathrm{t}}CS$. Let

$$D := \mathrm{diag}(\mu_1(q), \dots, \mu_n(q)), \qquad C_1 := S^{-\mathrm{t}}R^{\mathrm{t}}D^{-1}RS^{-1}$$

and $q_1 := q_{C_1}$. Then

$$\det(q_1) = \det(D)^{-1}\det(S^{-\mathrm{t}}R^{\mathrm{t}}RS^{-1}) = \frac{\det(q)}{\mu_1(q)\dots\mu_n(q)}.$$

Let $x \in \mathbb{Z}^{n\times 1} \setminus \{0\}$, $y := S^{-1}x$ and $z := Ry$. Let $k := \max\{1 \le i \le n : y_i \ne 0\}$. Then $x = Sy = y_1s_1 + \dots + y_ks_k$ is linearly independent of $s_1, \dots, s_{k-1}$. In particular, $q(x) \ge \mu_k(q)$. Since R is an upper triangular matrix, it also holds that $z_{k+1} = \dots = z_n = 0$. Therefore

$$q_1(x) = y^{\mathrm{t}}R^{\mathrm{t}}D^{-1}Ry = z^{\mathrm{t}}D^{-1}z = \frac{z_1^2}{\mu_1(q)} + \dots + \frac{z_k^2}{\mu_k(q)} \ge \frac{1}{\mu_k(q)}|z|^2 = \frac{1}{\mu_k(q)}y^{\mathrm{t}}R^{\mathrm{t}}Ry = \frac{1}{\mu_k(q)}q(x) \ge 1.$$

In particular, q_1 is positive and

$$1 \le (\min q_1)^n \le \left(\frac{4}{3}\right)^{n(n-1)/2}\det(q_1) = \left(\frac{4}{3}\right)^{n(n-1)/2}\frac{\det(q)}{\mu_1(q)\dots\mu_n(q)}$$

according to Hermite (Remark 20.50). □

Remark 20.67.

(a) The proof shows that one may replace $(4/3)^{(n-1)/2}$ in Minkowski's bound by the n-th Hermite constant γ_n (Example 20.30).

(b) For a δ-reduced lattice basis $v_1, \dots, v_n$, the same bound $|v_1|^2 \dots |v_n|^2 \le \left(\frac{4}{3}\right)^{n(n-1)/2}\det(q)$ holds asymptotically as $\delta \to 1$ according to Theorem 20.45.

(c) An integral positive quadratic form q is called *universal*, if $q\colon \mathbb{Z}^n \to \mathbb{N}_0$ is surjective. A theorem of Lagrange from number theory states that the quadratic form

$$q(x) = x1_4x^{\mathrm{t}} = x_1^2 + x_2^2 + x_3^2 + x_4^2$$

is universal, i.e., every natural number is the sum of four squares.[8] The 15-*Theorem* of CONWAY states that an integral positive quadratic form is already universal if it takes the values 1, 2, 3, 5, 6, 7, 10, 14, and 15.

[8] see Number Theory notes

Exercises

Exercise III.1 (Binary Exponentiation)**.** Let K be a field and $A \in K^{n\times n}$.

(a) Design an algorithm for the efficient calculation of A^k by iterated squaring.

(b) Show that in this way one can get by with at most $2\lfloor \log_2(k) \rfloor$ matrix multiplications (where $\lfloor \log_2(k) \rfloor$ is the largest integer z with $2^z \leq k$).

Exercise III.2 (Heron's method)**.** Let $a \in \mathbb{R}$ with $a \geq 1$. Show that the sequence

$$x_0 := 1, \qquad x_{n+1} := \frac{1}{2}\Big(x_n + \frac{a}{x_n}\Big)$$

converges *quadratically* to $\sqrt{a}$, i. e. $\lim_{n\to\infty} x_n = \sqrt{a}$ and $|x_{n+1} - \sqrt{a}| \leq \frac{1}{2}|x_n - \sqrt{a}|^2$ for $n \in \mathbb{N}$. Explain why the number of correct decimal places of x_n doubles in each iteration step.

Exercise III.3. Let K be a field and $A \in K^{n\times n}$ of rank k. Show that there exist $B, C \in K^{n\times k}$ with $A = BC^{\mathrm{t}}$. How can this be used to speed up calculations and reduce memory requirements if $k \ll n$.

Exercise III.4. Prove or disprove: Similar matrices in $\mathbb{C}^{n\times n}$ have the same condition number.

Exercise III.5. Show that $A \in \mathbb{C}^{n\times m}$ has the same rank as the pseudoinverse A^+.

Exercise III.6. Prove or disprove: $(AB)^+ = B^+A^+$ for all $A \in \mathbb{C}^{n\times m}$ and $B \in \mathbb{C}^{m\times k}$.

Exercise III.7. Show: If the system $Ax = b$ with $A \in \mathbb{C}^{n\times m}$ has at least one solution, then all solutions have the form $A^+b + (1_m - A^+A)y$ with $y \in \mathbb{C}^m$.

Exercise III.8. Determine the largest (finite) number that can be represented with the data type `float` (Remark 17.29).

Exercise III.9. Let $A \in K^{n\times n}$ with principal minors $\det(A_k) \neq 0$ for $k = 1, \ldots, n-1$ (see Remark 12.43). Show that a unique LU decomposition in the form $A = LU$ exists (see Theorem 17.35).

Exercise III.10 (Bruhat decomposition)**.** Show that for every matrix $A \in \mathrm{GL}(n, K)$ there exist permutation matrices P, Q, lower triangular matrices L_1, L_2 and upper triangular matrices U_1, U_2 with $A = L_1PL_2 = U_1QU_2$.

Exercise III.11 (Polar decomposition)**.** Show that for every matrix $A \in \mathbb{C}^{n\times n}$ there exist uniquely determined matrices $U \in \mathrm{U}(n, \mathbb{C})$ and $P \in \mathbb{C}^{n\times n}$ with $A = UP$, where P is positive semidefinite.

Exercise III.12 (HADAMARD inequality)**.** Let $A \in \mathbb{C}^{n\times n}$ with columns $s_1, \ldots, s_n$. Show $|\det(A)| \leq |s_1| \ldots |s_n|$ with equality if and only if $s_1, \ldots, s_n$ are pairwise orthogonal.
Hint: QR decomposition.

Exercise III.13. Let $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, $s, x, y \in \mathbb{R}_{\geq 0}$ and $v, w \in \mathbb{C}^n$. Show:

(a) (BERNOULLI inequality) $1 + sx \leq (1 + x)^s$ for $s \geq 1$.
Hint: One needs the continuity of the power function.

(b) (YOUNG inequality) $xy \leq \frac{x^p}{p} + \frac{y^q}{q}$.

(c) (HÖLDER inequality) $\sum_{i=1}^n |v_i w_i| \leq \|v\|_p \|w\|_q$.
Hint: This generalizes the Cauchy-Schwarz inequality.

(d) (MINKOWSKI inequality) $\|v + w\|_p \leq \|v\|_p + \|w\|_p$.

(e) The p-norm is indeed a norm on $\mathbb{C}^n$.

Exercise III.14. Let $\|.\|$ be a norm defined on $\mathbb{C}^{n\times 1}$ for all n. Show that

$$\|A\| = \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

defines a matrix norm on $\mathbb{C}^{n\times m}$.

Exercise III.15. Formulate and prove the orthonormalization procedure with Householder transformations and Givens rotations from Remark 17.87 for $A \in \mathrm{GL}(n, \mathbb{C})$.

Exercise III.16. Show $\lim_{k\to\infty} \left(1_n + \frac{1}{k}A\right)^k = \exp(A)$ for all $A \in \mathbb{C}^{n\times n}$.

Exercise III.17. Show that $\exp(A)$ is positive definite if $A \in \mathbb{C}^{n\times n}$ is Hermitian.

Exercise III.18. Show:

(a) The number of permutations in S_n with exactly k cycles (including 1-cycles) is

$$\begin{bmatrix} n \\ k \end{bmatrix} := \frac{n!}{k!} \sum_{\substack{1\leq l_1,\ldots,l_k\leq n \\ l_1+\ldots+l_k=n}} \frac{1}{l_1 \ldots l_k}.$$

Hint: $\begin{bmatrix} n \\ k \end{bmatrix}$ is called the *Stirling number* of the first kind.

(b) For $n \geq 2$ there are as many permutations in S_n with an even number of cycles as with an odd number of cycles, i. e. $\sum_{k=1}^n (-1)^k \begin{bmatrix} n \\ k \end{bmatrix} = 0$.

(c) Let $N \in \mathbb{C}^{n\times n}$ be nilpotent and

$$A := -\sum_{k=1}^{n} \frac{1}{k} N^k \in \mathbb{C}^{n\times n}.$$

Then $\exp(A) = 1_n - N$ holds.

Exercise III.19. Show that for $A \in \mathrm{SL}(n, \mathbb{C})$ there exists a $B \in \mathbb{C}^{n\times n}$ with $\exp(B) = A$ and $\mathrm{tr}(B) = 0$.

Exercise III.20. Let $A, B \in \mathbb{R}^{n\times n}$ be stochastic matrices. Show that AB is stochastic.

Exercise III.21. If $A \in \mathbb{R}^{n\times n}$ and A^{-1} are stochastic, then A is a permutation matrix.

Exercise III.22. Let $0 < p, q < 1$ and $A := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$. Calculate $\lim_{k\to\infty} A^k$.

Exercise III.23. Let $A \in \mathbb{R}^{n\times n}$ be non-negative and irreducible. Show $(1_n + A)^{n-1} > 0$.

Exercise III.24. Let V be a K-vector space of dimension n. A set $A \subseteq V$ is called *affinely dependent*, if pairwise distinct elements $a_1, \ldots, a_k \in A$ and $\lambda_1, \ldots, \lambda_k \in K^\times$ with $\lambda_1 a_1 + \ldots + \lambda_k a_k = 0$ and $\lambda_1 + \ldots + \lambda_k = 0$ exist. Otherwise, A is called *affinely independent*. Show:

(a) There exists an affinely independent set with $n + 1$ elements.

(b) Every set of at least $n + 2$ elements is affinely dependent.

(c) $A \subseteq V$ is affinely dependent if and only if $\{v - w : w \in A \setminus \{v\}\}$ is linearly dependent for some $v \in A$.

Exercise III.25. Let V be a K-vector space. An *affine* linear combination of $v_1, \ldots, v_k \in V$ is a sum of the form $\lambda_1 v_1 + \ldots + \lambda_k v_k$ with $\lambda_1, \ldots, \lambda_k \in K$ and $\lambda_1 + \ldots + \lambda_k = 1$.[9] The set of all affine linear combinations of elements from $\Delta \subseteq K^n$ is called the *affine hull* $\mathrm{aff}(\Delta)$.

(a) Describe the affine hull of two points in $\mathbb{R}^n$.

(b) Show that a set $\Delta \subseteq V$ is closed under affine linear combinations (i. e. $\mathrm{aff}(\Delta) = \Delta$) if and only if there exists an $x \in \Delta$ such that $\Delta - x$ is a subspace of V.

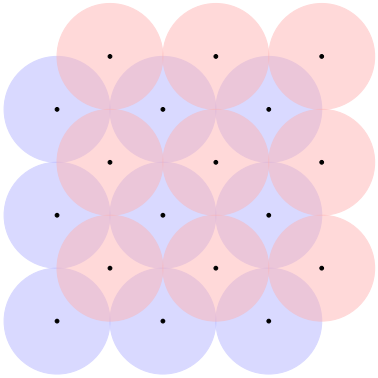
Exercise III.26. Let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^{n\times 1}$. Show the following variants of Corollary 19.13:

(a) There exists an x with $Ax \leq b$ if and only if $y^{\mathrm{t}} b \geq 0$ for all $y \geq 0$ with $y^{\mathrm{t}} A = 0$.
Hint: Consider $(1_n, A, -A)$.

(b) There exists an $x \geq 0$ with $Ax \leq b$ if and only if $y^{\mathrm{t}} b \geq 0$ for all $y \geq 0$ with $y^{\mathrm{t}} A \geq 0$.

Exercise III.27.

(a) Determine the lattice $\Delta \subseteq \mathbb{R}^3$ corresponding to the square arrangement of unit spheres (the spheres in each layer touch spheres in the layer below, see Remark 20.11):



[9] In contrast to convex combinations, the λ_i are not restricted.

(b) Calculate $\min \Delta$ and the density $\rho(\Delta)$.

(c) Explain why this arrangement is not a counterexample to Kepler's conjecture.

Exercise III.28. Let $\Delta \subseteq \mathbb{R}^n$ be a lattice with full rank and $\delta := \frac{1}{\sqrt{n}} \min \Delta$. Let $F \subseteq \mathbb{R}^n$ be the fundamental mesh of Δ (Example 20.9). Let $W := (-\delta/2, \delta/2)^n \subseteq \mathbb{R}^n$ be the open cube with center 0 and side length δ. Show:

(a) For distinct $x, y \in \Delta$, $(x + W) \cap (y + W) = \varnothing$.

(b) There exist $x_1, \ldots, x_k \in \Delta$ with $W \subseteq \bigcup_{i=1}^{k} (x_i + F)$.

(c) For the volume, it holds that

$$\delta^n = \mathrm{vol}(W) = \mathrm{vol}\big(F \cap ((W - x_1) \cup \ldots \cup (W - x_k))\big) \leq \mathrm{vol}(F) = \mathrm{disc}(\Delta),$$

i. e. $\min \Delta \leq \sqrt{n} \sqrt[n]{\mathrm{disc}(\Delta)}$.

(d) For the dual lattice Δ^*, it holds that $(\min \Delta)(\min \Delta^*) \leq n$.

(e) For every positive quadratic form q of rank n, it holds that $\min q \leq n \det(q)$.

Exercise III.29. Let q_C be a reduced positive quadratic form of rank 2. Show for the successive minimum $\mu_2(q) = c_{22}$.

Exercise III.30. Let q be an *indefinite* quadratic form of rank 2, i. e. there exist $x, y \in \mathbb{Z}^2$ with $q(x) < 0 < q(y)$. Let $d := -\det(q)$ not be a square number. Show:

(a) $d > 0$.

(b) q is equivalent to a form q_C with $0 < c_{12} < \sqrt{d} < |c_{11}| + c_{12} \leq |c_{22}| + c_{12}$.

(c) Up to equivalence, there are only finitely many integral indefinite quadratic forms of rank 2 and determinant $-d$ (even if d is a square).

(d) Determine up to equivalence all integral indefinite quadratic forms q of rank 2 with $\det(q) \geq -4$.

Remark: In contrast to positive binary quadratic forms, there is no canonical normal form for indefinite forms.

Exercise III.31. Show that a positive quadratic form q_C of rank 3 is reduced if and only if:

$$0 \leq 2c_{12} \leq c_{11} \leq c_{22}, \qquad 2|c_{13}| \leq c_{11},$$
$$0 \leq 2c_{23} \leq c_{22} \leq c_{33}, \qquad 2(c_{12} + c_{23} - c_{13}) \leq c_{11} + c_{22}.$$

Exercise III.32. Determine all reduced integral quadratic forms of rank 3 with determinant ≤ 5. Which of them are equivalent?

Exercise III.33. For $A \in K^{n\times m}$ and $B \in K^{s\times t}$ we define the *Kronecker product*[10]

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nm}B \end{pmatrix} \in K^{ns\times mt}$$

as a block matrix. Show for matrices A, B, C, D with suitable format and $\lambda \in K$:

(a) $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.

(b) $(A \otimes B)^{\mathrm{t}} = A^{\mathrm{t}} \otimes B^{\mathrm{t}}$.

(c) $A \otimes (B + C) = A \otimes B + A \otimes C$.

(d) $(A + B) \otimes C = A \otimes C + B \otimes C$.

(e) $\lambda(A \otimes B) = (\lambda A) \otimes B = A \otimes (\lambda B)$.

(f) $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$.

(g) $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A)\operatorname{tr}(B)$.

(h) $\operatorname{rk}(A \otimes B) = \operatorname{rk}(A)\operatorname{rk}(B)$.

(i) $\det(A \otimes B) = \det(A)^m \det(B)^n$ if $A \in K^{n\times n}$ and $B \in K^{m\times m}$.

Exercise III.34. Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$. Show:

(a) If A and B are orthogonal, then so is $A \otimes B$.

(b) If A and B are positive definite, then so is $A \otimes B$.

(c) If A and B are stochastic, then so is $A \otimes B$.

Exercise III.35. Let $A, B \in \mathbb{R}^{n\times n}$ be positive definite. Show $\min q_{A\otimes B} \leq (\min q_A)(\min q_B)$.
Remark: KITAOKA has proven that for quadratic forms with rank ≤ 43 equality holds. For rank 292 there exist examples for the strict inequality.

[10] also called *tensor product*

Appendix

Convex Optimization

Remark A.68. In this section, we investigate optimization problems with a convex (or concave) objective function under linear constraints. Since the simplex algorithm does not work in this concept, there is no reason to bring the constraints into standard form by introducing slack variables. We therefore consider convex sets of the form

$$M := \{x \in \mathbb{R}^m : Ax \leq b\}$$

with $A \in \mathbb{R}^{n\times m}$, $\operatorname{rk}(A) = m < n$ and $b \in \mathbb{R}^n$ (Remark 19.3). As usual, $x \in M$ is called a *vertex*, if $M \setminus \{x\}$ is convex. For $I \subseteq \{1, \ldots, n\}$ let $A_I := (a_{ij} : i \in I, j = 1, \ldots, m)$. We call I a *basis set*, if A_I is invertible. If applicable, $|I| = m$.

Theorem A.69. *Let $M := \{x \in \mathbb{R}^m : Ax \leq b\}$ with $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^m$. $x \in M$ is a vertex if and only if there exists a basis set I with $A_I x = b_I$.*

Proof. Let I be a basis set with $A_I x = b_I$. Let $y, z \in M$ and $0 \leq \lambda \leq 1$ with $x = \lambda y + (1-\lambda)z$. From

$$b_I = \lambda A_I y + (1-\lambda)A_I z \leq \lambda b_I + (1-\lambda)b_I = b_I$$

it follows that $A_I y = b_I$ and $A_I z = b_I$. Since A_I is invertible, $y = x = z$ holds. Thus x is a vertex.

Conversely, let x be a vertex. Let I be the set of indices i with $\sum_{j=1}^m a_{ij}x_j = b_i$. Assume $\operatorname{rk}(A_I) < m$. Then there exists a $y \in \mathbb{R}^m \setminus \{0\}$ with $A_I y = 0$. For $i \notin I$, $\sum_{j=1}^m a_{ij}x_j < b_i$ holds. Therefore, there exists an $\epsilon > 0$ with $x \pm \epsilon y \in M$. Because of $x = \frac{1}{2}(x + \epsilon y) + \frac{1}{2}(x - \epsilon y)$, x cannot be a vertex. Thus $\operatorname{rk}(A_I) = m$. Therefore, there exists a basis set $J \subseteq I$ with $A_J x = b_J$. □

Definition A.70. Let $M \subseteq \mathbb{R}^n$ convex. A function $f\colon M \to \mathbb{R}$ is called

- *convex*, if
$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda)f(y)$$
holds for all $x, y \in M$ and $0 < \lambda < 1$. If the strict inequality holds in the case $x \neq y$, then f is called *strictly* convex.
- (*strictly*) *concave*, if $-f$ is (strictly) convex, i. e.
$$f\big(\lambda x + (1-\lambda)y\big) \geq \lambda f(x) + (1-\lambda)f(y).$$

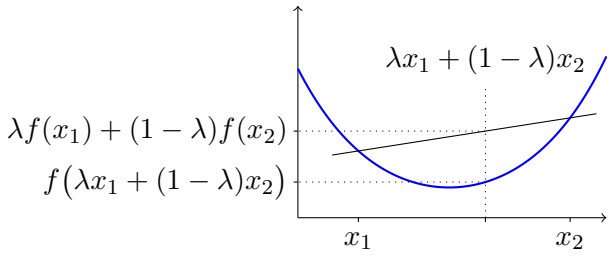
Remark A.71.

(a) Every linear function is convex and concave.

(b) If $f\colon M \to \mathbb{R}$ is convex, then $N := \{(x, y) \in M \times \mathbb{R} : f(x) \leq y\} \subseteq \mathbb{R}^{n+1}$ is convex, because for $f(x_i) \leq y_i$ and $0 \leq \lambda \leq 1$ it holds that

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \leq \lambda y_1 + (1-\lambda)y_2.$$

(c) In analysis, it is shown that f is (strictly) convex if and only if the Hessian matrix of the second derivatives $(\frac{\partial^2 f}{\partial x_i \partial x_j})$ is positive semidefinite (definite). For $n = 1$, this means $f''(x) \geq 0$ for all $x \in M$, i.e., the slope of the graph increases continuously.



Theorem A.72. *Let $M = \{x \in \mathbb{R}^m : Ax \leq b\} \neq \varnothing$ be convex and $f\colon M \to \mathbb{R}$ be convex (resp. concave). If f has a maximum (resp. minimum) on M, then it is attained at a vertex.*

Proof. We argue as in the proof of Theorem 19.24. Let $x \in M$ be a maximum of f. Let $I(x) = I$ be the set of indices i with $\sum_{j=1}^m a_{ij}x_j = b_i$. Assume $\operatorname{rk}(A_I) < m$. As in the proof of Theorem A.69, there exist $y \in \mathbb{R}^m$ and $\epsilon > 0$ with $x_\pm := x \pm \epsilon y \in M$. Since f is convex, it holds that

$$f(x) = f\Big(\frac{1}{2}x_+ + \frac{1}{2}x_-\Big) \leq \frac{1}{2}f(x_+) + \frac{1}{2}f(x_-) \leq \max\big\{f(x_+), f(x_-)\big\} \leq f(x)$$

and $f(x_+) = f(x) = f(x_-)$. Since we always assume that A has full rank, $Ay \neq 0$ holds. We can therefore choose ϵ such that $I(x_+) \supsetneq I$ or $I(x_-) \supsetneq I$ holds. Subsequently, we replace x by x_+ (resp. x_-) and repeat the argument. After finitely many steps, one reaches $\operatorname{rk}(A_I) = m$. Then x is a vertex. The proof for concave functions proceeds analogously. □

Remark A.73. If f is *strictly* convex (or concave), then the maximum (or *minimum*) can *only* be attained at vertices (for this, Theorem A.69 is not needed).

Theorem A.74 (Inequality of the harmonic, geometric and arithmetic mean)**.** *For all positive $x_1, \ldots, x_n \in \mathbb{R}$, it holds that*

$$\frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}} \leq \sqrt[n]{x_1 \ldots x_n} \leq \frac{x_1 + \ldots + x_n}{n}$$

with equality if and only if $x_1 = \ldots = x_n$.

Proof (Cauchy). The first inequality follows from the second by replacing x_i with $1/x_i$. We show the second inequality by an unusual induction on n. For $n = 1$, equality holds. For $n = 2$, we have

$$\frac{(x_1 + x_2)^2}{4} - x_1x_2 = \frac{(x_1 - x_2)^2}{4} \geq 0$$

with equality if and only if $x_1 = x_2$. Now let the statement be satisfied for n. Then

$$x_1 \ldots x_{2n} \leq \Big(\sum_{i=1}^n \frac{x_i}{n} \sum_{i=n+1}^{2n} \frac{x_i}{n}\Big)^n \leq \Big(\sum_{i=1}^{2n} \frac{x_i}{2n}\Big)^{2n}$$

with equality if and only if $x_1 = \ldots = x_{2n}$. Thus the claim holds for $2n$. Now let $A := \sum_{i=1}^{n-1} \frac{x_i}{n-1}$. Then

$$x_1 \ldots x_{n-1} A \leq \Big(\sum_{i=1}^{n-1} \frac{x_i}{n} + \frac{A}{n}\Big)^n = \Big(\frac{A(n-1)}{n} + \frac{A}{n}\Big)^n = A^n$$

and $x_1 \ldots x_{n-1} \leq A^{n-1}$ with equality if and only if $x_1 = \ldots = x_{n-1}$. Thus the claim also holds for $n-1$. □

Theorem A.75. *For every positive semidefinite matrix $A \in \mathbb{C}^{n\times n}$, it holds that $\sqrt[n]{\det(A)} \leq \operatorname{tr}(A)/n$ with equality if and only if A is a scalar matrix.*

Proof. According to Exercise II.18, A has non-negative eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. According to Remark 10.35, it holds that

$$\sqrt[n]{\det(A)} = \sqrt[n]{\lambda_1 \ldots \lambda_n} \leq \frac{1}{n}(\lambda_1 + \ldots + \lambda_n) = \frac{1}{n}\operatorname{tr}(A)$$

with equality if and only if $\lambda_1 = \ldots = \lambda_n$. Since A is diagonalizable according to the spectral theorem, A must then be a scalar matrix. □

Lemma A.76. *For $x_1, \ldots, x_n, y_1, \ldots, y_n \geq 0$, it holds that*

$$\sqrt[n]{(x_1 + y_1) \ldots (x_n + y_n)} \geq \sqrt[n]{x_1 \ldots x_n} + \sqrt[n]{y_1 \ldots y_n}$$

with equality if and only if $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are linearly dependent.

Proof. Wlog. let $x_i + y_i > 0$ for $i = 1, \ldots, n$. According to the inequality between the arithmetic and geometric mean, it holds that

$$1 = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{x_i + y_i} + \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i + y_i} \geq \sqrt[n]{\prod_{i=1}^{n} \frac{x_i}{x_i + y_i}} + \sqrt[n]{\prod_{i=1}^{n} \frac{y_i}{x_i + y_i}} = \frac{\sqrt[n]{x_1 \ldots x_n} + \sqrt[n]{y_1 \ldots y_n}}{\sqrt[n]{(x_1 + y_1) \ldots (x_n + y_n)}}$$

with equality if and only if $\frac{x_i}{x_i+y_i} = \frac{x_j}{x_j+y_j}$ and $\frac{y_i}{x_i+y_i} = \frac{y_j}{x_j+y_j}$ for all i, j. This means $x_1 y_i = x_i y_1$ for $i = 1, \ldots, n$. If x or y is the zero vector, then the vectors are linearly dependent. Otherwise, wlog. $x_1 \neq 0$ and $y_i = \frac{y_1}{x_1} x_i$ for $i = 1, \ldots, n$. Thus x and y are linearly dependent. □

Lemma A.77 (Simultaneous Diagonalization)**.** *Let $A \in \mathbb{C}^{n\times n}$ be positive definite and $B \in \mathbb{C}^{n\times n}$ be positive semidefinite. Then there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with $S^* A S = 1_n$ and $S^* B S = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.*

Proof. According to Exercise II.18, there exists a Hermitian root $\sqrt{A}$ of A. Clearly, $C := \sqrt{A}^{-1} B \sqrt{A}^{-1}$ is also Hermitian. By the spectral theorem, there exists a $U \in \mathrm{U}(n, \mathbb{C})$ with $U^* C U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. The claim now holds for $S := \sqrt{A}^{-1} U$. □

Theorem A.78 (Minkowski Inequality)**.** *Let $A \in \mathbb{C}^{n\times n}$ be positive definite and $B \in \mathbb{C}^{n\times n}$ be positive semidefinite. Then*

$$\sqrt[n]{\det(A + B)} \geq \sqrt[n]{\det(A)} + \sqrt[n]{\det(B)}$$

with equality if and only if $B = \lambda A$ for some $\lambda \geq 0$. In particular, $\det(A + B) \geq \det(A) + \det(B)$.

Proof. According to Lemma A.77, there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with $S^*AS = 1_n$ and

$$D := S^*BS = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

By multiplying from the left and right by $\sqrt[n]{\det(S^*)}$ and $\sqrt[n]{\det(S)}$ respectively, we can assume $A = 1_n$ and $B = D$. By assumption, $\lambda_1, \ldots, \lambda_n \geq 0$. From Lemma A.76 it follows that

$$\sqrt[n]{\det(1_n + D)} = \sqrt[n]{(1+\lambda_1)\ldots(1+\lambda_n)} \geq 1 + \sqrt[n]{\lambda_1 \ldots \lambda_n} = \sqrt[n]{\det(1_n)} + \sqrt[n]{\det(D)}$$

with equality if and only if $\lambda := \lambda_1 = \ldots = \lambda_n$. For the original matrix, this means $B = S^{-*}DS^{-1} = \lambda A$.

The second claim follows from

$$\begin{aligned}\det(A+B) &= \sqrt[n]{\det(A+B)}^n \geq \left(\sqrt[n]{\det(A)} + \sqrt[n]{\det(B)}\right)^n \\ &= \sum_{k=1}^{n} \binom{n}{k} \sqrt[n]{\det(A)}^k \sqrt[n]{\det(B)}^{n-k} \geq \det(A) + \det(B).\end{aligned}$$

□

Theorem A.79. *Let $v \in \mathbb{R}^n$. The set $\mathcal{M} \subseteq \mathbb{R}^{n\times n}$ of all positive definite matrices with main diagonal v is convex. The map $\mathcal{M} \to \mathbb{R}$, $A \mapsto \sqrt[n]{\det(A)}$ is strictly concave.*

Proof. Let $A, B \in \mathcal{M}$ and $0 < \lambda < 1$. Obviously $C := \lambda A + (1-\lambda)B$ has main diagonal v. For $x \in \mathbb{R}^n \setminus \{0\}$ we have

$$xCx^{\mathrm{t}} = \lambda xAx^{\mathrm{t}} + (1-\lambda)xBx^{\mathrm{t}} > 0.$$

This shows $C \in \mathcal{M}$. According to the Minkowski inequality, it holds that

$$\sqrt[n]{\det(C)} \geq \sqrt[n]{\det(\lambda A)} + \sqrt[n]{\det((1-\lambda)B)} = \lambda\sqrt[n]{\det(A)} + (1-\lambda)\sqrt[n]{\det(B)}$$

with equality if and only if $\lambda\mu A = (1-\lambda)B$ for some $\mu > 0$. Since A and B have the same main diagonals, it follows that $\lambda\mu = (1-\lambda)$ and $A = B$. Thus $A \mapsto \sqrt[n]{A}$ is strictly concave. □

Theorem A.80 (Oppenheim)**.** *Let q_C be a reduced positive quadratic form of rank 3. Then*

$$c_{11}c_{22}c_{33} \leq c_{11}c_{22}c_{33} + \frac{1}{2}c_{11}c_{22}(c_{33}-c_{22}) + \frac{1}{2}c_{11}c_{33}(c_{22}-c_{11}) \leq 2\det(q).$$

If $c_{11}c_{22}c_{33} = 2\det(q)$, then q is equivalent to the quadratic form with Gram matrix

$$\frac{c_{11}}{2}\begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{pmatrix}.$$

Proof (Barnes). Let $\mathcal{M} \subseteq \mathbb{R}^{3\times 3}$ be the set of reduced positive definite matrices with main diagonal (a, b, c), where $a \leq b \leq c$. For

$$C := \begin{pmatrix}a & x & z\\ x & b & y\\ z & y & c\end{pmatrix} \in \mathcal{M}$$

it holds that $0 \le x \le a/2$, $0 \le y \le b/2$, $z \le a/2$, $-z \le a/2$ and $x + y - z \le (a+b)/2$ according to Exercise III.31. One can therefore identify $\mathcal{M}$ with the convex set $M \subseteq \mathbb{R}^3$ of all (x, y, z) with

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \le \frac{1}{2} \begin{pmatrix} a \\ 0 \\ b \\ 0 \\ a \\ a \\ a+b \end{pmatrix} =: v.$$

According to Sarrus' rule, it holds that

$$\det(C) = abc + 2xyz - ay^2 - bz^2 - cx^2.$$

According to Theorem A.79 and Remark A.73, the minimum of $C \mapsto \sqrt[3]{\det(C)}$ (and thus also the minimum of det) on M is attained only at vertices. According to Theorem A.69, the vertices are obtained by choosing three linearly independent rows of A and solving the corresponding system of equations. There are eight possibilities to choose three linearly independent rows from the first six rows of A. However, the choice $(x, y, z) = (a, b, -a)$ does not satisfy the last inequality. If one chooses the seventh row as equality, then $z = \frac{1}{2}(a+b) - x - y$. This yields two further vertices:

I	$2(x,y,z)$	$\det(C)$	I	$2(x,y,z)$	$\det(C)$
$\{1,3,5\}$	(a,b,a)	$abc - \frac{1}{4}(ab^2 + a^2c)$	$\{1,3,7\}$	$(a,b,0)$	$abc - \frac{1}{4}(ab^2 + a^2c)$
$\{1,4,5\}$	$(a,0,a)$	$abc - \frac{1}{4}(ab^2 + a^2c)$	$\{1,4,7\}$	$(a,0,-b)$	as $I = \{1,4,6\}$ with $a = b$
$\{1,4,6\}$	$(a,0,-a)$	$abc - \frac{1}{4}(ab^2 + a^2c)$	$\{1,6,7\}$	$(a,b-a,-a)$	$abc - \frac{1}{4}(ab^2 + a^2c)$
$\{2,3,5\}$	$(0,b,a)$	$abc - \frac{1}{4}(ab^2 + a^2b)$	$\{2,3,7\}$	$(0,b,-a)$	as $I = \{2,3,6\}$
$\{2,3,6\}$	$(0,b,-a)$	$abc - \frac{1}{4}(ab^2 + a^2b)$	$\{2,6,7\}$	$(0,b,-a)$	
$\{2,4,5\}$	$(0,0,a)$	$abc - \frac{1}{4}a^2b$	$\{3,6,7\}$	$(0,b,-a)$	
$\{2,4,6\}$	$(0,0,-a)$	$abc - \frac{1}{4}a^2b$	$\{4,6,7\}$	$(b,0,-a)$	as $I = \{1,4,6\}$ with $a = b$

Apparently $\det(C) \ge abc - \frac{1}{4}(ab^2 + a^2c)$ and

$$abc \le abc + \frac{1}{2}ab(c-b) + \frac{1}{2}ac(b-a) \le 2\det(C).$$

Equality holds only if $a = b = c$ and $2(x, y, z) \in \{(a,a,a), (a,a,0), (a,0,\pm a), (0,a,\pm a)\}$. According to Example 20.52, $(a,a,a) \sim (a,a,0)$ holds. By permutation of x, y, z and sign changes, one sees $(a,a,0) \sim (a,0,\pm a) \sim (0,a,\pm a)$. □

Theorem A.81 (Gauss)**.** *For every lattice $\Delta \subseteq \mathbb{R}^3$, it holds that $\rho(\Delta) \le \frac{\pi}{3\sqrt{2}}$. If equality holds, then the Gram matrix of Δ has the form*

$$c \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

for a suitable choice of basis.

Proof. Let C be the Gram matrix of Δ. By a change of basis, we can assume that q_C is reduced. As in Remark 20.11, we arrange unit spheres with centers in Δ. Since the spheres do not overlap, $c_{11} = \min(\Delta)^2 \ge 4$ holds. From Oppenheim, it follows that $\det(C) \ge 2^5$. This shows

$$\rho(\Delta) = \frac{4\pi}{3\,\mathrm{disc}(\Delta)} = \frac{4\pi}{3\sqrt{\det(C)}} \le \frac{\pi}{3\sqrt{2}}.$$

If equality holds, then according to Oppenheim, one can construct a basis of Δ such that C has the specified form. □

Theorem A.82 (FISHER inequality)**.** *Let* $M = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right) \in \mathbb{C}^{n\times n}$ *be positive definite. Then* $\det(M) \leq \det(A)\det(C)$ *holds with equality if and only if* $B = 0$*.*

Proof. According to the Sylvester criterion, A, C as well as A^{-1}, C^{-1} are positive definite. With Exercise II.18, it holds that

$$D := \operatorname{diag}(\sqrt{A}, \sqrt{C})^{-1} M \operatorname{diag}(\sqrt{A}, \sqrt{C})^{-1} = \begin{pmatrix} 1 & \sqrt{A}^{-1} B \sqrt{C}^{-1} \\ \sqrt{C}^{-1} B^* \sqrt{A}^{-1} & 1 \end{pmatrix}.$$

From Theorem A.75, it follows that

$$\det(A)^{-1} \det(C)^{-1} \det(M) = \det(D) \leq (\operatorname{tr}(D)/n)^n = 1,$$

thus $\det(M) \leq \det(A)\det(C)$. Equality holds if and only if D is a scalar matrix, i.e. $\sqrt{A}^{-1} B \sqrt{C}^{-1} = 0 = B$. □

Corollary A.83. *Let* $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ *be positive definite and* $A^{-1} = (b_{ij})$*. Then* $a_{ii} b_{ii} \geq 1$ *for* $i = 1, \ldots, n$*.*

Proof. According to the formula for the complementary matrix $\tilde{A}$, $b_{ii} = \det(A_{ii})/\det(A)$. By swapping the first and i-th row/column, A is similar to $\left(\begin{smallmatrix} a_{ii} & B \\ B^* & A_{ii} \end{smallmatrix}\right)$. From Fisher, it follows that $\det(A) \leq a_{ii} \det(A_{ii}) = a_{ii} b_{ii} \det(A)$. □

Index

A

absolute value, 99
    complex number, 108
adjugate, 73
affine hull, 247
affine space, 28
affinely (in)dependent, 247
AlphaEvolve, 175
associative law, 21, 23
axiom, 13
axiom of choice, 13, 17

B

backward substitution, 185
Banach space, 186
Banach's Fixed Point Theorem, 186
Banach-Tarski paradox, 13
Barnes, 254
basic transposition, 83
basis
    dual, 59
    of a lattice, 223
    of a vector space, 30
basis change matrix, 54
Basis Extension Theorem, 32
basis set, 218, 251
    feasible, 218
Bauer-Fike, 192
Bernoulli inequality, 246
bidual space, 59
bijection, 19
bilinear form, 113
    alternating, 113
    antisymmetric, 113
    degenerate, 113
    indefinite, 120
    index, 119
    negative (semi)definite, 120
    positive (semi)definite, 120
    symmetric, 113
binary exponentiation, 245
Binet formula, 94
Blichfeldt, 234
Bruhat decomposition, 245
bubble sort, 83
Bézout, 147

C

cancellation, 180
Cantor, 13, 21
Carathéodory, 215
Cartan-Dieudonné, 112
cartesian product, 17
Cauchy, 252
Cauchy's interlacing theorem, 200
Cauchy-Binet formula, 72
Cauchy-Schwarz inequality, 100, 125
Cayley-Hamilton, 96
centralizer, *154*, 170
Ceres, 184
Chan-Li, 132
Change of basis, 57
characteristic polynomial
    of a map, 92
    of a matrix, 92
Chinese Remainder Theorem, 159
Cholesky decomposition, 184
Cholesky method, 195
circular reasoning, 31
codomain, 19
coefficient, 86
    leading, 86
coefficient matrix, 41
    augmented, 41
Collatz-Wielandt, 209
column operation, 43
column vector, 35
commutative law, 23
companion matrix, 150
complement, 32
    dual, 60
    orthogonal, 103, 116
complex conjugation, 108
composition, 19
condition number, 176, 190
congruence
    of matrices, 116
    of polynomials, 158
constant term, 86
continuum hypothesis, 13, 20
contraction, 186
contraposition, 12
converse, 12
convex combination, 215

convex hull, 215
convolution theorem, 173
Conway, 244
coordinate representation, 31
cosine, 101
Courant-Fischer, 200
Cramer's rule, 74
cross product, 104
cycle, 75
    disjoint, 75

D
De Morgan's law, 12, 15
decimal fractions, 14
Dedekind identity, 81
derivative, 159
determinant
    of a map, 92
    of a matrix, 68
    of a quadratic form, 238
determinant theorem, 71
diagonal matrix, 36
diagonalizability, 64
    simultaneous, 135
        for bilinear forms, 253
diagonalization arguments, 20
difference, 14
differential equation, 206
dimension, 33
dimension formula, 34
direct product, 24
discriminant, 225
distributive law, 12, 15, 24
divisor, 88
    common, 228
        greatest, 228
domain, 19
dual space, 59
Duality Theorem, 217

E
Eichler, 234
eigenspace, 63
eigenvalue, 63
eigenvector, 63
elementary divisor, 230
elementary matrix, 43
endomorphism, 63
equivalence class, 18
equivalence relation, 18
Euclidean algorithm, 146
Euclidean division, 88
Euclidean space, 99
Euler, 111
exchange lemma, 32
exponential function, 22, 203

F
Farkas' lemma, 216
FFT, *see* Fourier transform
Fibonacci numbers, 94
field, 24
    of complex numbers, 107
    of rational functions, 160
    ordered, 108
    with four elements, 157
    with three elements, 81
    with two elements, 24
Fillmore, 58
Fisher inequality, 256
Fitting, 136
floating-point numbers, 180
forward substitution, 184
Fourier matrix, 173
Fourier transform
    continuous, 174
    discrete, 173
    fast, 174
Francis algorithm, 194
Frobenius, 155
Frobenius inequality, 82
Frobenius norm, 189
Frobenius normal form, 151
function, *see* map
functional, 59
functional analysis, 33
functional equation, 204
fundamental mesh, 225
fundamental theorem
    of algebra, 109
    linear programming, 220

G
Gauss, 241, 255
Gauss algorithm, 44
    with pivoting, 181
Gauss-Seidel method, 187
general expansion theorem, 83
generalized eigenspace, 136
Generalized Eigenspace Decomposition, 137
generating set, 30
generator matrix, 223
geometric series, 201
Gershgorin, 196
Givens rotation, 198
Goldbach's conjecture, 11
golden ratio, 94
Golden-Thompson inequality, 205
Google matrix, 211
Gram matrix
    of a bilinear form, 114
    of a lattice, 223
    of a quadratic form, 238

Gram-Schmidt process, 102
    modified, 197
group, 23
    abelian, 23
    affine, 82
    alternating, 77
    general linear, 39
    orthogonal, 104
    special linear, 71
    special orthogonal, 105
    special unitary, 128
    symmetric, 75
    unitary, 127, 128
Gödels incompleteness theorems, 13

H
Hadamard inequality, 246
Hales, 226
Hamiltonian skew field, 170
Harriot, 226
Harvey-van der Hoeven, 174
Hermite, 232
Hermite constant, 234
Hermite normal form, 229
Heron's method, 245
Hessenberg matrix, 195
Hessian matrix, 121
Hilbert matrix, 177
Hölder inequality, 246
homeomorphism, 50
homogeneity, 100, 125, 185
homomorphism, 50
homomorphism theorem, 54
Horner scheme, 88
Householder transformation, 167, 198
hypercube, 67
hyperplane, 33

I
identity, 19
identity element, 23
identity matrix, 35
image, 19
inclusion map, 19
index, 119
Inequality harm., geom., arithm. mean, 252
interpolation, 89
intersection, 14
invariant, 67
inverse element, 23
inverse function, 21
isomorphism, 50
isomorphism theorem
    first, 53
    second, 53

J
Jacob, 151
Jacobi, 204
Jordan block, 138
    generalized, 162
Jordan normal form, 140
Jordan-Chevalley decomposition, 163

K
Karatsuba algorithm, 172
Kepler, 226
kernel, 51
Kitaoka, 249
Korkine-Zolotarev, 241
Kronecker delta, 35
Kronecker product, 249
Kronecker-Capelli, 41

L
Lagrange polynomial, 90
Laplace expansion, 72
large language model, 175
lattice, 223
    basis
        δ-reduced, 236
    (un)decomposable, 234
    dual, 224
    E_8, 227
    integral, 225
    self-dual, 224
Law of cosines, 167
law of excluded middle, 12
law of non-contradiction, 12
Law of sines, 167
Least Squares Method, 183
Leibniz formula, 77
length
    of a cycle, 75
Lights Out, 83
linear (in)dependence, 30
linear combination, 25
    affine, 247
    convex, 215
linear factor, 90
linear program, 214
    dual, 217
    infeasible, 214
    solvable, 214
    unbounded, 214
linear system, 41
    (in)homogeneous, 41
    overdetermined, 42
    solvable, 41
    underdetermined, 42
LLL algorithm, 236
logarithm tables, 173

Lovász condition, 236
LU decomposition, 181
LWE, 225

M

main diagonal, 36
mantissa, 180
map, 19
  adjoint, 127
  affine, 51
  bijective, 19
  concave, 251
    strict, 251
  convex, 251
    strict, 251
  diagonalizable, 64
    simultaneous, 135
  dual, 62
  hermitian, 127
  injective, 19
  linear, 50
  nilpotent, 138
  normal, 127
  orthogonal, 104
  semisimple, 161
  separable, 161
  surjective, 19
  symmetric, 104
  trigonalizable, 134
    simultaneous, 135
  unitary, 127
Markov chain, 206
mathematical induction, 16
matrix, 35
  adjoint, 128
  antisymmetric, 114
  block, 38
  block diagonal, 38
  commuting, 38
  complementary, 73
  congruent, 116
  converges, 191
  diagonal dominant, 196
  diagonalizable, 64
    simultaneously, 168
  equivalent, 48
  hermitian, 128
    positive (semi)definite, 168
  (in)decomposable, 206
  inverse, 38
  invertible, 38
  nilpotent, 138
  non-negative, 206
  normal, 128
  orthogonal, 105
  positive, 206
  quasi-converges, 194
  regular, 38
  row-equivalent, 43
  similar, 57
  singular, 38
  skew-symmetric, 114
  sparse, 72
  square, 35
  stochastic, 207
  symmetric, 36
  transpose, 36
  triangular, 66
    strict, 66
  tridiagonal, 240
  trigonalizable, 130, *134*
    simultaneously, 168
  unitary, 128
  well/ill-conditioned, 176
matrix inversion, 47
matrix norm, 189
  induced, 189
  submultiplicative, 189
Mazur-Ulam, 104
Mercator series, 206
Millennium problems, 12
Min-Max theorem, 200
minimal norm, 225
minimal polynomial, 95
Minkowski, 29, 240, 244
Minkowski inequality, 246, 253
Minkowski space, 119
Minkowski's lattice point theorem, 234
Minkowski's linear forms theorem, 233
Mirsky, 93
mnemonic, 37, 53, 54, 56, 65, 89, 110
module, 223
modus ponens, 12
Moore-Penrose, 179
Mordell, 240
multilinear form, 113
multiplicity
  algebraic, 90
  geometric, 63

N

natural logarithm, 22
Neumann series, 201
Newton's method, 109
norm, 99, 125, *185*
  equivalent, 186
normal equation, 184
normal form
  Frobenius, 151
  Hermite, 229
  Jordan, 140
  Smith, 230

Weierstraß, 151
numbers
cardinal, 20
complex, 107
integers, 14
natural, 14
rational, 14
real, 14

O

one-way function, 22
Oppenheim, 254
order relation, 18
lexicographical, 141
total, 18
orthogonal basis
wrt. a bilinear form, 117
orthogonal decomposition, 234
orthonormal basis, 102
wrt. a bilinear form, 117

P

Page rank, 211
paradox, 11
parallelogram law, 100
part
imaginary, 107
real, 107
partition, 14, 138
permutation, 75
permutation matrix, 76
Perron, 207
Perron-Frobenius, 208
Piazzi, 184
pivot, 181
Polar decomposition, 245
polar representation, 183
polarization, 114
polyhedron, 218
polynomial, 86
companion matrix, 150
congruent, 158
constant, 86
coprime, 146
derivative, 159
irreducible, 146
monic, 86
separable, 159
zero, 86
polytope, 218
Power Method, 193
power set, 15
predicate, 11
preimage, 19
Primary decomposition, 148
Prime factorization in $K[X]$, 147
Principal Axis Theorem, 110
principal minor, 123
Product rule, 160
projection, 51
proposition, 11
equivalent, 11
pseudoinverse, 179
Pythagoras, 100
trigonometric, 167

Q

QR algorithm, 194
QR decomposition, 182
quadratic form, 114, *238*
binary, 240
equivalent, 238
indefinite, 248
integral, 238
minimum, 238
non-degenerate, 238
positive, 238
reduced, 240
unimodular, 240
universal, 244
quantum mechanics, 206
quotient space, 28

R

Rado, 233
rank
of a lattice, 223
of a map, 51
of a matrix, 39
Rayleigh quotient, 200
reflection
in $\mathbb{R}^2$, 106
in $\mathbb{R}^n$, 112
relation, 17
antisymmetric, 18
asymmetric, 18
reflexive, 17
symmetric, 17
transitive, 18
trivial, 18
remainder, 88
representation matrix, 54
restriction, 19
right-hand rule, 104
ring, 37
root, 89
double, 91
of unity, 108
simple/multiple, 90
rotary reflections, 111
rotation, 106
row echelon form, 44

row operation, 43
row vector, 35
RSA, 225
Ruffini's rule, 88
Russell's antinomy, 13

S

Sarrus rule, 78
SAT problem, 12
scalar, 25
scalar matrix, 36
scalar product, 99, 125
Schatzman, 195
Schur decomposition, 130
    real, 168
Schur's Lemma, 156
Schur-Horn, 132
Schönhage-Strassen algorithm, 174
selection sort, 76
sesquilinear form, 125
set
    convex, 215
    countable, 20
    disjoint, 14
    empty, 13
    equinumerous, 19
    (in)finite, 13
    uncountable, 20
sign, 76
signum, 76, 83
Simplex algorithm, 221
Simplex criterion, 220
sine, 101
Singular Value Decomposition, 177
slack variable, 214
Smith normal form, 230
solution set, 41
span, 29
spectral radius, 201
Spectral Theorem, 129
spectrum, 129
splitting field, 157
standard basis, 30
standard inner product, 99
standard matrix, 36
standard scalar product, 125
standard simplex, 215
statement, 11
Steinitz, 32
Stirling number, 246
Strassen algorithm, 175
subgroup, 26
    proper, 26
subset, 14
    proper, 14
subspace, 26
    cyclic, 149
    F-invariant, 170
    f-invariant, 134
    proper, 27
successive minima, 243
sum of subspaces, 29, 64
    direct, 29
Sum rule, 160
support, 218
Sylvester criterion, 122
Sylvester inequality, 82
Sylvester's determinant formula, 166
Sylvester's Law of Inertia
    for hermitian matrices, 131
    for symmetric bilinear forms, 118
symplectic space, 120
system of representatives, 18

T

Taussky, 169
tensor product, 249
trace
    of a map, 58
    of a matrix, 58
transition matrix, 206
transitivity, 12
translation, 82
transposition, 75
triangle inequality, 100, 126, 185
trigonometric identities, 106
tripel, 17
tupel, 17

U

union, 13
    disjoint, 14
unitary space, 125

V

Vandermonde matrix, 73
variable, 86
vector, 25
    normalized, 99, 125
    orthogonal, 99, 125
    shortest, 225
vector space, 25
    euclidean, 99
    finite-dimensional, 33
    finitely generated, 30
    isomorphic, 50
    unitary, 125
Venn diagram, 14
vertex, 218, 251
Viazovska, 226
von Mises, 209

W
Weierstraß normal form, 151
Wielandt, 210
Wilkinson, 194

Y
Young inequality, 246

Z
Zassenhaus algorithm, 48
Zermelo-Fraenkel system, 13
zero matrix, 35
zero space, 25
zero vector, 25
Zorn's Lemma, 33