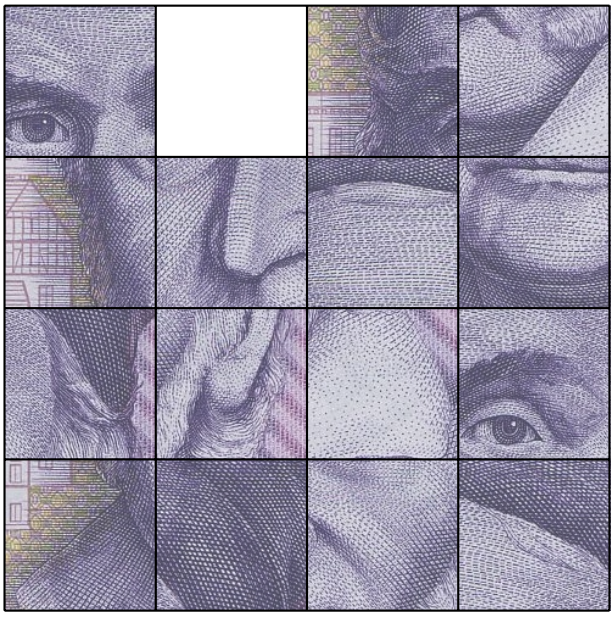
Linear Algebra

Benjamin Sambale
Leibniz University Hannover

Version: November 15, 2026



Contents

Introduction **5**
Preface 5
Notation 6
Conventions 9

Linear Algebra I **10**

1 Propositional Logic and Set Theory **11**
1.1 Statements 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 Spanning Sets 29
4.2 Characterization and Existence of Bases 30
4.3 Dimension 33

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

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

7 Linear Maps **50**
7.1 Homomorphisms and Isomorphisms 50
7.2 Matrices of Linear Maps 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 **85**

10 Polynomials **86**
10.1 The Vector Space of Polynomials . . . 86
10.2 Roots . . . 89
10.3 Characteristic Polynomials . . . 92
10.4 Minimal Polynomials . . . 95

11 Euclidean Geometry **99**
11.1 Inner 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 . . . 127
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 . . . 160
16.3 Generalized Jordan Blocks . . . 162

17 Kronecker Products **166**
17.1 Definition and Properties . . . 166
17.2 Linear Matrix Equations . . . 170
17.3 The Hadamard Product . . . 172

Exercises **174**

Linear Algebra III **180**

18 Numerical Methods **181**
18.1 Efficient Arithmetic . . . 181
18.2 The Condition Number . . . 185
18.3 Stable Variants of the Gaussian Algorithm . . . 189
18.4 Iterative Methods . . . 194
18.5 Matrix Norms . . . 198
18.6 Computing Eigenvalues . . . 201
18.7 Orthonormalization . . . 207

19 Analytic Aspects **210**
19.1 Eigenvalue Estimates . . . 210
19.2 The Spectral Radius . . . 211
19.3 The Exponential Function of a Matrix . . . 213
19.4 Nonnegative Matrices . . . 216
19.5 PageRank . . . 221

20 Linear Optimization **223**
20.1 Linear Programs . . . 223
20.2 Convex Sets . . . 224
20.3 The Simplex Algorithm . . . 228

21 Lattices and Quadratic Forms **233**
21.1 Lattices . . . 233
21.2 The Minimal Norm . . . 235
21.3 Integer Matrices . . . 238
21.4 Theorems of Hermite and Minkowski . . . 242
21.5 The LLL Algorithm . . . 244
21.6 Quadratic Forms . . . 248
21.7 Successive Minima . . . 253

Exercises **255**

Appendix **260**

Convex Optimization **261**

Index **267**

Introduction

Disclaimer. This English version is an automatic translation of the German original, produced by the AI models Claude Opus 5.5 and Claude Sonnet 5.5 (Anthropic). It may contain errors. In case of doubt, the German original on the author's homepage is authoritative.

Preface

This book discusses the contents of linear algebra that I consider most important, in as concise a form as possible. In contrast to many books for beginning students, it dispenses with lengthy explanations and motivating examples from other fields. In the following, I describe the contents and distinguishing features.

The first part begins with an introduction to elementary propositional logic and ends in the usual way with the Leibniz formula for determinants. The signum of a permutation is defined via the determinant of the permutation matrix. The little-known Theorem 7.36 of Fillmore shows that within the similarity class of a nonscalar matrix every diagonal compatible with the trace is realized. Moreover, the Cauchy-Binet formula and the uniqueness of the reduced row echelon form are proved (useful for grading exams). The Zassenhaus algorithm allows the simultaneous computation of the intersection and the sum of subspaces.

In the second part, we first introduce polynomials as a tool. Here we present Theorem 10.36 of Mirsky, which guarantees the existence of matrices with prescribed eigenvalues and main diagonal entries. Next, we deal with Euclidean geometry and unitary spaces. Besides the important spectral theorem for normal matrices, we also obtain the Theorem 13.29 of Schur-Horn, which, like Mirsky's theorem, solves an inverse problem. The Theorem 11.19 of Mazur-Ulam, rarely treated in this context, states that distance-preserving maps are affine. Then the Jordan normal form is constructed and used for various applications. Unlike other books, I additionally treat the Frobenius and Weierstrass normal forms in full generality. The proof avoids the theory of modules and instead uses the dual space, following a paper of Jacob. The companion matrices from the Frobenius normal form give rise to introducing splitting fields as centralizers of matrices. The highlight is a dimension formula of Frobenius for the centralizer of a matrix (Theorem 15.32). Splitting fields also play a role in the Jordan-Chevalley decomposition for separable matrices. The second part closes with an excursion on Kronecker products. Here we treat the theorems of Roth on the solvability of the matrix equation $AX - YB = C$.

The third and last part is devoted to practical aspects of linear algebra. It begins with numerical methods such as the fast Fourier transform and the Strassen algorithm. Afterwards, we turn to stable matrix algorithms such as the singular value, LU, QR and Cholesky decomposition. Among the iterative methods, the Francis algorithm and the Cholesky algorithm for computing eigenvalues deserve emphasis. Both proofs are hard to find in the literature in such brevity. Next, we study analytic topics such as estimates of the spectral radius, nonnegative matrices and the exponential function. Here a version of the principal branch of the matrix logarithm is discussed (Theorem 19.19). Another chapter describes the simplex algorithm from linear optimization. In the last chapter of the book, the theory of lattices and quadratic forms is presented in detail. Worth mentioning are the Smith normal form for integer

matrices, Minkowski's linear forms theorem and the LLL algorithm. For quadratic forms, Minkowski's reduction theory is at the center.

All three parts end with theory-oriented exercises, which should be supplemented by practical exercises in self-study. The appendix contains the somewhat technical solution of the Kepler problem for lattices in $\mathbb{R}^3$ due to Gauss and Barnes. Apart from lattices, the general theory of modules over arbitrary rings is not taken up in this book. Theorems that are interesting in context but cannot be proved within the scope of this book are stated as such in remarks.

The first version of the lecture notes was written in 2020/21 in the context of the lectures Linear Algebra A & B at Leibniz University Hannover. The German original is available on my homepage. This English translation was produced by the AI models Claude Opus 5.5 and Claude Sonnet 5.5 (Anthropic). I thank Annika Bartelt, Alexander Ivanov, Gereon Koßmann, Paulina Ligocka, Christoph Pegel, Dawid Sokolihs, Marcos Soriano Sola and Claude Opus for pointing out errors (further comments are welcome).

Hannover, November 15, 2026 Benjamin Sambale

Notation

cf.	compare
e. g.	for example
i. e.,	that is
resp.	respectively
w. r. t.	with respect to
w. l. o. g.	without loss of generality
$\mathbf{t}, \mathbf{f}$	true, false
$\wedge, \vee, \neg, \Rightarrow, \Leftrightarrow, \exists, \forall$	logical symbols
$:=$, $:\Leftrightarrow$	left-hand side is defined by right-hand 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
$\lvert A\rvert$	size 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
$\operatorname{Map}(A, B)$	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, nonnegative) and complex numbers
$\overline{\mathbb{Q}}$	algebraic 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 of f
$f(A), f^{-1}(B), \operatorname{Ker} f$	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), quotient 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	annihilators
$V \oplus W$, $V \times W$	direct sum/product of V and W
$V \otimes W$	tensor product of V 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 vector of v with respect to 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}	matrix unit 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}	$= (a_{ij} : i \neq s, j \neq t)$ (deleting a row/column)
A^{-1}, A^+	inverse and pseudoinverse of A
A^{t}, $A^{-\mathrm{t}}$	transpose and inverse transpose of A
$\widehat{A}$, $\widetilde{A}$	reduced row echelon form, adjugate of A
$\overline{A}$, A^*	complex conjugate, adjoint matrix of A
$\mathrm{rk}\, A$, $\mathrm{tr}\, A$, $\det A$	rank, trace, determinant of A
$\mathrm{vol}\, A$	volume (Jordan measure) of $A \subseteq \mathbb{R}^n$
${}_C[f]_B$, $[f]$, ${}_C\Delta_B$	matrix of a linear map, change-of-basis 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	signum, 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 with respect to 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 semicircular arc of radius 1
$\cos\varphi$, $\sin\varphi$	cosine, sine of φ
$S^\perp$	orthogonal complement of $S \subseteq V$
$v_1 \times \ldots \times v_k$	vector product/cross product
$D(\varphi)$, $S(\varphi)$	rotation, reflection in $\mathbb{R}^2$
$D_{st}(\phi)$	Givens rotation
S_v	reflection in the hyperplane $v^\perp$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$, $\overline{z}$	real part, imaginary part, complex conjugate 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 $f \in \mathrm{End}(V)$, $A \in K^{n\times n}$
$A \otimes B$	Kronecker product of A, B
$A \oplus B$	Kronecker sum of A, B
$A \odot B$	Hadamard product of A, B
$\mathrm{vec}(A)$	vectorization of A
W_n	nth Fourier matrix
H_n	nth 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
$\lVert v\rVert_\infty$, $\lVert A\rVert_\infty$	maximum norm, row sum norm
$\lambda_k(A)$	kth largest eigenvalue of $A = A^*$
$\rho(A)$	spectral radius of A
$\exp A$	exponential function of A
A_+	nonnegative matrix $(\lvert a_{ij}\rvert)_{ij}$
L^*	linear program dual to L
$\mathrm{con}\Delta$	convex hull of $\Delta \subseteq \mathbb{R}^n$
$\mathrm{supp}x$	support of $x \in \mathbb{R}^n$
$\lfloor x\rfloor$, $\lfloor x\rceil$	rounded down and rounded value of $x \in \mathbb{R}$
$\mathrm{disc}\,\Delta$	discriminant of the lattice Δ
Δ^*	lattice dual to Δ
E_8	integer lattice in $\mathbb{R}^8$ with discriminant 1
$\Delta \perp \Lambda$	orthogonal decomposition of lattices
γ_n	nth 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 CAPITALS at 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 Latin capital 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 one often uses "calligraphic" letters ($\mathcal{M}$, $\mathcal{P}$).
- For polynomials, scalars (field elements in the context of vector spaces) and bilinear forms we use Greek letters. The most common ones are:

α	β	γ, Γ	δ, Δ	ϵ, ε	ζ	η	$\theta, \vartheta, \Theta$	κ	λ, Λ	μ
alpha	beta	gamma	delta	epsilon	zeta	eta	theta	kappa	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 Statements

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

Definition 1.2.

- A *statement* A is an English sentence that takes either the *truth value true* (**t**) or *false* (**f**). One then says: A *holds* or A *does not hold*.
- For statements A and B, the following are also statements: $\neg A$ (*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), with the following truth values:

A	B	$\neg A$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftrightarrow B$
t	**t**	**f**	**t**	**t**	**t**	**t**
t	**f**	**f**	**f**	**t**	**f**	**f**
f	**t**	**t**	**f**	**t**	**t**	**f**
f	**f**	**t**	**f**	**f**	**t**	**t**

- Two statements 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 becomes a statement only after substituting an object for the variable x. In this case, $\forall x : A(x)$ (*for all* x, $A(x)$ holds) and $\exists x : A(x)$ (*there exists* an x such that $A(x)$ holds) are statements.

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

- All green cats can fly (**t**).
- $1 + 1 = 3$ (**f**).
- Every even number greater than 2 is the sum of two primes (?).[1]

By contrast, the following are not statements:

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

[1] Goldbach's conjecture

From the predicate $x > 42$ one can form the true statement $\exists x : x > 42$.

Remark 1.4.

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

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

(c) For statements $A_1, \ldots, A_n$ 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 statement A, one performs *equivalence transformations*, i. e., one replaces A by an equivalent statement. The following rules of inference are useful for this.[4]

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

(a) The following statements 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, and so are $A \vee B$ and $B \vee A$ (commutative law).

(c) We have $A \vee \neg A$ (law of excluded middle) *and $\neg(A \wedge \neg A)$* (law of noncontradiction).

(d) $A \wedge (B \vee C)$ and $(A \wedge B) \vee (A \wedge C)$ are equivalent, and so 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, and so 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) $A \wedge (A \Rightarrow B)$ implies B (modus ponens).

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

Proof. All claims can easily be verified by truth tables. For three variables one has to distinguish $2^3 = 8$ cases.[5] Alternatively, one can derive some of the claims from those already proved. For instance, the second De Morgan 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 of little significance on its own.

[5] In 2026 it was shown with the help of AI that one can check in subexponential time whether a propositional formula is always true. The so-called SAT *problem* is nevertheless NP-complete and therefore relevant for the *Millennium Prize Problem* 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 form all further expressions 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 into a whole of definite, well-distinguished objects x of our intuition or our thought.[7] One then says: x is an *element* of M and writes $x \in M$ and $M = \{x : x \in M\}$ (or $x \notin M$ for $\neg(x \in M)$). The number $|M|$ of elements of M is called the *cardinality* or *size* of M. If $|M| < \infty$, then M is called *finite* and otherwise *infinite*.

Remark 1.8.

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

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

Then the statement $M \in M$ can be neither true nor false. In modern mathematics one prevents such contradictions by introducing a *system of axioms*, i. e., one prescribes the truth value of as few "elementary" statements (axioms) as possible. Widely used is the Zermelo-Fraenkel system. 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. So we have $\{2, 1, 1, 2, 2\} = \{1, 2\}$.

(b) In some situations one additionally needs the so-called *axiom of choice* (see Example 2.3). It is accepted by most mathematicians, although it allows the construction of counterintuitive sets: For example, the Banach-Tarski *paradox* states that a ball of volume 1 can be decomposed into five pieces which, when put together differently, yield two balls of volume 1.

(c) By 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 statements whose truth value cannot be determined.[10] The best-known example is the *continuum hypothesis* (see Remark 2.11(f)).

Definition 1.9.

(a) For sets A and B let

$$\begin{aligned} \varnothing &:= \{\} && (\textit{empty set}),\\ A \cup B &:= \{x : x \in A \vee x \in B\} && (\textit{union}),\\ A \cap B &:= \{x : x \in A \wedge x \in B\} && (\textit{intersection}),^{11} \end{aligned}$$

[7] Cantor's wording
[8] The symbol := means that the left-hand side is defined by the right-hand side.
[9] See lecture notes on logic and set theory
[10] The idea of the proof is to formalize the statement "This sentence is not provable."

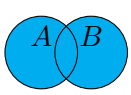
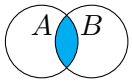
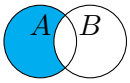
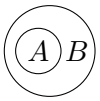
$$A \setminus B := \{x : x \in A \wedge x \notin B\} \qquad (\textit{set difference}).^{12}$$

(b) If $A \cup B = B$, then 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 $A \nsubseteq B$.

(c) A and B are called *disjoint* if $A \cap B = \varnothing$. In this case, one calls $A \dot\cup B := A \cup B$ a *disjoint union*.

Remark 1.10.

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

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

Caution: 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 (nonempty) 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 statement $(A \subseteq B) \wedge (B \subseteq A)$.

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] Caution: For some authors, $0 \in \mathbb{N}$.

(b) The set of *integers* $\mathbb{Z} := \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. We have $\mathbb{N} = \{n \in \mathbb{Z} : n > 0\}$. The integers of the form $2n$ (or $2n + 1$) with $n \in \mathbb{Z}$ are called *even* (or *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 expansions* 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 nonperiodic). In analysis, real numbers are defined as equivalence classes of rational Cauchy sequences. In what follows, we assume the usual rules for the 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] Some books write $A - B$ instead of $A \setminus B$.
[13] Unfortunately, the symbol $\subset$ is not used consistently in the literature.
[14] see `https://de.wikipedia.org/wiki/Mengendiagramm`
[15] Strictly axiomatically, one defines $0 := \varnothing$, $1 := \{\varnothing\}$ and in general $n + 1 := n \cup \{n\}$.

(e) We have $\mathbb{N} \subsetneq \mathbb{N}_0 \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}$. We prove 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 $b \neq 0$. Without loss of generality (for short: w. l. o. g.) we may assume that a and b are coprime (otherwise one can reduce $\frac{a}{b}$). 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 even as well. Hence 2 is a common divisor of a and b. This contradiction shows that the assumption was false. Hence $\mathbb{Q} \neq \mathbb{R}$.

(f) The elements of a set can very well be sets themselves. In such cases one often uses calligraphic 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 we have:*

(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).

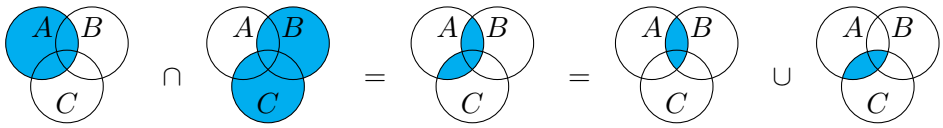
(c) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$ *and* $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$ (De Morgan's laws).

(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.*
- *Induction 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 is 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 induction step, $A(n-1) \Rightarrow A(n)$ holds. Hence $A(n)$ 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.
- Induction 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 $n = 1$ we have $1^2 = 1 = 1^3$.

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

Induction step: We have to prove the claim for $n + 1$. First, an auxiliary computation:

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

(Gauss realized this at the age of 9[16]). By the binomial formula we now have

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

□

[16] see Wikipedia

2 Cartesian Products and Functions

2.1 Pairs and Tuples

Remark 2.1. By 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

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

  We have $\boxed{|A \times B| = |A||B|,}$ provided that 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 $A := A_1 = \ldots = A_n$, 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 points of the 2-dimensional[2] plane.

(b) We have

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

(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 nonempty sets. The previously mentioned *axiom of choice* states $\bigtimes_{i \in I} A_i \neq \varnothing$, i. e., we can *simultaneously* choose an element from each set A_i.

Definition 2.4. A *relation* on a 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 $\forall a, b \in A : (a \sim b \Rightarrow b \sim a)$.
- *antisymmetric* if $\forall a, b \in A : ((a \sim b \wedge b \sim a) \Rightarrow a = b)$.[3]

[1] The formal definition of pairs can be reduced to sets: $(a, b) := \{\{a\}, \{a, b\}\}$.
[2] The *dimension* is defined in Definition 4.18.
[3] Caution: There is also the stronger property *asymmetric*: $\forall a, b \in A : (a \sim b \Rightarrow b \not\sim a)$.

- *transitive* if $\forall a, b, c \in A : ((a \sim b \wedge b \sim c) \Rightarrow a \sim c)$.
- an *equivalence relation* if R is reflexive, symmetric, and transitive.
- a (*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 a (boring) 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.[4] Many further equivalence relations can be reduced to equality. For example, let A be the set of all people and $a \sim b$ if $a, b \in A$ live in the same country. Then the equivalence classes correspond to the countries.

(c) Simple examples show 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 relation $\leq$ is an order relation. In addition, it 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 $A = \mathbb{N}$, the sets $\{1\}$ and $\{2\}$ are not related ($\{1\} \not\subseteq \{2\} \not\subseteq \{1\}$). In contrast to $\leq$, the relation $\subseteq$ 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 $b \sim c$. Since $\sim$ is symmetric, $c \sim b$. Since $\sim$ is transitive, $a \sim b$. Hence for every $d \in [b]$ we have $a \sim b \sim d$ and $a \sim d$. This shows $[b] \subseteq [a]$, and analogously we obtain $[a] \subseteq [b]$. It follows that $[a] = [b]$. Thus any two equivalence classes are either equal or disjoint. Moreover, every $a \in A$ lies in its "own" equivalence class $[a]$ by reflexivity. The existence of T now follows from the axiom of choice. □

Remark 2.7.

(a) In the situation of Lemma 2.6, one calls T 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). Hence partitions and equivalence relations correspond to each other.

Example 2.8. For the equivalence relation on the set of all people from Example 2.5, the presidents of the countries form a system of representatives.

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

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 $f(a) \in B$.[5] One then writes[6]

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

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

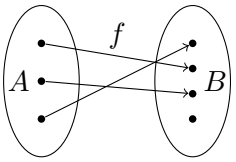
- One calls A the *domain* and B the *codomain* of f. Moreover, $f(a)$ is the *image* of a under f, and $f(A) := \{f(a) : a \in A\} \subseteq B$ is the *image* of f. For $B' \subseteq B$,

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

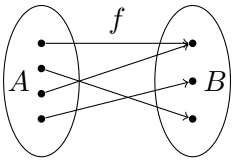
  is the *preimage* of B' under f.

- One calls $f\colon A \to B$

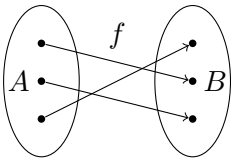
  - *injective* if $\forall a, a' \in A : (f(a) = f(a') \Longrightarrow a = a')$.



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



  - *bijective* (or a *bijection*) if f is injective and surjective. In this case, one calls A and B *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

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

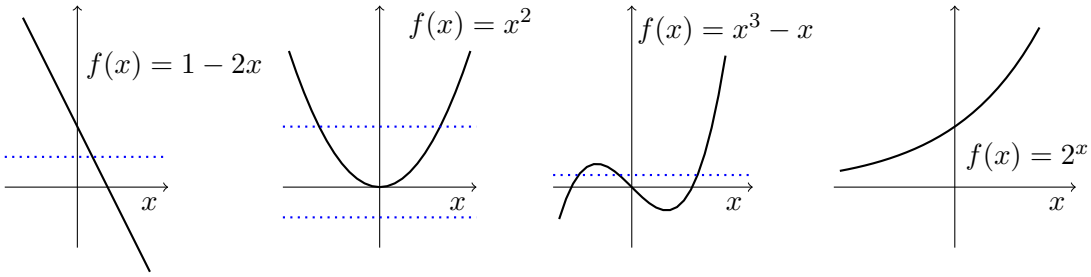
  is called the *composition* (or *composite*) of f and g.

Example 2.10.

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

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

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



Injective (respectively surjective) means that the graph of f intersects every horizontal line at most (respectively 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 $|\mathrm{Map}(A, B)| = |B|^{|A|}$, because for $f\colon A \to B$ and every $a \in A$ there are $|B|$ possibilities to choose $f(a) \in B$.

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

(c) Every function $f\colon A \to B$ can be made surjective by restricting the codomain to the image: $f\colon A \to f(A)$.

(d) For $f\colon A \to B$ we have

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

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

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

(f) Although $\mathbb{N}$ contains only "half as many" numbers as $\mathbb{Z}$, the sets $\mathbb{N}$ and $\mathbb{Z}$ are equinumerous via 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 equinumerous to $\mathbb{N}$ is called *countable*[7]. By Cantor's *diagonal 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 equinumerous to either $\mathbb{N}$ or $\mathbb{R}$. The following theorem shows that there are arbitrarily "large" sets (*cardinal numbers*), which are rarely encountered in practice, however.

[7] 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 map $M \to \mathcal{P}(M)$, but no bijection. If M is finite, then $|\mathcal{P}(M)| = 2^{|M|}$.*

Proof. The map $M \to \mathcal{P}(M)$, $a \mapsto \{a\}$ is clearly injective. Suppose that 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 $f(a) = A$. This yields the contradiction $a \in A = f(a) \iff a \notin f(a)$. For the second claim, let $M = \{x_1, \ldots, x_n\}$. Then the map

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

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

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 $f \circ g = \mathrm{id}_B$. In this case, 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 $(g \circ f)(a) = (g \circ f)(a')$ we have $g(f(a)) = g(f(a'))$, hence $f(a) = f(a')$ 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 $C = (g \circ f)(A) = g(f(A)) \subseteq g(B) \subseteq C$, hence $g(B) = C$.

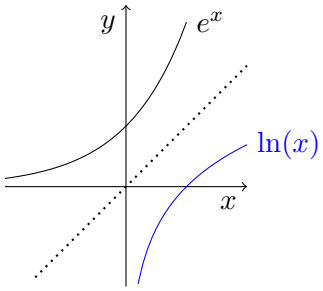
(f) If $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$, then f is injective by (d) and surjective by (e), hence also bijective. Conversely, let f be bijective. Then for every $b \in B$ there exists exactly one $g(b) \in A$ with $f(g(b)) = b$. Hence $g\colon B \to A$ is the only map with $f \circ g = \mathrm{id}_B$. 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 relationship between the two 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* $\ln\colon \mathbb{R}_{>0} \to \mathbb{R}$. One obtains the graph of ln by reflection in the line $y = x$:



Note that the mere existence of the inverse function is far from yielding a concrete formula for $f^{-1}(x)$. This fact 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 ($\mathbb{Q}, \mathbb{R}, \ldots$), one replaces number systems 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 studying linear problems which have proven their worth over time (just like metric spaces in analysis or Bohr's atomic model in chemistry).

Definition 3.2. A *binary 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*),
- $\exists e \in G : (\forall x \in G : e \cdot x = x = x \cdot e)$ (*neutral element*),
- $\forall x \in G : (\exists y \in G : y \cdot x = e = x \cdot y)$ (*inverse element*).

If in addition

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

then G is called *abelian*.[1]

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

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

(b) If $e' \in G$ is also a neutral element, then $e' = e' \cdot e = e$. Hence 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 inverse to $x \in G$. 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) Since $e \in G$, a group is never empty. On the other hand, there is the *trivial* group $G = \{e\}$.

[1] Named after N. Abel.

[2] In nonabelian groups the notation $\frac{x}{y}$ is problematic, since both xy^{-1} and $y^{-1}x$ could be meant.

(b) By the usual rules of arithmetic, $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, and $(\mathbb{R}, +)$ are abelian groups with neutral element 0. On the other hand, $(\mathbb{Z}, -)$ is *not* a group, since the associative law is violated:

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

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

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

(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 neutral 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 binary operations $+$ and $\cdot$ such that the following properties hold:

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

Remark 3.6. In what follows, let K always be a field.

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

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

(c) For $x, y, z \in K$ with $z \neq 0$, the *cancellation law* $xz = yz \Longrightarrow x = y$ holds, since

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

Example 3.7.

(a) By the familiar rules of arithmetic, $\mathbb{Q}$ and $\mathbb{R}$ are fields. Moreover, there are infinitely many fields "between" $\mathbb{Q}$ and $\mathbb{R}$ (cf. Exercise I.15). On the other hand, $(\mathbb{Z}, +, \cdot)$ is not a field, since $(\mathbb{Z} \setminus \{0\}, \cdot)$ is not a group.

(b) Every field has 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 computations in $\mathbb{F}_2$ are carried out 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 and Example 15.38).

[3] see algebra lecture notes

3.2 Vector Spaces and Subspaces

Definition 3.8. A *vector space* V over a field K (in short: K-vector space) is an abelian group with respect to $+$ 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 of K *scalars* (in this context). The neutral element 0_V in V is called the *zero vector*.

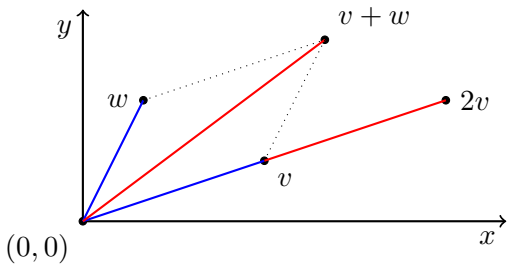
Remark 3.9. Note that $+$ denotes both the addition in K and in V. Likewise, $\cdot$ stands for the multiplication in K and for the scalar multiplication (this is imprecise, but quite common). In both cases we will often omit the symbol $\cdot$. If misunderstandings are excluded, we also write 0 instead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

Example 3.10.

(a) In every vector space we have $\lambda \cdot 0_V = 0_V$ for all $\lambda \in K$, since $\lambda \cdot 0_V = \lambda \cdot (0_V + 0_V) = \lambda \cdot 0_V + \lambda \cdot 0_V$. Similarly, $0_K \cdot v = 0_V$ for all $v \in V$. In particular, the *zero space* $V = \{0_V\}$ is a vector space.

(b) For K-vector spaces $V_1, \ldots, V_n$, the direct product (with respect to $+$) $V_1 \times \ldots \times V_n$ is also a vector space with componentwise 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) Clearly, K itself is a vector space in which the scalar multiplication coincides with the ordinary multiplication. By (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 in 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). For this one uses the summation sign:

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

If the index range is clear from the context, one writes more briefly $\sum \lambda_i v_i$. 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

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

called *nontrivial*. Sometimes the *empty sum* without summands occurs. It is 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 interchange the summation signs:

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

An advantage of algebra over analysis is that (almost) 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* $\mathrm{Map}(A, V)$ *is a K-vector space with the following operations:*

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

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

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

Hence $+$ is associative. In the same way, the remaining vector space axioms carry over from V to $\mathrm{Map}(A, V)$. □

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.,
  - $1_G \in H$,
  - $\forall g, h \in H : gh \in H$,
  - $\forall h \in H : h^{-1} \in H$.

  In this case, we write $H \leq G$. In the case $H \neq G$, one calls H a *proper* subgroup and writes $H < G$.
- 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 for subgroups. In the case $U \neq V$, one calls U a *proper* subspace and writes $U < V$.

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, respectively, are *well-defined*. The remaining group axioms and vector space axioms, respectively, need not be checked, since they already hold in the larger set G or V, respectively.

(b) In what follows, we restrict ourselves to the study of subspaces. Most statements hold analogously for (abelian) groups as well.

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

Example 3.14.

(a) Every vector space V has 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 follows that $U \leq W$.

(c) The intersection of arbitrarily many subspaces is again a subspace (verify).

(d) We prove $U := \big\{(x, 0) : x \in \mathbb{R}\big\} \leq \mathbb{R}^2$ using Remark 3.13: Since $(0, 0) \in U$, we have $U \neq \varnothing$. 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. Similarly, 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, since $(1, 1) \in U$, 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) Clearly, $U := \big\{(0, 0), (1, 0)\big\}$ 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 $V/U := \{v + U : v \in V\}$ with*

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

becomes a K-vector space.

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 $0 + U = U$. Clearly, $\overline{0v} = U = \overline{0v'}$. For $\lambda \in K^\times$ we have

$$\overline{\lambda v} = \lambda v + U = \{\lambda(v + \lambda^{-1}u) : u \in U\} = \{\lambda(v' + u') : u' \in U\} = \overline{\lambda v'}.$$

Hence the scalar multiplication is also well-defined. The vector space axioms for V/U follow immediately 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 called *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 Spanning Sets

Remark 4.1. To be able to compare infinitely large vector spaces, we introduce the dimension as a finer invariant. It will turn out 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 of S. We call $\langle S\rangle$ 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 we write $U \oplus W$ 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 of 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 $\langle S\rangle \leq V$. Since $0 \in U \cap W$, 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

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

Hence $U + W \leq V$ as well. □

Example 4.4.

(a) We have $\langle\varnothing\rangle = \{0\}$, since the empty sum is the only linear combination of elements of $\varnothing$.

(b) For $U \leq W \leq V$ we have $U + W = W$ and $\langle U\rangle = U = U \oplus \{0\}$.

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

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

- a *spanning set*, if $\langle S\rangle = V$. In the case $|S| < \infty$, V is called *finitely generated*.

[1] In some books, $\mathrm{Span}(S)$ is written instead of $\langle S\rangle$.
[2] This replaces the (disjoint) union, see Exercise I.13.

- *linearly dependent*, if 0_V is a nontrivial linear combination of elements of S.
- *linearly independent*, if not linearly dependent, i. e., for pairwise distinct elements $s_1, \ldots, s_n \in S$ and $\lambda_1, \ldots, \lambda_n \in K$ we have:

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

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

Remark 4.6. Since bases are sets, their elements have no fixed order. In fact, however, many theorems depend on the order of the basis elements. We therefore introduce the following convention: 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) Since $1_K \cdot 0_V = 0_V$, the zero vector is never an element of a linearly independent set. A single vector $v \neq 0$, on the other hand, is always linearly independent, because $\lambda v = 0$ with $\lambda \in K^\times$ yields 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 in K^n. Since every vector $v = (v_1, \ldots, v_n) \in K^n$ can be written in the form $v = \sum_{i=1}^{n} v_i e_i$, the set $\{e_1, \ldots, e_n\}$ is a spanning set of K^n. From $v = 0 \Longleftrightarrow v_1 = \ldots = v_n = 0$ we obtain the linear independence of $\{e_1, \ldots, e_n\}$. We call $e_1, \ldots, e_n$ the *standard basis* of K^n.

4.2 Characterization and Existence of Bases

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

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

is a bijection.

Proof. Since $V = \langle b_1, \ldots, b_n \rangle$, every $v \in V$ is a linear combination of the given form. Let $\lambda_1, \ldots, \lambda_n, \mu_1, \ldots, \mu_n \in K$ 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 $\{b_1, \ldots, b_n\}$ is linearly independent, we obtain $\lambda_i = \mu_i$ for $i = 1, \ldots, n$. □

Definition 4.9. In the situation of Theorem 4.8, ${}_B[v]$ is called the *coordinate vector* of v with respect to 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 spanning set, i. e., for all $b \in B$ the set $B \setminus \{b\}$ is not a spanning set.

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

Proof. We give a *cyclic chain of implications.*[3]

(1) ⇒ (2): Let B be a basis, hence in particular a spanning set of V. Suppose that $B \setminus \{b\}$ is also a spanning set for some $b \in B$. Then there exist $\lambda_1, \ldots, \lambda_n \in K$ and $b_1, \ldots, b_n \in B \setminus \{b\}$ 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 spanning set. Let $\sum_{i=1}^{n} \lambda_i b_i = 0$ for $\lambda_1, \ldots, \lambda_n \in K$ and pairwise distinct $b_1, \ldots, b_n \in B$. 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 spanning set. Hence $\lambda_1 = \ldots = \lambda_n = 0$ and B is linearly independent. Now let $v \in V \setminus B$. Since $\langle B \rangle = V$, there exist $\lambda_1, \ldots, \lambda_n \in K$ and $b_1, \ldots, b_n \in B$ 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 maximal linearly independent. We have to show $\langle B \rangle = V$. Let $v \in V$. In the case $v \in B$ we have $v \in \langle B \rangle$. So let $v \notin B$. Then $B \cup \{v\}$ is linearly dependent. Hence there exist $\lambda_1, \ldots, \lambda_n \in K^\times$, $\mu \in K$, $b_1, \ldots, b_n \in B$ with $\mu v + \sum_{i=1}^{n} \lambda_i b_i = 0$. Since B is linearly independent, we must have $\mu \neq 0$. This yields

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

Altogether, $V = \langle B \rangle$. □

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

[3] A *circular reasoning*, in contrast, is a faulty argument in which the claim is already assumed.

Proof. Let $E = \{s_1, \ldots, s_n\}$. In the case $E \subseteq \langle U \rangle$ we have $V = \langle E \rangle \subseteq \langle U \rangle$, i. e., U is already a basis. So let $E \nsubseteq \langle U \rangle$ and w. l. o. g. $s_1 \notin \langle U \rangle$. As in the last proof, $U_1 := U \cup \{s_1\}$ is then linearly independent. Now we can repeat the argument with U_1 in place of U. In the case $E \subseteq \langle U_1 \rangle$, U_1 is a basis, and otherwise we may assume $s_2 \notin \langle U_1 \rangle$. Then $U_2 := U_1 \cup \{s_2\}$ is linearly independent, and so on. Since E is finite, we obtain 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 by the standard basis vector e_3 (but not by e_1 or e_2).

Theorem 4.13 (Steinitz exchange lemma)**.** *Let V be a vector space with spanning set E. For every linearly independent subset $U \subseteq V$ we have $|U| \leq |E|$.*

Proof. W. l. o. g. let E be finite, say $E = \{s_1, \ldots, s_n\}$. Let $u_1, \ldots, u_m \in U$ be pairwise distinct. We have to show $m \leq n$. Since U is linearly independent, we have $0 \neq u_1 = \sum_{i=1}^{n} \lambda_i s_i$, where not all $\lambda_1, \ldots, \lambda_n \in K$ vanish. So w. l. o. g. $\lambda_1 \neq 0$ and hence

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

Consequently, $\{u_1, s_2, \ldots, s_n\}$ is also a spanning 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 $\mu_1, \ldots, \mu_n \in K$. Since $u_2 \notin \langle u_1 \rangle$, at least one μ_i with $i \geq 2$ must be nonzero. Say $\mu_2 \neq 0$. Because of

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

we can exchange s_2 for u_2 in the same way. Repeating this process, we finally obtain the spanning 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 spanning set of $\mathbb{R}^4$ (note that no computation is necessary).

Theorem 4.15. *Every finitely generated vector space has a finite basis, and any two bases are equinumerous.*

Proof. Let V be a vector space with finite spanning set E. By the basis extension theorem, we can extend the linearly independent set $\varnothing$ by elements of E to a basis B of V. In particular, $|B| \leq |E| < \infty$. Let C be another basis of V. By the exchange lemma, we have $|C| \leq |B| \leq |C|$, hence $|C| = |B|$. Since B and C are finite, they must be equinumerous by Remark 2.11(e). □

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

Proof. Let B be a basis of V and $S \subseteq U$ linearly independent. By the exchange lemma, we have $|S| \leq |B| < \infty$. In particular, U has a maximal linearly independent subset C. By Lemma 4.10, C is a (finite) basis of U. Hence U is finitely generated. By the basis extension theorem, C can be extended to a basis D of V. The second claim 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, we have

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

(b) Using the axiom of choice (more precisely Zorn's *lemma*), one can show that every vector space has a (possibly infinite) basis and any two bases are equinumerous. For example, $\mathbb{R}$ as a $\mathbb{Q}$-vector space has infinite bases, none of which can be given explicitly. In Theorem 6.12 we discuss how to compute bases of finitely generated vector spaces efficiently.

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.[4] By Theorem 4.15, d does not depend on the choice of B. Instead of "finitely generated" we 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 vector with respect to a basis shows

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

Remark 4.20.

(a) The theorems above imply some useful facts:

- For $U \leq V$ we have $\dim U \leq \dim V$ with equality if and only if $U = V$ (extend a basis of U to a basis of V).
- Every spanning set E of V contains a basis of V (reduce to a minimal spanning 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 in the foreground, 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).

[4] For omnipresent operators such as dim, one likes to omit the pair of parentheses around the argument for better readability if misunderstandings are ruled out.

Theorem 4.21 (Dimension formula)**.** *For subspaces U and W of a finite-dimensional vector space V we have*

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

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 it 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 the elements of the form $u + w$ with $u \in U$ and $w \in W$, $U + W$ is spanned by $b_1, \ldots, b_n, c_1, \ldots, c_s, d_1, \ldots, d_t$.

For the 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 $v + u = -w \in U \cap W$. Hence $v + u$ can be expressed as a linear combination of $b_1, \ldots, b_n$. On the other hand, the representation of $v + u$ with respect to the basis $\{b_1, \ldots, b_n, c_1, \ldots, c_s\}$ is unique by 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 claim follows. □

Example 4.22. Let

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

Clearly, $\dim U = 2$ and $\dim W = 1$. 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 $(\lambda, \lambda + 2\mu, \mu) = (\rho, \rho, \rho)$. Comparing coefficients yields $\lambda = \rho = \mu$ and $3\rho = \lambda + 2\mu = \rho$. This can only hold for $\rho = 0$. Hence $v = 0(1,1,1) = 0$ and $U \cap W = \{0\}$. We obtain $\dim(U + W) = \dim(U) + \dim(W) = 2 + 1 = 3$. Since $U + W \leq \mathbb{R}^3$, we therefore have $\mathbb{R}^3 = U \oplus W$.

Theorem 4.23. *For vector spaces $U \leq V$ we have* $\boxed{\dim V = \dim(U) + \dim(V/U).}$

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 of V has the form $v = w + u$ with $u \in U$ and $w \in W$. Since $v + U = w + U$, $\overline{B}$ is a spanning 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 $U \cap W = \{0\}$ we obtain $\lambda_b = 0$ for all $b \in B$. Hence $\overline{B}$ is linearly independent. □

5 Matrices

5.1 The Vector Space of Matrices

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

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

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

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. We can regard the vectors in K^n as $1 \times n$ matrices. One then speaks of *row vectors*. Accordingly, the $n \times 1$ matrices are called *column vectors*. If no confusion is possible, we use the standard basis $e_1, \ldots, e_n$ both as row vectors and as 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 no confusion is possible).

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

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

Here δ_{ij} is the KRONECKER *delta*:

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

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

(d) The entries $a_{11}, a_{22}, \ldots$ form the *main diagonal* of A. 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*. In the special case $\lambda_1 = \ldots = \lambda_n$ one speaks of *scalar matrices*.

(e) The $n \times m$ *matrix unit* E_{st} has a 1 at position (s, t) and only zeros elsewhere. With the Kronecker delta we have $E_{st} = (\delta_{is}\delta_{jt})_{i,j}$.

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

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

Clearly, $(A^{\mathrm{t}})^{\mathrm{t}} = A$. Transposition turns row vectors into column vectors and vice versa.

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

Lemma 5.4. *With componentwise operations, $K^{n\times m}$ becomes a 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 matrix units 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} if one arranges the vectors of K^{nm} as $n \times m$ matrices. 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 matrix units. □

Example 5.5.

(a) Caution: Only matrices of the same size 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 given 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} arises by "combining" 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 picture the following scheme:

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

(b) The multiplication of diagonal matrices is easy:

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

(c) For matrix units of suitable size we have $E_{ij}E_{kl} = \delta_{jk}E_{il}$, since the entry at position (s, t) is

$$\sum_{u=1}^{n} \delta_{si}\delta_{uj}\delta_{uk}\delta_{tl} = \delta_{jk}\delta_{si}\delta_{tl}.$$

(d) 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 of "suitable" size and $\lambda \in K$ we have*

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

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).[2] Then we have

$$(A1_m)_{ij} = \sum_{k=1}^{m} a_{ik}\delta_{kj} = a_{ij} = \sum_{k=1}^{n} \delta_{ik}a_{kj} = (1_nA)_{ij},$$

[1] One calls this weaker structure a *ring*.

[2] Caution: This notation changes in Definition 9.3.

$$
\begin{aligned}
\big((AB)^{\mathrm{t}}\big)_{ij} &= \sum_{k=1}^{m} a_{jk}b_{ki} = \sum_{k=1}^{m} (B^{\mathrm{t}})_{ik}(A^{\mathrm{t}})_{kj} = (B^{\mathrm{t}}A^{\mathrm{t}})_{ij},\\
\big(\lambda(AB)\big)_{ij} &= \lambda\sum_{k=1}^{m} a_{ik}b_{kj} = \sum_{k=1}^{m} (\lambda a_{ik})b_{kj} = \big((\lambda A)B\big)_{ij} = \sum_{k=1}^{m} a_{ik}(\lambda b_{kj}) = \big(A(\lambda B)\big)_{ij},\\
\big(A(BC)\big)_{ij} &= \sum_{k=1}^{m} a_{ik}(BC)_{kj} = \sum_{k=1}^{m} a_{ik}\sum_{l=1}^{n} b_{kl}c_{lj} = \sum_{k=1}^{m}\sum_{l=1}^{n} a_{ik}b_{kl}c_{lj}\\
&= \sum_{l=1}^{n}\Big(\sum_{k=1}^{m} a_{ik}b_{kl}\Big)c_{lj} = \sum_{l=1}^{n}(AB)_{il}c_{lj} = ((AB)C)_{ij},\\
\big(A(B+C)\big)_{ij} &= \sum_{k=1}^{m} a_{ik}(B+C)_{kj} = \sum_{k=1}^{m} a_{ik}(b_{kj}+c_{kj}) = \sum_{k=1}^{m}(a_{ik}b_{kj}+a_{ik}c_{kj})\\
&= \sum_{k=1}^{m} a_{ik}b_{kj} + \sum_{k=1}^{m} a_{ik}c_{kj} = (AB)_{ij} + (AC)_{ij},\\
\big((A+B)C\big)_{ij} &= \sum_{k=1}^{m}(a_{ik}+b_{ik})c_{kj} = \sum_{k=1}^{m} a_{ik}c_{kj} + \sum_{k=1}^{m} b_{ik}c_{kj} = (AC)_{ij} + (BC)_{ij}.
\end{aligned}
$$
□

Remark 5.9.

(a) For $A \in K^{n\times m}$ we clearly have $0_{k\times n}\cdot A = 0_{k\times m}$ and $A\cdot 0_{m\times k} = 0_{n\times k}$.

(b) We say that matrices $A, B \in K^{n\times n}$ *commute*, if $AB = BA$. For $n \geq 2$ this is usually not the case:

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

Hence $(K^{n\times n}, +, \cdot)$ is *not* a field. We also see that not every matrix (other than 0_n) has an inverse with respect to $\cdot$ (even the cancellation law does not hold for matrices).

(c) A square matrix $A \in K^{n\times n}$ is called *invertible* if there exists a matrix $B \in K^{n\times n}$ such that

$$
AB = 1_n = BA
$$

holds.[3] If also $AC = 1_n = CA$, then $C = C1_n = C(AB) = (CA)B = 1_nB = B$. Hence B is uniquely determined and, as for groups, one writes $A^{-1} := B$. One calls A^{-1} the *inverse* of A. We show in Lemma 5.15 that $AB = 1_n$ already implies invertibility, i. e., $BA = 1_n$ need not be checked.

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

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

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

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

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

[3] Invertible matrices are also called *nonsingular*, and noninvertible matrices *singular*.

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

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

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*.

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

Proof. The identity element 1_n is clearly invertible. Together with A, also A^{-1} is invertible. For invertible matrices A and B we have

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

and similarly $(B^{-1}A^{-1})(AB) = 1_n$. This shows that AB is invertible as well 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 $z_1, \ldots, z_n$ be the rows and $s_1, \ldots, s_m$ the columns of a matrix $A \in K^{n\times m}$. Then* $\dim\langle z_1, \ldots, z_n\rangle = \dim\langle s_1, \ldots, s_m\rangle$.

Proof. By Remark 4.20 there exists $I \subseteq \{1, \ldots, n\}$ such that $\{z_i : i \in I\}$ is a basis of $\langle z_1, \ldots, z_n\rangle$. Similarly, let $\{s_j : j \in J\}$ be a basis of $\langle s_1, \ldots, s_m\rangle$. 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. To this end let $\lambda_i \in K$ with $\sum_{i\in I} \lambda_i a_{ij} = 0$ for all $j \in J$. For $k \in \{1, \ldots, m\} \setminus J$ 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 $i \in \{1, \ldots, n\}$. It follows that

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

Hence $\sum_{i\in I} \lambda_i a_{ij} = 0$ for all $j \in \{1, \ldots, m\}$, 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 $i \in I$. Hence the rows of B are linearly independent. Since they lie in the $|J|$-dimensional vector space $K^{|J|}$, we have $|I| \le |J|$. The rows (resp. columns) of A are the columns (resp. rows) of A^{t}. Applying the above argument to A^{t} yields $|J| \le |I|$. Altogether, $\dim\langle z_1, \ldots, z_n\rangle = |I| = |J| = \dim\langle s_1, \ldots, s_m\rangle$. □

Definition 5.13. In the situation of Theorem 5.12 one calls

$$\operatorname{rk} A := \dim\langle z_1, \ldots, z_n\rangle = \dim\langle s_1, \ldots, s_m\rangle$$

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

Example 5.14.

(a) The identity matrix 1_n has (full) rank n, since its rows form the standard basis. We have $\operatorname{rk}\left(\begin{smallmatrix}1 & 2\\ 2 & 4\end{smallmatrix}\right) = 1$, since the second row is twice the first. In Remark 6.13 we will figure out how to compute the rank of an arbitrary matrix efficiently.

(b) For every matrix A we have $\operatorname{rk} A = \operatorname{rk}(A^{\mathrm{t}})$ and $\operatorname{rk} A = 0 \Longleftrightarrow A = 0$.

Lemma 5.15.

(a) For matrices A and B of "suitable" size we have $\operatorname{rk}(AB) \leq \min\{\operatorname{rk} A, \operatorname{rk} B\}$.

(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. Hence 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 we obtain

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

(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. Hence there exists $B \in K^{n\times n}$ with $AB = 1_n$. Since $\operatorname{rk}(A^{\mathrm{t}}) = \operatorname{rk} A$, there also exists $C \in K^{n\times n}$ with $A^{\mathrm{t}}C = 1_n$, i. e., $C^{\mathrm{t}}A = (A^{\mathrm{t}}C)^{\mathrm{t}} = 1_n^{\mathrm{t}} = 1_n$. Since $C^{\mathrm{t}} = C^{\mathrm{t}}(AB) = (C^{\mathrm{t}}A)B = B$, A is invertible. □

6 The Gaussian Algorithm

6.1 Systems of Linear Equations

Definition 6.1. A *system of linear 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. We seek the *solution set*

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

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

Example 6.2. The system of linear equations

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

corresponds to the matrix equation

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

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

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

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

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

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

□

[1] sometimes also named after Rouché and Frobenius

Remark 6.5. Let $A \in K^{n\times m}$ have 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. If $n = m$, then A is invertible by Lemma 5.15 and $x = A^{-1}b$ is the unique solution. If $n < m$, the system of linear equations $Ax = b$ is called *underdetermined*, i. e., there are fewer equations than unknowns. We show that there are then several 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 linear equations $Ax = 0$ is a subspace L_0 of $K^{m\times 1}$ of dimension $m - \operatorname{rk}(A)$.

(b) If the system of linear equations $Ax = b$ has a solution $\widetilde{x}$, then $\widetilde{x} + L_0$ is the solution set.

Proof.

(a) Since $0 \in L_0$, we have $L_0 \neq \varnothing$. 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, since 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 $K^{m\times 1}$, we obtain $\lambda_{k+1} = \ldots = \lambda_m = 0$ 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 linear equations $Ax = b$ has at most one solution. If $m < n$, one speaks of *overdetermined* systems of linear equations. In the following, we deal with the explicit construction of the solution set of an arbitrary system of linear equations.

(b) Since the map $L_0 \to \widetilde{x} + L_0$, $v \mapsto \widetilde{x} + v$ is a bijection, a solvable system of linear equations has as many solutions as the corresponding homogeneous system. 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 linear equations has no solution, exactly one solution, 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 multiplication by an *elementary matrix* of the form

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

  on A from the left.

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

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

  on A from the left.

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

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

  on A from the left.

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

Remark 6.9.

(a) All elementary row operations are reversible. Hence $A \sim B$ implies $B \sim A$. Moreover, the elementary matrices are invertible. By Lemma 5.10, the product of elementary matrices is invertible as well. Therefore $A \sim B$ implies the existence of a matrix $S \in \mathrm{GL}(n, K)$ with $SA = B$.

(b) By (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 multiplication by elementary matrices from the *right* on A (try it out). Column operations can also be realized by row operations on A^{t}. We also use the notation $A \sim B$ if B arises from A by column operations. In this case there exists an $S \in \mathrm{GL}(m, K)$ with $AS = B$.

Theorem 6.10 (GAUSSian 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.,*

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

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:
  - Interchange the i-th row with the z-th row. Afterwards $a_{zs} \neq 0$.
  - Divide the z-th row by a_{zs}. Afterwards $a_{zs} = 1$.
  - For $j = 1, \ldots, z-1, z+1, \ldots, n$ subtract a_{js} times the z-th row from the j-th row. Afterwards $a_{js} = 0$.
  - 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 also $B \sim C$ 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 $s_k = e_k$ for $k \leq \operatorname{rk} B$ by induction on k. In the case $k = 0$ we have $b_i \in \langle \varnothing \rangle = \{0\}$, so $b_i = 0$. Clearly then also $c_i = Sb_i = 0$. Now suppose the claim has already been proved 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 $b_i = e_k$ because of the reduced row echelon form. The columns of S are linearly independent, since S is invertible. This shows

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

Hence 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 $b_j \in \langle e_1, \ldots, e_k \rangle$ we now have $c_j = Sb_j = (e_1, \ldots, e_k, s_{k+1}, \ldots, s_n) b_j = b_j$. Thus $b_i = c_i$ for $i = 1, \ldots, m$, i. e., $B = C$. □

Example 6.11.

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

$$\sim \begin{pmatrix} 1 & 1/2 & 3/2 & 0 \\ 0 & 1 & 1 & 3 \\ 0 & -5/2 & -5/2 & 1 \end{pmatrix} \begin{matrix} + \\ -1/2 \quad 5/2 \\ + \end{matrix} \sim \begin{pmatrix} 1 & 0 & 1 & -3/2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 17/2 \end{pmatrix} \begin{matrix} \\ \\ \mid :17/2 \end{matrix}$$

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

[3] Every nonzero row contains a *leading one*. All entries to the left of, 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.

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

6.3 Applications

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

Proof. By elementary row operations, rows of A are replaced by linear combinations of rows. Hence the rows of $\widehat{A}$ generate a subspace $W \leq U$. Since all row operations are reversible, we even have $W = U$ and $\dim U = \operatorname{rk} A = \operatorname{rk} \widehat{A}$. Let k be the number of nonzero rows in $\widehat{A}$. Then $\operatorname{rk} \widehat{A} \leq k$. On the other hand, $\widehat{A}$ has the linearly independent columns $e_1, \dots, e_k$. 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 in the Gaussian algorithm (the nonzero rows are still linearly independent, since their number stays the same). Moreover, it is advisable to avoid divisions by interchanging 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 one may also use elementary column operations because of $\operatorname{rk} A = \operatorname{rk}(A^{\mathrm{t}})$. This can be useful if A has fewer columns than rows (there are many ways to reach the goal).

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

$$\begin{pmatrix} 1 & -1 & 0 \\ 2 & 0 & 2 \\ 3 & -1 & 2 \\ -1 & 2 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 2 \\ 3 & 2 & 2 \\ -1 & 1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 2 & 0 \\ -1 & 1 & 0 \end{pmatrix} \implies \operatorname{rk} A = 2$$

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

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

If so, the solution set is obtained as follows: Let $(1, n_1), \dots, (k, n_k)$ be the positions of the leading ones in $(\widehat{A}|c)$. The $n - k$ zero rows are deleted. For all $i \in \{1, \dots, m\} \setminus \{n_1, \dots, n_k\}$ 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 $n_1, \dots, n_k$ 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 $x \in K^{m\times 1}$ we have

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

If e_{m+1} is a row of $(\widehat{A}|c)$, then the equation $\widehat{A}x = c$ yields 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 in the following example

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

(here . stands for 0 for better readability). One easily sees that $\widehat{A}\widetilde{x} = c$. Hence also $A\widetilde{x} = b$. This proves (a) and (b). In the same way one sees that the columns s_1, s_3, and s_5 of M (marked red) lie in L_0. The different positions of the entries -1 in s_i imply the linear independence of $\{s_1, s_3, s_5\}$. By Theorem 6.6 and Theorem 6.12, on the other hand,

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

This shows (c). From Theorem 6.6 we obtain (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 reduced row echelon form of $(A|1_n) \in K^{n\times 2n}$. Then A is invertible if and only if $\widehat{A} = 1_n$. If so, $A^{-1} = B$.*

Proof. We have

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

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 $A \in \mathrm{GL}(n, K)$ we have $A \sim 1_n$ by Theorem 6.17. □

Example 6.19.

$$(A|1_3) = \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 & 1 & 0 \\ -1 & 1 & -2 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & -1 & 0 & 1 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & 0 & -3 & 2 & 1 & 1 \end{array}\right) \mid : (-3)$$

$$\sim \left(\begin{array}{ccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 1 \\ 0 & 0 & 1 & -2/3 & -1/3 & -1/3 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/3 & -1/3 & -1/3 \\ 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & -2/3 & -1/3 & -1/3 \end{array}\right)$$

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

Remark 6.20.

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

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

then the reduced row echelon form must have a zero row. The matrix cannot be invertible then and one can stop the algorithm early.

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

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

Here $r = \operatorname{rk} A$ is uniquely determined. In particular, A and B are equivalent if and only if $\operatorname{rk} A = \operatorname{rk} B$. Hence one does not need to introduce a new symbol for equivalence. The number of equivalence classes of $n \times n$ matrices is $n + 1$.

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

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. Since $U + W = \langle u_1, \ldots, u_s, w_1, \ldots, w_t\rangle$, the set $\{b_1, \ldots, b_k\}$ is a basis of $U + W$ by Theorem 6.12. Moreover, 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 $m = 1, \ldots, l$. Because of the reduced row echelon form, $\{c_1, \ldots, c_l\}$ is linearly independent. By elementary *column* operations, one transforms A into

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

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

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

Hence $\{c_1, \ldots, c_l\}$ 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 need to carry out 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 Homomorphisms and Isomorphisms

Remark 7.1. To relate different vector spaces V and W, we study maps $V \to W$ that "respect" addition and scalar multiplication. It will turn out 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 $\lambda \in K$ we have:

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

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

Remark 7.3.

(a) For $f \in \mathrm{Hom}(V, W)$ we have $f(0_V) = 0_W$, since

$$f(0_V) = f(1_K 0_V + 0_V) = 1_K f(0_V) + f(0_V) = f(0_V) + f(0_V).$$

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

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

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

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

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

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

(d) If $f\colon V \to W$ is an isomorphism, then so is $f^{-1}\colon W \to V$[1], since for $w, w' \in W$ and $\lambda \in K$ we have

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

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

[1] In analysis, the inverse of a bijective continuous function is not necessarily continuous. Hence there is the curious notion of a *homeomorphism* (not a typo).

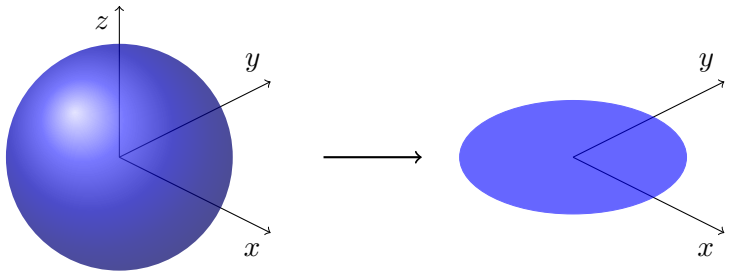
[2] However, the collection of all vector spaces is not a set, but a *class*.

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 the 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 $x \in \mathbb{R}$ we have $f(x) = f(x1) = xf(1) = ax$. Hence the graph of f describes a line through the origin. Caution: In school mathematics, functions of the form $f(x) = ax + b$ are sometimes also called "linear" (such maps are called *affine*,[3] see Exercise I.25).

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

(f) 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 $\mathrm{Map}(M,V) \to V^n$, $f \mapsto (f(x_1),\dots,f(x_n))$ is an isomorphism.

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

Proof. Since $0 = f(0) \in f(V_1)$, we have $f(V_1) \neq \varnothing$. For $u, v \in V_1$ and $\lambda \in K$ we have

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

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

Definition 7.6. In the situation of Lemma 7.5, we call $\operatorname{Ker} f$ the *kernel* of f and

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

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

Lemma 7.7. *A map $f \in \mathrm{Hom}(V,W)$ is injective if and only if $\operatorname{Ker} f = \{0\}$. 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 $v \in \operatorname{Ker} f$. From $f(v) = 0 = f(0)$ we obtain $v = 0$, i. e., $\operatorname{Ker} f = \{0\}$. Conversely, let $\operatorname{Ker} f = \{0\}$ and $u, v \in V$ with $f(u) = f(v)$. Then $f(u - v) = f(u) - f(v) = 0$, so $u - v \in \operatorname{Ker} f = \{0\}$. Hence $u = v$ 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 $i = 1, \ldots, n$. Moreover:*

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

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

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

Proof. Every $u \in V$ can be written uniquely 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 $\rho \in K$ we have

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

Hence f is linear with $f(b_i) = c_i$ for $i = 1, \ldots, n$. If also $g \in \operatorname{Hom}(V, W)$ satisfies $g(b_i) = c_i$ 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 $\sum_{i=1}^n \lambda_i b_i \in \operatorname{Ker} f = \{0\}$. Since $b_1, \ldots, b_n$ are linearly independent, we obtain $\lambda_1 = \ldots = \lambda_n = 0$. Hence $c_1, \ldots, c_n$ are linearly independent. Conversely, let $c_1, \ldots, c_n$ be linearly independent and $u := \sum_{i=1}^n \lambda_i b_i \in \operatorname{Ker} f$. Then $\sum_{i=1}^n \lambda_i c_i = f(u) = 0$ and we obtain $\lambda_1 = \ldots = \lambda_n = 0$. Hence $u = 0$ and $\operatorname{Ker} f = \{0\}$. By Lemma 7.7, f is injective.

(b) From $f(\sum \lambda_i b_i) = \sum \lambda_i c_i$ for all $\lambda_1, \ldots, \lambda_n \in K$ it follows that $f(V) = \langle c_1, \ldots, c_n \rangle$. This shows the claim.

(c) Follows from (a) and (b). □

Remark 7.9. Let $f \in \operatorname{Hom}(V, W)$. In the case $\dim V < \dim W$, f is not surjective, since the image of a basis of V cannot be a spanning set of W. In the case $\dim V > \dim W$, f is not injective, since the image of a basis cannot be linearly independent. For $\dim V = \dim W$ we obtain

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

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

Remark 7.11.

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

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

(c) Although $\mathbb{Q}$ and $\mathbb{Q}^2$ are equinumerous, we have $\mathbb{Q} \ncong \mathbb{Q}^2$ by Theorem 7.10.

(d) Let $U, V \leq W$ be vector spaces. Clearly, $V \leq U + V$ and $U \cap V \leq U$. 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 we obtain the so-called *first isomorphism theorem*[4]

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

(e) For vector spaces $U \leq V \leq W$ we clearly have $V/U \leq W/U$. From

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

we obtain the *second isomorphism theorem*[5]

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

(cf. Exercise I.11).

Theorem 7.12 (Homomorphism theorem)**.** *For $f \in \operatorname{Hom}(V, W)$, 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. Hence $\boxed{V/\operatorname{Ker}(f) \cong f(V)}$ *and* $\boxed{\dim V = \operatorname{rk}(f) + \dim \operatorname{Ker}(f).}$

Proof. For $v, w \in V$ we have

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

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

[4] Mnemonic: On one side there are two U, on the other side two V.

[5] Mnemonic: Canceling a double fraction.

7.2 Matrices of Linear Maps

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

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

$$(f+g)(\mu u + v) = f(\mu u + v) + g(\mu u + v) = \mu f(u) + f(v) + \mu g(u) + g(v) = \mu(f+g)(u) + (f+g)(v),$$
$$(\lambda f)(\mu u + v) = \lambda f(\mu u + v) = \lambda(\mu f(u) + f(v)) = \mu\lambda f(u) + \lambda f(v) = \mu(\lambda f)(u) + (\lambda f)(v)$$

the maps $f + g$ and λf are linear. □

Remark 7.14. If B is a basis of V, then the restriction map $\mathrm{Hom}(V,W) \to \mathrm{Map}(B,W)$, $f \mapsto f_{|B}$ is an isomorphism by 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) We call ${}_C[f]_B := (a_{ji}) \in K^{n \times m}$ the *matrix* of f with respect to B and C.

(b) In the case $V = W$ and $f = \mathrm{id}_V$, we call ${}_C\Delta_B := {}_C[\mathrm{id}_V]_B$ the *change-of-basis matrix* with respect to B and C.

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

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

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

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

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

$$f(1,-1) = (3,-1,-3) = -7(1,1,1) - 6(0,-1,1) + 10(1,0,1),$$
$$f(0,2) = (-2,2,0) = 4(1,1,1) + 2(0,-1,1) - 6(1,0,1)$$

we obtain

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

(if in doubt, you have to determine the entries by solving 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 $C = \{c_1, \dots, c_n\}$. By Theorem 7.8, ${}_C[.]_B$ is a bijection. Let $f, g \in \operatorname{Hom}(V,W)$ with ${}_C[f]_B = (a_{ij})$ and ${}_C[g]_B = (a'_{ij})$. For $i = 1, \dots, m$ and $\lambda \in K$ we have

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

This shows ${}_C[\lambda f + g]_B = \lambda\,{}_C[f]_B + {}_C[g]_B$. Hence ${}_C[.]_B$ 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

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

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

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

Remark 7.19.

(a) For $V = W$ and $f = \mathrm{id}_V$ we obtain ${}_C[v]^{\mathrm{t}} = {}_C\Delta_B\,{}_B[v]^{\mathrm{t}}$. For $f \colon K^n \to K^m$ we have $f(v)^{\mathrm{t}} = [f]v^{\mathrm{t}}$ with respect to the standard bases.

(b) Let $U \le V$. By 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 $\operatorname{Ker} f = U$. Hence every subspace is the kernel of a homomorphism. Let B and C be bases of V and W, respectively. For $A := {}_C[f]_B$ we have $\operatorname{Ker} f = \{v \in V : A\,{}_B[v]^{\mathrm{t}} = 0\}$ by Theorem 7.18. Hence U can be described as the solution set of a system of linear equations.

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

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

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

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 $A := {}_C[f]_B \in K^{n\times m}$ with respect to arbitrary bases. Then*

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

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

Proof.

(a) f injective $\overset{7.7}{\iff} \operatorname{Ker} f = \{0\} \overset{7.12}{\iff} \operatorname{rk} A = \operatorname{rk} f = m$.

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

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. By Remark 7.3, $g \circ f \in \mathrm{Hom}(U,W)$. Let $B = \{b_1,\dots,b_m\}$, $C = \{c_1,\dots,c_n\}$ and $D = \{d_1,\dots,d_k\}$. Let ${}_C[f]_B = (a_{ij})$ and ${}_D[g]_C = (a'_{ij})$. Then

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

Here $\sum_{j=1}^{n} a'_{lj}a_{ji}$ 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 "suitable" spaces, the distributive laws carry over from matrices:

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

Of course, one can also verify this directly.

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

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

i. e., $g\circ f = \mathrm{id}_{\mathbb{R}^2}$. Conversely, $f\circ g \neq \mathrm{id}_{\mathbb{R}^3}$, since 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 $V = W$ we have ${}_C\Delta_B^{-1} = {}_B\Delta_C$.*

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

$$ {}_C[f]_B\,{}_B[f^{-1}]_C = {}_C[\mathrm{id}_W]_C = {}_C\Delta_C = 1_{|C|} $$

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 to the matrices in $\mathrm{GL}(n,K)$, where $n = \dim V$. More precisely, ${}_B[.]_B\colon \mathrm{GL}(V) \to \mathrm{GL}(n,K)$ 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' bases of W. For $f \in \mathrm{Hom}(V,W)$ we have*

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

In the case $V = W$ we have

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

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 $C' = B'$ we obtain ${}_{B'}[f]_{B'} = {}_{B'}\Delta_B\,{}_B[f]_B\,{}_B\Delta_{B'} = {}_{B'}\Delta_B\,{}_B[f]_{B'}\Delta_B^{-1}$ with 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 $B = \big((1,-1),(0,2)\big)$ by $B' := \big((0,1),(1,1)\big)$. Because of $(0,1) = 0(1,-1) + \frac{1}{2}(0,2)$ and $(1,1) = (1,-1) + (0,2)$ we have

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

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

Remark 7.30.

(a) Clearly, $A \approx A$ (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 $C = TSAS^{-1}T^{-1} = (TS)A(TS)^{-1}$. Hence similarity of matrices is an equivalence relation.

(b) For bases B and C of V and $f \in \mathrm{Hom}(V,V)$, the matrices ${}_B[f]_B$ and ${}_C[f]_C$ are similar by Corollary 7.27. Conversely, if $S \in \mathrm{GL}(n,K)$ is given (with $n = \dim V$), then there exists $g \in \mathrm{GL}(V)$ with ${}_B[g]_B = S$ by Theorem 7.18. For $C := \{g(b) : b \in B\}$ we then have $S = {}_B[\mathrm{id}_V]_C = {}_B\Delta_C$. Hence f determines a similarity class of such matrices in $K^{n\times n}$. This enables concrete computations with abstract maps. Later we construct special bases such that these matrices have the simplest possible form (for example diagonal matrices). This speeds up computations.

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 (sum of the main diagonal entries).

Lemma 7.32. *The map* $\operatorname{tr}\colon K^{n\times n} \to K$ *is linear with* $\operatorname{tr}(A^{\mathrm{t}}) = \operatorname{tr} A$. *For* $A \in K^{n\times m}$ *and* $B \in K^{m\times n}$ *we have* $\boxed{\operatorname{tr}(AB) = \operatorname{tr}(BA).}$

Proof. For $A = (a_{ij})$, $B = (b_{ij})$ in $K^{n\times n}$ and $\lambda \in K$ we have

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

Hence tr is linear. Clearly, $\operatorname{tr}(A^{\mathrm{t}}) = \operatorname{tr} A$. For $A \in K^{n\times m}$ and $B \in K^{m\times n}$ we have

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

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

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

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

Remark 7.35. The trace map can be characterized in a "coordinate-free" way by the property $\operatorname{tr} f = \operatorname{rk} f$ for all $f \in \operatorname{Hom}(V, V)$ with $f^2 = f$ (Exercise II.6).

Theorem 7.36 (Fillmore)**.** *Let* $A \in K^{n\times n} \setminus K1_n$ *and* $d_1, \ldots, d_n \in K$ *with* $\operatorname{tr} A = d_1 + \ldots + d_n$. *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 we have

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

In any case, there exists $b_1 \in K^n$ 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. With respect to 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 done, since $\operatorname{tr} A_1 = \operatorname{tr}(A) - d_1 = d_2$.

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

$$A(b_1 + b_3) = d_1b_1 + b_2 + Ab_3 \in b_2 + \langle b_1, b_3\rangle.$$

By replacing b_3 with $b_1 + b_3$, we achieve $a'_{12} = 1$ and $A_1 \notin K1_{n-1}$ (this does not change $\operatorname{tr} A_1$). 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 $d_1, \ldots, d_n$. $\square$

Example 7.37. 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

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

by Corollary 7.27.

7.3 Dual Spaces

Definition 7.38. For a K-vector space V we call $V^* := \operatorname{Hom}(V, K)$ the *dual space* of V. Its elements are called (*linear*) *functionals*.

Lemma 7.39. *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 $0 = f(b_i) = \lambda_i$. Hence $b_1^*, \ldots, b_n^*$ are linearly independent. The claim follows from $\dim V^* = \dim V$ (Theorem 7.18). □

Example 7.40. 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 7.41.

(a) In the situation of Lemma 7.39, $b_1^*, \ldots, b_n^*$ is called the *dual basis* of $b_1, \ldots, b_n$.

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

(c) By Theorem 7.8 there exists an isomorphism $V \to V^*$ that maps a basis to the corresponding dual basis. However, there is no *canonical* isomorphism, one that does not depend on a choice of basis (cf. Theorem 12.7). The next theorem shows that the situation between V and the *double dual space* $V^{**} := (V^*)^*$ is better.

Theorem 7.42. *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 $v, w \in V$ we have

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

i. e., F is linear. Now let $F_v = 0$. In the case $v \neq 0$ we 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. Because of $\dim V^{**} = \dim V^* = \dim V < \infty$, F must also be surjective. □

Definition 7.43. 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 (or W_0) the *annihilator* of U (or W).

Lemma 7.44. *For $U, \tilde{U} \leq V$ and $W, \tilde{W} \leq V^*$ we have*

(a) $U \subseteq \tilde{U} \Rightarrow \tilde{U}^0 \subseteq U^0$ *and* $W \subseteq \tilde{W} \Rightarrow \tilde{W}_0 \subseteq W_0$.

(b) $U^0 \leq V^*$, $W_0 \leq V$ *and*

$$\dim V = \dim(U) + \dim(U^0) = \dim(W) + \dim(W_0).$$

(c) The maps $U \mapsto U^0$ and $W \mapsto W_0$ are inverse to each other.[6]

(d)

$$\boxed{\begin{array}{ll} (U + \tilde{U})^0 = U^0 \cap \tilde{U}^0, & (W + \tilde{W})_0 = W_0 \cap \tilde{W}_0, \\ (U \cap \tilde{U})^0 = U^0 + \tilde{U}^0, & (W \cap \tilde{W})_0 = W_0 + \tilde{W}_0. \end{array}}$$

(e) $V = U \oplus \tilde{U} \iff V^* = U^0 \oplus \tilde{U}^0$.

Proof.

(a) Follows from the definition.

(b) The restriction $\Gamma\colon V^* \to U^*$, $f \mapsto f_{|U}$ is a homomorphism with kernel $U^0 \leq V^*$. Since every functional in U^* can be extended to V (basis extension), Γ is surjective. By the homomorphism theorem we obtain

$$\dim V = \dim V^* = \dim(U^0) + \dim(U^*) = \dim(U^0) + \dim(U).$$

With W in place of U we get

$$W^0 = \{f \in V^{**} : f(W) = \{0\}\} \leq V^{**}$$

and $\dim V = \dim(W) + \dim(W^0)$. For the isomorphism $F\colon V \to V^{**}$, $v \mapsto F_v$ from Theorem 7.42 we have

$$v \in W_0 \iff F_v(W) = \{0\} \iff F_v \in W^0.$$

Hence $W_0 = F^{-1}(W^0) \leq V$ with $\dim W_0 = \dim W^0$.

(c) By definition, $U \leq (U^0)_0$ and $W \leq (W_0)^0$. From (b) it follows that $\dim U = \dim(U^0)_0$ and $\dim W = \dim(W_0)^0$.

[6] In this situation one speaks of a Galois *connection*.

(d) From the definition we obtain

$$\begin{aligned}(U+\tilde{U})^0 &= \{f \in V^* : f(U+\tilde{U}) = \{0\}\}\\ &= \{f \in V^* : f(U) = \{0\} = f(\tilde{U})\} = U^0 \cap \tilde{U}^0,\\ (W+\tilde{W})_0 &= \{v \in V : \forall f \in W+\tilde{W} : f(v) = 0\} = W_0 \cap \tilde{W}_0.\end{aligned}$$

With (c) we obtain

$$\begin{aligned}(U \cap \tilde{U})^0 &= ((U^0)_0 \cap (\tilde{U}^0)_0)^0 = ((U^0+\tilde{U}^0)_0)^0 = U^0+\tilde{U}^0,\\ (W \cap \tilde{W})_0 &= ((W_0)^0 \cap (\tilde{W}_0)^0)_0 = ((W_0+\tilde{W}_0)^0)_0 = W_0+\tilde{W}_0.\end{aligned}$$

(e) Since $V^0 = \{0\}$ and $\{0\}^0 = V^*$, we have

$$\begin{aligned}U+\tilde{U} = V &\overset{\text{(c)}}{\iff} (U+\tilde{U})^0 = \{0\} \overset{\text{(d)}}{\iff} U^0 \cap \tilde{U}^0 = \{0\},\\ U \cap \tilde{U} = \{0\} &\iff (U \cap \tilde{U})^0 = V^* \iff U^0+\tilde{U}^0 = V^*.\end{aligned}$$

This shows $V = U \oplus \tilde{U} \iff V^* = U^0 \oplus \tilde{U}^0$. □

Corollary 7.45. *Let V be an n-dimensional vector space and $0 \le k \le n$. 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. By Lemma 7.44, $U \mapsto F^{-1}(U^0)$ is a bijection with the desired properties. □

Theorem 7.46. *Let V be an n-dimensional vector space over a field K with $q < \infty$ elements. For $1 \le k \le n$, 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 $U \le V$ 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. Because of linear independence, v_2 must not lie in $\langle v_1\rangle \cong K$. Hence 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 $v_i \in V \setminus \langle v_1,\dots,v_{i-1}\rangle$. Altogether 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 them span the same space.

The same argument with U in place 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. So exactly $D(k,k)$ of the k-tuples in V span the same subspace. Hence the number of k-dimensional subspaces is $\frac{D(n,k)}{D(k,k)}$. The claim follows by canceling all factors q. □

Remark 7.47. At first glance it is not clear why the formula given in Theorem 7.46 is an integer at all. So we have proved an arithmetic statement with the help of linear algebra. Because of

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

we obtain Corollary 7.45 for finite fields.

Example 7.48. 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 7.49. 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* of f.

Theorem 7.50. *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 injective $\iff$ f^* surjective.

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

Proof. For $g_1, g_2 \in W^*$ and $\lambda \in K$ we have

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

by Remark 7.23. This shows $f^* \in \operatorname{Hom}(W^*, V^*)$. For $g \in W^*$ and $f_1, f_2 \in \operatorname{Hom}(V,W)$ we similarly have

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

Hence $f \mapsto f^*$ 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 we have

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

for $j = 1, \dots, m$. It follows that $f^*(c_i^*) = \sum_{k=1}^m a_{ik} b_k^*$ for $i = 1, \dots, n$. This shows ${}_C[f]_B^{\mathrm{t}} = (a_{ji})^{\mathrm{t}} = (a_{ij}) = {}_{B^*}[f^*]_{C^*}$. By 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 7.51. For $f \in \operatorname{Hom}(V,W)$ and $g \in \operatorname{Hom}(W,U)$ we have $(g \circ f)^* = f^* \circ g^*$, 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 study homomorphisms $f\colon V \to V$ from a vector space to itself. One then speaks of *endomorphisms* and writes $\mathrm{End}(V) := \mathrm{Hom}(V, V)$. 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*[2] of f if the *eigenspace*

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

is not the zero space. In this case, $\dim E_\lambda(f)$ 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 $\lambda \in K$ we have

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

Hence the eigenspace $E_\lambda(f) = \mathrm{Ker}(f - \lambda\,\mathrm{id})$ is indeed a subspace of V. By 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 we have

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

Hence $E_\lambda(f)$ can be computed by solving the homogeneous system of linear 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 (w.r.t. different bases), they have the same eigenvalues (but not necessarily 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

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

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

[2] Contrary to what the name suggests, there was no mathematician named Eigen. The word comes from the German *eigen* ("own").

Subtracting $\lambda = 1$ on the main diagonal yields a matrix of rank 1 with three identical rows. Clearly, $b_1 := (1, 0, -1)$ and $b_2 := (0, 1, -1)$ form a basis of $E_1(f)$. In particular, $\lambda = 1$ has geometric 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 computes ${}_B[f \circ f \circ f]_B = {}_B[f]_B^3 = \operatorname{diag}(1, 1, 4)^3 = \operatorname{diag}(1, 1, 64)$.

8.2 Diagonalizability

Definition 8.5. We call $f \in \operatorname{End}(V)$ *diagonalizable* if there exists a basis B of V 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. Clearly, $f \in \operatorname{End}(V)$ is diagonalizable if and only if V has a basis 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 (Remark 7.30).

Example 8.7.

(a) The map from Example 8.4 is diagonalizable.

(b) Diagonal matrices are clearly 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. Since $E_0(A) = \langle (1, 0) \rangle$, there is 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

$$A - \lambda 1_2 = \begin{pmatrix} -\lambda & 2 \\ 1 & -\lambda \end{pmatrix} \begin{matrix} \leftarrow\!\!\rceil^{+} \\ \lrcorner_{\lambda} \end{matrix} \sim \begin{pmatrix} 0 & 2 - \lambda^2 \\ 1 & -\lambda \end{pmatrix}.$$

Since $\sqrt{2} \notin \mathbb{Q}$ (Example 1.11), we have $2 - \lambda^2 \neq 0$. So 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 $U + W$ (or $U \oplus W$) of two subspaces $U, W \leq V$. For $U_1, \ldots, U_n \leq V$ one defines inductively

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

Clearly, $U_1 + \ldots + U_n$ 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 written uniquely 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 $u_1 = \ldots = u_n = 0$.

Proof.

(1) $\Rightarrow$ (2): For $n = 1$, (2) is trivial. By induction we may assume that (2) already holds for $n - 1$. The dimension formula then implies

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

(2) $\Rightarrow$ (3): The given map is always linear and surjective. Since

$$\dim(U_1 \times \ldots \times U_n) \overset{7.11}{=} \dim(U_1) + \ldots + \dim(U_n) = \dim(U_1 + \ldots + U_n)$$

it must also be injective.

(3) $\Rightarrow$ (4): Follows from the injectivity of $U_1 \times \ldots \times U_n \to U_1 + \ldots + U_n$.

(4) $\Rightarrow$ (5): By (4), the two decompositions of the zero vector $u_1 + \ldots + u_n = 0 + \ldots + 0$ must be identical, i. e., $u_1 = \ldots = u_n = 0$.

(5) $\Rightarrow$ (1): For $n = 1$ there is nothing to show. By induction we may assume $U_1 + \ldots + U_{n-1} = U_1 \oplus \ldots \oplus U_{n-1}$, since the hypothesis (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 $u = 0$. Hence (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 suppose (8.1) has already been proved for $k - 1$. Let $v_i \in E_{\lambda_i}(f)$ with $v_1 + \ldots + v_k = 0$. Then we have

$$\begin{aligned} 0 &= f(v_1 + \ldots + v_k) - \lambda_k(v_1 + \ldots + v_k) = \lambda_1 v_1 + \ldots + \lambda_k v_k - \lambda_k v_1 - \ldots - \lambda_k v_k \\ &= (\lambda_1 - \lambda_k)v_1 + \ldots + (\lambda_{k-1} - \lambda_k)v_{k-1} \in E_{\lambda_1}(f) \oplus \ldots \oplus E_{\lambda_{k-1}}(f). \end{aligned}$$

Lemma 8.9 implies $(\lambda_i - \lambda_k)v_i = 0$ for $i = 1, \ldots, k - 1$. Since $\lambda_i \neq \lambda_k$, we have $v_1 = \ldots = v_{k-1} = 0$. Finally, $v_k = v_1 + \ldots + v_k = 0$ as well. Now (8.1) follows from Lemma 8.9. The last claim follows from $\dim E_{\lambda_i}(f) \geq 1$ for $i = 1, \ldots, k$. □

Remark 8.11. Mnemonic: Eigenvectors for 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. By Theorem 8.10, $K^{n \times 1}$ has a basis of eigenvectors of A (one for each eigenvalue). By Remark 8.6, A is diagonalizable. □

Example 8.13.

(a) The identity matrix shows that the converse of Corollary 8.12 is false. In Theorem 10.34 we derive an exact characterization of diagonalizability.

(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. Via a detour through the determinant, we show that the eigenvalues of a matrix are exactly the roots of a polynomial with coefficients in K (Theorem 10.32).

Definition 8.14. A matrix $A = (a_{ij}) \in K^{n\times n}$ is called an (*upper*[3]) *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 in addition $a_{ii} = 0$ for $i = 1, \ldots, n$, one speaks of a *strictly* (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 a shift of the rows occurs in 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 nonzero. Hence $A - \lambda 1_n$ 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.

[3] 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 $f \in \operatorname{End}(V)$) by simpler ones (such as $\operatorname{rk} f$ or $\operatorname{tr} f$) in order to make essential information visible. For example, we saw in Lemma 5.15 that $\operatorname{rk} f$ allows conclusions about the bijectivity of f, provided one knows $\dim V$. Such quantities are called *invariants* if they remain unchanged under "natural" transformations (such as a change of basis). In this chapter we define another invariant, the *determinant* $\det f$. We show that f is bijective if and only if $\det f \neq 0$ (in contrast to $\operatorname{rk} f$, this criterion no longer depends on $\dim V$).[1]

(b) In measure theory one tries to assign a "volume" $\operatorname{vol} S \in \mathbb{R}_{\geq 0}$ 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 when f is applied, i. e., $\operatorname{vol} f(S) = |\det f| \operatorname{vol}(S)$ holds whenever $\operatorname{vol} S$ is defined. The n-dimensional *hypercube*

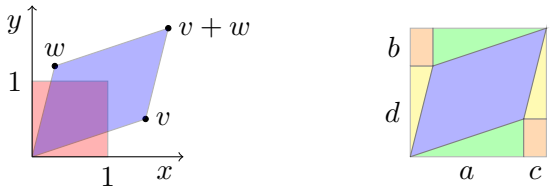
$$H := \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \forall i : 0 \leq x_i \leq 1\big\}$$

is assigned the volume $\operatorname{vol} H = 1$. It follows that

$$|\det f| = \operatorname{vol} f(H).$$

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., $[f] = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$. Das Bild des (roten) Quadrats $H = \{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1\}$ ist das von den Vektoren v und w aufgespannte (blaue) Parallelogramm $f(H)$:



We read off the area of $f(H)$ from the drawing:

$$\operatorname{vol} f(H) = (a + c)(b + d) - 2bc - ab - cd = ad - bc.$$

The area is 0 if and only if v and w lie on a line through the origin, i.e., are linearly dependent. This is equivalent to $\operatorname{rk} f \leq 1$.

[1] From real life: If you have trouble remembering a person's age, remember the year of birth instead, since this invariant does not change every year.

Definition 9.3. Let $A = (a_{ij}) \in K^{n\times n}$ and $1 \le s, t \le n$. Deleting the s-th row and t-th column of A yields 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 $n = 2$ we obtain

$$\det \begin{pmatrix} a & c \\ b & d \end{pmatrix} = a \det(A_{11}) - b \det(A_{21}) = ad - bc$$

(cf. Example 9.2).

(b) For every upper triangular matrix $A = (a_{ij})$ we have $\det A = a_{11} \dots a_{nn}$. This is clear for $n = 1$. Suppose inductively that the claim has already been proved for $n - 1$. 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 every matrix can be transformed into an upper triangular matrix by 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 every row, i. e., for* $a_1, \dots, a_n, b \in K^n$, $\lambda \in K$ *and* $1 \le k \le n$ *we have*

$$\det \begin{pmatrix} a_1 \\ \vdots \\ \lambda a_k + b \\ \vdots \\ a_n \end{pmatrix} = \lambda \det \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \\ a_n \end{pmatrix} + \det \begin{pmatrix} a_1 \\ \vdots \\ b \\ \vdots \\ a_n \end{pmatrix}.$$

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 suppose that the claim has already been proved for $n - 1$. 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 $\det A$. However, this can be confused with a matrix norm (Example 18.61).

$$+\sum_{i\neq k}(-1)^{i+1}a_{i1}\big(\lambda\det(M(a_k)_{i1})+\det(M(b)_{i1})\big)$$
$$=\lambda\sum_{i=1}^{n}(-1)^{i+1}a_{i1}\det(M(a_k)_{i1})+(-1)^{k+1}b_1\det(M(b)_{k1})+\sum_{i\neq k}(-1)^{i+1}a_{i1}\det(M(b)_{i1})$$
$$=\lambda\det(M(a_k))+\det(M(b)).$$ □

Theorem 9.6. *For $A\in K^{n\times n}$ we have:*

(a) Multiplying a row of A by $\lambda\in K$ also multiplies $\det A$ *by λ.*

(b) Swapping two rows of A changes the sign of $\det A$.[3]

(c) Adding a multiple of one row to another row of A does not change $\det A$.

Proof. Let $a_1,\dots,a_n$ be the rows of A.

(a) If we first set $\lambda=1$ and $b=0$ in Lemma 9.5, we see that the determinant vanishes if A has a zero row. The claim now follows by choosing λ arbitrary and $b=0$ in Lemma 9.5.

(b) We first show $\det A=0$ if $a_k=a_{k+1}$ for some k. For $n=1$ there is nothing to show. Let $n\geq 2$. For $i\notin\{k,k+1\}$ the matrix A_{i1} has two adjacent identical rows. Moreover, $a_{k1}=a_{k+1,1}$ and $A_{k1}=A_{k+1,1}$. By induction on n it follows that

$$\det A=\sum_{i=1}^{n}(-1)^{i+1}a_{i1}\det(A_{i1})=a_{k1}\det(A_{k1})-a_{k1}\det(A_{k1})=0.$$

For arbitrary $a_1,\dots,a_n$ we obtain

$$0=\det\begin{pmatrix}\vdots\\ a_k+a_{k+1}\\ a_k+a_{k+1}\\ \vdots\end{pmatrix}\overset{9.5}{=}\det\begin{pmatrix}\vdots\\ a_k\\ a_k\\ \vdots\end{pmatrix}+\det\begin{pmatrix}\vdots\\ a_k\\ a_{k+1}\\ \vdots\end{pmatrix}+\det\begin{pmatrix}\vdots\\ a_{k+1}\\ a_k\\ \vdots\end{pmatrix}+\det\begin{pmatrix}\vdots\\ a_{k+1}\\ a_{k+1}\\ \vdots\end{pmatrix}$$
$$=\det\begin{pmatrix}\vdots\\ a_k\\ a_{k+1}\\ \vdots\end{pmatrix}+\det\begin{pmatrix}\vdots\\ a_{k+1}\\ a_k\\ \vdots\end{pmatrix},$$

i. e., swapping adjacent rows changes the sign of $\det A$. Swapping arbitrary rows a_s and a_t with $s<t$ can be reduced to $2(t-s)-1$ swaps of adjacent rows: $a_s\leftrightarrow a_{s+1}$, $a_s\leftrightarrow a_{s+2},\dots,a_s\leftrightarrow a_t$, $a_t\leftrightarrow a_{t-1},\dots,a_t\leftrightarrow a_{s+1}$.

(c) We add λa_k to row a_l with $k\neq l$ and obtain

$$\det\begin{pmatrix}a_1\\ \vdots\\ a_l+\lambda a_k\\ \vdots\\ a_n\end{pmatrix}\overset{9.5}{=}\det(A)+\lambda\det\begin{pmatrix}\vdots\\ a_k\\ \vdots\\ a_k\\ \vdots\end{pmatrix}.$$

[3]Note: For $K=\mathbb{F}_2$ there are no signs, since $1=-1$.

Hence it suffices to show $\det A = 0$ if A has two identical rows. By (b) we may assume that these rows are adjacent. This case was already dealt with in the proof of (b). $\square$

Example 9.7.

$$\det\begin{pmatrix} -4 & -2 & -2 \\ 6 & 3 & 2 \\ 8 & 7 & 6 \end{pmatrix} \begin{matrix} \mid :(-2) \\ \\ \\ \end{matrix} = -2\det\begin{pmatrix} 2 & 1 & 1 \\ 6 & 3 & 2 \\ 8 & 7 & 6 \end{pmatrix} \begin{matrix} -3 \quad -4 \\ + \\ + \end{matrix}$$

$$= -2\det\begin{pmatrix} 2 & 1 & 1 \\ 0 & 0 & -1 \\ 0 & 3 & 2 \end{pmatrix} = 2\det\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 2 \\ 0 & 0 & -1 \end{pmatrix} \overset{9.4}{=} -12$$

Remark 9.8. For $A \in K^{n\times n}$ and $\lambda \in K$ we have $\boxed{\det(\lambda A) = \lambda^n \det(A),}$ since each of the n rows is multiplied by λ.

Lemma 9.9. *Sei* $A = \left(\begin{smallmatrix} A_1 & A_2 \\ 0 & A_3 \end{smallmatrix}\right) \in K^{(n+m)\times(n+m)}$ *with* $A_1 \in K^{n\times n}$, $A_2 \in K^{n\times m}$ *and* $A_3 \in K^{m\times m}$. *Then* $\det A = \det(A_1)\det(A_3)$.

Proof. If we apply the Gaussian algorithm to A, then first A_1 and then A_3 is 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* $A \in K^{n\times n}$ *we have*

$$A \text{ invertible} \iff \operatorname{rk} A = n \iff \det A \neq 0.$$

Proof. The first equivalence follows from Lemma 5.15. Since the reduced row echelon form $\widehat{A}$ is an upper triangular matrix, we have

$$\det A \neq 0 \overset{9.6}{\iff} \det \widehat{A} \neq 0 \overset{9.4}{\iff} \widehat{A} = 1_n \iff \operatorname{rk} A = \operatorname{rk} \widehat{A} = n. \qquad \square$$

Theorem 9.11 (Multiplicativity of the determinant)**.** *For* $A, B \in K^{n\times n}$ *we have* $\boxed{\det(AB) = \det(A)\det(B).}$

Proof. If $\det A = 0$, then $\operatorname{rk}(AB) \leq \operatorname{rk} A < n$ by Lemma 5.15. Hence $\det(AB) = 0$ by Theorem 9.10. So we may assume $A \in \mathrm{GL}(n, K)$. By 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 we have in each case

$$\det(A_i M) = \left\{\begin{matrix} \lambda\det(M) \\ -\det(M) \\ \det(M) \end{matrix}\right\} = \det(A_i 1_n)\det(M) = \det(A_i)\det(M)$$

by Theorem 9.6. Altogether 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 $A \in K^{n\times n}$ we have $\boxed{\det(A^{\mathrm{t}}) = \det A.}$

(b) For $A \in \mathrm{GL}(n, K)$ we have $\boxed{\det(A^{-1}) = \det(A)^{-1}.}$

(c) Similar matrices have the same determinant.

Proof.

(a) Since $\operatorname{rk} A = \operatorname{rk}(A^{\mathrm{t}})$, we may 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 we have $A_i^{\mathrm{t}} = A_i$. For the third row operation we have $\det A_i = 1 = \det(A_i^{\mathrm{t}})$. 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.

- By the multiplicativity of the determinant 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. We have $\mathrm{SL}(n, \mathbb{F}_2) = \mathrm{GL}(n, \mathbb{F}_2)$.
- Let V be a vector space with basis B. For $f \in \mathrm{End}(V)$ one defines $\det f := \det({}_B[f]_B)$.

Remark 9.14.

(a) Since $\det(A^{\mathrm{t}}) = \det A$, one may also use elementary column operations when computing $\det A$.

(b) By Corollary 7.27 and Corollary 9.12, $\det f$ does not depend on the choice of the basis of V. By Theorem 9.10, f is an isomorphism if and only if $\det f \neq 0$.

(c) The following statement generalizes the multiplicativity of the determinant.

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 $\det(AB^{\mathrm{t}}) = 0$ by Lemma 5.15. So let $n \leq m$. We decompose the i-th row of A as $a_i = a_i' + a_i''$. If we replace a_i by a_i' or a_i'', we obtain matrices A' and A'', respectively, with $\det A_I = \det(A_I') + \det(A_I'')$ for all $I \subseteq \{1, \dots, m\}$ with $|I| = n$ by Lemma 9.5. The i-th row of AB^{t} is $a_i'B^{\mathrm{t}} + a_i''B^{\mathrm{t}}$. Hence also $\det(AB^{\mathrm{t}}) = \det(A'B^{\mathrm{t}}) + \det(A''B^{\mathrm{t}})$. So it suffices to prove the claim for A' and A'', respectively, instead of A. In this way we can ensure that each row of A has at most one nonzero entry. Choose $I \subseteq \{1, \dots, m\}$ with $|I| = n$ and $a_{ij} = 0$ for $1 \leq i \leq n$ and $j \notin I$. For $I' \neq I$ the matrix $A_{I'}$ has a zero column and $\det A_{I'} = 0$. We have $AB^{\mathrm{t}} = A_I(B_I)^{\mathrm{t}}$ and $\det(AB^{\mathrm{t}}) = \det(A_I)\det(B_I)$ by Corollary 9.12. $\square$

9.3 Laplace Expansion

Theorem 9.16 (Laplace expansion)**.** *Let $n \geq 2$ and $A \in K^{n\times n}$. For $1 \leq k \leq n$ 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 $a_1, \ldots, a_n$ be the columns of A. By Remark 9.14 we have

$$\begin{aligned}\det\big(\cdots \overset{\downarrow}{a_{k-1}} \overset{\downarrow}{a_k} \cdots\big) &= -\det\big(\cdots \overset{\downarrow}{a_{k-2}} \overset{\downarrow}{a_k} a_{k-1} \cdots\big) = \ldots \\ &= (-1)^{k-1} \det\big(a_k \quad a_1 \quad \cdots \quad a_{k-1} \quad a_{k+1} \quad \cdots\big) = (-1)^{k-1} \sum_{i=1}^{n} (-1)^{i+1} a_{ik} \det(A_{ik}).\end{aligned}$$

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 the *expansion along the k-th column/row*. The signs $(-1)^{i+k}$ are arranged in a checkerboard pattern:

$$\begin{pmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Example 9.18. Like most recursive methods, the Laplace expansion is usually inefficient. However, it is well suited for so-called *sparse* matrices, i. e., matrices with many zero entries. 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 $a_1, \ldots, a_n \in K$ one defines the product $\prod_{i=1}^{n} a_i = a_1 \cdot \ldots \cdot a_n$ analogously to the summation symbol $\sum$.

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] *We have*

$$\det A = \prod_{1 \leq i < j \leq n} (x_j - x_i).$$

In particular, A is invertible if and only if $x_1, \ldots, x_n$ are pairwise distinct.

[4] In some books one considers the transposed matrix.

Proof. Induction on n: In the case $n = 1$ we have $A = 1_1$ 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 $n \geq 2$. We subtract x_1 times the penultimate column from the last column. Then we subtract x_1 times the $(n-2)$-th column from the penultimate 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 (Remark 9.14). By expanding along the first row, we can pass to the smaller matrix

$$\begin{pmatrix} x_2 - x_1 & (x_2 - x_1)x_2 & \cdots & (x_2 - x_1)x_2^{n-2} \\ \vdots & \vdots & & \vdots \\ x_n - x_1 & (x_n - x_1)x_n & \cdots & (x_n - x_1)x_n^{n-2} \end{pmatrix}$$

instead. 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 $A \in K^{n\times n}$ the matrix

$$\widetilde{A} := \begin{cases} 1_1 & \text{if } n = 1 \\ \big((-1)^{i+j} \det(A_{ji})\big)_{i,j} & \text{if } n > 1 \end{cases} \qquad \in K^{n\times n}$$

is called the *adjugate* of A.[5]

Theorem 9.22. *For all $A \in K^{n\times n}$ we have* $\boxed{A\widetilde{A} = \det(A)1_n = \widetilde{A}A.}$ *In particular, $A^{-1} = \frac{1}{\det A}\widetilde{A}$ if $A \in \mathrm{GL}(n, K)$.*

Proof. For $n = 1$ the claim is clear. So let $n \geq 2$. Let $A_{(k,l)}$ be the matrix obtained from A by replacing the l-th row by the k-th row. For $k \neq l$ the matrix $A_{(k,l)}$ has two identical rows and hence $\det A_{(k,l)} = 0$. On the other hand, $A_{(k,k)} = A$. Let $A\widetilde{A} = (c_{ij})$. Expansion along the l-th row of $A_{(k,l)}$ yields

$$\delta_{kl} \det(A) = \det A_{(k,l)} = \sum_{i=1}^{n} a_{ki}(-1)^{i+l} \det(A_{li}) = c_{kl}.$$

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 $A := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ we obtain

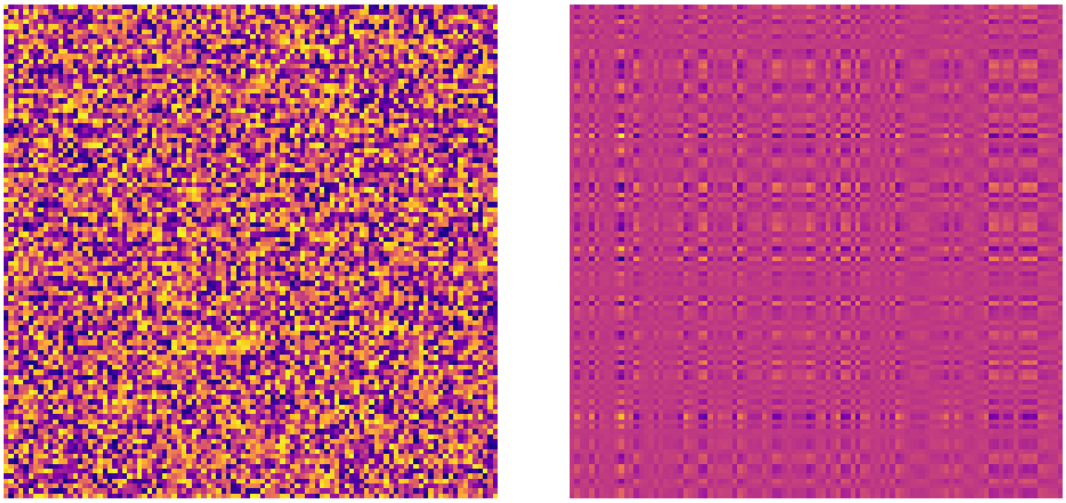
$$A^{-1} = \frac{1}{\det A}\widetilde{A} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

[5] also called the *classical adjoint*; not to be confused with the *adjoint* matrix A^* from Definition 13.12

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 $s = j$ or $t = i$. Hence the inverse of a randomly chosen 100×100 matrix exhibits clear structures:[6]

A $\qquad\qquad$ A^{-1}



(b) For "large" n this formula is of more theoretical than practical significance (use Theorem 6.17 to compute A^{-1}). For example, if $A \in \mathbb{Z}^{n\times n}$, then also $\det(A)A^{-1} = \widetilde{A} \in \mathbb{Z}^{n\times n}$. In particular, $A^{-1} \in \mathbb{Z}^{n\times n}$ if $\det A = \pm 1$. This observation is not apparent from the Gaussian algorithm. Theorem 9.26 gives a similar statement for systems of linear equations.

Theorem 9.25. *For $A, B \in K^{n\times n}$ we have $\widetilde{AB} = \widetilde{B}\widetilde{A}$.*

Proof. W. l. o. g. we may assume $n \geq 2$. Let $A = (a_{ij})$, $B = (b_{ij})$, $\widetilde{AB} = (c_{ij})$ and $\widetilde{B}\widetilde{A} = (d_{ij})$. Deleting the j-th row of A and the i-th column of B, respectively, yields the matrices $A_{j.}$ and $B_{.i}$. Clearly $(AB)_{ji} = A_{j.}B_{.i}$. The Cauchy-Binet formula yields

$$\begin{aligned} c_{ij} &= (-1)^{i+j}\det(A_{j.}B_{.i}) = (-1)^{i+j}\sum_{k=1}^{n}\det(A_{jk})\det(B_{ki}) \\ &= \sum_{k=1}^{n}(-1)^{i+k}\det(B_{ki})(-1)^{k+j}\det(A_{jk}) = d_{ij}. \end{aligned}$$

□

Theorem 9.26 (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 obtained from A by replacing the k-th column by b. For the unique solution $x = (x_1,\ldots,x_n)^{\mathrm{t}}$ of the system of equations $Ax = b$ we then have $x_k = \frac{\det A_k}{\det A}$ 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.

Example 9.27. For the unique solution x of the system

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

we have (expansion along the third row)

$$x_1 = \det \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 1 & 0 & 1 \end{pmatrix} = 1 - 1 = 0.$$

9.4 The Leibniz Formula

Remark 9.28. If one reduces 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 9.29. 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 9.30.

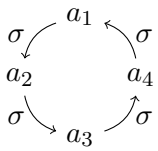
(a) Analogously to $\mathrm{GL}(V)$, S_n is also a group with respect to composition of maps. We will therefore often omit the composition symbol $\circ$. S_n is called the *symmetric group of degree n*.

(b) Let $\sigma \in S_n$. There are n possibilities for the choice of $\sigma(1) \in \{1, \ldots, n\}$. Since σ is injective, we have $\sigma(2) \neq \sigma(1)$. Hence there remain $n-1$ possibilities for the choice of $\sigma(2)$, and so on. In total there are $n!$ possibilities to define a permutation, i. e., $|S_n| = n!$.

Example 9.31.

(a) For $k \geq 2$, an element $\sigma \in S_n$ is called a (k-)*cycle* (or cycle of *length* k) if there exist pairwise distinct $1 \leq a_1, \ldots, a_k \leq n$ 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.,

$$\sigma = (a_2, \ldots, a_k, a_1) = \ldots = (a_k, a_1, \ldots, a_{k-1}).$$

As usual for maps, the composition of cycles is carried out from right to left:

$$(1,3,4,5) \circ (3,4,5) \circ (1,3,2) = (1,5,4)(2,3).$$

Moreover, $(a_1, \ldots, a_k)^{-1} = (a_k, a_{k-1}, \ldots, a_1)$.

(b) Cycles of length 2 are called *transpositions*. Thus a transposition swaps two elements and fixes all other elements. Every k-cycle is a composition of $k-1$ transpositions:

$$(a_1, \dots, a_k) = (a_1, a_2)(a_2, a_3) \dots (a_{k-1}, a_k).$$

However, there are many possibilities for such a composition.

(c) In S_3, every nontrivial element is a cycle: $S_3 = \{\mathrm{id}, (1,2), (1,3), (2,3), (1,2,3), (1,3,2)\}$.

(d) Cycles $\sigma = (a_1, \dots, a_k)$ and $\tau = (b_1, \dots, b_l)$ are called *disjoint* if

$$\{a_1, \dots, a_k\} \cap \{b_1, \dots, b_l\} = \varnothing.$$

In that case $\sigma\tau = \tau\sigma$.

Theorem 9.32. *Every permutation $\sigma \in S_n$ is a composition of pairwise disjoint cycles $\sigma = \sigma_1 \dots \sigma_k$. Here $\sigma_1, \dots, \sigma_k$ are uniquely determined up to 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 $i \geq 2$. Since $n < \infty$, there exist $i < j$ with $a_i = a_j$, i. e., $\sigma^{j-i}(a_1) = a_1$. Hence there exists

$$s := \min\{1 \leq i \leq n : \sigma^i(a_1) = a_1\}.$$

The elements $a_1, \dots, a_s$ are pairwise distinct. Thus $\sigma_1 := (a_1, \dots, a_s)$ is an s-cycle. For $\rho := \sigma\sigma_1^{-1}$ we have

$$\rho(x) = \begin{cases} x & \text{if } x \in \{a_1, \dots, a_s\}, \\ \sigma(x) & \text{otherwise.} \end{cases}$$

In the case $\rho = \mathrm{id}$ we have $\sigma = \sigma_1$ and we are done. Otherwise we can repeat the procedure with ρ instead of σ. Since the number of fixed points of ρ grows with each repetition, after finitely many steps one reaches $\sigma = \sigma_1 \dots \sigma_k$ with pairwise disjoint cycles $\sigma_1, \dots, \sigma_k$.

Let also $\sigma = \tau_1 \dots \tau_l$ be a composition of pairwise disjoint cycles $\tau_1, \dots, \tau_l$. Then there exists an i with $\tau_i(a_1) \neq a_1$. Since the cycles are disjoint, we have $\tau_i(a_1) = \sigma(a_1) = \sigma_1(a_1) = a_2$, $\tau_i(a_2) = \sigma_1(a_2) = a_3$ and so on. Hence $\tau_i = \sigma_1$ and $\sigma_2 \dots \sigma_k = \tau_1 \dots \tau_{i-1}\tau_{i+1} \dots \tau_l$. The uniqueness of the σ_i now follows by induction on k. □

Remark 9.33.

(a) Since every cycle is a composition of transpositions by Example 9.31, every permutation is even a composition of (in general not disjoint) transpositions.

(b) One can make the notation in disjoint cycles

$$\sigma = (a_1, \dots, a_s)(b_1, \dots, b_t) \dots$$

completely unique by requiring $a_1 = \min\{a_1, \dots, a_s\} < b_1 = \min\{b_1, \dots, b_t\} < \dots$.

Definition 9.34. For $\sigma \in S_n$ the matrix

$$P_\sigma := (\delta_{i\sigma(j)})_{i,j} \in \mathbb{Q}^{n \times n}$$

is called the *permutation matrix* of σ. Moreover, $\operatorname{sgn}\sigma := \det P_\sigma$ is called the *signum* or *sign* of σ.

Remark 9.35. The permutation matrix of σ is obtained by permuting the rows of the identity matrix (that is, the standard basis $e_1, \ldots, e_n$) according to σ. With the Gaussian algorithm, this permutation can be realized by finitely many row interchanges (this corresponds to the sorting algorithm *selection sort*). Since $\det 1_n = 1$, $\operatorname{sgn}\sigma \in \{\pm 1\}$ is indeed a "sign".

Example 9.36. We determine the permutation matrix and the signum for the permutations in S_3:

σ	id	$(1,2)$	$(1,3)$	$(2,3)$	$(1,2,3)$	$(1,3,2)$
P_σ	1_3	$\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix}$	$\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}$	$\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}$	$\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}$	$\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}$
$\operatorname{sgn}\sigma$	1	-1	-1	-1	1	1

Theorem 9.37. *For $\sigma, \tau \in S_n$ we have $P_{\sigma\circ\tau} = P_\sigma P_\tau$ 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 multiplicativity of the determinant. □

Remark 9.38.

(a) The permutation matrix of a transposition is precisely the elementary matrix for interchanging rows. In particular, every transposition has signum -1. By Theorem 9.37, a product of an even number of transpositions is never a product of an odd number of transpositions.

(b) By Example 9.31, every l-cycle has signum $(-1)^{l-1}$. 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 fixed points as cycles of length 1, the formula simplifies to $\operatorname{sgn}\sigma = (-1)^{n-k}$.

(c) By Theorem 9.37, the permutations with signum 1 form a subgroup $A_n \leq S_n$. A_n is called the *alternating* group of degree n. In a certain sense, A_n is to S_n as $\mathrm{SL}(n,K)$ is to $\mathrm{GL}(n,K)$.

Theorem 9.39 (Leibniz formula)**.** *For $A = (a_{ij}) \in K^{n\times n}$ we have*

$$\boxed{\det A = \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)}a_{2\sigma(2)}\ldots a_{n\sigma(n)}.}$$

Proof. The rows $a_1, \dots, a_n$ of A can be expressed as linear combinations of the standard basis: $a_i = \sum_{j=1}^{n} a_{ij} e_j$. Since det is linear in each row (Lemma 9.5), we have

$$\det A = \sum_{i_1=1}^{n} a_{1i_1} \det \begin{pmatrix} e_{i_1} \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \sum_{i_1=1}^{n} a_{1i_1} \sum_{i_2=1}^{n} a_{2i_2} \det \begin{pmatrix} e_{i_1} \\ e_{i_2} \\ a_3 \\ \vdots \end{pmatrix} = \dots$$

$$= \sum_{1 \le i_1, \dots, i_n \le n} a_{1i_1} \dots a_{ni_n} \det \begin{pmatrix} e_{i_1} \\ \vdots \\ e_{i_n} \end{pmatrix}.$$

If there exist $s \neq t$ with $i_s = i_t$, the corresponding determinant vanishes. So we only need to sum over the tuples $(i_1, \dots, i_n)$ with pairwise distinct entries. Each such tuple describes a permutation $\sigma \in S_n$ with $\sigma(j) = i_j$ for $j = 1, \dots, n$. It follows that

$$\det A = \sum_{\sigma \in S_n} a_{1\sigma(1)} \dots a_{n\sigma(n)} \det(P_\sigma^{\mathrm{t}}) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)} \dots a_{n\sigma(n)}. \qquad \square$$

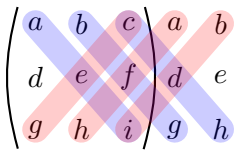
Corollary 9.40 (Rule of Sarrus)**.** *For 3×3 matrices we have*

$$\boxed{\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = aei + bfg + cdh - gec - hfa - idb.}$$

Proof. Use the Leibniz formula with Example 9.36. $\square$

Remark 9.41.

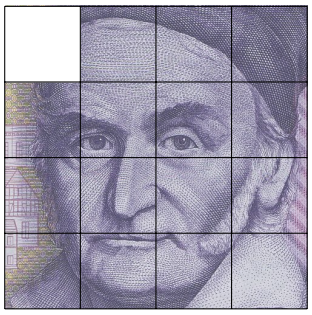
(a) One can remember the rule of Sarrus with the following scheme:

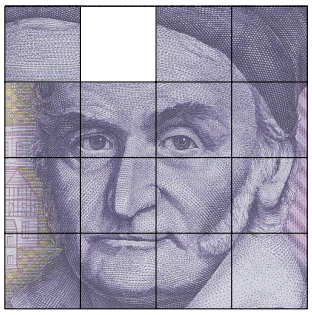


(b) Caution: The rule of Sarrus holds *only* for 3×3 matrices (for 4×4 matrices one needs $4! = 24$ summands).

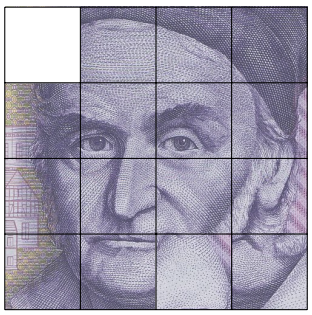
(c) Let $A \in K^{n \times n}$ with eigenvalue λ. Then $E_\lambda(A) = \operatorname{Ker}(A - \lambda 1_n) \neq \{0\}$ and $\det(A - \lambda 1_n) = 0$. By the Leibniz formula, $\det(A - \lambda 1_n)$ is a polynomial in λ. In this way we will compute all eigenvalues of A.

Example 9.42. The following *sliding puzzle* consists of 15 movable squares and one empty space. A square that borders the empty space horizontally or vertically may be slid onto it (cf. the cover):





Sam Loyd offered a prize of 1000 $ to anyone who could transform the following configuration into the initial state:[7]



Every move corresponds to a transposition in S_{16}. If we take a checkerboard pattern as a basis, the empty square moves from black to white or vice versa with every move. Since the empty square lies in its initial position in Loyd's configuration, an even number of moves is needed for a solution. On the other hand, Loyd's configuration differs from the initial state by only one transposition. By Remark 9.38, this configuration is therefore unsolvable and Loyd never had to pay out the prize money.

[7] see [J. Slocum and D. Sonneveld, *The 15 puzzle*, The Slocum Puzzle Foundation, Beverly Hills, 2006]

Exercises

Exercise I.1. Which of the following statements 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. They know about each other's knowledge. The two have 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 we have $|A \cup B| = |A| + |B| - |A \cap B|$ by 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. 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.6. Prove that $\mathbb{N} \times \mathbb{N}$ is countable.

Exercise I.7. Let $U := \{2z + 1 : z \in \mathbb{Z}\} \cup \{0\}$. Determine 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 nonempty subset H of a group G is a subgroup if and only if $xy^{-1} \in H$ for all $x, y \in H$.

(b) A nonempty subset U of a K-vector space V is a subspace if and only if $\lambda u + v \in U$ for all $u, v \in U$ and $\lambda \in K$.

Exercise I.11 (Correspondence theorem)**.** Let $U \leq V$ be vector spaces.

(a) Show that every subspace of V/U has the form W/U with $U \leq W \leq V$.

(b) Show that the map $W \mapsto W/U$ is a bijection between $\{W : U \leq W \leq V\}$ and the set of subspaces of V/U.

Exercise I.12 (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.13. 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.14. Let u, v, w be vectors in 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.15. Clearly $\mathbb{R}$ is a $\mathbb{Q}$-vector space in which 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.16. Let $n = n_1 + \ldots + n_k$ and let $\lambda_1, \ldots, \lambda_k \in K$ 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) We have $AB = BA$ 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) We have $AB = BA$ for all $B \in K^{n \times n}$ if and only if A is a scalar matrix (i. e., $k = 1$).

Exercise I.17. Hannover wants to found as many clubs as possible because of the prospect of funding. To always guarantee a majority in votes, every club shall have an odd number of members. Two distinct clubs, on the other hand, shall always have an even number of common members. Show:

(a) There cannot be more clubs than residents.
Hint: Construct $A \in \mathbb{F}_2^{n \times m}$ with $AA^{\mathrm{t}} = 1_n$.

(b) It is always possible to found as many clubs as there are residents. Note that by §56 BGB a club needs at least seven members.
Hint: Construct $A = (1_m + v^{\mathrm{t}}v)(1_m + w^{\mathrm{t}}w)$ with $v, w \in \mathbb{F}_2^m$.

Exercise I.18. Show:

(a) Interchanging 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.19. Let $A \in \mathrm{GL}(n, K)$. Show that the matrix $\left(\begin{smallmatrix} A \\ 1_n \end{smallmatrix}\right) \in K^{2n \times n}$ can be transformed by elementary column operations into the form $\left(\begin{smallmatrix} 1_n \\ B \end{smallmatrix}\right)$. Here $B = A^{-1}$.

Exercise I.20. Let $A \in K^{n \times m}$. Show that $\operatorname{rk} A$ is the smallest number r such that there exist vectors $v_1, \ldots, v_r \in K^n$ and $w_1, \ldots, w_r \in K^m$ with $A = v_1^{\mathrm{t}} w_1 + \ldots + v_r^{\mathrm{t}} w_r$.

Exercise I.21. Let $A \in \mathbb{Q}^{n \times m}$ and $b \in \mathbb{Q}^{n \times 1}$. 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) The solution sets of $Ax = b$ over $\mathbb{Q}$ and $\mathbb{R}$ can be different.

Exercise I.22. Let V, W be vector spaces and $f \in \operatorname{Hom}(V, W)$. Show:

(a) f is injective if and only if there exists a $g \in \operatorname{Hom}(W, V)$ with $g \circ f = \mathrm{id}_V$.

(b) f is surjective if and only if there exists a $g \in \operatorname{Hom}(W, V)$ with $f \circ g = \mathrm{id}_W$.

(c) Are the maps g uniquely determined in each case?

Exercise I.23. Let U, V, W, X be vector spaces and let $f\colon U \to V$, $g\colon V \to W$ 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 $g(V) = g(f(U)) \oplus Y$ we have $h(g(V)) = h\big(g(f(U))\big) + h(Y)$.

(b) Deduce Lemma 5.15(a) from part (a).

(c) Deduce $\operatorname{rk}(A) + \operatorname{rk}(B) \leq \operatorname{rk}(AB) + n$ for $A \in K^{m \times n}$ and $B \in K^{n \times k}$ from part (a) (Sylvester inequality).

Exercise I.24. Let K be a field and $n \in \mathbb{N}$. Show:

(a) The sum and the product of (strictly) upper (or lower) triangular matrices in $K^{n \times n}$ are again (strictly) upper (or lower) triangular matrices (Definition 8.14).

(b) The upper (or lower) triangular matrices form a subspace U of $K^{n \times n}$. Compute $\dim U$.

(c) The set of invertible upper (or lower) triangular matrices in $K^{n \times n}$ is a subgroup of $\mathrm{GL}(n, K)$.

Exercise I.25. 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)
$$\mathrm{Aff}(V) := \{f_{\varphi,v} : \varphi \in \mathrm{GL}(V), v \in V\} \subseteq \mathrm{Map}(V,V)$$
is a group with respect to 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.26. 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.27. Let $A \in K^{n\times n}$ be diagonalizable. Show that A^{t} is also diagonalizable and has the same eigenvalues. Do the eigenspaces also coincide?

Exercise I.28. 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: Add columns 2 to n to the first one and then perform row operations.

Exercise I.29 (Generalized Laplace expansion)**.** Let $A \in K^{r\times s}$, $B \in K^{s\times t}$, $I \subseteq \{1,\dots,r\}$ and $J \subseteq \{1,\dots,t\}$ with $|I| = |J| = k$. For a matrix $M = (m_{ij})$ let $M_{IJ} := (m_{ij})_{i\in I, j\in J}$. Show:
$$\det\big((AB)_{IJ}\big) = \sum_{\substack{L\subseteq\{1,\dots,s\} \\ |L|=k}} \det(A_{IL}) \det(B_{LJ}).$$
Remark: This generalizes matrix multiplication, the multiplicativity of the determinant, and the Cauchy-Binet formula.

Exercise I.30. Show for $A \in K^{n\times n}$:

(a) $\det \widetilde{A} = \det(A)^{n-1}$, where $0^0 = 1$.

(b) $\widetilde{A^{\mathrm{t}}} = (\widetilde{A})^{\mathrm{t}}$.

(c) $\widetilde{\widetilde{A}} = \det(A)^{n-2} A$ for $n \geq 2$ (even if $\det A = 0$).
Hint: Case distinction on $\mathrm{rk}\, A$.

Exercise I.31. Transpositions of the form $(a, a+1) \in S_n$ are called *adjacent transpositions*. Show that every permutation is a product of adjacent transpositions.
Remark: This is the basis of *bubble sort*.

Exercise I.32. Show $|A_n| = \frac{n!}{2}$ 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.33. Let $\sigma, \tau \in S_n$ with $\operatorname{sgn}\sigma \neq \operatorname{sgn}\tau$. Show $\det(P_\sigma + P_\tau) = 0$.
Hint: $P_\sigma + P_\tau = P_\sigma(P_\sigma^{-1} + P_\tau^{-1})P_\tau$.

Exercise I.34. Let $\sigma \in S_n$. Show

$$\operatorname{sgn}\sigma = \prod_{1 \leq i < j \leq n} \frac{\sigma(j) - \sigma(i)}{j - i}.$$

Hint: First consider transpositions σ.

Exercise I.35. Is the sliding puzzle on the cover solvable?

Exercise I.36. In the game *Lights Out*, a 5×5 grid of lights is given, each of which can be on or off. Touching a light toggles the state of that light together with its horizontal and vertical neighbors (above/below, right/left) (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 has to be touched.

(b) Model the solution of the game as a system of equations with coefficient matrix in $\mathbb{F}_2^{25 \times 25}$.

(c) Check (with 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 lit light are solvable?

(g) Develop an easy-to-learn algorithm for solving the game that needs no 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 9.41 we hinted that the eigenvalues of a matrix A are solutions of certain (nonlinear) polynomial equations. In this chapter we define polynomials with coefficients in an arbitrary field and study 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

$$\alpha = \sum_{k=0}^{d} a_k X^k = a_0 + a_1 X + \ldots + a_n X^n$$

with *coefficients* $a_0, \ldots, a_n \in K$.

- One calls a_0 the *constant term* of α.
- If not all coefficients are 0, one calls

  $$d := \deg \alpha := \max\{k \in \mathbb{N}_0 : a_k \neq 0\}$$

  the *degree* of α and a_d the *leading* coefficient.[1] If $a_d = 1$, then 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 not known, 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 nonzero.

(b) Polynomials are regarded as 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, we have $0, 1 \in K \subseteq K[X]$.

Example 10.4. The polynomial $\alpha = X^2 - 3X + 1 \in \mathbb{Q}[X]$ is monic of degree 2 with constant term 1.

[1] also called *lead coefficient*

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. Formally, a polynomial is nothing but a map $\mathbb{N}_0 \to K$, $k \mapsto a_k$ with $|\{k \in \mathbb{N}_0 : a_k \neq 0\}| < \infty$. One easily sees that sums and scalar multiples of polynomials are again polynomials. With the given operations, $K[X]$ therefore becomes a subspace of $\mathrm{Map}(\mathbb{N}_0, K)$. By definition, every polynomial is a finite linear combination of $1, X, X^2, \ldots$, i. e., $K[X] = \langle 1, X, X^2, \ldots \rangle$. The uniqueness of the coefficients (Remark 10.3) implies the linear independence of $\{1, X, X^2, \ldots\}$. □

Remark 10.6.

(a) For $\alpha, \beta \in K[X]$ and $\lambda \in K$ we clearly have $\deg(\alpha + \beta) \leq \max\{\deg \alpha, \deg \beta\}$ and $\deg(\lambda\alpha) \leq \deg \alpha$. Hence 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 rules of computation hold:

$$\alpha\beta = \beta\alpha, \qquad \alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma, \qquad \alpha(\beta\gamma) = (\alpha\beta)\gamma.$$

Proof. If $\alpha = 0$ or $\beta = 0$, then $\alpha\beta = 0$ 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 $\gamma = \sum c_k X^k$ we have

$$\alpha\beta = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l b_{k-l} \Big) X^k = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} b_l a_{k-l} \Big) X^k = \beta\alpha$$
$$\alpha(\beta + \gamma) = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l (b_{k-l} + c_{k-l}) \Big) X^k = \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l b_{k-l} \Big) X^k + \sum_{k=0}^{\infty} \Big(\sum_{l=0}^{k} a_l c_{k-l} \Big) X^k = \alpha\beta + \alpha\gamma.$$

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 $(\alpha\beta)\gamma$. Hence $\alpha(\beta\gamma) = (\alpha\beta)\gamma$. □

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 with respect to multiplication. For example, there is no $\alpha \in K[X]$ with $X \cdot \alpha = 1$. Nevertheless, the cancellation law holds: $\alpha\beta = \alpha\gamma \Rightarrow \beta = \gamma$ if $\alpha \neq 0$. This follows from

$$\deg(\beta - \gamma) \leq \deg(\alpha) + \deg(\beta - \gamma) = \deg(\alpha(\beta - \gamma)) = \deg 0 = -\infty.$$

Thus one can compute in $K[X]$ as in $\mathbb{Z}$.

Theorem 10.9 (Division with remainder)**.** *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 contradicts the choice of γ. Hence $d < e$ and $\alpha = \beta\gamma + \delta$.

Uniqueness: Now let $\alpha = \beta\tilde{\gamma} + \tilde{\delta}$ with $\tilde{\gamma}, \tilde{\delta} \in K[X]$ and $\deg \tilde{\delta} < e$. By Theorem 10.7 we have

$$e + \deg(\tilde{\gamma} - \gamma) = \deg(\beta) + \deg(\tilde{\gamma} - \gamma) = \deg(\beta(\tilde{\gamma} - \gamma)) = \deg(\delta - \tilde{\delta}) \leq \max\{\deg \delta, \deg \tilde{\delta}\} < e.$$

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. If $\delta = 0$, then β is called a *divisor* of α, and one writes $\beta \mid \alpha$ (in addition, let $0 \mid 0$). In this case, one also says "β *divides* α" or "α is *divisible* by β".

Example 10.11.

$$\begin{array}{rrll} (2X^3 & -X^2 & +5X & +1) : (X^2 + 3) = 2X - 1 =: \gamma \\ -(2X^3 & & +6X) & \\ & -X^2 & -X & +1 \\ & -(-X^2 & & -3) \\ & & -X & +4 =: \delta \end{array}$$

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 done efficiently with 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 compute $c_n := 0$, $c_k := a_{k+1} + b c_{k+1}$ for $k = n-1, \ldots, 0$ and $d := a_0 + b c_0$:

$$\begin{array}{rccccc} & a_n & a_{n-1} & a_{n-2} & \cdots & a_0 \\ + & 0 & bc_{n-1} & bc_{n-2} & \cdots & bc_0 \\ \hline & c_{n-1} & c_{n-2} & \cdots & c_0 & d \end{array}$$

[2] also called RUFFINI's *rule*

For $\gamma := c_{n-1}X^{n-1} + \ldots + c_0$ we now have

$$\begin{aligned}\beta\gamma + d &= c_{n-1}X^n + (c_{n-2} - bc_{n-1})X^{n-1} + \ldots + (c_0 - bc_1)X - bc_0 + d \\ &= a_nX^n + a_{n-1}X^{n-1} + \ldots + a_0 = \alpha.\end{aligned}$$

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:

$$\alpha(x) := \sum_{k=0}^{d} a_kx^k \in K.^3$$

One calls x a *root* of α if $\alpha(x) = 0$.

Lemma 10.15. *For $\alpha, \beta \in K[X]$ and $x \in K$ we have*

$$\begin{aligned}(\alpha + \beta)(x) &= \alpha(x) + \beta(x), \\ (\alpha\beta)(x) &= \alpha(x)\beta(x).\end{aligned}$$

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 substitute, or first substitute and then add/multiply. Caution: Neither $\alpha(x + y) = \alpha(x) + \alpha(y)$ nor $\alpha(xy) = \alpha(x)\alpha(y)$ holds 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 $y_1, \ldots, y_n \in K$ 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:

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

[3] We agree that $0^0 := 1$.

The coefficient matrix of this system of equations is a Vandermonde matrix. Since the x_i are pairwise distinct, the matrix is invertible by Theorem 9.20. Hence there exists exactly one solution $(a_0, \ldots, a_{n-1})$. □

Remark 10.18. An explicit solution of the interpolation problem is given by the LAGRANGE *polynomial*

$$\alpha := \sum_{i=1}^{n} y_i \prod_{j \neq i} \frac{X - x_j}{x_i - x_j} \in K[X]$$

(check by computation). One can read off $\deg \alpha < n$ directly.

Example 10.19.

(a) For $K^n = \mathbb{R}^2$, Theorem 10.17 corresponds to the geometric statement that two distinct points in $\mathbb{R}^2$ are connected by exactly one line (vertical lines are, however, not described by polynomials).

(b) We look 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 $(1, 1/2, 3/2)$, i. e., $\alpha = 1 + \frac{1}{2}X + \frac{3}{2}X^2$.

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 that α has pairwise distinct roots $x_1, \ldots, x_{d+1} \in K$. By Theorem 10.17 with $y_1 = \ldots = y_{d+1} = 0$, α is the only polynomial of degree $\leq d$ with these roots. On the other hand, the zero polynomial also has these roots. Hence $\alpha = 0$ and $d = -\infty$, contradicting the hypothesis.

(b) Since $|K| = \infty$, $\alpha - \beta$ has infinitely many roots. By (a) we obtain $\alpha = \beta$. □

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 polynomial and function. Over finite fields K one would lose information in doing so, because there are only finitely many maps $K \to K$, 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. *An element $x \in K$ is a root of $\alpha \in K[X]$ if and only if $(X - x) \mid \alpha$.*

Proof. Division with remainder 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 $\alpha \neq 0$. Then $X - x$ is called 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. If $e = 1$, 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. For $n = 1$, 1 (the empty product with $k = 0$) and $X - x_1$ are the only monic divisors. So let $n \geq 2$ and assume that the claim has already been proved for $n - 1$. Let $\alpha, \beta \in K[X]$ be monic with $(X - x_1) \ldots (X - x_n) = \alpha\beta$. Then $\alpha(x_n)\beta(x_n) = (\alpha\beta)(x_n) = 0$, w. l. o. g. let $\alpha(x_n) = 0$. By Lemma 10.22 we have $\alpha = (X - x_n)\gamma$ for some $\gamma \in K[X]$. By Remark 10.8 we may cancel $X - x_n$ and obtain $(X - x_1) \ldots (X - x_{n-1}) = \gamma\beta$. The claim 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 of the equation

$$x^3 + x^2 - 5x + 3 = 0.$$

Even though there are solution formulas for such equations (of degree three and four[4]), they are laborious in practice. We will therefore choose our examples (and exercises) such that one can guess "small" integer roots. Suppose that there is a root $x \in \mathbb{Z}$. Since $x(x^2 + x - 5) = -3$, 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 we now have $\alpha(y) = (y - 1)\gamma(y) = 0$, i. e., y is also a root of α. With the quadratic formula $\frac{1}{2}(-p \pm \sqrt{p^2 - 4q})$ for quadratic equations one obtains the roots of γ:

$$x_2 = \frac{1}{2}\big(-2 + \sqrt{4 + 12}\big) = 1, \qquad x_3 = \frac{1}{2}\big(-2 - \sqrt{4 + 12}\big) = -3.$$

Hence $x_1 = x_2 = 1$ is a root of α with algebraic multiplicity 2 (a *double* root). Moreover, α *splits* into linear factors $\alpha = (X - 1)^2(X + 3)$.

(b) Clearly, $\alpha(0)$ is the constant term of $\alpha \in K[X]$. Hence $x = 0$ 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. Later we construct a "larger" field over which this polynomial also splits into linear factors (Remark 11.33).

(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 possible.

[4] see the algebra lecture notes

10.3 Characteristic Polynomials

Remark 10.26. In the following we consider matrices with entries in $K[X]$. By the rules for polynomials (Theorem 10.7) it is easy to see that the usual rules (Lemma 5.8) for matrices also hold in $K[X]^{n\times n}$. Moreover, we can even apply the definition of the determinant to matrices in $K[X]^{n\times n}$ (here matrix entries are only added and multiplied, but never divided). Likewise, the multiplicativity of the determinant, Laplace expansion, the Leibniz formula, and Theorem 9.22 on the adjugate remain true in this greater generality. On the other hand, the Gaussian algorithm does not work in $K[X]^{n\times n}$, since division is required here.

Definition 10.27. For $A = (a_{ij}) \in K^{n\times n}$ we consider the matrix

$$X1_n - A = \begin{pmatrix} X - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & X - a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -a_{n-1,n} \\ -a_{n1} & \cdots & -a_{n,n-1} & X - a_{nn} \end{pmatrix} \in K[X]^{n\times n}.$$

We call $\chi_A := \det(X1_n - A) \in K[X]$ the *characteristic polynomial* of A.[5]

Lemma 10.28. *Similar matrices have the same characteristic polynomial.*

Proof. For $A \in K^{n\times n}$ and $S \in \mathrm{GL}(n, K)$ we have

$$\chi_{SAS^{-1}} = \det(X1_n - SAS^{-1}) = \det(S(X1_n - A)S^{-1}) \overset{9.12}{=} \det(X1_n - A) = \chi_A. \qquad \square$$

Definition 10.29. Let V be a vector space with basis B. For $f \in \mathrm{End}(V)$ one defines $\chi_f := \chi_{B[f]_B}$. By Corollary 7.27 and Lemma 10.28, χ_f does not depend on the choice of B. In the following theorems we can therefore replace matrices by endomorphisms (and vice versa).

Example 10.30. The characteristic polynomial of $A := \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right) \in \mathbb{Q}^{2\times 2}$ 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 $A \in K^{n\times n}$ we have $\chi_A = X^n - \mathrm{tr}(A)X^{n-1} + \ldots + (-1)^n \det(A)$. In particular, χ_A is monic of degree n.*

Proof. Let $A = (a_{ij})$. By the Leibniz formula we have

$$\begin{aligned} \chi_A = \det(X1_n - A) &= (X - a_{11})(X - a_{22}) \ldots (X - a_{nn}) \\ &+ \sum_{\sigma \in S_n \setminus \{\mathrm{id}\}} \mathrm{sgn}(\sigma)(\delta_{1\sigma(1)}X - a_{1\sigma(1)}) \ldots (\delta_{n\sigma(n)}X - a_{n\sigma(n)}). \end{aligned}$$

[5] In some books, χ_A is defined by $\det(A - X1_n) = (-1)^n \det(X1_n - A)$. This makes no big difference, but has the drawback that χ_A is not monic if n is odd.

For $\sigma \in S_n \setminus \{\mathrm{id}\}$ there exists $k \in \{1, \ldots, n\}$ with $l := \sigma(k) \neq k$. Since σ is injective, we have $\sigma(l) \neq \sigma(k) = l$. Hence $\delta_{k\sigma(k)} = 0 = \delta_{l\sigma(l)}$ 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 $\leq n-2$. Altogether,

$$\chi_A = (X - a_{11})(X - a_{22}) \ldots (X - a_{nn}) + \alpha$$

with $\deg \alpha \leq n-2$. Expanding shows $\chi_A = X^n - (a_{11} + \ldots + a_{nn})X^{n-1} + \ldots = X^n - \mathrm{tr}(A)X^{n-1} + \ldots$. To compute the constant term, we substitute $X = 0$ and obtain $\chi_A(0) \overset{10.15}{=} \det(-A) = (-1)^n \det(A)$ from Remark 9.8 (for $n = 1$ we have $\det A = \mathrm{tr}\, A$). □

Theorem 10.32. *The eigenvalues of $A \in K^{n\times n}$ are the roots of χ_A.*

Proof. We have

$$\mathrm{Ker}(A - \lambda 1_n) \neq \{0\} \iff \det(A - \lambda 1_n) = 0 \overset{9.8}{\iff} \det(\lambda 1_n - A) = 0 \overset{10.15}{\iff} \chi_A(\lambda) = 0. \qquad \square$$

Lemma 10.33. *Let $\lambda \in K$ be an eigenvalue of $f \in \mathrm{End}(V)$. Then the geometric multiplicity of λ is at most the algebraic multiplicity of λ as a root of χ_f.*

Proof. Extend a basis $b_1, \ldots, b_e$ of $E_\lambda(f)$ to a basis $B := \{b_1, \ldots, b_n\}$ of V. Then

$$\chi_f = \det(X1_n - {}_B[f]_B) = \det \begin{pmatrix} (X - \lambda)1_e & * \\ 0 & * \end{pmatrix}.$$

Lemma 9.9 shows $\chi_f = (X - \lambda)^e \beta$ for some $\beta \in K[X]$. Hence the algebraic multiplicity of λ as a root of χ_f is at least e. □

Theorem 10.34. *Then $f \in \mathrm{End}(V)$ 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 $m_1, \ldots, m_k$. By successively splitting off linear factors we obtain a monic $\alpha \in K[X]$ 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. By Lemma 10.33 we have

$$\dim V = \deg \chi_f \overset{10.7}{=} m_1 + \ldots + m_k + \deg(\alpha) \geq m_1' + \ldots + m_k'$$

with equality if and only if V has a basis of eigenvectors (Theorem 8.10). 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 $i = 1, \ldots, k$). □

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:

$$\boxed{\begin{aligned} \mathrm{tr}\, A &= \lambda_1 + \ldots + \lambda_n, \\ \det A &= \lambda_1 \ldots \lambda_n, \end{aligned}}$$

i. e., the trace is the sum of the eigenvalues and the determinant is the product of the eigenvalues (provided these exist). If $\lambda_1, \ldots, \lambda_{n-1}$ have already been determined, then $\lambda_n = \mathrm{tr}(A) - \lambda_1 - \ldots - \lambda_{n-1}$.

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 let $A \in K^{n\times n} \setminus K1_n$ be a triangular matrix with main diagonal $\lambda_1, \ldots, \lambda_n$. By Example 8.15, A has eigenvalues $\lambda_1, \ldots, \lambda_n$. By 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. □

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

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

Changing 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

$$\begin{pmatrix} F_{k+1} \\ F_k \end{pmatrix} = A \begin{pmatrix} F_k \\ F_{k-1} \end{pmatrix} = A^2 \begin{pmatrix} F_{k-1} \\ F_{k-2} \end{pmatrix} = \ldots = A^k \begin{pmatrix} F_1 \\ F_0 \end{pmatrix} = A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Because of

$$\chi_A = (X-1)X - 1 = X^2 - X - 1$$

the matrix A has the eigenvalues $\varphi := \frac{1+\sqrt{5}}{2}$ and $\psi := \frac{1-\sqrt{5}}{2}$ (one calls $\varphi = 1.618\ldots$ the *golden ratio*). Clearly, $v := (\varphi, 1)^{\mathrm{t}}$ is an eigenvector for φ and $w := (\psi, 1)^{\mathrm{t}}$ is an eigenvector for ψ. We have

$$\begin{pmatrix} F_{k+1} \\ F_k \end{pmatrix} = A^k \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{5}} A^k (v - w) = \frac{1}{\sqrt{5}} (\varphi^k v - \psi^k w) = \frac{1}{\sqrt{5}} \begin{pmatrix} * \\ \varphi^k - \psi^k \end{pmatrix}.$$

We obtain

$$\boxed{F_k = \frac{1}{\sqrt{5}} (\varphi^k - \psi^k).} \qquad \text{(BINET } \textit{formula}^6\text{)}$$

Since $|\psi^k| \approx 0.6^k \to 0$, we have $F_k \approx \frac{1}{\sqrt{5}} \varphi^k$, i. e., F_k grows exponentially.

Lemma 10.38. *For $A \in K^{m\times n}$ and $B \in K^{n\times m}$ we have $X^n \chi_{AB} = X^m \chi_{BA}$. In particular, AB and BA have the same nonzero eigenvalues.*

[6] One can also prove the formula by induction, provided one has guessed it beforehand.

Proof. By the rules for block matrices (Remark 5.9, Lemma 9.9) and the multiplicativity of the determinant we have

$$\begin{aligned} X^n \chi_{AB} &= \det\left(\begin{pmatrix} 1_m & 0 \\ -B & X1_n \end{pmatrix}\begin{pmatrix} X1_m - AB & A \\ 0 & 1_n \end{pmatrix}\begin{pmatrix} 1_m & 0 \\ B & 1_n \end{pmatrix}\right) \\ &= \det\left(\begin{pmatrix} 1_m & 0 \\ -B & X1_n \end{pmatrix}\begin{pmatrix} X1_m & A \\ B & 1_n \end{pmatrix}\right) = \det\begin{pmatrix} X1_m & A \\ 0 & X1_n - BA \end{pmatrix} = X^m \chi_{BA}. \end{aligned}$$

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 $n = m$ we even have $\chi_{AB} = \chi_{BA}$ in the situation of Lemma 10.38.

10.4 Minimal Polynomials

Remark 10.40. We have already substituted polynomials into matrices. Now, conversely, we 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. Moreover, $\alpha(A)^{\mathrm{t}} = \alpha(A^{\mathrm{t}})$.

Theorem 10.41. *For $A \in K^{n\times n}$ there exists exactly one monic polynomial $\mu_A \in K[X]$ with $\mu_A(A) = 0_n$ and $\deg \mu_A$ minimal.*

Proof. Since $\dim K^{n\times n} = n^2$ (Lemma 5.4), the powers $1_n = A^0, A, A^2, \ldots, A^{n^2}$ are linearly dependent in $K^{n\times n}$. Hence 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 $\alpha = \sum a_k X^k \in K[X]$ we thus have $\alpha(A) = 0$. Dividing by the leading coefficient of α, we may 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. We call 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, we have $\deg \mu_A \geq 2$. We make 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 $x, y \in \mathbb{Q}$. Comparing the matrix entries at position $(1, 2)$ shows $x = -1$. In fact, the equation now holds for $y = 2$. Hence $\mu_A = X^2 - X + 2$.

Lemma 10.44. *Let $A \in K^{n\times n}$ and $\alpha \in K[X]$ with $\alpha(A) = 0_n$. Then $\mu_A \mid \alpha$.*

Proof. We divide with remainder: $\alpha = \mu_A\gamma + \delta$ with $\gamma, \delta \in K[X]$ and $\deg\delta < \deg\mu_A$. Then

$$\delta(A) = (\alpha - \mu_A\gamma)(A) = \alpha(A) - \mu_A(A)\gamma(A) = 0_n.$$

The minimality of $\deg\mu_A$ implies $\delta = 0$ 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

$$\mu_A(SAS^{-1}) = \sum a_k(SAS^{-1})^k = \sum a_k SA^kS^{-1} = S\Big(\sum a_k A^k\Big)S^{-1} = S\mu_A(A)S^{-1} = 0_n.$$

Lemma 10.44 implies $\mu_{SAS^{-1}} \mid \mu_A$. Since similarity is a symmetric relation, we also have $\mu_A \mid \mu_{SAS^{-1}}$. Since both minimal polynomials are monic, they must be equal. □

Definition 10.46. Let V be a vector space with basis B. For $f \in \mathrm{End}(V)$ let $\mu_f := \mu_{B[f]_B}$ as usual. By Corollary 7.27 and Lemma 10.45, μ_f does not depend on the choice of B. The following theorems about matrices hold analogously for endomorphisms.

Remark 10.47. From the proof of Theorem 10.41 we obtain $\deg\mu_A \le n^2$. The next theorem implies $\deg\mu_A \le n$.

Theorem 10.48 (Cayley-Hamilton)**.** *For $A \in K^{n\times n}$ we have $\chi_A(A) = 0_n$ and $\mu_A \mid \chi_A$.*

Proof. Let $B := X1_n - A \in K[X]^{n\times n}$ and let $\widetilde{B} \in K[X]^{n\times n}$ be the adjugate of B. From every entry of $\widetilde{B}$ we extract the coefficient of X^k and form the matrix $B_k \in K^{n\times n}$ from these. Now we have

$$\widetilde{B} = \sum_{k=0}^{\infty} B_k X^k,$$

where only finitely many of the B_k are nonzero. Let $\chi_A = \sum a_k X^k$. By Theorem 9.22 we have

$$\sum_{k=0}^{\infty} a_k 1_n X^k = \chi_A 1_n = \det(B)1_n = \widetilde{B}B = \sum_{k=0}^{\infty} B_k X^k(X1_n - A) = \sum_{k=0}^{\infty}(B_{k-1} - B_kA)X^k,$$

where $B_{-1} := 0_n$. Comparing coefficients yields $a_k 1_n = B_{k-1} - B_kA$ for $k = 0, 1, \ldots$. Hence

$$\chi_A(A) = \sum_{k=0}^{\infty} a_k A^k = \sum_{k=0}^{\infty}(B_{k-1}A^k - B_kA^{k+1}) = \sum_{k=0}^{\infty} B_{k-1}A^k - \sum_{k=0}^{\infty} B_kA^{k+1} = 0_n.$$

The second claim follows from Lemma 10.44. □

Example 10.49.

(a) For $A \in K^{2\times 2}$ we have $A^2 - \mathrm{tr}(A)A + \det(A)1_2 = 0$ by Lemma 10.31.

(b) Let $A \in \mathrm{GL}(n, K)$. By Lemma 10.31, $(-1)^n \det(A)$ is the constant term of χ_A. Hence there exists a monic polynomial $\beta \in K[X]$ with $\chi_A - (-1)^n \det(A) = X\beta$. Since $\chi_A(A) = 0_n$, we have $A\beta(A) = (-1)^{n+1} \det(A) 1_n$. Multiplying by A^{-1} shows

$$A^{-1} = \frac{(-1)^{n+1}}{\det A} \beta(A).$$

Theorem 9.22 implies $\beta(A) = (-1)^{n+1} \widetilde{A}$. In particular for $n = 2$:

$$A^{-1} \overset{\text{(a)}}{=} \frac{1}{\det A} \big(\mathrm{tr}(A) 1_2 - A \big)$$

(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 hence 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 $k \in \mathbb{N}_0$ we have $A^k v = A^{k-1} \lambda v = \ldots = \lambda^k v$. 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 $v \neq 0$, we have $\mu_A(\lambda) = 0$, i. e., λ is a root of μ_A. □

Remark 10.51. Since $\deg \mu_A \leq \deg \chi_A$, 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. *Then $A \in K^{n \times n}$ is diagonalizable if and only if μ_A splits into pairwise distinct linear factors.*

Proof. Let $A \in K^{n \times n}$ be diagonalizable. By Lemma 10.45 we may assume $A = \mathrm{diag}(d_1, \ldots, d_n)$. Let $\lambda_1, \ldots, \lambda_k \in \{d_1, \ldots, d_n\}$ be the pairwise distinct eigenvalues of A. For $\alpha := (X - \lambda_1) \ldots (X - \lambda_k)$ we have

$$\alpha(A) = \mathrm{diag}(\alpha(d_1), \ldots, \alpha(d_n)) = 0_n$$

and $\mu_A \mid \alpha$ by Lemma 10.44. On the other hand, $\lambda_1, \ldots, \lambda_k$ are eigenvalues and hence roots of μ_A. This shows $\mu_A = \alpha$ (Lemma 10.24).

Conversely, suppose $\mu_A = (X - \lambda_1) \ldots (X - \lambda_k)$ with pairwise distinct $\lambda_1, \ldots, \lambda_k$. For $i = 1, \ldots, k$ let $\gamma_i := \prod_{j \neq i} (X - \lambda_j) \in K[X]$. The Lagrange polynomial (Remark 10.18) for the pairs $(x_i, y_i) = (\lambda_i, 1)$ has the form $\gamma = \sum_{i=1}^k c_i \gamma_i$ with $c_i \in K$. Since γ has degree $< k$ and coincides with the constant polynomial 1 at $\lambda_1, \ldots, \lambda_k$, we have $\gamma = 1$ by Theorem 10.17. 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., $\gamma_i(A) v \in E_{\lambda_i}(A)$. On the other hand,

$$v = 1_n v = \gamma(A) v = c_1 \gamma_1(A) v + \ldots + c_k \gamma_k(A) v.$$

This shows $K^{n \times 1} = E_{\lambda_1}(A) + \ldots + E_{\lambda_k}(A)$. By 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 $\deg \mu_A \geq 3$. By Theorem 10.52, A is not diagonalizable.

Remark 10.54. If χ_A splits into linear factors, then we have decompositions $\chi_A = (X-\lambda_1)^{m_1} \dots (X-\lambda_k)^{m_k}$ and $\mu_A = (X-\lambda_1)^{l_1} \dots (X-\lambda_k)^{l_k}$ with pairwise distinct eigenvalues $\lambda_1, \dots, \lambda_k$. By Theorem 10.50 we have $l_1, \dots, l_k \geq 1$. Since the algebraic multiplicity of an eigenvalue is at most n, it follows that $\chi_A \mid \mu_A^n$. The following theorem yields this statement without any assumption.

Theorem 10.55. *For $A \in K^{n\times n}$ we have $\chi_A \mid \mu_A^n$.*

Proof. Let $\mu_A = \sum a_k X^k$. For $k \geq 1$ we have

$$(X^k 1_n - A^k) = (X1_n - A) \sum_{i=1}^{k} X^{k-i} A^{i-1}.$$

It follows that

$$\mu_A 1_n = \mu_A 1_n - \mu_A(A) = \sum_{k\geq 1} a_k (X^k 1_n - A^k) = (X1_n - A)B$$

for some $B \in K[X]^{n\times n}$. Taking the determinant on both sides, we obtain $\mu_A^n = \chi_A \det(B)$. □

11 Euclidean Geometry

11.1 Inner Products

Remark 11.1. In this chapter we consider $K = \mathbb{R}$. In contrast to arbitrary fields, one can divide $\mathbb{R}$ 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 that every nonnegative real number has exactly one positive square root. Thus we have $|x| = \sqrt{x^2}$ 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 an *inner 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 an inner product, V becomes a *Euclidean space*.
- Vectors $v, w \in V$ are called *orthogonal*, if $[v, w] = 0$. In this case, one writes $v \perp w$.
- One calls $|v| := \sqrt{[v, v]} \geq 0$ the *norm* of v.[2] In the case $|v| = 1$ one calls v *normalized*.

Remark 11.3.

(a) The symmetry of the inner 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 $\lambda \in \mathbb{R}$. Hence 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 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 (for $V \neq \{0\}$) *not* linear, so not a functional, since $[v, v] > 0 = [0, v] + [v, 0]$ for $v \neq 0$.

[1] The notation $[v, w]$ is not uniform in the literature. One also finds $\langle v, w \rangle$ (confusion with the span), $(v \mid w)$ and similar ones.

[2] In a more general context (Definition 18.48) we use $\|v\|$ instead of $|v|$.

(b) For $v, w \in V$ we have

$$|v \pm w|^2 = [v, v] \pm 2[v, w] + [w, w] = |v|^2 \pm 2[v, w] + |w|^2.$$

In particular, the *parallelogram law* holds:

$$|v + w|^2 + |v - w|^2 = 2|v|^2 + 2|w|^2.$$

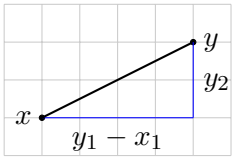
(c) Every subspace of a Euclidean space is itself a Euclidean space with the restricted inner product.

Example 11.4.

(a) The most important example of a Euclidean space is $V = \mathbb{R}^n$ with the *standard inner product*

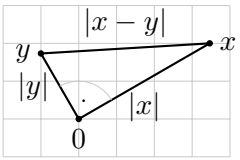
$$\boxed{[x, y] := xy^{\mathrm{t}} = \sum_{i=1}^{n} x_i y_i} \qquad (x, y \in \mathbb{R}^n).$$

One easily checks the three properties (positive definite, symmetric and bilinear). In the case $n = 1$, $|x| = \sqrt{x_1^2}$ is the ordinary absolute value (this justifies the use of the absolute value bars). By the Pythagorean theorem[3] 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 $|x - y|^2 = |x|^2 + |y|^2$ follows from Remark 11.3. By the converse of the Pythagorean theorem, x and y form a right angle, i. e., they are perpendicular to each other:



(b) Vectors in $\mathbb{R}^n$ are "discrete" functions $\{1, \ldots, n\} \to \mathbb{R}$. 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 $V \leq \mathrm{Map}([0, 1], \mathbb{R})$ 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).

[3] One could also *define* the distance between x and y by $|y - x|$ and thereby *prove* the Pythagorean theorem.

Proof.

(a) $|\lambda v| = \sqrt{[\lambda v, \lambda v]} = \sqrt{\lambda^2 [v,v]} = \sqrt{\lambda^2}\sqrt{[v,v]} = |\lambda||v|$.

(b) W. l. o. g. let $w \neq 0$. Let $\lambda := \frac{[v,w]}{[w,w]}$. By the properties of the inner product we have

$$0 \leq |v - \lambda w|^2 = [v - \lambda w, v - \lambda w] = [v,v] - 2\lambda[v,w] + \lambda^2 [w,w] = |v|^2 - \frac{[v,w]^2}{|w|^2}.$$

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 $\mu \in \mathbb{R}$ with $v = \mu w$ and $|[v,w]| = |\mu||w|^2 \overset{\text{(a)}}{=} |\mu w||w| = |v||w|$.

(c) First, we have

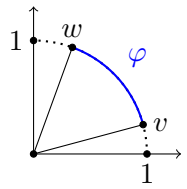
$$|v+w|^2 = [v+w, v+w] = [v,v] + 2[v,w] + [w,w] \overset{\text{(b)}}{\leq} |v|^2 + 2|v||w| + |w|^2 = (|v| + |w|)^2$$

and $|v+w| \leq |v| + |w|$. It follows that $|v| = |v + w - w| \leq |v+w| + |w|$ and $|v| - |w| \leq |v+w|$. Interchanging v and w yields $-(|v| - |w|) = |w| - |v| \leq |v+w|$, so $||v| - |w|| \leq |v+w|$. □

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 says that the sum of any two sides is greater than the third side (the proof above shows that equality cannot occur).

(b) For normalized $v, w \in \mathbb{R}^n$ one defines the *angle* φ (in radians[4]) as the length of the shortest arc on the unit circle[5] between v and w:



The length of the semicircular arc (the largest possible angle) is called π and one computes $\pi \approx 3.14$ (Exercise II.11).

(c) The *cosine* of φ is defined by $\cos\varphi := [v,w]$.[6] The Cauchy-Schwarz inequality implies $-1 \leq \cos\varphi \leq 1$. Since v and $-w$ form the angle $\pi - \varphi$, we have $\cos(\pi - \varphi) = -[v,w] = -\cos\varphi$. In particular, we have

$$\cos 0 = [e_1, e_1] = 1, \ \cos(\pi/2) = [e_1, e_2] = 0, \ \cos\pi = [e_1, -e_1] = -1.$$

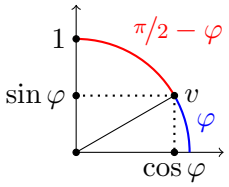
By $\cos(-\varphi) := \cos(\varphi)$ and $\cos(\varphi + 2k\pi) := \cos\varphi$ for $k \in \mathbb{Z}$, cos extends periodically to all of $\mathbb{R}$.

[4] Using *degree measure* with $\pi \mathrel{\hat{=}} 180°$ brings no advantages in mathematics.
[5] in the plane $\langle v, w \rangle$
[6] One can show that this definition does not depend on the choice of v, w and agrees with the analytic definition as a power series.

(d) 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 a normalized vector $v = (x, y)$ with $y \geq 0$ we have $x = [e_1, v] = \cos\varphi$ and $y = [e_2, v] = \cos(\pi/2 - \varphi) = \sin\varphi$:



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 linear independence. Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with $\lambda_1 b_1 + \ldots + \lambda_n b_n = 0$. Then we have

$$\lambda_i = \sum_{j=1}^{n} \lambda_j [b_j, b_i] = \Big[\sum_{j=1}^{n} \lambda_j b_j, b_i\Big] = [0, b_i] = 0$$

for $i = 1, \ldots, n$.

Example 11.9. The standard basis $e_1, \ldots, e_n \in \mathbb{R}^n$ is an orthonormal basis with respect to 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 $k = 1$ we have $b_1 = v_1 \neq 0$. Now let $k \geq 2$ and suppose the claim has already been proved 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 $1 \leq i < j \leq k - 1$. In particular, $b_1, \ldots, b_{k-1}$ are linearly independent. Since $\sum_{i=1}^{k-1} \frac{[v_k, b_i]}{[b_i, b_i]} b_i \in \langle b_1, \ldots, b_{k-1}\rangle$, we have

$$\langle v_1, \ldots, v_k\rangle = \langle b_1, \ldots, b_{k-1}, v_k\rangle = \langle b_1, \ldots, b_k\rangle.$$

Moreover, for $i = 1, \ldots, k - 1$ we have

$$[b_k, b_i] = [v_k, b_i] - \sum_{j=1}^{k-1} \frac{[v_k, b_j]}{[b_j, b_j]} [b_j, b_i] = [v_k, b_i] - [v_k, b_i] = 0.$$

This proves the first claim. Since $|\frac{1}{|b_i|}b_i| = \frac{1}{|b_i|}|b_i| = 1$, the second claim follows. □

Corollary 11.11. *Every Euclidean space has (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 $\mathbb{R}^3$. With respect to the standard inner product we obtain

$$\begin{aligned} b_1 &:= v_1 = (1,0,1), \\ b_2 &:= v_2 - \frac{[v_2,b_1]}{[b_1,b_1]}b_1 = (0,1,1) - \frac{1}{2}(1,0,1) = \frac{1}{2}(-1,2,1), \\ b_3 &:= v_3 - \frac{[v_3,b_1]}{[b_1,b_1]}b_1 - \frac{[v_3,b_2]}{[b_2,b_2]}b_2 = (-1,2,0) + \frac{1}{2}(1,0,1) - \frac{5}{6}(-1,2,1) = \frac{1}{3}(1,1,-1). \end{aligned}$$

Note that scaling factors play no role in this computation. Thus one can compute b_3 somewhat more conveniently with $b_2 = (-1,2,1)$. 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 apply the Gaussian algorithm first 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

$$S^\perp := \{v \in V : \forall s \in S : s \perp v\}$$

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 $\lambda \in \mathbb{R}$ we have $[\lambda v + w, s] = \lambda[v,s] + [w,s] = 0$ for all $s \in S$, i. e., $\lambda v + w \in S^\perp$. Hence $S^\perp$ is a subspace, even if S is not a subspace. Moreover, $S^\perp = \langle S \rangle^\perp$.

Lemma 11.15. *For subspaces U, W of a Euclidean space V we have*

(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 $v \in U \cap U^\perp$ we have $|v|^2 = [v,v] = 0$ and $v = 0$. Hence $U \cap U^\perp = \{0\}$ 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, $U \subseteq (U^\perp)^\perp$. By (a), $\dim(U^\perp)^\perp = \dim(V) - \dim(U^\perp) = \dim U$. Hence $U = (U^\perp)^\perp$.

(c) By definition we have

$$U \subseteq W \implies W^\perp \subseteq U^\perp \overset{\text{(b)}}{\implies} U = (U^\perp)^\perp \subseteq (W^\perp)^\perp = W. \qquad \square$$

Example 11.16.

(a) In $V = \mathbb{R}^n$ the orthogonal complement of $U \leq V$ with respect to the standard inner product can be determined with the Gaussian algorithm: One writes the vectors of a spanning set of U as rows in a matrix $A \in \mathbb{R}^{k\times n}$. Now $U^\perp$ is the solution set of the system of equations $Ax = 0$.

(b) For linearly independent vectors $v_1, \ldots, v_{n-1} \in \mathbb{R}^n$ let $\langle v_1, \ldots, v_{n-1}\rangle^\perp = \langle w\rangle$. We determine an explicit formula for the coordinates of w. Let A be the matrix with rows $w, v_1, \ldots, v_{n-1}$. Then $Aw = |w|^2 e_1$. Cramer's rule and Laplace expansion along the i-th column show

$$w_i = \frac{\det A_i}{\det A} = (-1)^{i+1}|w|^2 \frac{\det A_{1i}}{\det A}$$

for $i = 1, \ldots, n$. If we scale w by the factor $\frac{\det A}{|w|^2}$, the equation simplifies to $w_i = (-1)^{i+1}\det(A_{1i})$, where A_{1i} does not depend on w. This w is called the *vector product* of $v_1, \ldots, v_{n-1}$ and one writes $w = v_1 \times \ldots \times v_{n-1}$. For linearly dependent vectors let $v_1 \times \ldots \times v_{n-1} = 0$, in accordance with $\det A_{1i} = 0$. In the case $n = 2$ we obtain $(x, y)^\perp = \langle (y, -x)\rangle$. In the case $n = 3$ one speaks of the *cross product*. Here we have

$$a \times b = \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_3b_1 - a_1b_3 \\ a_1b_2 - a_2b_1 \end{pmatrix}.$$

The direction of $a \times b$ can be determined with the *right-hand rule*: If the thumb points in the direction of a and the index finger in the direction of b, then the middle finger of the right hand points in the direction of $a \times b$ (Exercise II.12).

11.3 Symmetric and Orthogonal Maps

Definition 11.17. Let V be a Euclidean space and $f \in \mathrm{End}(V)$. We call f

- *symmetric*,[7] if $[f(v), w] = [v, f(w)]$ for all $v, w \in V$.
- *orthogonal*,[8] 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 clearly $\lambda f + g$ is symmetric as well. Hence the symmetric maps form a subspace of $\mathrm{End}(V)$.

(b) Orthogonal maps $f \in \mathrm{End}(V)$ are isomorphisms, since $v \in \operatorname{Ker} f$ implies $|v|^2 = [v, v] = [f(v), f(v)] = 0$, so $v = 0$. Because of $|f(v - w)| = |v - w|$ and

$$\frac{[f(v), f(w)]}{|f(v)||f(w)|} = \frac{[v, w]}{|v||w|}$$

the map f preserves distances and angles (Remark 11.6). In particular, f maps orthonormal bases to orthonormal bases. Together with f, g, also $f \circ g$ and f^{-1} are orthogonal. Hence the set of orthogonal maps forms a subgroup $\mathrm{O}(V)$ of $\mathrm{GL}(V)$. We call $\mathrm{O}(V)$ the *orthogonal group* of V.

[7] or *self-adjoint*, see Section 13.2
[8] or *isometric*

(c) For an eigenvalue $\lambda \in \mathbb{R}$ of $f \in \mathrm{O}(V)$ with eigenvector v we have $|\lambda||v| = |\lambda v| = |f(v)| = |v|$, i. e., $\lambda = \pm 1$.

(d) The following theorem states that distance-preserving maps can be written as the composition of a translation and an orthogonal map (cf. Exercise I.25).

Theorem 11.19 (Mazur-Ulam)**.** *Let V be a Euclidean space and $f \in \mathrm{Map}(V, V)$ with $|f(v) - f(w)| = |v - w|$ for all $v, w \in V$. Then there exist $x \in V$ and $g \in \mathrm{O}(V)$ with $f(v) = g(v) + x$ for all $v \in V$.*

Proof. Let $x := f(0)$ and $g(v) := f(v) - x$ for $v \in V$. Then $g(0) = 0$ and $|g(v) - g(w)| = |f(v) - f(w)| = |v - w|$. It follows that $|g(v)| = |g(v) - g(0)| = |v|$ and

$$[g(v), g(w)] = \frac{1}{2}\big(|g(v)|^2 + |g(w)|^2 - |g(v) - g(w)|^2\big) = \frac{1}{2}\big(|v|^2 + |w|^2 - |v - w|^2\big) = [v, w].$$

This implies

$$\begin{aligned} |g(v + w) - g(v) - g(w)|^2 &= |g(v + w)|^2 - 2[g(v + w), g(v)] \pm \ldots = |v + w|^2 - 2[v + w, v] \pm \ldots \\ &= |v + w - v - w|^2 = 0 \end{aligned}$$

and similarly $|g(\lambda v) - \lambda g(v)| = |\lambda v - \lambda v| = 0$ for $\lambda \in \mathbb{R}$. □

Lemma 11.20. *Let V be Euclidean with orthonormal basis B, $f \in \mathrm{End}(V)$ and $A := {}_B[f]_B$. Then:*

(a) f symmetric $\iff$ $A^{\mathrm{t}} = A$.

(b) f 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 $1 \le i, j \le n$. Hence $A = A^{\mathrm{t}}$. Conversely, let $A = A^{\mathrm{t}}$. Then $[f(b_i), b_j] = [b_i, f(b_j)]$ for $1 \le i, j \le n$. For arbitrary $v = \sum \lambda_i b_i$ and $w = \sum \mu_j b_j$ in V we have

$$[f(v), w] = \sum_{i,j=1}^n \lambda_i \mu_j [f(b_i), b_j] = \sum_{i,j=1}^n \lambda_i \mu_j [b_i, f(b_j)] = [v, f(w)].$$

Hence f is symmetric.

(b) If f is orthogonal, then

$$\sum_{k=1}^n a_{ki} a_{kj} = \Big[\sum_{k=1}^n a_{ki} b_k, \sum_{k=1}^n a_{kj} b_k\Big] = [f(b_i), f(b_j)] = [b_i, b_j] = \delta_{ij}.$$

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 one calls matrices $A \in \mathrm{GL}(n, K)$ *orthogonal* if $A^{\mathrm{t}} = A^{-1}$. One easily shows 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 $\mathrm{O}(V)$ by choosing an orthonormal basis (just as $\mathrm{GL}(n, K)$ and $\mathrm{GL}(V)$ correspond to each other). For every orthogonal matrix A we have

$$\det(A)^2 \overset{9.12}{=} \det(A)\det(A^{\mathrm{t}}) = \det(AA^{\mathrm{t}}) = \det 1_n = 1$$

and $\det A = \pm 1$. One calls

$$\mathrm{SO}(n, K) := \mathrm{O}(n, K) \cap \mathrm{SL}(n, K) \leq \mathrm{O}(n, K) \leq \mathrm{GL}(n, K)$$

the *special orthogonal group* of degree n over K.

(b) For real matrices, the equation $A^{\mathrm{t}}A = 1_n = AA^{\mathrm{t}}$ means that the columns (or rows) of A form an orthonormal basis of $\mathbb{R}^n$ with respect to the standard inner product.

Example 11.22.

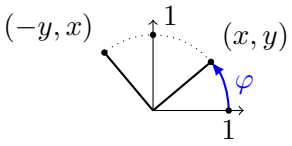
(a) Every permutation matrix is orthogonal, since the rows form an orthonormal basis (namely a permutation of the standard basis).

(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* counterclockwise by the *oriented* angle φ from e_1 to (x, y):



  In the case $y < 0$ we have $\varphi = 2\pi - \varphi'$ with the unoriented angle φ'. We obtain $\cos\varphi = \cos\varphi'$, $\sin\varphi = -\sin\varphi'$ and

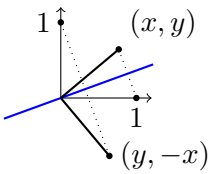
$$A = D(\varphi) := \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

  For two oriented angles φ and ψ we obtain $D(\varphi+\psi) = D(\varphi)D(\psi)$, from which the well-known *addition formulas* follow:

$$\begin{aligned} \cos(\varphi + \psi) &= \cos(\varphi)\cos(\psi) - \sin(\varphi)\sin(\psi), \\ \sin(\varphi + \psi) &= \sin(\varphi)\cos(\psi) + \sin(\psi)\cos(\varphi). \end{aligned}$$

  Clearly, $D(\varphi)$ has (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* in 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 that $\det A$ is the product of the eigenvalues). By 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 we obtain (cf. Exercise II.10):

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
$S(\varphi)$	$\frac{1}{2}\begin{pmatrix} \sqrt{3} & 1 \\ 1 & -\sqrt{3} \end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$

This also allows us to compute

$$D(\pi/12) = D(\pi/3 - \pi/4) = D(\pi/3) D(\pi/4)^{-1} = D(\pi/3) D(\pi/4)^{\mathrm{t}}$$

in this way.

Remark 11.23. To show that symmetric endomorphisms are diagonalizable, we have to temporarily leave the real numbers.

11.4 Complex Numbers

Lemma 11.24. *With the multiplication*

$$(a, b) \cdot (c, d) := (ac - bd,\ ad + bc) \qquad (a, b, c, d \in \mathbb{R})$$

the $\mathbb{R}$-vector space $\mathbb{C} := \mathbb{R}^2$ becomes a field.

Proof. As a vector space, $(\mathbb{C}, +)$ is already an abelian group. We reduce the multiplication in $\mathbb{C}$ to matrix multiplication.[9] To this end, 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 $x, y, z \in \mathbb{C}$ we have $\Gamma(x) + \Gamma(y) = \Gamma(x + y)$ and $\Gamma(x)\Gamma(y) = \Gamma(x \cdot y)$ (verify this). 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 distributive law is proved similarly. The commutative law is better verified directly from the definition (since matrices are not commutative in general). Because of

[9] One can also verify the axioms directly using the definition.

$\Gamma(1,0) = 1_2$, the element $(1,0)$ is the identity element in $\mathbb{C}$. It remains to show that $z \neq 0$ is invertible. We have $\det(\Gamma(z)) = a^2 + b^2 > 0$ 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. We call $\mathbb{C}$ the field of *complex numbers*.

- Via 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.
- We call $\mathrm{i} := (0,1) \in \mathbb{C} \setminus \mathbb{R}$ the *imaginary unit*. We have $\mathrm{i}^2 = (-1,0) = -1$. Since $1, \mathrm{i}$ form an $\mathbb{R}$-basis of $\mathbb{C}$, every complex number can be written uniquely 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.
- We call $|z| := \sqrt{\mathrm{Re}(z)^2 + \mathrm{Im}(z)^2} \geq 0$ 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 total 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 $a, b, c \in \mathbb{C}$. Indeed, if $\mathrm{i} > 0$, then $-1 = \mathrm{i}^2 > 0$ and $1 = (-1)^2 > 0$. Then we obtain 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 $a^2 + b^2 = |z_0|^2 = 1$ we have $\Gamma(z_0) = D(\varphi) \in \mathrm{O}(2, \mathbb{R})$ for an oriented 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

$$D(\varphi_k)^n = D(n\varphi_k) = D(\varphi + 2k\pi) = D(\varphi) = \Gamma(z_0).$$

Hence the numbers $z_k := \cos\varphi_k + \mathrm{i}\sin\varphi_k \in \mathbb{C}$ satisfy $z_k^n = z_0$. Because of $|z| > 0$, the root $\sqrt[n]{|z|} \in \mathbb{R}_{>0}$ exists (analysis). For $\zeta_k := \sqrt[n]{|z|}z_k$ we obtain $\zeta_k^n = |z|z_0 = z$ 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 $k = l$. Hence the n-th roots $\zeta_1, \ldots, \zeta_n$ are pairwise distinct. $\square$

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 $k \geq 2$ with $A^k = A$. Then A is diagonalizable.*

Proof. The minimal polynomial μ_A divides $X^k - X = X(X^{k-1} - 1)$ by Lemma 10.44. The roots of $X^k - X$ are the $(k-1)$-th roots of unity and 0. By Lemma 10.24, μ_A splits into pairwise distinct linear factors. The claim follows from Theorem 10.52. $\square$

Example 11.30. Let $k \in \mathbb{N}$ and $A := D(2\pi/k)$. Because of $A^{k+1} = AA^k = AD(2\pi) = A$, the matrix 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 $z, w \in \mathbb{C}$ we have $|z|^2 = z\overline{z}$, $\overline{z + w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z} \cdot \overline{w}$.*

Proof. For $z = a + b\mathrm{i}$ and $w = c + d\mathrm{i}$ we have

$$\begin{aligned} |z|^2 &= a^2 + b^2 = (a + b\mathrm{i})(a - b\mathrm{i}) = z\overline{z}, \\ \overline{z + w} &= a + c - (b + d)\mathrm{i} = a - b\mathrm{i} + c - d\mathrm{i} = \overline{z} + \overline{w}, \\ \overline{zw} &= ac - bd - (ad + bc)\mathrm{i} = (a - b\mathrm{i})(c - d\mathrm{i}) = \overline{z} \cdot \overline{w}. \end{aligned}$$

□

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 $B = (b_{ij}) \in \mathbb{C}^{m \times k}$ we have

$$\overline{AB} = \Big(\overline{\sum_{l=1}^{m} a_{il} b_{lj}}\Big)_{ij} \overset{11.32}{=} \Big(\sum_{l=1}^{m} \overline{a_{il} b_{lj}}\Big)_{ij} \overset{11.32}{=} \Big(\sum_{l=1}^{m} \overline{a_{il}}\,\overline{b_{lj}}\Big)_{ij} = \overline{A} \cdot \overline{B}.$$

(b) By Lemma 11.27, the polynomial $X^n - z$ has a root for every $z \in \mathbb{C}$ (in particular, i is a root of $X^2 + 1$).

Theorem 11.34 (Fundamental theorem of algebra)**.** *Every polynomial $\alpha \in \mathbb{C}[X] \setminus \mathbb{C}$ has a root in $\mathbb{C}$.*[10]

Proof. The proof uses analysis (more precisely, the completeness of $\mathbb{R}$) and is too difficult for this course. See the algebra lecture notes. □

Corollary 11.35. *Every monic polynomial $\alpha \in \mathbb{C}[X] \setminus \mathbb{C}$ splits into linear factors.*[11]

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 $d \geq 2$. By the fundamental theorem, α has a root $x \in \mathbb{C}$. By 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 be monic as well. By induction, β splits into linear factors, and hence so does α. □

Example 11.36. Let $\alpha \in \mathbb{R}[X]$ have odd degree. To show that α has a root, we may assume that α is monic. Then $\lim_{x \to \pm\infty} \alpha(x) = \pm\infty$. By 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 computed approximately. Let $\alpha := X^5 - 4X + 2$ be concrete. From 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 *Newton's method* one computes the recursively defined sequence

$$x_{n+1} := x_n - \frac{\alpha(x_n)}{\alpha'(x_n)} \qquad (n \geq 0),$$

so $x_1 = 0.50847\ldots$, $x_2 = 0.50849948\ldots$ etc. If x_0 is chosen "well" (as it is here), then the sequence $(x_n)_n$ converges quadratically to a root, i. e., with each iteration the number of correct decimal places doubles (cf. Exercise III.2). In fact, all the given decimal places of x_2 are already correct.

[10] One says: $\mathbb{C}$ is *algebraically closed*.
[11] cf. Theorem 15.39

Remark 11.37. Let $x \in \mathbb{C}$ be a root of $\alpha = \sum a_k X^k \in \mathbb{C}[X]$. We define $\overline{\alpha} := \sum \overline{a_k} X^k \in \mathbb{C}[X]$. Then

$$\overline{\alpha}(\overline{x}) = \sum \overline{a_k x^k} = \overline{\alpha(x)} = \overline{0} = 0,$$

i. e., $\overline{x}$ is a root of $\overline{\alpha}$. In the case $\alpha \in \mathbb{R}[X]$ we thus have $\alpha(x) = 0 \iff \alpha(\overline{x}) = 0$. For $x \notin \mathbb{R}$,

$$(X - x)(X - \overline{x}) = X^2 - (x + \overline{x})X + x\overline{x} = X^2 - 2\mathrm{Re}(x)X + |x|^2 \in \mathbb{R}[X]$$

is a divisor of α.

11.5 The Principal Axis Theorem

Theorem 11.38 (Principal axis theorem)**.** *Let V be a Euclidean space and $f \in \mathrm{End}(V)$. Then f is symmetric if and only if V has 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 hence symmetric. By 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 as an orthonormal basis of eigenvectors. Let $n \geq 2$ and suppose the claim has already been proved for $n - 1$. First let B be an arbitrary orthonormal basis of V. Then we can also regard the symmetric matrix $A := {}_B[f]_B$ as a complex matrix. By the fundamental theorem of algebra, $\chi_A \in \mathbb{C}[X] \setminus \mathbb{C}$ has a root $\lambda \in \mathbb{C}$. Hence 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

$$\lambda \sum_{i=1}^{n} |v_i|^2 = \lambda \overline{v}^{\mathrm{t}} v = \overline{v}^{\mathrm{t}} A v = \overline{v}^{\mathrm{t}} \overline{A}^{\mathrm{t}} v = (\overline{Av})^{\mathrm{t}} v = \overline{\lambda} \overline{v}^{\mathrm{t}} v = \overline{\lambda} \sum_{i=1}^{n} |v_i|^2$$

we have $\lambda = \overline{\lambda} \in \mathbb{R}$. 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 $U := b_1^{\perp}$ we have $V = \langle b_1 \rangle \oplus U$ by Lemma 11.15. For $u \in U$ we have

$$[f(u), b_1] = [u, f(b_1)] = \lambda [u, b_1] = 0,$$

i. e., $f(u) \in U$. Therefore the restriction $g := f_{|U}$ lies in $\mathrm{End}(U)$. Clearly, g is also symmetric. Since $\dim U = n - 1$, by induction U has an orthonormal basis $b_2, \ldots, b_n \in U$ of eigenvectors of g. Since $f(b_i) = g(b_i)$, the vectors $b_2, \ldots, b_n$ are also eigenvectors of f. Altogether, $b_1, \ldots, b_n$ is an orthonormal basis of V of eigenvectors of f. □

Corollary 11.39. *A matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if and only if there exists a matrix $S \in \mathrm{O}(n, \mathbb{R})$ with $S^{\mathrm{t}} A S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. In this case, $\lambda_1, \ldots, \lambda_n$ 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. By Lemma 11.20, f is symmetric. By 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, we have $S \in \mathrm{O}(n, \mathbb{R})$. 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 $v \perp w$. Mnemonic: Eigenvectors for distinct eigenvalues are orthogonal. Hence one can compute the desired orthonormal basis of V by applying the Gram-Schmidt process to each eigenspace.

Example 11.41. In Example 8.4 we studied the map $f \in \mathrm{End}(\mathbb{R}^3)$ with symmetric matrix

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

We have

$$E_1(A) = \langle \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \rangle, \qquad E_4(A) = \langle \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \rangle$$

(note that $\operatorname{tr} A = 6 = 1 + 1 + 4$). 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. Altogether we obtain the orthonormal basis $\frac{1}{\sqrt{2}}(1,0,-1)$, $\frac{1}{\sqrt{6}}(-1,2,-1)$, $\frac{1}{\sqrt{3}}(1,1,1)$ of eigenvectors of f. For the orthogonal matrix

$$S := \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ -1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \end{pmatrix}$$

we have $S^{\mathrm{t}}AS = S^{-1}AS = \operatorname{diag}(1,1,4)$. One can use this result to take roots of matrices. For $\sqrt{A} := S \operatorname{diag}(1,1,2) S^{\mathrm{t}}$ we have namely

$$\sqrt{A}^2 = S \operatorname{diag}(1,1,2) S^{\mathrm{t}} S \operatorname{diag}(1,1,2) S^{\mathrm{t}} = S \operatorname{diag}(1,1,2)^2 S^{\mathrm{t}} = S \operatorname{diag}(1,1,4) S^{\mathrm{t}} = A.$$

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 exist 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 $\lambda \in \mathbb{R}$ by Example 11.36 (or Remark 11.37). By Remark 11.18, $\lambda \in \{\pm 1\}$. Let b_1 be a corresponding normalized eigenvector and $U := b_1^{\perp}$. For $u \in U$ we have

$$[f(u), b_1] = [f(u), \lambda^2 b_1] = \lambda [f(u), f(b_1)] = \lambda [u, b_1] = 0$$

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 we have ${}_C[g]_C \in \mathrm{O}(2,\mathbb{R})$ by Lemma 11.20. In the case $\det g = 1$ we have ${}_C[g]_C = D(\varphi)$ by Example 11.22 and the claim follows with $B := \{b_1, b_2, b_3\}$. Now let $\det g = -1$. By 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. Afterwards $\det g = 1$ and the claim follows as before. □

Remark 11.43. The matrices in Theorem 11.42 with determinant 1 (i. e., of the form $\operatorname{diag}(1, D(\varphi))$) correspond to rotations by the angle φ, where the axis of rotation is spanned by the first basis vector. By the multiplicativity of the determinant, the composition of rotations is again a rotation (in the 2-dimensional space this is clear because of $D(\varphi)D(\psi) = D(\varphi + \psi)$). Caution: 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 soccer game there is a point on the surface of the ball that is at exactly the same location at two different times. Reason: The center of the ball lies on the center spot 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 on the center spot is also described by a rotation f. The axis of rotation of f intersects the surface of the ball in two points. Hence these remain fixed.

Definition 11.45. Let V be a Euclidean space and $v \in V \setminus \{0\}$. The map

$$S_v \colon V \to V, \qquad w \mapsto w - 2\frac{[w,v]}{[v,v]}v$$

is called the *reflection* in $v^\perp$.

Remark 11.46. For $\lambda \in \mathbb{R}^\times$ we have $S_{\lambda v} = S_v$. Hence we may assume that v is normalized. Then

$$S_v(w) = w - 2[w,v]v$$

for all $w \in V$. Clearly, S_v is linear and $S_v(w) = w$ holds if and only if $w \in v^\perp$. Moreover, $S_v(v) = -v$. Geometrically, S_v thus corresponds to a reflection in the hyperplane $v^\perp$. With respect to a suitable orthonormal basis, S_v has the matrix $\operatorname{diag}(-1, 1, \ldots, 1)$ (cf. Exercise II.13). In particular, S_v is orthogonal and $\det S_v = -1$. Moreover, $S_v \circ S_v = \mathrm{id}_V$.

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. So let $f \neq \mathrm{id}_V$ and $w \in V$ with $f(w) \neq w$. In the case $n = 1$ we have $f = -\mathrm{id}_V = S_w$. Now let $n \geq 2$ and suppose the claim has already been proved for $n - 1$. For $v := f(w) - w \neq 0$ we have

$$[f(w) + w, v] = [f(w) + w, f(w) - w] = [f(w), f(w)] - [w, w] = 0,$$

i. e., $f(w) + w \in v^\perp$. Hence

$$(S_v \circ f)(w) = \frac{1}{2}\Big(S_v(f(w) + w) + S_v(v)\Big) = \frac{1}{2}\big(f(w) + w - v\big) = w.$$

Let $U := w^\perp$. Since $g := S_v \circ f \in \mathrm{O}(V)$, we have $g(U) = U$. By induction, $g_{|U}$ is a composition of at most $n - 1$ reflections $S_{w_1}, \ldots, S_{w_k}$ with $w_1, \ldots, w_k \in U$. We can regard S_{w_i} with the same formula also as a reflection of V. Since $w \in U^\perp$, we have $S_{w_i}(w) = w$ for $i = 1, \ldots, k$. Thus $f = S_v \circ g$ is a product of at most n reflections. □

Remark 11.48. By Remark 11.46, $f \in \mathrm{SO}(V)$ is a product of an even number of reflections. In particular, in the case $\dim V = 2k + 1$ one needs only $2k$ reflections.

12 Bilinear Forms

12.1 Gram Matrices

Remark 12.1. In this chapter we generalize parts of Euclidean geometry to arbitrary fields. Throughout, let V 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 both 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 $\lambda \in K$.

- We call β
  - *symmetric* if $\beta(v, w) = \beta(w, v)$ for all $v, w \in V$.
  - *antisymmetric* if $\beta(v, w) = -\beta(w, v)$ for all $v, w \in V$.
  - *alternating* if $\beta(v, v) = 0$ 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 nondegenerate bilinear forms.

(b) An inner product $\beta(v, w) = [v, w]$ on a Euclidean space V is a nondegenerate symmetric bilinear form, since $[v, v] = |v|^2 > 0$ for $v \in V \setminus \{0\}$.

(c) The cross product $(v, w) \mapsto v \times w$ on $\mathbb{R}^3$ is bilinear and antisymmetric, but not a bilinear form, since it does not map to $\mathbb{R}$.

(d) 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 rules of matrix arithmetic. We will see in Theorem 12.9 that every bilinear form has this form.

(e) Choosing $A = 1_n$ in (d) yields the symmetric bilinear form

$$\beta(v, w) = vw^{\mathrm{t}} = \sum_{i=1}^{n} v_i w_i.$$

For $K = \mathbb{R}$ we obtain the standard inner product.

(f) Let $V := K^2$ and $\beta(v, w) := \det\left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right)$ for $v, w \in V$. By Lemma 9.5, β is an alternating bilinear form. In general, det is a *multilinear form* (Exercise II.15).

Remark 12.4.

(a) Every alternating bilinear form β is antisymmetric, since 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, since $\beta(v,v) = -\beta(v,v)$ implies $\beta(v,v) = 0$. For $K = \mathbb{F}_2$ this reasoning is invalid, since $1+1=0$. Here symmetric and antisymmetric are even equivalent. In particular, the bilinear form $\beta(v,w) = vw^{\mathrm{t}}$ on $V = \mathbb{F}_2^n$ is (anti)symmetric, but not alternating. To avoid this case, we will often assume $1+1 \neq 0$ in what follows.[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 one can recover β 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* $\mathrm{Map}(V \times V, K)$.

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 both components (however, $\mathrm{Bil}(V) \nsubseteq \mathrm{Hom}(V \times V, K)$). This shows $\beta + \gamma, \lambda\beta \in \mathrm{Bil}(V)$. □

Remark 12.6. Clearly, the symmetric (antisymmetric, alternating) bilinear forms each form a subspace of $\mathrm{Bil}(V)$. Theorem 12.9 gives 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.*[2] *Moreover,* β *is nondegenerate if and only if* F *is an isomorphism.*

Proof. Because of the bilinearity of β we have $F_v \in V^*$ and F is a homomorphism. Clearly, β is nondegenerate if and only if F is injective. Since $\dim V^* = \dim V$, 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. We call

$${}_B[\beta]_B := \big(\beta(b_i, b_j)\big)_{i,j=1}^n \in K^{n \times n}$$

the *Gram matrix* of β with respect to 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 $A := {}_B[\beta]_B$ *we have:*

(a) β *symmetric* $\iff$ A *symmetric.*

(b) β *antisymmetric* $\iff$ $A^{\mathrm{t}} = -A$.[3]

[1] Besides $\mathbb{F}_2$ there are many (also infinite) fields K with $1+1=0$. In this case one says that K has *characteristic* 2.
[2] In functional analysis one likes to write $\langle f, v\rangle = f(v)$ for $f \in V^*$ and $v \in V$ to emphasize the bilinearity.
[3] One calls A *antisymmetric* or *skew-symmetric*.

(c) β nondegenerate $\iff$ A invertible.

Proof. For $\beta, \gamma \in \mathrm{Bil}(V)$ and $\lambda \in K$ we have

$$ {}_B[\lambda\beta + \gamma]_B = \big((\lambda\beta + \gamma)(b_i, b_j)\big)_{i,j} = \lambda(\beta(b_i, b_j)) + (\gamma(b_i, b_j)) = \lambda\, {}_B[\beta]_B + {}_B[\gamma]_B, $$

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 $\lambda_i, \mu_j \in K$. Hence $\beta = 0$ and ${}_B[.]_B$ is injective. For surjectivity let $A \in K^{n\times n}$ be given. From the linearity of the coordinate vector $v \mapsto {}_B[v]$ (see the proof of Theorem 7.10) it follows that

$$ \beta(v, w) := {}_B[v] A\, {}_B[w]^{\mathrm{t}} $$

defines a bilinear form. Since $\beta(b_i, b_j) = e_i A e_j^{\mathrm{t}} = a_{ij}$, we have ${}_B[\beta]_B = A$. Hence ${}_B[.]_B$ is an isomorphism.

(a) Clearly, $\beta'(v, w) := \beta(w, v)$ defines a bilinear form with ${}_B[\beta']_B = A^{\mathrm{t}}$. By the first part of the proof we have

$$ \beta \text{ symmetric} \iff \beta = \beta' \iff A = A^{\mathrm{t}}. $$

(b) As in (a) with $\beta'(v, w) := -\beta(w, v)$.

(c) Let $B^* := \{b_1^*, \ldots, b_n^*\} \in V^*$ be the basis dual to B. For the map F from Theorem 12.7 we have

$$ F_{b_i}(b_j) = \beta(b_i, b_j) = a_{ij} = \sum_{k=1}^{n} a_{ik} b_k^*(b_j), $$

i. e., ${}_{B^*}[F]_B = A^{\mathrm{t}}$. By Theorem 12.7, β is nondegenerate if and only if A^{t} (and hence A) is invertible. □

Example 12.10. Let $1 + 1 \neq 0$ in K and $n := \dim V$ odd. Let $\beta \in \mathrm{Bil}(V)$ be antisymmetric with Gram matrix A with respect to an arbitrary basis. By Theorem 12.9 we have

$$ \det A = \det(A^{\mathrm{t}}) = \det(-A) \overset{9.8}{=} (-1)^n \det(A) = -\det(A) $$

and $\det A = 0$. Hence β must be degenerate.

Theorem 12.11. *Let B and C be bases of V. For all $\beta \in \mathrm{Bil}(V)$ we have*

$$ \boxed{{}_C[\beta]_C = {}_B\Delta_C^{\mathrm{t}}\, {}_B[\beta]_B\, {}_B\Delta_C.} $$

Proof. Let $B = \{b_1, \ldots, b_n\}$, $C = \{c_1, \ldots, c_n\}$ and $S = (s_{ij}) = {}_B\Delta_C$. For $1 \leq i, j \leq n$ we have

$$ \beta(c_i, c_j) = \sum_{k,l=1}^{n} s_{ki} \beta(b_k, b_l) s_{lj} = \sum_{k=1}^{n} s_{ki} ({}_B[\beta]_B S)_{kj} = (S^{\mathrm{t}}\, {}_B[\beta]_B S)_{ij}. $$ □

Definition 12.12. Matrices $A, B \in K^{n\times n}$ are called *congruent*, if there exists $S \in \mathrm{GL}(n, K)$ with $B = SAS^{\mathrm{t}}$.[4]

[4] In contrast to similarity, we do not introduce a symbol for congruence of matrices.

Remark 12.13.

(a) As in Remark 7.30 one shows that congruence of matrices is an equivalence relation. By Theorem 12.9 and Theorem 12.11, every bilinear form on V determines a congruence class of Gram matrices. We show in the next section that, similar to endomorphisms, bilinear forms can be diagonalized.

(b) By Lemma 5.15 we have

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

By symmetry it follows that $\operatorname{rk} B = \operatorname{rk} A$, i. e., congruent matrices have the same rank. On the other hand, the example $A = 1_1$ and $B = (2)A(2) = (4)$ shows that in general congruent matrices have neither the same trace, determinant, nor the same eigenvalues.

(c) By the principal axis theorem, every real symmetric matrix is congruent and similar to a diagonal matrix.

12.2 Sylvester's Law of Inertia

Definition 12.14. Let $\beta \in \operatorname{Bil}(V)$ and $S \subseteq V$. As in Euclidean spaces we define the *orthogonal complement* of S by

$$S^{\perp} := \{v \in V : \forall s \in S : \beta(v, s) = 0\} \leq V.$$

Remark 12.15.

(a) In the definition of $S^{\perp}$ one would strictly speaking have to distinguish between "left-orthogonal" and "right-orthogonal" ($\forall s \in S : \beta(s, v) = 0$). As a rule, we will assume that β is (anti)symmetric, so that both versions are equivalent.

(b) β is degenerate if and only if $V^{\perp} \neq \{0\}$.

Theorem 12.16. *Let $\beta \in \operatorname{Bil}(V)$ be nondegenerate. For $U \leq V$ we have* $\dim V = \dim(U) + \dim(U^{\perp})$.

Proof. Let $F\colon V \to V^*$, $v \mapsto F_v$ be the isomorphism from Theorem 12.7. Then we have

$$F_v \in F(U^{\perp}) \iff \forall u \in U : F_v(u) = \beta(v, u) = 0 \iff F_v \in U^0,$$

i. e., $F(U^{\perp}) = U^0$. The claim follows from Lemma 7.44. □

Remark 12.17. Caution: In contrast to Euclidean spaces, in general $U \cap U^{\perp} \neq \{0\}$. For example, for $U = \langle (1, 1) \rangle \leq V = \mathbb{F}_2^2$ with $[\beta] = 1_2$ we have $U = U^{\perp}$.

Definition 12.18. Let $\beta \in \operatorname{Bil}(V)$ be symmetric. A basis $B = \{b_1, \ldots, b_n\}$ of V is called an *orthogonal basis* with respect to β if $\beta(b_i, b_j) = 0$ for $i \neq j$. If in addition $\beta(b_i, b_i) = 1$, then one calls B an *orthonormal basis* with respect to β.

Remark 12.19.

(a) Clearly, B is an orthogonal basis or orthonormal basis, respectively, with respect to β if and only if ${}_B[\beta]_B$ is a diagonal matrix or identity matrix, respectively.

(b) By Corollary 11.11, every inner product has an orthonormal basis.

Theorem 12.20. *Let $1+1 \neq 0$ in K. Then every symmetric bilinear form on V has 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 some $v, w \in V$. 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 thus exists $b_1 \in V \setminus \{0\}$ with $\beta(b_1, b_1) \neq 0$. Let $U := b_1^\perp$. For all $v \in V$ we have $v - \frac{\beta(v,b_1)}{\beta(b_1,b_1)} b_1 \in U$. This shows $V = \langle b_1 \rangle \oplus U$. The restriction of β to $U \times U$ is a symmetric bilinear form. By induction, U has 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 reads: Every symmetric matrix is congruent to a diagonal matrix (if $1+1 \neq 0$). We will see that the diagonal entries can be chosen in a special way.

(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, since $\beta(v,v) = 0$ for all $v \in K^2$.

(c) The Gram-Schmidt process for computing orthogonal bases does not always work in this generality, since one has to divide by $\beta(b_i, b_i)$, but this value can be 0. Instead, we modify the Gaussian algorithm as follows:

(1) Let $A := [\beta]$. Suppose inductively that rows and columns $1, \ldots, k-1$ of A already have the desired diagonal form (initially let $k = 1$).

(2) First suppose $a_{kk} \neq 0$. Then one can achieve $a_{ik} = 0$ for $i = k+1, \ldots, n$ by adding, as usual, a multiple of the k-th row to the rows below it. This corresponds to multiplication by elementary matrices $S_1, \ldots, S_{n-k}$ from the left. Afterwards one sets $a_{ki} = 0$ for $i = k+1, \ldots, n$. This corresponds to multiplication by $S_1^{\mathrm{t}}, \ldots, S_{n-k}^{\mathrm{t}}$ from the right (in any order). Altogether, A becomes SAS^{t}, where $S = S_1 \ldots S_{n-k}$. Since $(SAS^{\mathrm{t}})^{\mathrm{t}} = (S^{\mathrm{t}})^{\mathrm{t}} A^{\mathrm{t}} S^{\mathrm{t}} = SAS^{\mathrm{t}}$, A remains symmetric under this procedure. In the case $k = n$ we are done. Otherwise we increase k.

(3) Now let $a_{kk} = 0$. If there exists $l > k$ with $a_{ll} \neq 0$, then swap rows k and l and afterwards columns k and l. This interchanges a_{kk} and a_{ll}, and A remains symmetric. We are now in situation (2).

(4) Finally, let $a_{ii} = 0$ for $i = k, \ldots, n$. If $a_{ij} = 0$ for all $k \leq i, j \leq n$, then we are done. So let $a_{ij} \neq 0$ for some $k \leq i < j \leq n$. We add the i-th row to the j-th row and afterwards the i-th column to the j-th column. The entry at position (j,j) thereby becomes $2a_{ij} \neq 0$ (note $1+1 \neq 0$). Hence we are in situation (3).

(5) To determine the orthogonal basis B, perform all row operations (but not the column operations) on the identity matrix. This yields the matrix $S \in \mathrm{GL}(n, K)$ such that SAS^{t} is a diagonal matrix (cf. Theorem 6.17). By Theorem 12.11, the rows of S are the vectors of B.

Example 12.22.

$$(A|1_3) = \left(\begin{array}{ccc|ccc} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc|ccc} 2 & 1 & 2 & 1 & 1 & 0 \\ 0 & -1/2 & 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & -2 & -1 & -1 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 2 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1/2 & 0 & -1/2 & 1/2 & 0 \\ 0 & 0 & -2 & -1 & -1 & 1 \end{array}\right)$$

$$S = \begin{pmatrix} 1 & 1 & 0 \\ -1/2 & 1/2 & 0 \\ -1 & -1 & 1 \end{pmatrix}$$

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*

$$_B[\beta]_B = \operatorname{diag}(1_r, -1_s, 0_t).$$

The numbers r, s, t do not depend on B.

Proof. By Theorem 12.20 there exists an orthogonal basis $B = \{b_1, \dots, b_n\}$ of V. After reordering we may 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 $i = r+1, \dots, r+s$. This proves the existence of B.

Since $V^\perp = \langle b_{r+s+1}, \dots, b_n \rangle$, the number $t = \dim V^\perp$ 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 $\lambda_i \in \mathbb{R}$ we have $\beta(v, v) = \sum_{i=1}^r \lambda_i^2 \geq 0$ with equality if and only if $v = 0$. Let $B' = \{b'_1, \dots, b'_n\}$ be another basis of V with

$$_{B'}[\beta]_{B'} = \operatorname{diag}(1_{r'}, -1_{s'}, 0_t).$$

Let $V'_{\leq 0} := \langle b'_{r'+1}, \dots, b'_n \rangle$. For $v \in V_+ \cap V'_{\leq 0}$ we have on the one hand $\beta(v, v) \geq 0$ and on the other hand $\beta(v, v) \leq 0$. This shows $\beta(v, v) = 0$ and $v = 0$. Hence $V_+ \cap V'_{\leq 0} = \{0\}$ and

$$r + s' + t = \dim(V_+ \oplus V'_{\leq 0}) \leq \dim V = n = r + s + t.$$

It follows that $s' \leq s$. By symmetry, $s' = s$ and $r' = r$. □

Definition 12.24. In the situation of Theorem 12.23, one calls $\operatorname{ind}\beta := (r, s, t)$ the *index* of β.[5] For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ we define $\operatorname{ind} A := \operatorname{ind}\beta$ via the bilinear form β with $[\beta] = A$.

Example 12.25.

(a) Every inner product on $\mathbb{R}^n$ has index $(n, 0, 0)$.

[5] This notion is not uniform in the literature! Some authors call $r - s$ the *signature* of β. The index can be determined from the dimension, rank, and signature.

(b) The matrix A from Example 12.22 has index $(1, 2, 0)$.

(c) If $\mathrm{ind}\beta = (r, s, t)$, then β is degenerate if and only if $t > 0$.

(d) In the theory of relativity one considers *Minkowski space* $\mathbb{R}^4$ with respect to a bilinear form β with index $(3, 1, 0)$. The fourth dimension describes time.

Theorem 12.26. *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. By the matrix version of the principal axis theorem, A is congruent to $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A (the algebraic multiplicity of each eigenvalue agrees with its geometric multiplicity). One can apply the argument from the proof of Theorem 12.23 to this matrix. The positive λ_i are thereby transformed to 1 and the negative ones to -1. □

Remark 12.27. To construct a basis for Sylvester's law of inertia, one first starts with an arbitrary orthogonal basis $b_1, \ldots, b_n$ with respect to β (for example with the modified Gaussian algorithm or the principal axis theorem). The basis vectors are sorted according to the diagonal entries $\lambda := \beta(b_i, b_i)$ (first positive, then negative, then zeros). For $\lambda \neq 0$ one divides b_i by $\sqrt{|\lambda|}$. For the matrix from Example 12.22 one obtains the basis

$$\frac{1}{\sqrt{2}}(1, 1, 0),\ \frac{1}{\sqrt{2}}(-1, 1, 0),\ \frac{1}{\sqrt{2}}(-1, -1, 1).$$

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) $\mathrm{ind}A = \mathrm{ind}B$.

(3) A and B have the same number of positive eigenvalues and the same number of negative eigenvalues.[6]

Proof. Since congruent matrices describe the same bilinear form, we have (1)⇒(2). From Theorem 12.26 we obtain (2)⇒(3). If (3) holds, then A and B have the same index by Theorem 12.26. Hence A and B are congruent, i. e., (1) holds. □

Remark 12.29.

(a) For every $m \leq n$ there are exactly $m + 1$ congruence classes of symmetric $n \times n$ matrices of rank m (namely with index $(i, m - i, n - m)$ for $i = 0, \ldots, m$). Hence there are

$$\sum_{m=0}^{n}(m + 1) = \sum_{m=1}^{n+1} m = \frac{(n + 1)(n + 2)}{2}$$

symmetric bilinear forms on $\mathbb{R}^n$ up to choice of basis (cf. Example 1.16).

[6] This describes the *inertia* in Sylvester's theorem.

(b) If one works over $\mathbb{C}$ instead of $\mathbb{R}$, one can always achieve ${}_B[\beta]_B = \operatorname{diag}(1_r, 0_t)$. Indeed, one can divide every basis vector b with $w := \beta(b,b) \neq 0$ by $\sqrt{w} \in \mathbb{C}$.[7] Hence 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 nondegenerate. Then there exists a basis B of V with ${}_B[\beta]_B = \left(\begin{smallmatrix} 0_n & 1_n \\ -1_n & 0_n \end{smallmatrix}\right)$. In particular, $\dim V = 2n$ is even.*[8]

Proof. Induction on $\dim V$. Let $b_1 \in V \setminus \{0\}$. Since β is nondegenerate, 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 $v = \lambda b_1 + \mu c_1 \in U$ we have $\beta(v,b_1) = \mu\beta(c_1,b_1) = -\mu$ and $\beta(v,c_1) = \lambda$. This shows $U \cap U^\perp = \{0\}$. By Theorem 12.16 we have $V = U \oplus U^\perp$. For $u \in U^\perp$ there exists $v \in V$ with $\beta(u,v) \neq 0$. Writing $v = v_1 + v_2$ with $v_1 \in U$ and $v_2 \in U^\perp$, we obtain $\beta(u,v_2) = \beta(u,v) \neq 0$. Hence the restriction β' of β to $U^\perp \times U^\perp$ is also nondegenerate 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, nondegenerate bilinear form is called a *symplectic space*. We do not go into this any further.

12.3 Positive Definite Matrices

Remark 12.32. In this section we generalize the positive definiteness of inner products in Euclidean spaces. Throughout, let V be an $\mathbb{R}$-vector space.

Definition 12.33. A symmetric bilinear form $\beta \in \operatorname{Bil}(V)$ is called

- *positive (semi)definite* if $\beta(v,v) > 0$ (respectively $\beta(v,v) \geq 0$) for all $v \in V \setminus \{0\}$,
- *negative (semi)definite* if $\beta(v,v) < 0$ (respectively $\beta(v,v) \leq 0$) for all $v \in V \setminus \{0\}$,
- *indefinite* if $\exists v, w \in V : \beta(v,v) > 0 > \beta(w,w)$.

These notions carry over to symmetric matrices A by choosing a β with $[\beta] = A$.

Remark 12.34.

(a) Clearly, every positive definite bilinear form is also positive semidefinite. If β is positive (semi)definite, then $-\beta$ is negative (semi)definite. Hence one can usually restrict to the positive property.

(b) If β is positive (semi)definite, then the associated quadratic form from Remark 12.4 is positive (respectively nonnegative) on $V \setminus \{0\}$.

[7] In $\mathbb{C}$ there is no canonical choice of the square root.
[8] cf. Example 12.10

(c) The sum of positive (semi)definite bilinear forms and matrices is again positive (semi)definite. One can regard positive semidefinite matrices as a multidimensional generalization of nonnegative real numbers (see Theorem 12.46, Example 19.20 and Exercise III.11). However, the product of positive (semi)definite matrices is usually not symmetric.

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 $x \in \mathbb{R}^n$ we have

$$xAx^{\mathrm{t}} = 2\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n-1} x_i x_{i+1} = x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + x_n^2 \geq 0.$$

In the case $xAx^{\mathrm{t}} = 0$ we have $x_1 = 0$ and $x_i = x_{i+1}$ for $i = 1, \ldots, n-1$. Hence $x = 0$ and 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 is positive semidefinite at local minima of f.

Theorem 12.36. *For every symmetric bilinear form $\beta \in \mathrm{Bil}(V)$ with index (r, s, t) we have*

(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 $s = 0$ (respectively $s = t = 0$), then $\beta(b, b) \geq 0$ (respectively $\beta(b, b) > 0$) for all $b \in B$. Since B is an orthogonal basis, it follows easily that β is positive (semi)definite. Moreover, 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 (respectively nonnegative).*

Proof. As is well known, all eigenvalues of A are real. The claim follows from Theorem 12.26. □

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. Since $\operatorname{rk}(A - 1_n) = 1$, the number 1 is an eigenvalue of A with (geometric) multiplicity $n-1$ (cf. Exercise I.28). The missing eigenvalue must be $\operatorname{tr}(A) - (n-1) = 2n - n + 1 = n + 1$. Hence A is positive definite.

Remark 12.39. Since eigenvalues are hard to compute 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 matrix $A \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if there exists a matrix $S \in \mathbb{R}^{n\times n}$ with $A = SS^{\mathrm{t}}$. Moreover, 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 $v \in \mathbb{R}^n$ we have $vAv^{\mathrm{t}} = vS(vS)^{\mathrm{t}} = |vS|^2 \geq 0$. Hence 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 $S = UD$ we have $SS^{\mathrm{t}} = A$ because of $D^2 = D$. If A is positive definite, then $S = U$ is invertible. □

Theorem 12.41 (Sylvester's criterion)**.** *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be symmetric and $A_k := (a_{ij})_{i,j=1}^k$ for $1 \leq k \leq n$. Then 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 $\det A_1 = a_{11} > 0$. Now let $n \geq 2$ and $1 \leq k \leq n$. Let A be positive definite and $v \in \mathbb{R}^k \setminus \{0\}$. For $w := (v_1, \ldots, v_k, 0, \ldots, 0) \in \mathbb{R}^n$ we have $vA_kv^{\mathrm{t}} = wAw^{\mathrm{t}} > 0$. Hence 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, suppose $\det A_k > 0$ 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. Appending a 0 yields the vectors $b_i := (b'_i, 0) \in \mathbb{R}^n$ with

$$\beta(b_i, b_j) = b_i A b_j^{\mathrm{t}} = b'_i A_{n-1} (b'_j)^{\mathrm{t}} = \delta_{ij}$$

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, we have $V_1 \cap V_1^\perp = \{0\}$. 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 $\lambda = \det D = \det(S)^2 \det(A_n) > 0$. Together with D, 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

$$A := \begin{pmatrix} 2 & -1 & . & . & -1 \\ -1 & 2 & -1 & . & . \\ . & -1 & 2 & -1 & . \\ . & . & -1 & 2 & -1 \\ -1 & . & . & -1 & 2 \end{pmatrix}.$$

By Example 12.35 we have $\det A_k > 0$ 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) The numbers $\det A_1, \ldots, \det A_n$ are called the *leading principal minors* of A.

(b) Caution: $\det A_k \geq 0$ for $k = 1, \ldots, n$ does not imply that A is positive semidefinite (consider for example $A = \operatorname{diag}(0, -1)$).

(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 $\lambda \leq 0$ we then have

$$\chi_A(\lambda) = (-1)^n \big(|\lambda|^n + |a_1||\lambda|^{n-1} + \ldots + |a_n|\big) \neq 0.$$

Hence A has only positive eigenvalues. By Corollary 12.37, A is positive definite.

Conversely, let A be positive definite with eigenvalues $\lambda_1, \ldots, \lambda_n > 0$. Then

$$\chi_A = \prod_{i=1}^{n} (X - \lambda_i) = X^n - \Big(\sum_{i=1}^{n} \lambda_i\Big) X^{n-1} + \Big(\sum_{i<j} \lambda_i \lambda_j\Big) X^{n-2} - \ldots + (-1)^n \lambda_1 \ldots \lambda_n$$

and

$$a_k (-1)^k = \sum_{i_1 < \ldots < i_k} \lambda_{i_1} \ldots \lambda_{i_k} > 0$$

for $k = 1, \ldots, n$. □

Example 12.45. For

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

we have

$$\chi_A = (X-2)^2(X-1) + 2 - 2(X-2) - (X-1) = X^3 - 5X^2 + 5X + 3.$$

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 has exactly one positive (semi)definite k-th root $W \in \mathbb{R}^{n \times n}$, i. e., $W^k = A$.*

Proof. By 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, we have $\lambda_1, \ldots, \lambda_n \geq 0$ by 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 of $(S^{\mathrm{t}}AS)^i = S^{\mathrm{t}}A^iS$ for $i \in \mathbb{N}$ we have

$$\alpha(A) = S\alpha(S^{\mathrm{t}}AS)S^{\mathrm{t}} = W.$$

Let $B \in \mathbb{R}^{n\times n}$ also be positive (semi)definite with $B^k = A$. Again there exists 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 $B^k = A$. Hence 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 inner product, we now have a "skew-symmetric" map that is linear only in the first coordinate. Symmetric and orthogonal maps are replaced by Hermitian and unitary maps, respectively. The counterpart of the principal axis theorem is the spectral theorem.

Definition 13.2. An *inner product* on a $\mathbb{C}$-vector space V is a map $V \times V \to \mathbb{C}$, $(v, w) \mapsto [v, w]$ 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*),
- $[v, w] = \overline{[w, v]}$ (*Hermitian*),
- $[\lambda u + v, w] = \lambda[u, w] + [v, w]$.

Together with an inner product, V becomes a *unitary space*. Vectors $v, w \in V$ are called *orthogonal*, written $v \perp w$, if $[v, w] = 0$. One calls $|v| := \sqrt{[v, v]} \geq 0$ the *norm* of v. In the case $|v| = 1$ one calls v *normalized*.

Example 13.3. The *standard inner 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 $v, w \in V$. This extends Example 11.4 from $\mathbb{R}^n$ to $\mathbb{C}^n$. One easily checks that the properties of the inner product are satisfied.

Remark 13.4.

(a) For $u, v, w \in V$ and $\lambda \in \mathbb{C}$ we have

$$[u, \lambda v + w] = \overline{[\lambda v + w, u]} = \overline{\lambda[v, u] + [w, u]} = \overline{\lambda}[u, v] + [u, w],$$

i. e., the inner product is not linear in the second component. Therefore one speaks of a *sesquilinear form*.[1]

(b) Most of the theorems proved in Section 11.1 do not depend essentially on the bilinearity of the inner product and can therefore be carried over 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).

[1] *Sesqui* is Latin for one and a half.

(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) $|\lambda v| = \sqrt{[\lambda v, \lambda v]} = \sqrt{\lambda\overline{\lambda}[v,v]} = \sqrt{|\lambda|^2}\sqrt{[v,v]} = |\lambda||v|$.

(b) W. l. o. g. let $w \neq 0$ and $\lambda := \frac{[v,w]}{[w,w]}$. Since $\overline{\lambda} = \frac{[w,v]}{[w,w]}$, we have

$$0 \leq |v-\lambda w|^2 = [v-\lambda w, v-\lambda w] = [v,v] - \lambda[w,v] - \overline{\lambda}[v,w] + |\lambda|^2[w,w] = |v|^2 - \frac{|[v,w]|^2}{|w|^2}.$$

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 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

$$[v,w] + [w,v] = [v,w] + \overline{[v,w]} = 2a \leq 2\sqrt{a^2+b^2} = 2|[v,w]|.$$

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 $|v+w| \leq |v|+|w|$. Hence $|v| = |v+w-w| \leq |v+w|+|w|$ and $|v|-|w| \leq |v+w|$. Interchanging 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 $1 \leq i,j \leq n$. The Gram-Schmidt process for computing an orthonormal basis works unchanged for unitary spaces. In particular, every unitary space has an orthonormal basis.

(b) For $U \leq V$ one defines as usual $U^\perp := \{v \in V : \forall u \in U : u \perp v\} \leq V$. 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 a unique $f^* \in \mathrm{End}(V)$ with $[f(v), w] = [v, f^*(w)]$ for all $v, w \in V$.*

Proof. Let $b_1, \ldots, b_n$ be an orthonormal basis of V. We define $f^* \in \mathrm{End}(V)$ by

$$f^*(b_j) := \sum_{i=1}^{n} [b_j, f(b_i)] b_i$$

for $j = 1, \ldots, n$. For $v = \sum \lambda_i b_i$ and $w = \sum \mu_j b_j$ we have

$$[f(v), w] = \sum_{i,j=1}^{n} \lambda_i \overline{\mu_j} [f(b_i), b_j] = \sum_{i,j=1}^{n} \lambda_i \overline{\mu_j} [b_i, [b_j, f(b_i)] b_i] = \sum_{i,j=1}^{n} \lambda_i \overline{\mu_j} [b_i, f^*(b_j)] = [v, f^*(w)].$$

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 $v := f^*(w) - f_1(w)$ we obtain $f^*(w) = f_1(w)$ for all $w \in V$. This shows $f_1 = f^*$. □

Definition 13.8. In the situation of Theorem 13.7, one calls f^* the *adjoint* map of 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) Caution: For the adjoint map of f we use the same notation as for the dual map of f. Although the dual map lies in $\mathrm{End}(V^*)$, there is a close connection (Exercise II.19).

Definition 13.10. Let V be a unitary space. One calls $f \in \mathrm{End}(V)$

- *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 $v \in \operatorname{Ker} f$ we have $|v|^2 = [v, v] = [f(v), f(v)] = 0$, i. e., $v = 0$. Hence f is an isomorphism. Since $[f(v), w] = [v, f^{-1}(w)]$ for all $v, w \in V$, it follows moreover that $f^* = f^{-1}$. If $f, g \in \mathrm{End}(V)$ are unitary, then so are $f \circ g$ and f^{-1}. Hence the unitary maps form a subgroup $\mathrm{U}(V)$ of $\mathrm{GL}(V)$. One calls $\mathrm{U}(V)$ the *unitary group* of V.

(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) For normal $f \in \mathrm{End}(V)$ and $v \in V$ we have

$$|f^*(v)|^2 = [f^*(v), f^*(v)] = [f(f^*(v)), v] = [f^*(f(v)), v] = [f(v), f(v)] = |f(v)|^2,$$

so $|f^*(v)| = |f(v)|$.

(e) Hermitian and unitary maps are clearly normal.

Definition 13.12. For $A \in \mathbb{C}^{n\times m}$ let $A^* := \overline{A}^{\mathrm{t}} = \overline{A^{\mathrm{t}}}$ be the *adjoint* matrix of A. For $A \in \mathrm{GL}(n, \mathbb{C})$ we use the abbreviation $A^{-*} := (A^{-1})^* = (A^*)^{-1}$.

Lemma 13.13. *Let V be unitary with orthonormal basis B and $f \in \mathrm{End}(V)$. Then ${}_B[f^*]_B = {}_B[f]_B^*$.*

Proof. Let $B = \{b_1, \ldots, b_n\}$, $f(b_i) = \sum_{j=1}^n a_{ji} b_j$ and $f^*(b_i) = \sum_{j=1}^n a_{ji}^* b_j$ for $i = 1, \ldots, n$. The claim follows from

$$\overline{a_{ji}^*} = [b_j, f^*(b_i)] = [f(b_j), b_i] = a_{ij}. \qquad \square$$

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. We have

$$f \text{ Hermitian} \iff f^* = f \overset{13.13}{\iff} A^* = A \iff \overline{A^*} = \overline{A} \iff A^{\mathrm{t}} = \overline{A}.$$

The other statements are proved 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 (respectively unitary) if and only if they are symmetric (respectively orthogonal).

(c) Inverting, transposing and complex conjugating are commuting operators on $\mathbb{C}^{n\times n}$ (cf. Remark 5.9). If A is Hermitian (respectively unitary, normal), then so are $\overline{A}$, A^{t} and A^* (verify this).

Remark 13.17. The unitary matrices form a subgroup $\mathrm{U}(n,\mathbb{C})$ of $\mathrm{GL}(n,\mathbb{C})$, which corresponds to $\mathrm{U}(V)$ as usual. For $A \in \mathrm{U}(n,\mathbb{C})$ we have $|\det A|^2 = \det(A)\overline{\det(A)} = \det(A^{\mathrm{t}}\overline{A}) = 1$. By 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 $f \in \mathrm{End}(V)$. Then 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. Clearly, A^* is also a diagonal matrix. In particular, A and A^* commute. By 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 yields an orthonormal basis of eigenvectors. Let $n \geq 2$. By the fundamental theorem of algebra, f has an eigenvalue $\lambda \in \mathbb{C}$ with eigenvector $b_1 \in V$. W. l. o. g. let $|b_1| = 1$. Together with f, also $g := f - \lambda \mathrm{id}_V$ is normal (verify this). From Remark 13.11 it follows that

$$|f^*(b_1) - \overline{\lambda} b_1| = |g^*(b_1)| = |g(b_1)| = 0,$$

i. e., b_1 is an eigenvector of f^* for the eigenvalue $\overline{\lambda}$. For $U := \langle b_1 \rangle$ we have $V = U \oplus U^\perp$. For $u \in U^\perp$ we have $[f(u), b_1] = [u, f^*(b_1)] = \lambda[u, b_1] = 0$, so $f(u) \in U^\perp$ and analogously $f^*(u) \in U^\perp$. Hence the restriction of f to $U^\perp$ is a normal endomorphism. By induction there exists an orthonormal basis $b_2, \ldots, b_n$ of $U^\perp$ of eigenvectors of f. Now $b_1, \ldots, b_n$ is an orthonormal basis of V 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 $S^*AS = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. In contrast to the principal axis theorem, one does not obtain that S is orthogonal (and hence real).

Corollary 13.20. *Let V be unitary and $f \in \mathrm{End}(V)$ normal with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. Then*

(a) f Hermitian $\iff$ $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.

(b) f unitary $\iff$ $|\lambda_1| = \ldots = |\lambda_n| = 1$.

Proof. By the spectral theorem there exists an orthonormal basis B of V with $D := {}_B[f]_B = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. By Corollary 13.14, f is Hermitian (respectively unitary) if and only if $\overline{D} = D^{\mathrm{t}} = D$ (respectively $1_n = D\overline{D} = \mathrm{diag}(|\lambda_1|^2, \ldots, |\lambda_n|^2)$). This proves the claim. □

[2]The set of eigenvalues of $f \in \mathrm{End}(V)$ is called the *spectrum* of f.

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 $[v,w] = 0$. Hence eigenvectors for distinct eigenvalues are orthogonal (as for symmetric matrices, see Remark 11.40). The orthonormal basis in the spectral theorem can therefore be determined with the Gram-Schmidt process, as in the case of the principal axis theorem.

Example 13.22. For

$$A := \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 4 & 5\mathrm{i} & 2\mathrm{i} \\ -2 & 2\mathrm{i} & 8\mathrm{i} \end{pmatrix}$$

one computes

$$\chi_A = (X-5\mathrm{i})^2(X-8\mathrm{i}) - 32\mathrm{i} + 8(X-5\mathrm{i}) + 16(X-8\mathrm{i}) = X^3 - 18\mathrm{i}X^2 - 81X = X(X-9\mathrm{i})^2.$$

Since

$$A \sim \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 0 & {}^{9\mathrm{i}}/_{5} & {}^{18\mathrm{i}}/_{5} \\ 0 & {}^{18\mathrm{i}}/_{5} & {}^{36\mathrm{i}}/_{5} \end{pmatrix} \sim \begin{pmatrix} 5\mathrm{i} & -4 & 2 \\ 0 & {}^{9\mathrm{i}}/_{5} & {}^{18\mathrm{i}}/_{5} \\ 0 & 0 & 0 \end{pmatrix}$$

the vector $(2\mathrm{i}, -2, 1)$ 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 $9\mathrm{i}$ is spanned by $c_2 := (0,1,2)$ and $c_3 := (\mathrm{i},1,0)$. We observe that these vectors are orthogonal to b_1. Hence A must be normal. The Gram-Schmidt process yields:

$$\begin{aligned} b_2 &:= \frac{1}{\sqrt{5}}(0,1,2), \\ c_3' &:= c_3 - [c_3,b_2]b_2 = (\mathrm{i}, {}^{4}/_{5}, {}^{-2}/_{5}), \\ b_3 &:= \frac{1}{3\sqrt{5}}(5\mathrm{i},4,-2). \end{aligned}$$

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. Moreover, A is normal if and only if S^*AS is a diagonal matrix.*

Proof. W. l. o. g. 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 we have $A = \left(\begin{smallmatrix} \lambda & * \\ 0 & A_1 \end{smallmatrix}\right)$ with $A_1 \in \mathbb{C}^{(n-1)\times(n-1)}$. Inductively there exists an $S_1 \in \mathrm{U}(n-1,\mathbb{C})$ such that $S_1^* A_1 S_1$ is an upper triangular matrix. Now $S := \mathrm{diag}(1_1, S_1) \in \mathrm{U}(n,\mathbb{C})$ 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. Since

$$\sum_{j\leq i} |d_{ji}|^2 = (D^*D)_{ii} = (DD^*)_{ii} = \sum_{j\geq i} |d_{ij}|^2$$

we inductively have $d_{ij} = 0$ for $j > i$ and $i = 1, \ldots, n$. This shows that D is a diagonal matrix. Conversely, if D is a diagonal matrix, then A is normal by the spectral theorem. □

Example 13.24. The matrix

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

clearly has the eigenvector e_3 for the eigenvalue 1. Passing 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 we obtain

$$A \approx \frac{1}{3}\begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1-\mathrm{i} & -1 \\ 0 & 1 & 1+\mathrm{i} \end{pmatrix}\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}\begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1+\mathrm{i} & 1 \\ 0 & -1 & 1-\mathrm{i} \end{pmatrix} = \begin{pmatrix} 1 & (2+2\mathrm{i})/\sqrt{3} & 2/\sqrt{3} \\ 0 & \mathrm{i} & 1-2\mathrm{i} \\ 0 & 0 & -\mathrm{i} \end{pmatrix}.$$

Theorem 13.25 (Law of inertia 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. By the spectral theorem we may assume $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. By Corollary 13.20, the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A are real. Let w. l. o. g. $\lambda_1, \ldots, \lambda_r > 0$, $\lambda_{r+1}, \ldots, \lambda_{r+s} < 0$ and $\lambda_{r+s+1} = \ldots = \lambda_n = 0$. The equation now holds with

$$S := \mathrm{diag}\big(\sqrt{|\lambda_1|}^{-1}, \ldots, \sqrt{|\lambda_{r+s}|}^{-1}, 1_t\big) \in \mathrm{GL}(n, \mathbb{C}).$$

Since $\operatorname{rk} A = r + s$, t is uniquely determined. For all $v \in \langle e_1, \ldots, e_r\rangle \setminus \{0\}$ we have $vAv^* > 0$. Conversely, let $U \leq \mathbb{C}^n$ with $uAu^* > 0$ for all $u \in U \setminus \{0\}$. For $u \in U \cap \langle e_{r+1}, \ldots, e_n\rangle$ we have on the one hand $uAu^* \geq 0$ and on the other hand $uAu^* \leq 0$. 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 maximal dimension of a subspace $U \leq \mathbb{C}^n$ with $uAu^* > 0$ for all $u \in U \setminus \{0\}$. In this way, r does not depend on the choice of S. Now $s = n - r - t$ is also uniquely determined. □

Example 13.26. The numbers r, s, t in Theorem 13.25 can be determined with the modified Gaussian algorithm from Remark 12.21 (note that $1 + 1 \neq 0$ in $\mathbb{C}$). For example:

$$\begin{aligned}
(A|1_3) &= \left(\begin{array}{ccc|ccc} 1 & \mathrm{i} & 2 & 1 & 0 & 0 \\ -\mathrm{i} & 1 & -1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & \mathrm{i} & 2 & 1 & 0 & 0 \\ 0 & 0 & 2\mathrm{i}-1 & \mathrm{i} & 1 & 0 \\ 0 & -2\mathrm{i}-1 & -1 & -2 & 0 & 1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -2\mathrm{i}-1 & -2 & 0 & 1 \\ 0 & 2\mathrm{i}-1 & 0 & \mathrm{i} & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & -2 & 0 & 1 \\ 0 & 0 & 5 & 2-3\mathrm{i} & 1 & 2\mathrm{i}-1 \end{array}\right) \\
&\sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & (2-3\mathrm{i})/\sqrt{5} & 1/\sqrt{5} & (2\mathrm{i}-1)/\sqrt{5} \\ 0 & 0 & -1 & -2 & 0 & 1 \end{array}\right)
\end{aligned}$$

$$(r, s, t) = (2, 1, 0), \qquad S^* = \frac{1}{\sqrt{5}}\begin{pmatrix} \sqrt{5} & 0 & 0 \\ 2-3\mathrm{i} & 1 & 2\mathrm{i}-1 \\ -2\sqrt{5} & 0 & \sqrt{5} \end{pmatrix}$$

Remark 13.27. Mirsky's theorem gives a necessary and sufficient condition for the existence of matrices with prescribed eigenvalues and main diagonal entries (whose sums must be equal). For Hermitian matrices, stronger restrictions apply to these numbers.

Lemma 13.28. *Let $\lambda_1 \geq d \geq \lambda_2$ be real. Then there exists an $S \in \mathrm{O}(2,\mathbb{R})$ such that $S^{\mathrm{t}} \operatorname{diag}(\lambda_1,\lambda_2)S$ has first entry d.*

Proof. In the case $\lambda_1 = \lambda_2$ we have $d = \lambda_1$ and $S := 1_2$ satisfies the claim. So let $\lambda_1 > \lambda_2$. Then

$$S = (s_{ij}) := \frac{1}{\sqrt{\lambda_1 - \lambda_2}} \begin{pmatrix} \sqrt{d-\lambda_2} & -\sqrt{\lambda_1 - d} \\ \sqrt{\lambda_1 - d} & \sqrt{d-\lambda_2} \end{pmatrix} \in \mathrm{O}(2,\mathbb{R}).$$

For $A = (a_{ij}) := S^{\mathrm{t}} \operatorname{diag}(\lambda_1,\lambda_2)S$ we have

$$a_{11} = \lambda_1 s_{11}^2 + \lambda_2 s_{21}^2 = \frac{1}{\lambda_1 - \lambda_2}\big((d-\lambda_2)\lambda_1 + (\lambda_1 - d)\lambda_2\big) = d. \qquad \square$$

Theorem 13.29 (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 $\lambda_1 \geq \ldots \geq \lambda_n$ with $\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{k} \lambda_i$ for $k = 1,\ldots,n$ and $\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \lambda_i$ are given, 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 $1 \leq k \leq n$ we have

$$\sum_{i=1}^{k} d_i = \sum_{i=1}^{k} a_{ii} = \sum_{i=1}^{k}\sum_{j=1}^{n} \overline{s_{ji}} s_{ji} \lambda_j = \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{k} |s_{ji}|^2. \tag{13.1}$$

Since S is unitary, $t_j := \sum_{i=1}^{k} |s_{ji}|^2 \leq 1$ with equality if $k = n$. It follows that

$$\begin{aligned} \sum_{j=1}^{n} t_j &= \sum_{i=1}^{k}\sum_{j=1}^{n} |s_{ji}|^2 = k, \\ \sum_{i=1}^{k}(d_i - \lambda_i) &= \sum_{j=1}^{n} \lambda_j t_j - \sum_{j=1}^{k} \lambda_j + \lambda_k \overbrace{\big(k - \sum_{i=1}^{n} t_i\big)}^{=0} \\ &= \sum_{j=1}^{k} \underbrace{(\lambda_j - \lambda_k)}_{\geq 0} \underbrace{(t_j - 1)}_{\leq 0} + \sum_{i=k+1}^{n} t_i \underbrace{(\lambda_i - \lambda_k)}_{\leq 0} \leq 0. \end{aligned}$$

Hence $\sum_{i=1}^{k} d_i \leq \sum_{i=1}^{k} \lambda_i$ 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 $n \geq 2$. Because of

$$\lambda_1 \geq d_1 \geq \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} \lambda_i \geq \lambda_n$$

there exists a k with $\lambda_k \geq d_1 \geq \lambda_{k+1}$. By Lemma 13.28 there exists a $T \in \mathrm{O}(2,\mathbb{R})$ such that $T^{\mathrm{t}} \operatorname{diag}(\lambda_k, \lambda_{k+1})T$ has main diagonal (d_1, λ'), where $\lambda' := \lambda_k + \lambda_{k+1} - d_1$ by (a). For

$$S_1 := \operatorname{diag}(1_{k-1}, T, 1_{n-k-1})P \in \mathrm{O}(n,\mathbb{R})$$

with the permutation matrix $P := P^{\mathrm{t}}_{(1,\ldots,k)}$ we have

$$A := S_1^{\mathrm{t}} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) S_1 = \begin{pmatrix} d_1 & * \\ * & \operatorname{diag}(\mu_1, \ldots, \mu_{n-1}) \end{pmatrix},$$

where

$$(\mu_1, \ldots, \mu_{n-1}) := (\lambda_1, \ldots, \lambda_{k-1}, \lambda', \lambda_{k+2}, \ldots, \lambda_n).$$

Since $\lambda_k \geq \lambda' \geq \lambda_{k+1} \geq \lambda_{k+2}$ and $\sum_{i=2}^{l} d_i \leq \sum_{i=1}^{l-1} \mu_i$, by induction there exists a $U \in \mathrm{O}(n-1,\mathbb{R})$ such that $U^{\mathrm{t}} \operatorname{diag}(\mu_1, \ldots, \mu_{n-1})U$ has main diagonal $(d_2, \ldots, d_n)$. For $S := S_1 \operatorname{diag}(1, U) \in \mathrm{O}(n,\mathbb{R})$, the matrix

$$S^{\mathrm{t}} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) S = \begin{pmatrix} 1 & 0 \\ 0 & U^{\mathrm{t}} \end{pmatrix} A \begin{pmatrix} 1 & 0 \\ 0 & U \end{pmatrix}$$

has main diagonal $(d_1, \ldots, d_n)$ and eigenvalues $\lambda_1, \ldots, \lambda_n$. □

Example 13.30. We construct a symmetric matrix with eigenvalues $6, 6, 0$ and main diagonal $5, 5, 2$. With the notation from the proof above we have $k = 2$, $\lambda' = 1$ and $(\mu_1, \mu_2) = (6, 1)$. Lemma 13.28 yields

$$T = \frac{1}{\sqrt{6}} \begin{pmatrix} \sqrt{5} & -1 \\ 1 & \sqrt{5} \end{pmatrix}, \qquad U = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}.$$

For $S = \operatorname{diag}(1, T)P_{(1,2)} \operatorname{diag}(1, U)$ we obtain

$$\begin{aligned} S^{\mathrm{t}} \operatorname{diag}(6,6,0) S &= \frac{1}{5} \begin{pmatrix} \sqrt{5} & . & . \\ . & 2 & 1 \\ . & -1 & 2 \end{pmatrix} \begin{pmatrix} . & 1 & . \\ 1 & . & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} 6 & . & . \\ . & 5 & -\sqrt{5} \\ . & -\sqrt{5} & 1 \end{pmatrix} \\ &\quad \cdot \begin{pmatrix} . & 1 & . \\ 1 & . & . \\ . & . & 1 \end{pmatrix} \begin{pmatrix} \sqrt{5} & . & . \\ . & 2 & -1 \\ . & 1 & 2 \end{pmatrix} \\ &= \frac{1}{5} \begin{pmatrix} \sqrt{5} & . & . \\ . & 2 & 1 \\ . & -1 & 2 \end{pmatrix} \begin{pmatrix} 5 & . & -\sqrt{5} \\ . & 6 & . \\ -\sqrt{5} & . & 1 \end{pmatrix} \begin{pmatrix} \sqrt{5} & . & . \\ . & 2 & -1 \\ . & 1 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 5 & -1 & -2 \\ -1 & 5 & -2 \\ -2 & -2 & 2 \end{pmatrix}. \end{aligned}$$

14 The Jordan Normal Form

14.1 Generalized Eigenspaces

Remark 14.1. By 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 by the fundamental theorem of algebra. To avoid the second problem, we replace eigenspaces by larger subspaces. In what follows, let V be a finite-dimensional K-vector space.

Definition 14.2. Let $f \in \operatorname{End}(V)$. A subspace $U \leq V$ is called *f-invariant* if $f(U) \subseteq U$.

Example 14.3. For $f \in \operatorname{End}(V)$, $\operatorname{Ker} f$ and $f(V)$ are always f-invariant subspaces, since $f(\operatorname{Ker}(f)) = \{0\} \subseteq \operatorname{Ker} f$ and $f(f(V)) \subseteq f(V)$.

Remark 14.4. Let $U \leq V$ be an f-invariant subspace.

(a) Clearly $g := f_{|U} \in \operatorname{End}(U)$ and $\mu_f(g) = 0$. By Lemma 10.44, $\mu_g \mid \mu_f$.

(b) For $v, w \in V$ we have

$$v + U = w + U \implies v - w \in U \implies f(v) - f(w) = f(v - w) \in U \implies f(v) + U = f(w) + U.$$

Hence $\overline{f} \colon V/U \to V/U$, $v + U \mapsto f(v) + U$ is a well-defined endomorphism. Because of $\mu_f(\overline{f}) = \overline{\mu_f(f)} = 0$ we have $\mu_{\overline{f}} \mid \mu_f$.

(c) Let B_1 be a basis of U and $B = B_1 \dot\cup B_2$ a basis of V. Then $\overline{B}_2 := \{b + U : b \in B_2\}$ is a basis of V/U (see the proof of Theorem 4.23). Moreover, ${}_B[f]_B = \left(\begin{smallmatrix} A & C \\ 0 & D \end{smallmatrix}\right)$, where $A = {}_{B_1}[f_{|U}]_{B_1}$ and $D = {}_{\overline{B}_2}[\overline{f}]_{\overline{B}_2}$. If U has an f-invariant complement W (i. e., $V = U \oplus W$), then one can achieve $C = 0$ by a suitable choice of basis. We therefore try to decompose V into f-invariant subspaces that are as small as possible.

Definition 14.5. We call $f \in \operatorname{End}(V)$ *triangularizable*, if there exists a basis B of V such that ${}_B[f]_B$ is a triangular matrix. Similarly, $A \in K^{n \times n}$ is called *triangularizable*, 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. So in Definition 14.5 one does not need to specify whether one considers upper or lower triangular matrices (cf. Theorem 15.25).

Example 14.7. By the Schur decomposition, every complex square matrix is triangularizable. The next theorem extends Theorem 10.52.

Theorem 14.8. *A map $f \in \mathrm{End}(V)$ is triangularizable if and only if μ_f (or χ_f) splits into linear factors.*

Proof. If ${}_B[f]_B$ (for some 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, suppose that μ_f splits into linear factors. W. l. o. g. we may assume $V \neq \{0\}$. By Theorem 10.50 there exists an eigenvector $v \in V$ of f. Clearly $U := \langle v \rangle$ 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 annihilators U^0, U_0 and the dual map f^* were defined in Section 7.3.

(a) Let $\gamma \in U^0$. For all $u \in U$ we have $f^*(\gamma)(u) = \gamma(f(u)) = 0$, i. e., $f^*(\gamma) \in U^0$.

(b) Let $v \in U_0$. For all $\gamma \in U$ we have $\gamma(f(v)) = f^*(\gamma)(v) = 0$, i. e., $f(v) \in U_0$. □

Lemma 14.10. *Let $f \in \mathrm{End}(V)$ be diagonalizable and $U \leq V$ an f-invariant subspace. Then the restriction $f_{|U}$ is also diagonalizable.*

Proof. For the minimal polynomial μ_1 of $f_{|U}$ we have $\mu_1 \mid \mu_f$ by Remark 14.4. By Theorem 10.52, μ_f splits into pairwise distinct linear factors. By Lemma 10.24 the same 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. Then $D_f D_g = D_g D_f$ implies $f \circ g = g \circ f$. 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 $g(v) \in U$. Hence U is a g-invariant subspace. By Lemma 14.10, $g_{|U}$ is diagonalizable. So one can choose a basis of V consisting of common eigenvectors of f and g. □

Lemma 14.12 (Simultaneous triangularization)**.** *Let $f, g \in \mathrm{End}(V)$ be commuting and triangularizable. Then f and g are* simultaneously *triangularizable, i. e., there exists a basis B of V such that ${}_B[f]_B$ and ${}_B[g]_B$ are lower triangular matrices.*

Proof. W. l. o. g. we may assume $V \neq \{0\}$. By 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. So 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 of $\mu_{\overline{f}} \mid \mu_f$ and $\mu_{\overline{g}} \mid \mu_g$ we may assume inductively that $\overline{f}$ and $\overline{g}$ are simultaneously triangularizable. 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 triangularizable, but do not commute.

Definition 14.14. Let $\dim V = n$ and $f \in \mathrm{End}(V)$ with eigenvalue $\lambda \in K$. We call

$$H_\lambda(f) := \mathrm{Ker}\big((f - \lambda \,\mathrm{id}_V)^n\big) \leq V$$

the *generalized eigenspace* of f for λ. If λ is an eigenvalue of a matrix $A \in K^{n\times n}$, we define 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 $(A - 1_2)^2 = 0$ we have $H_1(f) = K^2$, i. e., the generalized eigenspace is larger than the eigenspace.

Remark 14.16. For $v \in E_\lambda(f)$ we have

$$(f - \lambda \,\mathrm{id}_V)^n(v) = (f - \lambda \,\mathrm{id}_V)^{n-1}\big((f - \lambda \,\mathrm{id}_V)(v)\big) = (f - \lambda \,\mathrm{id}_V)^{n-1}(0) = 0.$$

This shows $E_\lambda(f) \subseteq H_\lambda(f)$. Clearly f commutes with $f - \lambda \,\mathrm{id}$ and $(f - \lambda \,\mathrm{id})^n$. For $v \in H_\lambda(f)$ we therefore have

$$(f - \lambda \,\mathrm{id}_V)^n(f(v)) = \big((f - \lambda \,\mathrm{id}_V)^n \circ f\big)(v) = f\big((f - \lambda \,\mathrm{id}_V)^n(v)\big) = f(0) = 0.$$

Hence $f(v) \in H_\lambda(f)$ and $H_\lambda(f)$ is f-invariant. Next, we show that $H_\lambda(f)$ has 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. For* $m \geq n$ *we have* $\mathrm{Ker}(f^m) = \mathrm{Ker}(f^n)$*.*

Proof. For $v \in \mathrm{Ker}(f^k)$ we have $f^{k+1}(v) = f(f^k(v)) = f(0) = 0$. 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 $v \in \mathrm{Ker}(f^{k+2})$ we have $f^{k+1}(f(v)) = f^{k+2}(v) = 0$, 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 $f^{2n}(w) = f^n(v) = 0$ we have $w \in \mathrm{Ker}(f^{2n}) = \mathrm{Ker}(f^n)$ and it follows $v = f^n(w) = 0$. So $\mathrm{Ker}(f^n) \cap f^n(V) = \{0\}$. 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. □

Remark 14.18. Let $\lambda, \lambda' \in K$ be distinct eigenvalues of $f \in \mathrm{End}(V)$. For $U := H_\lambda(f) \cap H_{\lambda'}(f)$ we have

$$(f - \lambda \,\mathrm{id})^n(U) = \{0\} = (f - \lambda' \,\mathrm{id})^n(U).$$

Therefore the minimal polynomial of the restriction $g := f_{|U} \in \mathrm{End}(U)$ divides $(X - \lambda)^n$ and $(X - \lambda')^n$ by Lemma 10.44. By Lemma 10.24 we obtain $\mu_g = 1$ and $U = \{0\}$.

Theorem 14.19 (Generalized eigenspace decomposition)**.** *Let $f \in \mathrm{End}(V)$ 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 suppose the claim has already been proved for smaller n. Since μ_f splits into linear factors, f has an eigenvalue $\lambda = \lambda_1 \in K$ by Theorem 10.50. For $g := f - \lambda \,\mathrm{id}_V$ we have

$$V = g^n(V) \oplus \mathrm{Ker}(g^n) = g^n(V) \oplus H_\lambda(f)$$

by Fitting's lemma. For $U := g^n(V)$ we have $f(U) = (g + \lambda \,\mathrm{id})(U) \subseteq g(U) + U \subseteq U$, i. e., U is f-invariant. Let $h := f_{|U} \in \mathrm{End}(U)$ and $d := \dim U$. By Remark 14.4, $\mu_h \mid \mu_f$ and μ_h splits into linear factors (Lemma 10.24). Because of $\dim H_\lambda(f) \geq \dim E_\lambda(f) \geq 1$ we have $k < n$. By induction we therefore have

$$U = H_{\lambda_2}(h) \oplus \ldots \oplus H_{\lambda_k}(h)$$

for the distinct eigenvalues $\lambda_2, \ldots, \lambda_k$ of h. Because of $E_{\lambda_i}(h) \subseteq E_{\lambda_i}(f)$, $\lambda_2, \ldots, \lambda_k$ are also eigenvalues of f. Because of $E_{\lambda_i}(h) \cap H_\lambda(f) \leq U \cap H_\lambda(f) = \{0\}$ we have $\lambda \neq \lambda_i$ for $i = 2, \ldots, k$. Conversely, let $\mu \neq \lambda$ be an eigenvalue of f. As usual, $H_\mu(f)$ is g-invariant. Because of

$$\mathrm{Ker}(g) \cap H_\mu(f) = E_\lambda(f) \cap H_\mu(f) \leq H_\lambda(f) \cap H_\mu(f) \overset{14.18}{=} \{0\}$$

the restriction of g to $H_\mu(f)$ is injective, hence also surjective. It follows

$$H_\mu(f) = g^n(H_\mu(f)) = U \cap H_\mu(f) = \mathrm{Ker}\big((h - \lambda_i \,\mathrm{id}_U)^n\big) \overset{14.17}{=} \mathrm{Ker}\big((h - \mu \,\mathrm{id}_U)^d\big) = H_\mu(h).$$

Because of $E_\mu(f) \leq U$, μ is an eigenvalue of h, i. e., $\mu = \lambda_i$ for some $i \geq 2$. Altogether we obtain

$$V = H_\lambda(f) \oplus U = H_\lambda(f) \oplus H_{\lambda_2}(h) \oplus \ldots \oplus H_{\lambda_k}(h) = H_{\lambda_1}(f) \oplus \ldots \oplus H_{\lambda_k}(f). \qquad \square$$

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 compute by 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 dimension 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. After the generalized eigenspace decomposition, we are interested in bases of generalized eigenspaces.

Definition 14.23. For $\lambda \in K$ and $n \geq 1$ we call

$$J_n(\lambda) := \begin{pmatrix} \lambda & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & \lambda \end{pmatrix} \in K^{n \times n}$$

a *Jordan block* for the eigenvalue λ.[1]

Remark 14.24. Clearly $A := J_n(\lambda)$ has 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 as far from being diagonalizable as possible (Theorem 10.34). By Cayley-Hamilton we have $\mu_A \mid \chi_A$. Because of

$$(A - \lambda 1_n)^{n-1} e_1 = J_n(0)^{n-1} e_1 = J_n(0)^{n-2} e_2 = \ldots = e_n \neq 0$$

we have $\mu_A = \chi_A$.

Definition 14.25. We call $f \in \mathrm{End}(V)$ (or $A \in K^{n \times n}$) *nilpotent*, if $f^m = f \circ \ldots \circ f = 0$ (or $A^m = 0_n$) for some $m \in \mathbb{N}$.

Example 14.26. Let A be a strictly (upper or lower) triangular matrix. Then $\chi_A = X^n$. By Cayley-Hamilton, $A^n = \chi_A(A) = 0_n$. So 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 $m \in \mathbb{N}$. Since $\deg \mu_A \leq n$, 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 $n_1 \geq \ldots \geq n_k \geq 1$. For example, $p(5) = 7$, since

$$5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1.$$

No simple formula for $p(n)$ is known.

[1] Some books use $J_n(\lambda)^{\mathrm{t}}$. This makes no essential difference (Remark 14.45).

Theorem 14.28. *Let $f \in \operatorname{End}(V)$ be nilpotent. Then there exists a basis B of V such that*

$$ {}_B[f]_B = \operatorname{diag}(J_{n_1}(0), \ldots, J_{n_s}(0)). $$

The numbers $n_1 \geq \ldots \geq n_s \geq 1$ are uniquely determined by the equations

$$ \boxed{\big|\{1 \leq i \leq s : n_i \geq k\}\big| = \operatorname{rk}(f^{k-1}) - \operatorname{rk}(f^k)} \qquad (k \in \mathbb{N}) \tag{14.1} $$

In particular, $s = \dim \operatorname{Ker}(f)$.

Proof. Let $m \in \mathbb{N}$ with $f^m = 0 \neq f^{m-1}$. For $k \in \mathbb{N}$ we have $f(\operatorname{Ker}(f^k)) \subseteq \operatorname{Ker}(f^{k-1})$. By Corollary 4.16 there exist subspaces $U_1, \ldots, U_m$ with

$$ \begin{aligned} V = \operatorname{Ker}(f^m) &= \operatorname{Ker}(f^{m-1}) \oplus U_1, \\ \operatorname{Ker}(f^{m-1}) &= \big(\operatorname{Ker}(f^{m-2}) + f(U_1)\big) \oplus U_2, \\ &\vdots \\ \operatorname{Ker} f &= \big(f^{m-1}(U_1) + \ldots + f(U_{m-1})\big) \oplus U_m. \end{aligned} $$

We show that all sums are direct. This is given for the first sum. Suppose inductively that we have already shown:

$$ \operatorname{Ker}(f^{m-k+1}) = \operatorname{Ker}(f^{m-k}) \oplus f^{k-1}(U_1) \oplus f^{k-2}(U_2) \oplus \ldots \oplus f(U_{k-1}) \oplus U_k. \tag{14.2} $$

Let $w + f^k(u_1) + \ldots + f(u_k) = 0$ with $w \in \operatorname{Ker}(f^{m-k-1})$ and $u_i \in U_i$ for $i = 1, \ldots, k$. Then it follows that

$$ \begin{aligned} f^{m-k}\big(f^{k-1}(u_1) + \ldots + u_k\big) &= f^{m-k-1}\big(f^k(u_1) + \ldots + f(u_k)\big) \\ &= f^{m-k-1}\big(w + f^k(u_1) + \ldots + f(u_k)\big) = f^{m-k-1}(0) = 0 \end{aligned} $$

and

$$ f^{k-1}(u_1) + \ldots + u_k \in \operatorname{Ker}(f^{m-k}) \cap \big(f^{k-1}(U_1) \oplus \ldots \oplus f(U_{k-1}) \oplus U_k\big) \overset{(14.2)}{=} \{0\}. $$

This shows $f^{k-1}(u_1) = \ldots = u_k = 0$ (Lemma 8.9) and it follows that $w = 0$. Hence

$$ \operatorname{Ker}(f^{m-k}) = \operatorname{Ker}(f^{m-k-1}) \oplus f^k(U_1) \oplus f^{k-1}(U_2) \oplus \ldots \oplus f(U_k) \oplus U_{k+1} $$

as desired.

Altogether we have

$$ V = U_1 \oplus f(U_1) \oplus \ldots \oplus f^{m-1}(U_1) \oplus U_2 \oplus f(U_2) \oplus \ldots \oplus f^{m-2}(U_2) \oplus \ldots \oplus U_m. $$

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). Since $\operatorname{Ker}(f^j) \cap U_i = \{0\}$, 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 $f^j(U_i)$. Hence

$$ B := \bigcup_{i=1}^{m} \bigcup_{j=1}^{k_i} \big\{b_{ij}, f(b_{ij}), \ldots, f^{m-i}(b_{ij})\big\} $$

is a basis of V. Because of $U_i \subseteq \operatorname{Ker}(f^{m-i+1})$ we have $f^{m-i+1}(b_{ij}) = 0$. 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 ${}_B[f]_B$. Altogether, ${}_B[f]_B$ now has the desired form. Moreover,

$$ k_1 + \ldots + k_l \;=\; \dim\big(f^{l-1}(U_1) \oplus f^{l-2}(U_2) \oplus \ldots \oplus U_l\big) = \dim \operatorname{Ker}(f^{m-l+1}) - \dim \operatorname{Ker}(f^{m-l}) $$

$$\overset{7.12}{=} \operatorname{rk}(f^{m-l}) - \operatorname{rk}(f^{m-l+1})$$

is the number of Jordan blocks $J_{n_i}(0)$ with $n_i \geq m-l+1$. Replacing $m-l+1$ by k, we obtain (14.1). The last claim follows with $k=1$ in (14.1). □

Remark 14.29. In the situation of Theorem 14.28 we have

$$\begin{aligned}\big|\{1 \leq i \leq s : n_i = k\}\big| &= \big|\{1 \leq i \leq s : n_i \geq k\}\big| - \big|\{1 \leq i \leq s : n_i \geq k+1\}\big| \\ &= \operatorname{rk}(f^{k-1}) + \operatorname{rk}(f^{k+1}) - 2\operatorname{rk}(f^k)\end{aligned}$$

for $k \in \mathbb{N}$. In particular, $\operatorname{rk}(f^{k-1}) - \operatorname{rk}(f^k) \geq \operatorname{rk}(f^k) - \operatorname{rk}(f^{k+1}) \geq 0$, i. e., the decrease of the sequence $\operatorname{rk} f$, $\operatorname{rk}(f^2), \ldots$ can only become smaller.[2]

Example 14.30. Let $f \in \operatorname{End}(\mathbb{R}^5)$ with

$$A := [f] = \begin{pmatrix} 2 & -3 & -1 & -1 & 2 \\ 1 & -2 & 0 & 0 & 1 \\ 0 & 2 & -1 & -1 & 0 \\ 1 & -4 & 1 & 1 & 1 \\ 0 & -1 & 1 & 1 & 0 \end{pmatrix}.$$

We compute

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

So $\operatorname{rk}(f^2) = 1$. 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 $U_2 := \langle b_2 \rangle$ we have $\operatorname{Ker}(f^2) = \operatorname{Ker}(f) \oplus f(U_1) \oplus U_2$. Finally, $\operatorname{Ker} f = f^2(U_1) \oplus f(U_2)$ (so $U_3 := \{0\}$). With respect to the basis $B := \{b_1, f(b_1), f^2(b_1), b_2, f(b_2)\}$, the matrix of f is $\operatorname{diag}\big(J_3(0), J_2(0)\big)$.

[2] A sequence of numbers with this property is called *convex*.

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*

$$ {}_B[f]_B = \operatorname{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s)), \qquad \text{(JORDAN normal form)} $$

where $\lambda_1, \ldots, \lambda_s \in K$ are the eigenvalues of f (possibly with repetitions). The Jordan blocks $J_{n_i}(\lambda_i)$ are uniquely determined up to order.

Proof. Let $\rho_1, \ldots, \rho_k \in K$ be the distinct eigenvalues of f and $H_i := H_{\rho_i}(f)$ for $i = 1, \ldots, k$. By the generalized eigenspace decomposition we have $V = H_1 \oplus \ldots \oplus H_k$. For $g_i := (f - \rho_i \operatorname{id}_V)_{|H_i}$ we have $g_i^n = 0$, where $n := \dim V$. By Theorem 14.28 there exists a basis B_i of H_i with

$$ {}_{B_i}[g_i]_{B_i} = \operatorname{diag}\big(J_{m_1}(0), \ldots, J_{m_{t_i}}(0)\big). $$

By Theorem 7.18 we have

$$ \begin{aligned} {}_{B_i}[f_{|H_i}]_{B_i} &= {}_{B_i}[g_i + \rho_i \operatorname{id}_{H_i}]_{B_i} = {}_{B_i}[g_i]_{B_i} + \rho_i {}_{B_i}[\operatorname{id}_{H_i}]_{B_i} = \operatorname{diag}\big(J_{m_1}(0), \ldots, J_{m_{t_i}}(0)\big) + \rho_i 1_{\dim H_i} \\ &= \operatorname{diag}\big(J_{m_1}(\rho_i), \ldots, J_{m_{t_i}}(\rho_i)\big). \end{aligned} $$

Hence $B := B_1 \cup \ldots \cup B_k$ is a suitable basis (ρ_i occurs exactly t_i times among $\lambda_1, \ldots, \lambda_s$). For a fixed pair (n_i, λ_i), the number of blocks $J_{n_i}(\lambda_i)$ follows from Remark 14.29 applied to g_i. □

Corollary 14.32. *Every matrix $A \in \mathbb{C}^{n \times n}$ has a Jordan normal form (more precisely: A is similar to a Jordan normal form).*

Proof. By Corollary 11.35, μ_A splits into linear factors. □

Remark 14.33.

(a) Since the eigenvalues $\lambda_1, \ldots, \lambda_s$ cannot in general be ordered in a meaningful way (cf. Remark 11.26), in contrast to Theorem 14.28 there is no canonical order of the Jordan blocks. However, one can define the *lexicographic* order

$$ x <_l y \iff \operatorname{Re}(x) < \operatorname{Re}(y) \vee \big(\operatorname{Re}(x) = \operatorname{Re}(y) \wedge \operatorname{Im}(x) < \operatorname{Im}(y)\big) $$

on $\mathbb{C}$ and sort the Jordan blocks first by size and then by eigenvalue with respect to $<_l$. In what follows, we speak somewhat imprecisely 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 we obtain

$$ \chi_f = \chi_A = (X + \mathrm{i})\big((X - 5)(X + 1) + 9\big) = (X + \mathrm{i})(X^2 - 4X + 4) = (X + \mathrm{i})(X - 2)^2. $$

So $\lambda_1 = 2$ and $\lambda_2 = -\mathrm{i}$ are the eigenvalues of f. Clearly $b_3 := e_2$ 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. So 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

$$b_2 := (f - 2\,\mathrm{id})(b_1) = (1, -1, -3)$$

we have $f(b_1) = 2b_1 + b_2$ and $f(b_2) = (f - 2\,\mathrm{id})(b_2) + 2b_2 = 2b_2$. For the basis $B := \{b_1, b_2, b_3\}$ we obtain

$${}_B[f]_B = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & -\mathrm{i} \end{pmatrix} = \operatorname{diag}\big(J_2(2), J_1(-\mathrm{i})\big).$$

14.3 Applications

Theorem 14.35. *Let $J := \operatorname{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_t}(\lambda_t))$ 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 $\operatorname{rk}(J - \rho_i 1_n) = n - a_i$, 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, we have

$$0 = \mu_A(J) = \operatorname{diag}\big(\mu_A(J_{n_1}(\lambda_1)), \ldots, \mu_A(J_{n_t}(\lambda_t))\big).$$

Remark 14.24 yields $\mu_A = (X - \rho_1)^{m_1} \ldots (X - \rho_k)^{m_k}$. The last claim holds by Theorem 10.52. □

Example 14.36. For $A = \operatorname{diag}(J_3(1), J_2(1), J_4(\mathrm{i}))$ we have $\chi_A = (X - 1)^5(X - \mathrm{i})^4$ and $\mu_A = (X - 1)^3(X - \mathrm{i})^4$.

Corollary 14.37. *Let $A, B \in \mathbb{C}^{n\times n}$. For $n \leq 3$ we have*

$$A \approx B \iff \chi_A = \chi_B,\ \mu_A = \mu_B.$$

For $n \leq 6$ 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 *agree with those of* B.

Proof. As is well known, $A \approx B$ implies each of the other statements. 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 we have $r = n_1 + \ldots + n_k$. Since n_1 is determined by μ_A, also k and $n_2, \ldots, n_k$ are uniquely determined by r. Hence A and B have the same Jordan normal form. It follows that $A \approx B$.

Now let $n \leq 6$ and suppose that (2) holds. Besides r and n_1, the geometric multiplicity k of λ is now also given. So we have to show that a partition of $r \leq 6$ is uniquely determined by its largest part n_1 and the number of parts k. For $n_1 \leq 2$ the number of ones and twos is clearly determined by r and k (even for all $r \in \mathbb{N}$). For $n_1 \geq 3$, $(n_2, \ldots, n_k)$ is a partition of $r - n_1 \leq 3$ into $k - 1$ parts. There is only one possibility for this. □

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 $n = 7$, respectively.

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. W. l. o. g. 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. Since A^k is a lower triangular matrix, we obtain

$$\operatorname{tr}(A^k) = m_1\lambda_1^k + \ldots + m_s\lambda_s^k$$

for $k \in \mathbb{N}$. If A is nilpotent, then $s = 1$, $\lambda_1 = 0$ and $\operatorname{tr}(A^k) = 0$ for $k \in \mathbb{N}$.

Conversely, let $\operatorname{tr}(A^k) = m_1\lambda_1^k + \ldots + m_s\lambda_s^k = 0$ for $k = 1, \ldots, n$. Suppose for contradiction that A is not nilpotent. W. l. o. g. let $\lambda_i \neq 0$ for $i = 1, \ldots, s$. Then $(m_1, \ldots, m_s)^{\mathrm{t}}$ is a solution of the homogeneous system of linear equations with coefficient matrix $V = (\lambda_j^i) \in \mathbb{C}^{s\times s}$. By Vandermonde, however, $\det V = \lambda_1 \ldots \lambda_s \det(\lambda_j^{i-1}) \neq 0$. 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* $k \in \mathbb{N}_0$ *we have*

$$J_n(\lambda)^k = \begin{pmatrix} \lambda^k & & & & 0 \\ k\lambda^{k-1} & \ddots & & & \\ \binom{k}{2}\lambda^{k-2} & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ \binom{k}{n-1}\lambda^{k-n+1} & \cdots & \binom{k}{2}\lambda^{k-2} & k\lambda^{k-1} & \lambda^k \end{pmatrix},$$

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 theorem[3]

$$J_n(\lambda)^k = (\lambda 1_n + J_n(0))^k = \sum_{l=0}^{k} \binom{k}{l} \lambda^{k-l} J_n(0)^l. \qquad \square$$

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. Since $(SWS^{-1})^k = SW^kS^{-1}$ for $S \in \mathrm{GL}(n, \mathbb{C})$, we may assume that A is a Jordan normal form. It suffices to construct a k-th root for every Jordan block. So let $A = J_n(\lambda)$ with $\lambda \in \mathbb{C}$. Since A is invertible, we have $\lambda \neq 0$. By Lemma 11.27 there exists a $\mu \in \mathbb{C}$ with $\mu^k = \lambda$. Let $J := J_n(\mu)$. By Lemma 14.41, J^k has entries $k\mu^{k-1} \neq 0$ directly below the main diagonal. Hence $\mu^k = \lambda$ is the only eigenvalue of J^k and the geometric multiplicity is 1. By 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 look for a cube root of $J_3(1)$. For

$$S := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 9 \end{pmatrix}$$

we have

$$S^{-1}J_3(1)^3S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & -1/9 & 1/9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 3 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 9 \end{pmatrix} = J_3(1).$$

For

$$W := S^{-1}J_3(1)S = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & -1/9 & 1/9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ -1/9 & 1/3 & 1 \end{pmatrix}$$

we have $W^3 = J_3(1)$. In Theorem 15.35 we determine the number of k-th roots of $J_n(\lambda)$ for all $\lambda \in \mathbb{C}^\times$ and $n, k \in \mathbb{N}$.

(b) The nilpotent matrix $J_2(0)$ has no square root, since for $A \in K^{2\times 2}$ with $A^2 = J_2(0)$ we would have $A^4 = 0$ and hence $A^2 = 0$ (Remark 14.27). On the other hand, 0_2 has infinitely many square roots, namely $\lambda J_2(0)$ for $\lambda \in \mathbb{C}$.

(c) One easily checks that $J_2(1) \in \mathrm{GL}(2, \mathbb{F}_2)$ has no square root either.

Theorem 14.44. *For all $A \in \mathbb{C}^{n\times n}$ we have $A \approx A^{\mathrm{t}}$.*

[3]The formula is applicable since 1_n and $J_n(0)$ commute.

Proof. As usual, we may 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 by direct computation:

$$\begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix}^{(-1)} J_n(\lambda) \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} = \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix} \begin{pmatrix} 0 & & & \lambda \\ & & \iddots & 1 \\ & \iddots & \iddots & \\ \lambda & 1 & & 0 \end{pmatrix} = J_n(\lambda)^{\mathrm{t}}.$$

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 which, 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 factor the minimal polynomial into polynomials that are as "small" 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, since $\deg(\alpha\beta) = \deg(\alpha) + \deg(\beta)$ for $\alpha, \beta \in K[X]$. Conversely, every irreducible polynomial in $\mathbb{C}[X]$ has degree 1 by the fundamental theorem of algebra and Lemma 10.22. In general, monic polynomials of degree 2 or 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 $\mathbb{R}[X]$, since $X^2 - 2 = (X + \sqrt{2})(X - \sqrt{2})$. Similarly, $X^2 + 1$ is irreducible in $\mathbb{R}[X]$, but not in $\mathbb{C}[X]$ because of $X^2 + 1 = (X + \mathrm{i})(X - \mathrm{i})$.

(c) Two distinct 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. Hence 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 $\deg \alpha \geq \deg \beta \geq 0$. The following algorithm determines whether α and β are coprime:*

(1) Set $\alpha_0 := \alpha$, $\alpha_1 := \beta$ and $i := 1$.

(2) Repeat:

- *Division with remainder: $\alpha_{i-1} = \alpha_i \gamma_i + \alpha_{i+1}$ with $\deg \alpha_{i+1} < \deg \alpha_i$.*
- *If $\alpha_{i+1} = 0$, then stop.*
- *Increase i by 1.*

(3) α and β are coprime if and only if $\alpha_i \in K$.

Proof. Because of $\alpha_{i-1} = \alpha_i\gamma_i + \alpha_{i+1}$, the polynomials α_{i-1} and α_i have the same common divisors as α_i and α_{i+1} for $i = 1, 2, \ldots$. Since $\deg \alpha_i$ becomes smaller in every iteration, the algorithm must stop after finitely many steps. At the end, $\alpha_{i+1} = 0 \neq \alpha_i$ holds, i. e., the common divisors of α and β are the divisors of α_i. □

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

$$\begin{aligned} X^5 + X^4 + X^3 + 1 &= (X^4 + X^3 + X + 1)X + X^3 + X^2 + X + 1 && \Longrightarrow \alpha_2 = X^3 + X^2 + X + 1, \\ X^4 + X^3 + X + 1 &= (X^3 + X^2 + X + 1)X + X^2 + 1 && \Longrightarrow \alpha_3 = X^2 + 1 \notin \mathbb{F}_2, \\ X^3 + X^2 + X + 1 &= (X^2 + 1)(X + 1) && \Longrightarrow \alpha_4 = 0. \end{aligned}$$

Hence $X^2 + 1$ 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. Hence there exist $\tilde{\alpha}, \tilde{\beta} \in K[X]$ such that $\rho := \alpha\tilde{\alpha} + \beta\tilde{\beta} \neq 0$ has minimal degree. Division with remainder yields $\gamma, \delta \in K[X]$ with $\alpha = \gamma\rho + \delta$ and $\deg \delta < \deg \rho$. Hence

$$\delta = \alpha - \gamma\rho = \alpha - \gamma\alpha\tilde{\alpha} - \gamma\beta\tilde{\beta} = \alpha(1 - \gamma\tilde{\alpha}) - \beta(\gamma\tilde{\beta}).$$

The choice of ρ shows $\delta = 0$. It follows that $\rho \mid \alpha$. Similarly, we obtain $\rho \mid \beta$. Since α and β are coprime, $\rho \in K^\times$. After normalization, we may assume $\rho = 1$. □

Remark 15.7. The polynomials $\tilde{\alpha}$ and $\tilde{\beta}$ in Lemma 15.6 can be computed by reading the Euclidean algorithm backwards

$$\alpha_i = \alpha_{i-2} - \alpha_{i-1}\gamma_{i-1} = \alpha_{i-2} - (\alpha_{i-3} - \alpha_{i-2}\gamma_{i-2})\gamma_{i-1} = \alpha_{i-2}(1 + \gamma_{i-2}\gamma_{i-1}) - \alpha_{i-3}\gamma_{i-1} = \ldots$$

and dividing by α_i.

Example 15.8. Let $\alpha := X^3 + X^2 + 1$ and $\beta := X^2 - X$ in $\mathbb{Q}[X]$. The Euclidean algorithm yields

$$\begin{aligned} X^3 + X^2 + 1 &= (X^2 - X)(X + 2) + 2X + 1 \\ X^2 - X &= (2X + 1)\frac{1}{4}(2X - 3) + \frac{3}{4}. \end{aligned}$$

Hence

$$\begin{aligned} \frac{3}{4} &= \beta - \frac{1}{4}(2X + 1)(2X - 3) = \beta - \frac{1}{4}\Big(\alpha - \beta(X + 2)\Big)(2X - 3) \\ &= -\frac{1}{4}(2X - 3)\alpha + \frac{1}{4}\Big(4 + (X + 2)(2X - 3)\Big)\beta \end{aligned}$$

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 order.*

Proof. Let $\alpha \in K[X]$ be monic. Induction on $d := \deg \alpha$. If α is irreducible (for example $d = 1$), we are done. Otherwise, $\alpha = \beta\gamma$ with $\deg \beta, \deg \gamma < d$. We may assume that β and γ are monic. By induction, β and γ are products of irreducible factors, and hence 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 $m = 1$, we have $n = 1$ and $\sigma_1 = \tau_1$. Now let $m \geq 2$. In the case $\sigma_1 \neq \tau_1$, the polynomials 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 that

$$\sigma_1(\tilde{\sigma}\tau_2 \dots \tau_m + \sigma_2 \dots \sigma_n\tilde{\tau}) = \tau_2 \dots \tau_m.$$

Inductively, we obtain $\sigma_1 = \tau_i$ 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 for natural numbers, in the prime factorization of polynomials we will combine equal irreducible factors into powers. For example,

$$X^6 + X^5 - X^4 - X^3 = X^3(X + 1)^2(X - 1).$$

In contrast to the prime factorization of numbers, efficient algorithms for factorization in $\mathbb{Q}[X]$ are known.

Corollary 15.11. *Let $\alpha, \beta \in K[X]$ and let γ be an irreducible divisor of $\alpha\beta$. Then $\gamma \mid \alpha$ or $\gamma \mid \beta$.*

Proof. Because of $\gamma \mid \alpha\beta$, γ must appear in the prime factorization of $\alpha\beta$. This prime factorization is obtained by combining the prime factorizations of α and β. Hence γ must occur in the factorization of α or β. □

Theorem 15.12 (Primary decomposition)**.** *Let $f \in \operatorname{End}(V)$ and let $\mu_f = \gamma_1^{a_1} \dots \gamma_k^{a_k}$ be the prime factorization of μ_f in $K[X]$. Then*

$$V = \operatorname{Ker}\big(\gamma_1^{a_1}(f)\big) \oplus \dots \oplus \operatorname{Ker}\big(\gamma_k^{a_k}(f)\big)$$

is a decomposition into f-invariant subspaces. Moreover, $\gamma_i^{a_i}$ is the minimal polynomial of the restriction of f to $\operatorname{Ker}\big(\gamma_i^{a_i}(f)\big)$ for $i = 1, \dots, k$.

Proof. For $1 \leq i \leq k$ let $V_i := \operatorname{Ker}(\gamma_i^{a_i}(f))$. 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., $f(v) \in V_i$. Hence V_i is f-invariant. To prove the direct decomposition, we argue by induction on k. For $k = 1$ we have $V_1 = \operatorname{Ker}(\mu_f(f)) = \operatorname{Ker} 0 = V$. 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 $V_\beta := \operatorname{Ker}(\beta(f))$. Because of $(\alpha\beta)(f) = 0 = (\beta\alpha)(f)$, we have $\beta(f)(V) \subseteq V_1$ and $\alpha(f)(V) \subseteq V_\beta$. It follows that

$$V = \operatorname{id}(V) = (\alpha\tilde{\alpha} + \beta\tilde{\beta})(f)(V) \subseteq \alpha(f)(V) + \beta(f)(V) \subseteq V_\beta + V_1.$$

On the other hand, for $v \in V_1 \cap V_\beta$ we have

$$v = (\alpha\tilde{\alpha} + \beta\tilde{\beta})(f)(v) = \tilde{\alpha}(f)\big(\alpha(f)(v)\big) + \tilde{\beta}(f)\big(\beta(f)(v)\big) = 0.$$

This shows $V = V_1 \oplus V_\beta$.

The minimal polynomial α_1 of the restriction $f_{|V_1}$ divides α, since $\alpha(f)(V_1) = 0$. In particular, $\deg \alpha_1 \leq \deg \alpha$. Similarly, β is divisible by the minimal polynomial β_1 of $f_\beta := f_{|V_\beta}$. Because of $V = V_1 \oplus V_\beta$, 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 of

$$\mathrm{Ker}(\gamma_i^{a_i}(f)) = \mathrm{Ker}\big((f - \lambda_i \,\mathrm{id})^{a_i}\big) \subseteq \mathrm{Ker}\big((f - \lambda_i \,\mathrm{id})^{n}\big) = H_{\lambda_i}(f)$$

the primary decomposition coincides with the generalized eigenspace decomposition from Theorem 14.19 (where $n = \dim V$).

Definition 15.14. For $v \in V$ and $f \in \mathrm{End}(V)$, we call $U := \langle f^i(v) : i \in \mathbb{N}_0 \rangle \leq V$ a *cyclic* subspace. Clearly, 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 $i \in \mathbb{N}_0$. Hence $\mu_v \mid \alpha$. 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 $v \in V_i$ we have $\langle f^j(v) : j \in \mathbb{N}_0 \rangle \subseteq V_i$. Hence 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 of Jordan blocks over arbitrary fields.

Definition 15.18. For $\alpha := X^n + a_{n-1}X^{n-1} + \ldots + a_0 \in K[X] \setminus K$, we call

$$B(\alpha) := \begin{pmatrix} 0 & & & -a_0 \\ 1 & \ddots & & \vdots \\ & \ddots & 0 & -a_{n-2} \\ 0 & & 1 & -a_{n-1} \end{pmatrix} \in K^{n\times n}$$

the *companion matrix* of α.

Example 15.19. Clearly, $B(X^n) = J_n(0)$ 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 $i = 1, \ldots, n-1$ we have $Be_i = e_{i+1}$. Hence $e_1, Be_1, \ldots, B^{n-1}e_1$ 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 $\mu_B \mid \alpha$. Altogether, $\mu_B = \alpha$. 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 $N := \alpha(A)$ we have $N^k = \alpha^k(A) = 0$ by (a). By 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 $i = 1, \ldots, d-1$. Hence the vectors

$$Ne_i = A^n e_i + a_{n-1}A^{n-1}e_i + \ldots + a_0 e_i \in e_{n+i} + \langle e_1, \ldots, e_{n+i-1}\rangle$$

for $i = 1, \ldots, d-n$ are linearly independent. 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 C of V with ${}_C[f]_C = B(\mu_f)$.*

Proof. Let $n := \dim V$. By Lemma 15.16, there exists a $v \in V$ with $\mu_f = \mu_v$. For $U := \langle f^i(v) : i \in \mathbb{N}_0\rangle \leq V$ we have

$$n = \deg \chi_f = \deg \mu_f = \deg \mu_v \leq \dim U.$$

Hence $U = V$. By Remark 15.15, $C := \{v, f(v), \ldots, f^{n-1}(v)\}$ is a basis of V, and ${}_C[f]_C = B(\mu_f)$. □

Theorem 15.22. *For every matrix $A \in K^{n\times n}$ we have:*

(a) There exist uniquely determined monic polynomials $\alpha_1, \ldots, \alpha_k \in K[X] \setminus K$ with $\alpha_k \mid \alpha_{k-1} \mid \ldots \mid \alpha_1$ and

$$A \approx \operatorname{diag}\big(B(\alpha_1), \ldots, B(\alpha_k)\big). \qquad \text{(Frobenius normal form}^1\text{)}$$

Moreover, $\mu_A = \alpha_1$ and $\chi_A = \alpha_1 \ldots \alpha_k$.

[1] Also called *rational canonical form*.

(b) There exist irreducible polynomials $\gamma_1, \dots, \gamma_s \in K[X]$ and $a_1, \dots, a_s \in \mathbb{N}$ with

$$A \approx \operatorname{diag}\big(B(\gamma_1^{a_1}), \dots, B(\gamma_s^{a_s})\big). \qquad \text{(Weierstrass normal form)}$$

Here the powers $\gamma_1^{a_1}, \dots, \gamma_s^{a_s}$ are uniquely determined up to order, and $\chi_A = \gamma_1^{a_1} \dots \gamma_s^{a_s}$.

Proof (Jacob).
Step 1: Existence of the Frobenius normal form.
Let $V := K^n$ and $f \in \operatorname{End}(V)$ with $[f] = A$. By Lemma 15.16, there exists $v \in V$ with $\alpha_1 := \mu_v = \mu_f$. Let $d := \deg \alpha_1$ and

$$U := \langle f^i(v) : i \in \mathbb{N}_0 \rangle \leq V.$$

By Remark 15.15, $B(\alpha_1)$ is the matrix of $f_{|U}$ with respect to $\{b_i := f^{i-1}(v) : i = 1, \dots, d\}$. In the case $U = V$ we are done. 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 $f^* \in \operatorname{End}(V^*)$ the dual map of f. By 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 of $[f^*] = A^{\mathrm{t}}$ (Theorem 7.50), $\mu_{b_d^*} \mid \mu_{f^*} = \mu_f$. In particular, $\dim L \leq d$. By Remark 7.51, $(f^*)^i = (f^i)^*$ for $i \in \mathbb{N}_0$. Let

$$v^* := \Big(\sum_{i=1}^{t} \lambda_i f^{i-1}\Big)^*(b_d^*) = \sum_{i=1}^{t} \lambda_i (f^*)^{i-1}(b_d^*) \in L$$

with $t \leq d$, $\lambda_1, \dots, \lambda_t \in K$ and $\lambda_t \neq 0$. Then

$$v^*(b_{d-t+1}) = b_d^*\Big(\sum_{i=1}^{t} \lambda_i f^{i-1}(b_{d-t+1})\Big) = b_d^*(\lambda_t b_d) = \lambda_t \neq 0$$

and $v^* \neq 0$. Hence $\{(f^*)^i(b_d^*) : i = 0, \dots, d-1\}$ is a basis of L and $L \cap U^0 = \{0\}$. From $\dim U^0 = n - \dim(U) = n - d$ it follows that $V^* = L \oplus U^0$. By Lemma 7.44,

$$V = L_0 \oplus U,$$

where L_0 is f-invariant by Lemma 14.9. The minimal polynomial of $g := f_{|L_0}$ divides α_1. By induction on n, g has a Frobenius normal form $\operatorname{diag}(B(\alpha_2), \dots, B(\alpha_k))$ with $\alpha_k \mid \alpha_{k-1} \mid \dots \mid \alpha_2 = \mu_g \mid \alpha_1$. Altogether, a Frobenius normal form exists for f. From the block diagonal form and Lemma 15.20 we obtain $\chi_A = \alpha_1 \dots \alpha_k$.

Step 2: Existence and uniqueness of the Weierstrass normal form.
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 $i = 1, \dots, l$. By the primary decomposition,

$$V = V_1 \oplus \dots \oplus V_l$$

and $\mu_{f_i} = \gamma_i^{c_i}$. A Frobenius normal form of f_i therefore has the form $\operatorname{diag}(B(\gamma_i^{a_1}), \dots, B(\gamma_i^{a_t}))$ with $c_i = a_1 \geq \dots \geq a_t \geq 1$. From this we obtain a Weierstrass normal form for f. By Lemma 15.20,

$$\operatorname{diag}\big(\underbrace{J_{a_1}(0), \dots, J_{a_1}(0)}_{\deg \gamma_i}, \dots, \underbrace{J_{a_t}(0), \dots, J_{a_t}(0)}_{\deg \gamma_i}\big)$$

is the Jordan normal form of $\gamma_i(f_i)$. Hence the numbers $a_1, \dots, a_t$ are uniquely determined. Conversely, let $B(\rho)$ with $\rho \in K[X]$ be a block of an arbitrary 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 $\rho \mid \mu_f$. By

Corollary 15.11, $\rho \mid \gamma_i^{c_i}$ for some i and $U \leq V_i$. The corresponding blocks of W therefore also yield a Weierstrass normal form of f_i. Since the numbers $a_1, \ldots, a_t$ are uniquely determined, the Weierstrass normal form of f is also uniquely determined (up to the order of the blocks).

Step 3: Uniqueness of the Frobenius normal form.
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 $\alpha_i = \gamma_1^{a_{i1}} \ldots \gamma_l^{a_{il}}$ be the factorization of the minimal polynomial of $f_i := f_{|U_i}$ with $a_{ij} \geq 0$. In the Weierstrass normal form of f_i, the blocks $B(\gamma_j^{a_{ij}})$ with $a_{ij} > 0$ must then occur (otherwise μ_{f_i} would be "too small"). Because of $\chi_{f_i} = \alpha_i$ (Lemma 15.20), no further blocks can occur. Together, these $B(\gamma_j^{a_{ij}})$ form the unique Weierstrass normal form of f. From $\alpha_k \mid \ldots \mid \alpha_1$ it follows that $a_{1j} \geq \ldots \geq a_{kj}$ for $j = 1, \ldots, l$. In this way, one can recover $\alpha_1, \ldots, \alpha_k$ from the Weierstrass normal form (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. Moreover, the Frobenius normal form does not depend on the factorization of the minimal polynomial. In particular, the Frobenius normal form does not change if one replaces K 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 $\beta_1, \ldots, \beta_k \in K[X]$ with $\beta_k \notin K$ are given, then one obtains the Frobenius normal form

$$\operatorname{diag}\big(B(\beta_1 \ldots \beta_k), B(\beta_2 \ldots \beta_k), \ldots, B(\beta_k)\big) \in K^{n \times n}$$

with

$$n = \deg(\beta_1) + 2\deg(\beta_2) + \ldots + k\deg(\beta_k).$$

In this way, the similarity classes of matrices can be enumerated systematically. If one wants to count only invertible matrices, then 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 $\neq 0$ if $d \geq 1$. 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

Altogether, 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 harder 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 tell 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 may split). A is diagonalizable if and only if the Weierstrass normal form is a diagonal matrix. In this case, the Weierstrass normal form therefore coincides with the Jordan normal form.

Example 15.24.

(a) Let $\alpha, \beta, \gamma \in K[X]$ be irreducible and pairwise distinct. The conversion between the Frobenius and Weierstrass normal forms 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) Computing 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.[2] 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

$$(A - 1_3)(A - 2\cdot 1_3) = \begin{pmatrix} 6 & -1 & -3 \\ 30 & -5 & -15 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 5 & -1 & -3 \\ 30 & -6 & -15 \\ 0 & 0 & -1 \end{pmatrix} = 0$$

we have $\mu_A = (X-1)(X-2)$. Hence

$$\begin{pmatrix} B(\mu_A) & 0 \\ 0 & B(X-1) \end{pmatrix} = \begin{pmatrix} 0 & -2 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

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 $\mathbb{F}_2^{3\times 3}$ by the sequence $(\beta_1, \ldots, \beta_k)$ from Remark 15.23:

$$\begin{array}{ccc} (1,1,X) = 0_3, & (1,1,X+1) = 1_3, & (X,X), \\ (X,X+1) \approx \operatorname{diag}(1,1,0), & (X+1,X) \approx \operatorname{diag}(1,0,0), & (X+1,X+1) = P_{(1,2)}, \\ (X^3) = J_3(0), & (X^3+1) = P_{(1,2,3)}, & (X^3+X), \\ (X^3+X+1), & (X^3+X^2), & (X^3+X^2+1), \\ (X^3+X^2+X), & (X^3+X^2+X+1). & \end{array}$$

Hence there are exactly 14 similarity classes, of which six belong to invertible matrices (which ones?).

[2] see [M. Geck, *On Jacob's construction of the rational canonical form of a matrix*, Electron. J. Linear Algebra 36 (2020), 177–182]

Theorem 15.25. *For all $A \in K^{n\times n}$ we have $A \approx A^{\mathrm{t}}$.*

Proof. By the Frobenius normal form, we may assume $A = B(\alpha)$ for a monic polynomial $\alpha \in K[X] \setminus K$. By Lemma 15.20 and Exercise II.4, $\chi_{A^{\mathrm{t}}} = \chi_A = \mu_A = \mu_{A^{\mathrm{t}}}$. 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.32.

15.3 Centralizers

Definition 15.27. For $f \in \mathrm{End}(V)$ and $A \in K^{n\times n}$ let

$$\mathrm{C}(f) := \{g \in \mathrm{End}(V) : f \circ g = g \circ f\} \subseteq \mathrm{End}(V),$$
$$\mathrm{C}(A) := \{B \in K^{n\times n} : AB = BA\} \subseteq K^{n\times n}$$

the *centralizer* of f and A, respectively.

Remark 15.28. Clearly, centralizers are subspaces. Moreover, $CD \in \mathrm{C}(A)$ for $C, D \in \mathrm{C}(A)$. For $S \in \mathrm{GL}(n, K)$ we have $\mathrm{C}(SAS^{-1}) = S\,\mathrm{C}(A)S^{-1}$. Hence $\dim \mathrm{C}(A)$ is an invariant of the similarity class of A. We have

$$\{\alpha(A) : \alpha \in K[X]\} = \langle A^i : i \in \mathbb{N}_0\rangle \subseteq \mathrm{C}(A).$$

We investigate when equality holds.

Example 15.29.

(a) Let $A \in K^{n\times n}$ be a diagonal matrix. After permuting the basis vectors, we may assume $A = \mathrm{diag}(\lambda_1 1_{d_1}, \ldots, \lambda_k 1_{d_k})$ with pairwise distinct $\lambda_1, \ldots, \lambda_k \in K$. By Exercise I.16 we have $\mathrm{C}(A) = \{\mathrm{diag}(A_1, \ldots, A_k) : \forall i : A_i \in K^{d_i\times d_i}\}$. In particular, $\dim \mathrm{C}(A) = \sum_{i=1}^k d_i^2$. This is generalized in Exercise II.34 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) By Exercise II.27 we have $\dim \mathrm{C}(J_n(\lambda)) = n$ for all $\lambda \in K$. This is 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*

$$H := \big\{h \in \mathrm{Hom}(V, W) : h \circ f = g \circ h\big\}$$

is a vector space with $\dim H = \dim V$.

Proof. Let $d := \dim V$ and $e := \dim W$. Since $\chi_f = \mu_f$ and $\chi_g = \mu_g$, there exist $v \in V$ and $w \in W$ with

$$V = \langle f^i(v) : i = 0, \dots, d-1 \rangle, \qquad W = \langle g^i(w) : i = 0, \dots, e-1 \rangle.$$

Let $h \in H$. Then there exists a unique polynomial $\alpha \in K[X]$ with $h(v) = \alpha(g)(w)$ and $\deg \alpha < e$. For $i = 1, \dots, d-1$ it follows that $h(f^i(v)) = g^i(h(v))$. Hence h is uniquely determined by α. By Lemma 10.15, the map $H \to K[X]$, $h \mapsto \alpha$ is linear and injective. Because of

$$0 = h\big(\mu_f(f)(v)\big) = \mu_f(g)(h(v)) = \mu_f(g)\big(\alpha(g)(w)\big) = (\alpha\mu_f)(g)(w)$$

we also have $\mu_g \mid \alpha\mu_f$, 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 $< d$. Since $\deg(\alpha/\tau) = \deg(\alpha) + d - e < d$,

$$\Gamma \colon H \to P_d, \qquad h \mapsto \alpha/\tau$$

is an injective linear map.

Conversely, let $\beta \in P_d$ be given and $\alpha := \beta\tau$. Then $\mu_g \mid \alpha\mu_f$. We define $h \in \operatorname{Hom}(V, W)$ by $h(v) = \alpha(g)(w)$ and $h(f^i(v)) = g^i(h(v))$ for $i = 1, \dots, d-1$. For $i = 0, \dots, d-2$ we then have

$$(h \circ f)(f^i(v)) = g^{i+1}(h(v)) = g(g^i(h(v))) = (g \circ h)(f^i(v)).$$

Let $\mu_f = X^d + a_{d-1}X^{d-1} + \ldots + a_0$. Since $\mu_f(f) = 0$, we have

$$\begin{aligned}(h \circ f)(f^{d-1}(v)) &= h(f^d(v)) = h\big(-a_{d-1}f^{d-1}(v) - \ldots - a_0 v\big) = -a_{d-1}g^{d-1}(h(v)) - \ldots - a_0 h(v) \\ &= g^d(h(v)) - \mu_f(g)(h(v)) = g^d(h(v)) - (\mu_f\alpha)(g)(w) = g^d(h(v)) = (g \circ h)(f^{d-1}(v)).\end{aligned}$$

This shows $h \in H$ with $\Gamma(h) = \beta$. Hence Γ is an isomorphism and $\dim H = \dim P_d = d = \dim V$. □

Example 15.31. Let $V = K$ and $f = \lambda \operatorname{id}_V$ for some $\lambda \in K$. For $h \in H$ we have $\lambda h(1) = h(f(1)) = g(h(1))$, i. e., $h(1)$ is an eigenvector of g for the eigenvalue λ. Moreover, $H \to E_\lambda(g)$, $h \mapsto h(1)$ is an isomorphism. Since $\dim H = 1$, λ has geometric multiplicity 1. Of course, one can also derive this 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* $\operatorname{diag}(B(\alpha_1), \dots, B(\alpha_k))$. *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 $i = 1, \dots, k$. By Remark 15.28 we may assume $A = \operatorname{diag}(A_1, \dots, A_k)$. 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 $i \leq j$ we have $\chi_{A_j} = \mu_{A_j} = \alpha_j \mid \alpha_i = \chi_{A_i} = \mu_{A_i}$ by Lemma 15.20. By Lemma 15.30, there are d_j linearly independent choices for C_{ij}. In the case $i > j$ we can consider

$$C_{ij}^{\mathrm{t}} A_i^{\mathrm{t}} = (A_iC_{ij})^{\mathrm{t}} = (C_{ij}A_j)^{\mathrm{t}} = A_j^{\mathrm{t}} C_{ij}^{\mathrm{t}}$$

Since $\chi_{A_i^{\mathrm{t}}} = \alpha_i = \mu_{A_i^{\mathrm{t}}}$ (Exercise II.4), there are d_i linearly independent choices for C_{ij}^{t} (and hence also for C_{ij}). Since the blocks C_{ij} in C can be chosen independently of each other, we obtain

$$\dim \mathrm{C}(A) = \sum_{i,j=1}^{k} d_{\max\{i,j\}} = d_1 + 3d_2 + 5d_3 + \ldots + (2k-1)d_k \geq d_1 + \ldots + d_k = n$$

with equality if and only if $k = 1$, i. e., $\mu_A = \alpha_1 = \chi_A$ (Theorem 15.22). □

Remark 15.33. In Remark 15.23 we described the Frobenius normal form by the polynomials $\beta_1 := \alpha_1/\alpha_2$, $\beta_2 := \alpha_2/\alpha_3, \ldots, \beta_k := \alpha_k$. We have $\deg \alpha_i = \deg(\beta_i \beta_{i+1} \ldots \beta_k) = \sum_{j=i}^{k} \deg \beta_j$ and $\sum_{i=1}^{m}(2i-1) = m^2$ by 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 $\dim \mathrm{C}(A) \geq \dim\langle A^i : i \in \mathbb{N}_0\rangle = n$. By Frobenius' theorem, on the other hand, $\dim \mathrm{C}(A) = n$. □

Theorem 15.35. *Let $A \in \mathrm{GL}(n, \mathbb{C})$ with exactly e pairwise distinct eigenvalues and $\chi_A = \mu_A$. For $k \in \mathbb{N}$ there exist exactly k^e matrices $W \in \mathbb{C}^{n\times n}$ with $W^k = A$.*

Proof. W. l. o. g. let $A = \mathrm{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$ be in Jordan normal form. By Theorem 14.35, $\lambda_1, \ldots, \lambda_s$ are pairwise distinct, i. e., $s = e$. Every $W \in \mathbb{C}^{n\times n}$ with $W^k = A$ commutes with A. By Corollary 15.34, W is a polynomial in A and therefore has the form $W = \mathrm{diag}(W_1, \ldots, W_e)$ with $W_i^k = J_{n_i}(\lambda_i)$ for $i = 1, \ldots, e$. Hence we may assume $e = 1$ and $\lambda := \lambda_1$. By Theorem 14.42, there exists at least one W with $W^k = A$. For every k-th root of unity $\zeta \in \mathbb{C}$ we also have $(\zeta W)^k = A$. By Lemma 11.27, there exist at least k roots of A.

Conversely, let $W \in \mathbb{C}^{n\times n}$ with $W^k = A$. As a polynomial in A, W is a lower triangular matrix with main diagonal $(\omega, \ldots, \omega)$, where $\omega^k = \lambda$. Together with A, W is also invertible. Let $Y \in \mathbb{C}^{n\times n}$ also be a k-th root of A with main diagonal $(\omega, \ldots, \omega)$. Then $S := W^{-1}Y$ is a lower triangular matrix with ones on the main diagonal. Since W and Y commute as polynomials in A, we have $S^k = W^{-k}Y^k = 1_n$. By Corollary 11.29, S is diagonalizable. Since 1 is the only eigenvalue of S, it follows that $S \approx 1_n$, i. e., $S = 1_n$ and $W = Y$. Consequently, W is uniquely determined by its main diagonal. Hence there can be at most k roots of A. □

15.4 Splitting Fields

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. By 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) we have $0 \neq \frac{\mu_f}{\gamma}(f)(v) \in U$. This shows $U = V$, and $\chi_f = \mu_f = \gamma$ is irreducible.

Conversely, let χ_f be irreducible and $\{0\} \neq U \leq V$ 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 $\chi_{f_{|U}} = \chi_f$. It follows that $U = V$. □

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, matrix multiplication in L is commutative. For $g \in \mathrm{C}(f) \setminus \{0\}$, $\operatorname{Ker} g \leq V$ is an f-invariant subspace as usual. By Schur's lemma, $\operatorname{Ker} g = \{0\}$ and $g \in \mathrm{GL}(V)$. Together with g we also have $g^{-1} \in \mathrm{C}(f)$. Hence every element of $L \setminus \{0_n\}$ has an inverse with respect to multiplication. The remaining field axioms for $\mathrm{C}(A)$ follow from the computation 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 $A := B(\gamma) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$. By 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 $\gamma = X^2 - 2 \in \mathbb{Q}[X]$, $\mathrm{C}(B(\gamma)) \cong \mathbb{Q}(\sqrt{2})$ is the field from Exercise I.15.

(c) For $\gamma := X^2 + X + 1 \in \mathbb{F}_2[X]$ we have

$$ \mathrm{C}(B(\gamma)) = \left\{ 0_2, 1_2, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right\} $$

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 $d := \deg \gamma$. By Theorem 15.37, $L_1 := \mathrm{C}(B(\gamma))$ is a field. We can identify K with the scalar matrices $\{\lambda 1_d \in L_1 : \lambda \in K\}$ (the computation 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. By Lemma 10.22 we have $\alpha = (X - x)\beta$ 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. Clearly, 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. In this section we decompose square matrices 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 that α is *congruent* to β *modulo* δ if $\delta \mid \alpha - \beta$. In this case 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 $\alpha_i \equiv \beta_i \pmod{\delta}$ $(i = 1, 2)$ we have $\alpha_1 + \alpha_2 \equiv \beta_1 + \beta_2 \pmod{\delta}$ and $\alpha_1\alpha_2 \equiv \beta_1\beta_2 \pmod{\delta}$.*

Proof. Since $\delta \mid 0 = \alpha - \alpha$, 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 $\alpha \equiv \gamma \pmod{\delta}$. Hence $\equiv$ 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 claim. □

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) We have $\alpha \equiv \beta \pmod{X}$ if and only if α and β have the same constant term.

(c) For every $\alpha \in K[X]$ there exists a $\beta \in K[X]$ with $\alpha \equiv \beta \pmod{\delta}$ and $\deg \beta < \deg \delta$. This follows from division with remainder.

(d) Lemma 16.3 simplifies many divisibility considerations: To check whether

$$\alpha = (X^2 + X + 2)^4 + (X^5 - 1)(X^3 + X + 10) \in \mathbb{Q}[X]$$

is divisible by $\delta = X + 1 \in \mathbb{Q}[X]$, one can reduce each term individually modulo δ. Because of

$$\begin{aligned} X^2 + X + 2 &= X(X + 1) + 2 \equiv 2 \pmod{\delta}, \\ X^5 - 1 &= (X + 1)(X^4 - X^3 + X^2 - X + 1) - 2 \equiv -2 \pmod{\delta}, \\ X^3 + X + 10 &= (X + 1)(X^2 - X + 2) + 8 \equiv 8 \pmod{\delta} \end{aligned}$$

we have $\alpha \equiv 2^4 + (-2)8 \equiv 0 \pmod{\delta}$, 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 let $\beta_1, \ldots, \beta_n \in K[X]$ 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 $\alpha_i' := \prod_{j \neq i} \alpha_j$. By Corollary 15.11, every irreducible divisor of α_i' must divide some 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 $\gamma \equiv \alpha_i' \beta_i \gamma_i \equiv \beta_i \pmod{\alpha_i}$ for $i = 1, \ldots, n$. □

Remark 16.6. Together with γ, also $\gamma + \rho \prod_{i=1}^{n} \alpha_i$ satisfies the congruences from the Chinese remainder theorem for all $\rho \in K[X]$. Hence 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. We look 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 has to resort to 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 of 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 $\alpha' \neq 0$ 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, $(X^2)' = 0$ in $\mathbb{F}_2[X]$. In general, let $\alpha = \sum a_k X^k \in \mathbb{F}_2[X]$ with $\alpha' = \sum k a_k X^{k-1} = 0$. For odd k it follows that $a_k X^{k-1} = k a_k X^{k-1} = 0$, so $a_k = 0$. This shows

$$\alpha = a_0 + a_2 X^2 + \ldots + a_{2n} X^{2n}$$

for some $n \in \mathbb{N}$. Since $(x + y)^2 = x^2 + 2xy + y^2 = x + y$ 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). Hence 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. Let in particular $K := \mathbb{F}_2(X)$. One can show that $\alpha := Y^2 + X \in K[Y]$ 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

$$\begin{aligned}(\alpha + \beta)' &= \Big(\sum (a_k + b_k) X^k\Big)' = \sum k(a_k + b_k) X^{k-1} \\ &= \sum k a_k X^{k-1} + \sum k b_k X^{k-1} = \alpha' + \beta'.\end{aligned}$$

This implies

$$\begin{aligned}(\alpha\beta)' &= \Big(\sum_{k,l=0}^{\infty} a_k b_l X^{k+l}\Big)' = \sum_{k,l} a_k b_l \big(X^{k+l}\big)' = \sum_{k,l} (k+l) a_k b_l X^{k+l-1} \\ &= \sum_{k,l} k a_k b_l X^{k-1+l} + \sum_{k,l} l a_k b_l X^{l-1+k} = \alpha'\beta + \alpha\beta'.\end{aligned}$$

□

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 $\gamma = (X - \lambda)\rho$. The product rule yields

$$\gamma' = \rho + (X - \lambda)\rho'$$

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 $\rho(\lambda) = \gamma'(\lambda) \neq 0$. Hence λ is a simple root. Since λ was arbitrary, all roots are simple. □

16.2 Separable and Semisimple Maps

Definition 16.12. We call $f \in \mathrm{End}(V)$ (or $A \in K^{n\times n}$)

- *separable*, if μ_f (or μ_A) is separable.
- *semisimple*, if μ_f (or μ_A) splits into pairwise distinct irreducible polynomials.

Example 16.13.

(a) Over $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{F}_2$, all endomorphisms are separable by Example 16.9.

(b) Let $K = \mathbb{C}$. By 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. *An endomorphism $f \in \mathrm{End}(V)$ 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 yields a decomposition $V = V_1 \oplus \dots \oplus V_s$ into f-invariant subspaces such that the matrix of f on V_i is given by $B(\gamma_j)$ for some $1 \leq j \leq k$ (note that $k \leq s$). By Lemma 15.20 and Schur's lemma, V_i has no proper nontrivial 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 (possibly $W = \{0\}$). Suppose that $U \oplus W < V$. Then there exists an i with $V_i \not\subseteq U + W$. Clearly,

$$L := V_i \cap (U + W) < V_i$$

is f-invariant. Schur's lemma yields $L = \{0\}$. 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 $u = w \in U \cap W = \{0\}$. Hence $U \cap (W + V_i) = \{0\}$, contradicting the choice of W. This shows $V = U \oplus W$.

Conversely, suppose that every f-invariant subspace of V has an f-invariant complement. We assume indirectly that $\mu_f = \gamma^2\delta$ for some irreducible polynomial γ. By assumption, $U := \mathrm{Ker}(\gamma(f)) \leq V$ has an f-invariant complement W. For $w \in W$ we have

$$(\gamma\delta)(f)(w) \in U \cap W = \{0\}.$$

Since $(\gamma\delta)(f)(U) = \{0\}$, we even have $(\gamma\delta)(f)(v) = 0$ for all $v \in V$. But then $\mu_f \mid \gamma\delta$, a contradiction. □

Lemma 16.15. *Let $A \in K^{n\times n}$ be separable. Then 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. By Lemma 16.11, every 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. Hence γ_i and γ_j have no common roots in L. Altogether, μ_A splits into pairwise distinct linear factors in $L[X]$. By Theorem 10.52, A is diagonalizable in $L^{n\times n}$. Conversely, if A is diagonalizable over a field $L \supseteq K$, then μ_A splits in $L[X]$ into pairwise distinct linear factors. Thus μ_A cannot have multiple irreducible divisors in $K[X]$. □

Lemma 16.16. *If $A, B \in K^{n\times n}$ commute and are separable and semisimple, then $A + B$ is also semisimple.*

Proof. Let L be a splitting field of $\mu_A\mu_B$. By (the proof of) Lemma 16.15 and Lemma 14.11, A and B are simultaneously diagonalizable in $L^{n\times n}$. Hence $A + B$ is also diagonalizable in $L^{n\times n}$. By Theorem 10.52, μ_{A+B} has no multiple roots in L. By Lemma 16.11, $A + B$ is separable and by Lemma 16.15 semisimple. □

16.3 Generalized Jordan Blocks

Remark 16.17. If $\gamma \in K[X]$ is irreducible, one can regard $\lambda := B(\gamma)$ as an element of the field $L := \mathrm{C}(B(\gamma))$. In the following theorem 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 let $k \in \mathbb{N}$. 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 $J := J_k(\gamma)$. As in Lemma 14.41 we have

$$J^m = \begin{pmatrix} B(\gamma)^m & & & 0 \\ mB(\gamma)^{m-1} & \ddots & & \\ & \ddots & \ddots & \\ * & & mB(\gamma)^{m-1} & B(\gamma)^m \end{pmatrix}$$

for $m \in \mathbb{N}_0$ (cf. Theorem 14.42). It follows that

$$\gamma(J) = \begin{pmatrix} 0 & & & 0 \\ \gamma'(B(\gamma)) & \ddots & & \\ & \ddots & \ddots & \\ * & & \gamma'(B(\gamma)) & 0 \end{pmatrix}.$$

In particular, $\gamma(J)$ is nilpotent and $\mu_J \mid \gamma^k$. Similarly,

$$\gamma(J)^2 = \begin{pmatrix} 0 & & & & 0 \\ 0 & \ddots & & & \\ \gamma'(B(\gamma))^2 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ * & & \gamma'(B(\gamma))^2 & 0 & 0 \end{pmatrix}, \ldots, \gamma(J)^{k-1} = \begin{pmatrix} 0 & & & 0 \\ \vdots & \ddots & & \\ 0 & & \ddots & \\ \gamma'(B(\gamma))^{k-1} & 0 & \cdots & 0 \end{pmatrix}.$$

Since γ is separable, we have $0 \le \deg \gamma' < \deg \gamma$. In particular, $\gamma \nmid \gamma'$ and $\gamma \nmid (\gamma')^{k-1}$ by Corollary 15.11. Lemma 15.20 yields $\gamma'(B(\gamma))^{k-1} = (\gamma')^{k-1}(B(\gamma)) \neq 0$. 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.

Moreover, there exists an $\alpha \in K[X]$ with $\alpha(A) = D$.

Proof.
Step 1: Existence of D and N.
The properties of D and N as well as the existence of α are invariant under change of basis. We may therefore transform A into Weierstrass normal form. Since μ_A is separable, we can then transform every block $B(\gamma^a)$ in the Weierstrass normal form into $J_a(\gamma)$ (Theorem 16.18). Then

$$A = \operatorname{diag}(J_{a_1}(\gamma_1), \dots, J_{a_s}(\gamma_s)).$$

For $i = 1, \dots, s$ let $d_i := \deg \gamma_i$ and

$$D_i := \operatorname{diag}(B(\gamma_i), \dots, B(\gamma_i)) \in K^{a_i d_i \times a_i d_i}, \qquad N_i := J_{a_i}(\gamma_i) - D_i,$$
$$D := \operatorname{diag}(D_1, \dots, D_s), \qquad N := \operatorname{diag}(N_1, \dots, N_s).$$

We have $A = D + N$. W. l. o. g. let $\gamma_1, \dots, \gamma_k$ be the *distinct* prime divisors of μ_A. For $1 \le i \le k$ and $1 \le j \le s$ with $\gamma_i = \gamma_j$ let moreover $a_i \ge a_j$. By Lemma 15.20, $\mu_{D_i} = \gamma_i$ and $\mu_D = \gamma_1 \dots \gamma_k$, i. e., D is semisimple. On the other hand, N is nilpotent as a strictly lower triangular matrix (Example 14.26). A computation as in Theorem 16.18 shows $D_i N_i = N_i D_i$ and $DN = ND$.

Step 2: Existence of α.
Since $J_i := J_{a_i}(\gamma_i) = D_i + N_i$, we have $D_i \in \mathrm{C}(J_i)$. By Theorem 16.18 and Corollary 15.34, there exists an $\alpha_i \in K[X]$ with $\alpha_i(J_i) = D_i$ for $i = 1, \dots, k$. Let $1 \le j \le s$ with $\gamma_j = \gamma_i$. Then $a_j \le a_i$ 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 $\alpha_i(J_j) = D_j$. By the Chinese remainder theorem, there exists an $\alpha \in K[X]$ with $\alpha \equiv \alpha_i \pmod{\gamma_i^{a_i}}$ for $i = 1, \dots, k$. From $\gamma_i^{a_i}(J_i) = 0$ it follows that

$$\alpha(A) = \operatorname{diag}\big(\alpha(J_1), \dots, \alpha(J_s)\big) = \operatorname{diag}(D_1, \dots, D_s) = D.$$

Since μ_D has the same prime divisors as μ_A, D is separable.

Step 3: Uniqueness of D and N.
Let $A = \check{D} + \check{N}$ with the same properties. We first show that $\check{D}$ is separable. Over a splitting field, $\check{D}$ and $\check{N}$ are simultaneously triangularizable by Theorem 14.8 and Lemma 14.12. With respect to a suitable basis, A is then a triangular matrix. Being a nilpotent matrix, $\check{N}$ must be a strictly triangular matrix with respect to this basis. Hence A and $\check{D}$ have the same main diagonal and thus the same characteristic polynomial (note that χ_A does not depend on the splitting field). Since A is separable, $\check{D}$ must also be separable. By assumption, $\check{D}$ commutes with A and with $\alpha(A) = D$. By Lemma 16.16, $D - \check{D}$ is semisimple. Similarly, $\check{N}$ and N commute as well. Multiplying out $(\check{N} - N)^{2n}$, we obtain summands of the form $\check{N}^i N^{2n-i}$ with $0 \le i \le 2n$. In the case $i \ge n$ we have $\check{N}^i = 0$. Otherwise $2n - i \ge n$ and $N^{2n-i} = 0$. In any case, $(\check{N} - N)^{2n} = 0$. Altogether, $D - \check{D} = \check{N} - N$ is semisimple and nilpotent. This is only possible with minimal polynomial X, i. e., $\check{D} = D$ and $\check{N} = N$. □

Remark 16.20. (a) In the case $n \le 3$, the separability of A in Theorem 16.19 can be dropped. Indeed, if A is inseparable, then μ_A has an irreducible divisor of degree ≥ 2. In this case $A = D + 0_n$ is semisimple. Moreover, $\chi_A = \mu_A$ and $\mathrm{C}(A) = \langle A^i : i \in \mathbb{N}_0 \rangle$ by Corollary 15.34. For an arbitrary Jordan-Chevalley decomposition $A = D + N$ we have $N = \alpha(A) \in \mathrm{C}(A)$ for some $\alpha \in K[X]$. Since $\alpha^n(A) = N^n = 0$, we have $\mu_A \mid \alpha^n$. By Corollary 15.11, every prime divisor of μ_A divides α. Since A is semisimple, we must already have $\mu_A \mid \alpha$, i. e., $N = 0$. Hence the decomposition $A = D + 0_n$ is unique.

(b) For $n \geq 4$, by Exercise II.36 neither the existence nor the uniqueness of the Jordan-Chevalley decomposition is guaranteed for inseparable matrices.

(c) Chevalley constructed D and N with a version of Newton's method (Example 11.36) for matrices.[1]

Corollary 16.21. *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. By Example 16.13, A is separable and D is semisimple if and only if D is diagonalizable. □

Remark 16.22. If $A \in \mathbb{C}^{n\times n}$ is given in Jordan normal form with main diagonal $(\lambda_1, \ldots, \lambda_n)$, then the Jordan-Chevalley decomposition is obtained by $D := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $N := A - D$. For arbitrary fields K one can regard A over a splitting field L of μ_A and use the Jordan normal form there. If A is separable, the Jordan-Chevalley decompositions over K and L coincide by Lemma 16.15.

Example 16.23. Let $\gamma := X^2 + 1 \in \mathbb{Q}[X]$ and

$$A := B(\gamma^2) = \begin{pmatrix} . & . & . & -1 \\ 1 & . & . & . \\ . & 1 & . & -2 \\ . & . & 1 & . \end{pmatrix} \in \mathbb{Q}^{4\times 4}.$$

Avoiding the detour via $\mathbb{C}$ described in Remark 16.22 (or $\mathbb{Q}(\mathrm{i})$, cf. Exercise I.15), we determine the Jordan-Chevalley decomposition of A. By Theorem 16.18 we have $A \approx J_2(\gamma)$. To realize this change of basis, we need a basis vector $b_3 \in \mathbb{Q}^4$ with $\mu_{b_3} = \gamma$. A suitable choice is

$$b_3 := \gamma(A)e_1 = e_1 + e_3 = (1, 0, 1, 0)^{\mathrm{t}}.$$

Then $b_4 := Ab_3 = (0, 1, 0, 1)^{\mathrm{t}}$. For b_1 and b_2 we require: $Ab_1 = b_2 + b_3$ and $Ab_2 = -b_1 + b_4$. This gives 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

$$S := \begin{pmatrix} . & -1 & 1 & . \\ 2 & . & . & 1 \\ . & 1 & 1 & . \\ . & . & . & 1 \end{pmatrix} \in \mathrm{GL}(4, \mathbb{Q})$$

we have $S^{-1}AS = J_2(\gamma)$. This yields the Jordan-Chevalley decomposition $A = D + N$ with

$$D := S \operatorname{diag}(B(\gamma), B(\gamma))S^{-1} = \frac{1}{2}\begin{pmatrix} . & -1 & . & -1 \\ 3 & . & -1 & . \\ . & 1 & . & -3 \\ 1 & . & 1 & . \end{pmatrix},$$

$$N := S\begin{pmatrix} 0_2 & 0_2 \\ 1_2 & 0_2 \end{pmatrix}S^{-1} = \frac{1}{2}\begin{pmatrix} . & 1 & . & -1 \\ -1 & . & 1 & . \\ . & 1 & . & -1 \\ -1 & . & 1 & . \end{pmatrix}.$$

[1] see [M. Geck, *On the Jordan–Chevalley decomposition of a matrix*, arXiv:2205.05432]

Theorem 16.24. *Let $A \in K^{n\times n}$ be separable with Jordan-Chevalley decomposition $A = D + N$. Then* $\mathrm{C}(A) = \mathrm{C}(D) \cap \mathrm{C}(N)$.

Proof. In general, $\mathrm{C}(D) \cap \mathrm{C}(N) \subseteq \mathrm{C}(D + N) = \mathrm{C}(A)$. Since D and $N = A - D$ are polynomials in A, every matrix in $\mathrm{C}(A)$ also commutes with D and N, i. e., $\mathrm{C}(A) \subseteq \mathrm{C}(D) \cap \mathrm{C}(N)$. □

17 Kronecker Products

17.1 Definition and Properties

Remark 17.1. Determining the centralizer of a matrix A leads to the matrix equation $AX = XA$. In this chapter we solve, more generally, equations of the form $AX - XB = C$ and $AXB = C$ for given matrices A, B and C.

Definition 17.2. For $A = (a_{ij}) \in K^{n\times m}$ and $B \in K^{s\times t}$, the block matrix

$$A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{nm}B \end{pmatrix} \in K^{ns\times mt}$$

is called the *Kronecker product* of A and B. We number the rows of $A \otimes B$ with pairs (i,k) $(1 \le i \le n,$ $1 \le k \le s)$ and the columns with pairs (j,l) $(1 \le j \le m,\ 1 \le l \le t)$, in each case in lexicographic order. Then $(A \otimes B)_{(i,k),(j,l)} = a_{ij}b_{kl}$.

Remark 17.3.

(a) Let V and W be vector spaces with bases $B = \{b_1, \ldots, b_n\}$ and $C = \{c_1, \ldots, c_m\}$, respectively. The *tensor product* of V and W is an abstract vector space $V \otimes W$ with basis

$$T = \{b_i \otimes c_j : i = 1, \ldots, n,\ j = 1, \ldots, m\}.$$

For $v = \sum \lambda_i b_i$ and $w = \sum \mu_j c_j$ one writes

$$v \otimes w := \sum_{i,j} \lambda_i \mu_j (b_i \otimes c_j).$$

Let V' and W' be further vector spaces with bases B' and C', respectively. For $f \in \mathrm{Hom}(V, V')$ and $g \in \mathrm{Hom}(W, W')$ one easily shows that there exists exactly one map $f \otimes g \in \mathrm{Hom}(V \otimes W, V' \otimes W')$ with $(f \otimes g)(v \otimes w) = f(v) \otimes g(w)$ for all $v \in V$ and $w \in W$. Let $T' := \{b' \otimes c' : b' \in B',\ c' \in C'\}$ be the basis of $V' \otimes W'$ analogous to T. For the matrices of these linear maps we have

$${}_{T'}[f \otimes g]_T = {}_{B'}[f]_B \otimes {}_{C'}[g]_C.$$

If one chooses $V = K^n$ and $W = K^m$, then $e_i \otimes e_j$ can be identified with the Kronecker product $e_i^{\mathrm{t}} \otimes e_j = E_{ij} \in K^{n\times m}$. Then $V \otimes W = K^{n\times m}$. Since vector spaces are uniquely determined up to isomorphism by their dimension anyway, we restrict ourselves to the study of Kronecker products.

(b) It gets more interesting when one studies tensor products $V := V_1 \otimes \ldots \otimes V_k$ or $K^{d_1 \times \ldots \times d_k} := K^{d_1} \otimes \ldots \otimes K^{d_k}$ with more than two factors (so-called *higher-order tensors*). Every element of V has the form

$$t = \sum_{i=1}^{r} v_{i1} \otimes \ldots \otimes v_{ik}$$

with $v_{ij} \in V_j$. The smallest possible number r for which such a representation exists is called the *rank* of t. For $k = 2$, by Exercise I.20, $\operatorname{rk} t$ is the rank of the matrix from (a). While the rank of a matrix can be computed efficiently with the Gaussian algorithm, determining $\operatorname{rk} t$ for $k \geq 3$ is an NP-hard problem according to HÅSTAD. For $K = \mathbb{Q}$, not even an algorithm for determining $\operatorname{rk} t$ is known. By Exercise II.38, $\operatorname{rk} t \leq 3$ for all $t \in K^{2\times 2\times 2}$.

Lemma 17.4. *For matrices A, B, C, D of suitable sizes we have*

$$\begin{aligned} A \otimes (B \otimes C) &= (A \otimes B) \otimes C, & (A \otimes B)(C \otimes D) &= (AC) \otimes (BD), \\ A \otimes (B + C) &= A \otimes B + A \otimes C, & (A + B) \otimes C &= A \otimes C + B \otimes C, \\ (A \otimes B)^{\mathrm{t}} &= A^{\mathrm{t}} \otimes B^{\mathrm{t}}, & (A \otimes B)^* &= A^* \otimes B^* \ (K = \mathbb{C}). \end{aligned}$$

Moreover:

(a) If A and B are invertible, then so is $A \otimes B$, with $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.

(b) $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$.

(c) $\operatorname{rk}(A \otimes B) = \operatorname{rk}(A) \operatorname{rk}(B)$.

(d) $\det(A \otimes B) = \det(A)^m \det(B)^n$ if $A \in K^{n\times n}$ and $B \in K^{m\times m}$.

Proof. Let $C = (c_{xy}) \in K^{u\times v}$. The rows of $(A \otimes B) \otimes C$ are numbered with pairs $((i,k),x)$ and those of $A \otimes (B \otimes C)$ with pairs $(i,(k,x))$. In both cases this corresponds to the lexicographic order of the triples (i,k,x). Similarly, we number the columns with triples (j,l,y). Then

$$\big((A \otimes B) \otimes C\big)_{(i,k,x),(j,l,y)} = a_{ij} b_{kl} c_{xy} = \big(A \otimes (B \otimes C)\big)_{(i,k,x),(j,l,y)}.$$

For the distributive laws, the entry at position $((i,k),(j,l))$ is $a_{ij}(b_{kl} + c_{kl}) = a_{ij}b_{kl} + a_{ij}c_{kl}$ and $(a_{ij} + b_{ij})c_{kl} = a_{ij}c_{kl} + b_{ij}c_{kl}$, respectively, on both sides. For the product we have

$$\begin{aligned} \big((A \otimes B)(C \otimes D)\big)_{(i,k),(j,l)} &= \sum_{x,y} a_{ix} b_{ky} c_{xj} d_{yl} \\ &= \Big(\sum_x a_{ix} c_{xj}\Big)\Big(\sum_y b_{ky} d_{yl}\Big) \\ &= (AC)_{ij} (BD)_{kl}. \end{aligned}$$

Finally,

$$(A \otimes B)^{\mathrm{t}}_{(i,k),(j,l)} = (A \otimes B)_{(j,l),(i,k)} = a_{ji} b_{lk} = (A^{\mathrm{t}})_{ij} (B^{\mathrm{t}})_{kl}.$$

For A^* and B^* one additionally conjugates the entries.

(a) We have $(A \otimes B)(A^{-1} \otimes B^{-1}) = (AA^{-1}) \otimes (BB^{-1}) = 1_n \otimes 1_s = 1_{ns}$.

(b) For $A \in K^{n\times n}$ and $B \in K^{s\times s}$ we have

$$\operatorname{tr}(A \otimes B) = \sum_{i=1}^{n} \sum_{k=1}^{s} a_{ii} b_{kk} = \operatorname{tr}(A) \operatorname{tr}(B).$$

(c) Let $a := \operatorname{rk} A$ and $b := \operatorname{rk} B$. By Remark 6.20 there exist invertible matrices S, T, S', T' with $SAT = \left(\begin{smallmatrix} 1_a & 0 \\ 0 & 0 \end{smallmatrix}\right) =: D$ and $S'BT' = \left(\begin{smallmatrix} 1_b & 0 \\ 0 & 0 \end{smallmatrix}\right) =: D'$. By the computation rules and (a),

$$(S \otimes S')(A \otimes B)(T \otimes T') = D \otimes D' \sim \begin{pmatrix} 1_{ab} & 0 \\ 0 & 0 \end{pmatrix}$$

with invertible matrices $S \otimes S'$ and $T \otimes T'$. This shows $\operatorname{rk}(A \otimes B) = \operatorname{rk}(1_{ab}) = ab$.

(d) The Gaussian algorithm transforms A into an upper triangular matrix $R = (r_{ij})$ solely by adding multiples of rows to other rows (cf. Exercise I.18). By Theorem 9.6, $\det A = \det R = r_{11} \dots r_{nn}$. If one performs the same operations blockwise on $A \otimes B$, one obtains $R \otimes B$ without changing the determinant. By Lemma 9.9 and Remark 9.8, we therefore have

$$\begin{aligned} \det(A \otimes B) &= \det(r_{11}B) \dots \det(r_{nn}B) \\ &= r_{11}^m \dots r_{nn}^m \det(B)^n = \det(A)^m \det(B)^n. \end{aligned}$$

□

Remark 17.5. In general, $A \otimes B \neq B \otimes A$. Because of

$$(A \otimes B)_{(i,k),(j,l)} = a_{ij} b_{kl} = (B \otimes A)_{(k,i),(l,j)}$$

the matrix $B \otimes A$ arises from $A \otimes B$ by reordering the rows and columns. For square matrices $A \in K^{n \times n}$ and $B \in K^{m \times m}$ we therefore have $A \otimes B = P(B \otimes A)P^{-1} \approx B \otimes A$ for a permutation matrix P that depends only on n and m.

Definition 17.6. For $A \in K^{n \times m}$ with columns $a_1, \dots, a_m$ let $\operatorname{vec}(A) := (a_1^{\mathrm{t}}, \dots, a_m^{\mathrm{t}})^{\mathrm{t}} \in K^{nm}$ be the vector obtained by writing the columns one below the other. One calls $\operatorname{vec}(A)$ the *vectorization* of A. Clearly, $\operatorname{vec}\colon K^{n \times m} \to K^{nm}$ is an isomorphism.

Lemma 17.7. *For $A \in K^{s \times n}$, $X \in K^{n \times m}$ and $B = (b_{lj}) \in K^{m \times t}$ we have*

$$\boxed{\operatorname{vec}(AXB) = (B^{\mathrm{t}} \otimes A) \operatorname{vec}(X).}$$

Proof. Let $x_1, \dots, x_m$ be the columns of X. The j-th column of XB is $\sum_{l=1}^m b_{lj} x_l$. Hence the j-th column of AXB is equal to $\sum_{l=1}^m b_{lj} A x_l$. On the other hand, the (j,l)-block of $B^{\mathrm{t}} \otimes A$ is equal to $b_{lj} A$. Therefore the j-th block of $(B^{\mathrm{t}} \otimes A) \operatorname{vec}(X)$ is also $\sum_{l=1}^m b_{lj} A x_l$. □

Theorem 17.8. *Let $A \in K^{s \times n}$, $B \in K^{m \times t}$ and $C \in K^{s \times t}$.*

(a) The solutions of the homogeneous equation $AXB = 0$ form a vector space $L_0 \leq K^{n \times m}$ of dimension $nm - \operatorname{rk}(A) \operatorname{rk}(B)$.

(b) If $AXB = C$ has a solution $Y \in K^{n \times m}$, then $Y + L_0$ is the solution set.

Proof. By Lemma 17.7, $AXB = C$ is equivalent to the linear system of equations $(B^{\mathrm{t}} \otimes A) \operatorname{vec}(X) = \operatorname{vec}(C)$ with nm unknowns. By Lemma 17.4(c), $\operatorname{rk}(B^{\mathrm{t}} \otimes A) = \operatorname{rk}(B^{\mathrm{t}}) \operatorname{rk}(A) = \operatorname{rk}(A) \operatorname{rk}(B)$. Since vec is an isomorphism, the claim follows from Theorem 6.6. □

Example 17.9. Let

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad C = AB = \begin{pmatrix} 1 & 3 \\ 2 & 6 \end{pmatrix}.$$

Clearly, 1_2 is a solution of $AXB = C$. Since $\operatorname{rk}(A) = 1$ and $\operatorname{rk}(B) = 2$, L_0 has dimension $4 - 2 = 2$ by Theorem 17.8. Since B is invertible, we have $L_0 = \{X : AX = 0\}$. Hence the columns of every matrix in L_0 lie in $\operatorname{Ker} A = \langle(-2, 1)^{\mathrm{t}}\rangle$. For dimension reasons,

$$L_0 = \left\{ \begin{pmatrix} -2s & -2t \\ s & t \end{pmatrix} : s, t \in K \right\},$$

and $1_2 + L_0$ is the solution set.

Definition 17.10. For $A \in K^{n\times n}$ and $B \in K^{m\times m}$ one calls

$$A \oplus B := A \otimes 1_m + 1_n \otimes B \in K^{nm\times nm}$$

the *Kronecker sum* of A and B.

Theorem 17.11. *Let $A \in K^{n\times n}$ and $B \in K^{m\times m}$ such that $\chi_A = (X - \lambda_1) \dots (X - \lambda_n)$ and $\chi_B = (X - \mu_1) \dots (X - \mu_m)$ split into linear factors. Then we have*

$$\chi_{A\otimes B} = \prod_{i=1}^{n} \prod_{j=1}^{m} (X - \lambda_i \mu_j), \qquad \chi_{A\oplus B} = \prod_{i=1}^{n} \prod_{j=1}^{m} (X - \lambda_i - \mu_j).$$

Proof. By Theorem 14.8 there exist $S \in \mathrm{GL}(n, K)$ and $T \in \mathrm{GL}(m, K)$ such that SAS^{-1} and TBT^{-1} are upper triangular matrices. By Lemma 17.4 we have

$$(S \otimes T)(A \otimes B)(S \otimes T)^{-1} = SAS^{-1} \otimes TBT^{-1},$$
$$(S \otimes T)(A \oplus B)(S \otimes T)^{-1} = SAS^{-1} \oplus TBT^{-1}.$$

Since similar matrices have the same characteristic polynomial, we may assume that A and B are upper triangular matrices with main diagonals $(\lambda_1, \dots, \lambda_n)$ and $(\mu_1, \dots, \mu_m)$, respectively. Then $A \otimes B$ and $A \oplus B$ are clearly upper triangular matrices with diagonal entries $\lambda_i \mu_j$ and $\lambda_i + \mu_j$, respectively. The claim follows. □

Remark 17.12.

(a) If A and B are diagonalizable, then so are $A \otimes B$ and $A \oplus B$ (same proof with diagonal matrices instead of triangular matrices).

(b) For arbitrary A and B, Theorem 17.11 can be applied over a splitting field of $\chi_A \chi_B$ (Theorem 15.39). We obtain

$$\det(A \otimes B) = \prod_{i,j} \lambda_i \mu_j = \det(A)^m \det(B)^n,$$
$$\operatorname{tr}(A \otimes B) = \sum_{i,j} \lambda_i \mu_j = \operatorname{tr}(A) \operatorname{tr}(B)$$

(cf. Lemma 17.4).

Example 17.13. With the help of Theorem 17.11 one can show that the eigenvalues of all rational matrices form a field $\overline{\mathbb{Q}} \subseteq \mathbb{C}$. If $x, y \in \overline{\mathbb{Q}}$ are eigenvalues of A and B, respectively, then xy is an eigenvalue of $A \otimes B$ and $x + y$ is an eigenvalue of $A \oplus B$. One calls $\overline{\mathbb{Q}}$ the field of *algebraic numbers* (Exercise II.37). By Lemma 15.20, $\sqrt{2} + \sqrt{3}$ is, for example, an eigenvalue of

$$B(X^2 - 2) \oplus B(X^2 - 3) = \begin{pmatrix} 0 & 3 & 2 & 0 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 3 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

17.2 Linear Matrix Equations

Remark 17.14. The following theorem gives a sufficient condition for the solvability of the matrix equation $AX - XB = C$.

Theorem 17.15 (Sylvester equation)**.** *For $A \in K^{n\times n}$ and $B \in K^{m\times m}$ the following are equivalent:*

(1) For every $C \in K^{n\times m}$ there exists (exactly) one $X \in K^{n\times m}$ with $AX - XB = C$.

(2) From $AX = XB$ with $X \in K^{n\times m}$ it follows that $X = 0$.

(3) χ_A and χ_B are coprime.

(4) μ_A and μ_B are coprime.

Proof.

(1) $\Leftrightarrow$ (2): The linear map $K^{n\times m} \to K^{n\times m}$, $X \mapsto AX - XB$ is surjective (bijective) if and only if it is injective.

(2) $\Rightarrow$ (3): Suppose that χ_A and χ_B are not coprime. Then there exists a common divisor $\gamma \in K[X]$ with a root x in a splitting field L of $\chi_A\chi_B$ (Theorem 15.39). Now x is an eigenvalue of A and of B^{t} (as matrices over L). By Theorem 17.11, $x - x = 0$ is an eigenvalue of $P := (-B^{\mathrm{t}}) \oplus A$. In particular, $\det P = 0$. By Lemma 17.7,

$$\operatorname{vec}(AX - XB) = \operatorname{vec}(AX1_m) - \operatorname{vec}(1_nXB) = P \operatorname{vec}(X).$$

Hence there exists an $X \in K^{n\times m} \setminus \{0\}$ with $AX = XB$.

(3) $\Rightarrow$ (4): By the Cayley-Hamilton theorem (Theorem 10.48), $\mu_A \mid \chi_A$ and $\mu_B \mid \chi_B$. Hence every common divisor of μ_A and μ_B is a common divisor of χ_A and χ_B.

(4) $\Rightarrow$ (2): Let $AX = XB$. By induction, $A^kX = A^{k-1}XB = XB^k$ for $k \in \mathbb{N}$ and hence $\alpha(A)X = X\alpha(B)$ for all $\alpha \in K[X]$. With $\alpha = \mu_B$ we obtain $\mu_B(A)X = X\mu_B(B) = 0$. By Bézout (Lemma 15.6) there exist $\tilde{\alpha}, \tilde{\beta} \in K[X]$ with $\tilde{\alpha}\mu_A + \tilde{\beta}\mu_B = 1$. Since $\mu_A(A) = 0$, we have $\tilde{\beta}(A)\mu_B(A) = 1_n$ and $X = \tilde{\beta}(A)\mu_B(A)X = 0$ (cf. Exercise II.34). □

Example 17.16. Let $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 2 & 0 \\ 1 & 2 \end{smallmatrix}\right)$. Then $\chi_A = (X-1)^2$ and $\chi_B = (X-2)^2$ are coprime. By Theorem 17.15, $AX - XB = 1_2$ therefore has exactly one solution. The coefficient matrix is

$$P = (-B^{\mathrm{t}}) \oplus A = \begin{pmatrix} -1 & 1 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

By back substitution one obtains from $P\operatorname{vec}(X) = (1,0,0,1)^{\mathrm{t}}$ the solution $\operatorname{vec}(X) = (1,1,-1,-1)^{\mathrm{t}}$, hence

$$X = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

Remark 17.17. If χ_A and χ_B are not coprime, then the solvability of $AX - XB = C$ depends on C as follows.

Theorem 17.18 (Roth)**.** *Let $A \in K^{n\times n}$, $B \in K^{m\times m}$ and $C \in K^{n\times m}$. There exists an $X \in K^{n\times m}$ with $AX - XB = C$ if and only if $\left(\begin{smallmatrix} A & C \\ 0 & B \end{smallmatrix}\right) \approx \operatorname{diag}(A,B)$.*

Proof (Flanders-Wimmer). Let $k := n+m$, $D := \operatorname{diag}(A,B)$ and $M := \left(\begin{smallmatrix} A & C \\ 0 & B \end{smallmatrix}\right)$ in $K^{k\times k}$. For $X \in K^{n\times m}$ with $AX - XB = C$, the matrix $S := \left(\begin{smallmatrix} 1_n & X \\ 0 & 1_m \end{smallmatrix}\right)$ is invertible with $S^{-1} = \left(\begin{smallmatrix} 1_n & -X \\ 0 & 1_m \end{smallmatrix}\right)$. We have

$$S^{-1}DS = \begin{pmatrix} A & -XB \\ 0 & B \end{pmatrix}\begin{pmatrix} 1_n & X \\ 0 & 1_m \end{pmatrix} = \begin{pmatrix} A & AX - XB \\ 0 & B \end{pmatrix} = M.$$

Conversely, let $S \in \mathrm{GL}(k,K)$ with $M = S^{-1}DS$. We consider the subspaces

$$U := \{Y \in K^{k\times m} : MY = YB\},$$
$$V := \{Y \in K^{k\times m} : DY = YB\}.$$

Because $MY = YB \iff D(SY) = (SY)B$, the map $U \to V$, $Y \mapsto SY$ is an isomorphism and $\dim U = \dim V$. For $Y = \left(\begin{smallmatrix} P \\ Q \end{smallmatrix}\right)$ with $P \in K^{n\times m}$ and $Q \in K^{m\times m}$ we have

$$MY - YB = \begin{pmatrix} AP + CQ - PB \\ BQ - QB \end{pmatrix}, \quad DY - YB = \begin{pmatrix} AP - PB \\ BQ - QB \end{pmatrix}. \tag{17.1}$$

For $\pi\colon K^{k\times m} \to K^{m\times m}$, $Y \mapsto Q$ it follows that $\pi(U), \pi(V) \subseteq \mathrm{C}(B)$. Conversely, for $Q \in \mathrm{C}(B)$ we have $\left(\begin{smallmatrix} 0 \\ Q \end{smallmatrix}\right) \in V$, i. e., $\pi(V) = \mathrm{C}(B)$. In the case $Q = 0$ the right-hand sides in (17.1) are identical, i. e., $U \cap \operatorname{Ker}\pi = V \cap \operatorname{Ker}\pi$. The homomorphism theorem for $\pi_{|U}$ and $\pi_{|V}$ yields

$$\begin{aligned} \dim \pi(U) &= \dim(U) - \dim(U \cap \operatorname{Ker}\pi) \\ &= \dim(V) - \dim(V \cap \operatorname{Ker}\pi) = \dim \mathrm{C}(B). \end{aligned}$$

From $\pi(U) \subseteq \mathrm{C}(B)$ it therefore follows that $\pi(U) = \mathrm{C}(B) \ni -1_m$. Hence there exists an $X \in K^{n\times m}$ with $\left(\begin{smallmatrix} X \\ -1_m \end{smallmatrix}\right) \in U$. Now $AX - XB = C$ holds by (17.1). □

Example 17.19. Let $A = B = \operatorname{diag}(1,2)$. For $X = (x_{ij}) \in K^{2\times 2}$ we have $AX - XA = \left(\begin{smallmatrix} 0 & -x_{12} \\ x_{21} & 0 \end{smallmatrix}\right)$. Hence $AX - XA = C$ is solvable if and only if $c_{11} = c_{22} = 0$.

Remark 17.20. If one replaces similarity in Theorem 17.18 by equivalence (Remark 6.20) of matrices, one obtains a characterization of the solvability of the equation $AX - YB = C$ with two unknowns.

Theorem 17.21 (Roth)**.** *Let $A \in K^{n\times s}$, $B \in K^{t\times m}$ and $C \in K^{n\times m}$. There exist $X \in K^{s\times m}$ and $Y \in K^{n\times t}$ with $AX - YB = C$ if and only if*

$$\operatorname{rk}\begin{pmatrix} A & C \\ 0 & B \end{pmatrix} = \operatorname{rk}(A) + \operatorname{rk}(B).$$

Proof. Let $S_1 \in \mathrm{GL}(n, K)$, $T_1 \in \mathrm{GL}(s, K)$, $S_2 \in \mathrm{GL}(t, K)$ and $T_2 \in \mathrm{GL}(m, K)$. Because of

$$S_1(AX - YB)T_2 = (S_1AT_1)(T_1^{-1}XT_2) - (S_1YS_2^{-1})(S_2BT_2)$$

the equation $AX - YB = C$ is solvable if and only if $A'X' - Y'B' = C'$ is solvable, where $A' := S_1AT_1$, $B' := S_2BT_2$ and $C' := S_1CT_2$. Moreover,

$$\begin{pmatrix} A' & C' \\ 0 & B' \end{pmatrix} = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} A & C \\ 0 & B \end{pmatrix} \begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix}$$

and $\operatorname{rk}(A') = \operatorname{rk} A$, $\operatorname{rk}(B') = \operatorname{rk} B$. By Remark 6.20 we may therefore assume $A = \left(\begin{smallmatrix} 1_a & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 1_b & 0 \\ 0 & 0 \end{smallmatrix}\right)$ with $a = \operatorname{rk} A$ and $b = \operatorname{rk} B$. We partition $C = \left(\begin{smallmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{smallmatrix}\right)$ with $C_{11} \in K^{a\times b}$. Then $AX - YB = C$ is solvable if and only if $C_{22} = 0$. On the other hand,

$$\begin{pmatrix} A & C \\ 0 & B \end{pmatrix} = \begin{pmatrix} 1_a & 0 & C_{11} & C_{12} \\ 0 & 0 & C_{21} & C_{22} \\ 0 & 0 & 1_b & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1_a & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{22} \\ 0 & 0 & 1_b & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix on the right-hand side has rank $a + b + \operatorname{rk}(C_{22}) \geq a + b$, with equality if and only if $C_{22} = 0$. □

17.3 The Hadamard Product

Definition 17.22. For $A, B \in K^{n\times m}$ one calls $A \odot B := (a_{ij}b_{ij}) \in K^{n\times m}$ the *Hadamard product* (or *Schur product*) of A and B.

Remark 17.23.

(a) The Hadamard product corresponds to componentwise multiplication in K^{nm}. Hence the usual computation rules hold:

$$A \odot B = B \odot A, \qquad (A \odot B) \odot C = A \odot (B \odot C)$$
$$(A \odot B)^{\mathrm{t}} = (A^{\mathrm{t}}) \odot (B^{\mathrm{t}}), \qquad (A + B) \odot C = (A \odot C) + (B \odot C).$$

(b) Clearly, $A \odot B$ is the submatrix of $A \otimes B$ with the rows (i, i) and the columns (j, j). In particular, $\operatorname{rk}(A \odot B) \leq \operatorname{rk}(A \otimes B) = \operatorname{rk}(A)\operatorname{rk}(B)$ by Lemma 17.4.

Theorem 17.24 (Schur)**.** *Let $A, B \in \mathbb{C}^{n\times n}$ be positive semidefinite. Then $A \odot B$ is also positive semidefinite. If A is positive definite and $b_{11}, \ldots, b_{nn} > 0$, then $A \odot B$ is positive definite.*[1]

[1] This holds in particular if A and B are positive definite.

Proof. By Exercise II.23, A and B are Hermitian and have Hermitian square roots $\sqrt{A}$ and $\sqrt{B}$. By Lemma 17.4, $R := \sqrt{A} \otimes \sqrt{B}$ is Hermitian with $R^2 = A \otimes B$. For $v \in \mathbb{C}^{n^2}$ we therefore have $v(A \otimes B)v^* = (vR)(vR)^* = |vR|^2 \geq 0$. Let $w \in \mathbb{C}^n$ and $v := \sum_{i=1}^n w_i(e_i \otimes e_i) \in \mathbb{C}^{n^2}$. By Lemma 17.4, $(e_i \otimes e_i)(A \otimes B)(e_j \otimes e_j)^* = (e_i A e_j^{\mathrm{t}}) \otimes (e_i B e_j^{\mathrm{t}}) = a_{ij}b_{ij}$ for $i, j = 1, \ldots, n$. Hence

$$w(A \odot B)w^* = \sum_{i,j=1}^n w_i a_{ij} b_{ij} \overline{w_j} = v(A \otimes B)v^* \geq 0.$$

Now let A be positive definite and $b_{11}, \ldots, b_{nn} > 0$. For $\sqrt{B} = (r_{ij}) = (\overline{r_{ji}})$ we have $b_{ij} = \sum_{k=1}^n r_{ik}\overline{r_{jk}}$. Let $w \in \mathbb{C}^n \setminus \{0\}$, say $w_l \neq 0$. For $k = 1, \ldots, n$ let $v_k := (w_1 r_{1k}, \ldots, w_n r_{nk}) \in \mathbb{C}^n$. Since $\sum_k |r_{lk}|^2 = b_{ll} > 0$, we have $v_k \neq 0$ for some k. It follows that

$$w(A \odot B)w^* = \sum_{i,j=1}^n w_i a_{ij} b_{ij} \overline{w_j} = \sum_{k=1}^n \sum_{i,j=1}^n w_i r_{ik} a_{ij} r_{kj} \overline{w_j} = \sum_{k=1}^n v_k A v_k^* > 0. \qquad \square$$

Theorem 17.25 (Oppenheim inequality)**.** *For positive semidefinite matrices $A, B \in \mathbb{C}^{n\times n}$ we have*

$$\det(A \odot B) \geq \det(A) b_{11} \ldots b_{nn}.$$

Proof. If A is not positive definite, then 0 is an eigenvalue of A by Exercise II.23 and $\det A = 0$. In this case the claim follows from Theorem 17.24. So let A be positive definite. We argue by induction on n. In the case $n = 1$ equality holds. Now let $n \geq 2$.

Because $(0, u)A(0, u)^* = uA_{11}u^*$ for $u \in \mathbb{C}^{n-1}$, A_{11} is positive definite. In particular, $\det(A_{11}) > 0$. Let $c := \det(A)/\det(A_{11})$ and $A' := A - cE_{11}$. Expansion along the first row yields

$$\det(A') = \det(A) - c \det(A_{11}) = 0.$$

Since A is Hermitian, we have $A' = \left(\begin{smallmatrix} x & u \\ u^* & A_{11} \end{smallmatrix}\right)$ with $x \in \mathbb{R}$ and $u \in \mathbb{C}^{n-1}$. For $U := \left(\begin{smallmatrix} 1 & -uA_{11}^{-1} \\ 0 & 1_{n-1} \end{smallmatrix}\right)$ we have

$$\begin{aligned} UA'U^* &= \begin{pmatrix} x - uA_{11}^{-1}u^* & 0 \\ u^* & A_{11} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -A_{11}^{-1}u^* & 1_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} x - uA_{11}^{-1}u^* & 0 \\ 0 & A_{11} \end{pmatrix}. \end{aligned}$$

From $0 = \det(A') = (x - uA_{11}^{-1}u^*)\det(A_{11})$ it follows that $x = uA_{11}^{-1}u^*$. Because $UA'U^* = \operatorname{diag}(0, A_{11})$, A' is also positive semidefinite. By Theorem 17.24, $A' \odot B = A \odot B - cb_{11}E_{11}$ is positive semidefinite. Expansion along the first row and $(A \odot B)_{11} = A_{11} \odot B_{11}$ yield

$$0 \leq \det(A \odot B - cb_{11}E_{11}) = \det(A \odot B) - cb_{11} \det(A_{11} \odot B_{11}).$$

Clearly, B_{11} is positive semidefinite. By induction, $\det(A_{11} \odot B_{11}) \geq \det(A_{11})b_{22} \ldots b_{nn}$. Since $cb_{11} \geq 0$, it follows that

$$\det(A \odot B) \geq cb_{11} \det(A_{11})b_{22} \ldots b_{nn} = \det(A)b_{11} \ldots b_{nn}. \qquad \square$$

Corollary 17.26. *For positive semidefinite matrices $A, B \in \mathbb{C}^{n\times n}$ we have* $\det(A \odot B) \geq \det(A)\det(B)$.

Proof. By Exercise II.23 and Remark 10.35, $\det A \geq 0$. By Theorem 17.25 (applied to B and 1_n) we have $b_{11} \ldots b_{nn} = \det(B \odot 1_n) \geq \det B$. With Theorem 17.25 it follows that

$$\det(A \odot B) \geq \det(A)b_{11} \ldots b_{nn} \geq \det(A)\det(B). \qquad \square$$

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 $1 + 1 \neq 0$ in K. Show $\chi_A = X^3 - \operatorname{tr}(A)X^2 + \frac{1}{2}\big(\operatorname{tr}(A)^2 - \operatorname{tr}(A^2)\big)X - \det(A)$ for all $A \in K^{3\times 3}$.

Exercise II.3. Let $A \in K^{n\times m}$ and $B \in K^{m\times n}$. Prove SYLVESTER's *determinant theorem*

$$\det(1_n + AB) = \det(1_m + BA).$$

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 computed from χ_A and μ_A if A is invertible?

Exercise II.5. Let $n \in \mathbb{N}$. Show:

(a) $A \in \mathbb{F}_2^{n\times n}$ is diagonalizable if and only if $A^2 = A$.
Hint: Theorem 10.52.

(b) $A \in \mathrm{GL}(n, \mathbb{F}_2)$ is diagonalizable if and only if $A = 1_n$.

(c) $|\mathrm{GL}(n, \mathbb{F}_2)| = (2^n - 1)(2^n - 2)\dots(2^n - 2^{n-1})$.
Hint: Theorem 7.46.

Exercise II.6. Show that for $f \in (K^{n\times n})^*$ the following statements are equivalent:

(1) $f(E_{11}) = 1$ and $f(AB) = f(BA)$ for all $A, B \in K^{n\times n}$.

(2) $f(E_{11}) = 1$ and $f(S^{-1}AS) = f(A)$ for all $A \in K^{n\times n}$ and $S \in \mathrm{GL}(n, K)$.

(3) $f(P) = \operatorname{rk} P$ for all $P \in K^{n\times n}$ with $P^2 = P$.

(4) $f = \operatorname{tr}$.

Hint: $E_{ij}E_{kl} = \delta_{jk}E_{il}$.

Exercise II.7. 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 may use Lemma 7.44.

Exercise II.8. 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.9. 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.10. For $\zeta := \cos(\pi/5) + \mathrm{i}\sin(\pi/5) \in \mathbb{C}$ we have $\zeta^5 = -1$ (see the 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: $X^5 + 1 = (X+1)(X^4 - X^3 + X^2 - X + 1)$.

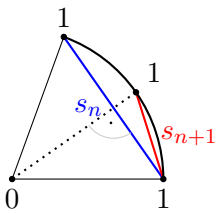
(b) $\omega = \frac{1}{2}(\sqrt{5}+1)$.

(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.11. We approximate the semicircle arc length π by half the perimeter 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[2]
$$s_{n+1} = \sqrt{2 - \sqrt{4 - s_n^2}} = \frac{s_n}{\sqrt{2 + \sqrt{4 - s_n^2}}}$$
by applying the Pythagorean theorem twice:



(c) Justify $\lim_{n\to\infty} 2^n s_{n+1} = \pi$. Compute $2^{10} s_{11}$.

[2]The first formula is numerically *unstable*, see Remark 18.29.

Exercise II.12. Show for all $v, w \in \mathbb{R}^3 \setminus \{0\}$:

(a) $|v|^2|w|^2 = |v \times w|^2 + |[v, w]|^2$.
Hint: One may use the Cauchy–Binet formula.

(b) For the angle φ between v and w we have $\sin\varphi = \frac{|v \times w|}{|v||w|}$.

(c) The right-hand rule holds (Example 11.16).
Hint: After a rotation we may assume that v lies on the x-axis and w in the xy-plane.

Exercise II.13. Let $V := \mathbb{R}^n$ be the Euclidean space with respect to the standard inner product. Let $v \in V$ be normalized and $S_v \in \mathrm{O}(V)$ the reflection in $v^\perp$. Show $[S_v] = 1_n - 2v^{\mathrm{t}}v$.
Remark: Such matrices are called *Householder transformations*.

Exercise II.14. Let $\alpha \in \mathbb{R}[X]$. Show that there exist polynomials $\gamma_1, \ldots, \gamma_k \in \mathbb{R}[X]$ with $\alpha = \gamma_1 \ldots \gamma_k$ and $\deg \gamma_i \leq 2$ for $i = 1, \ldots, k$.
Hint: Remark 11.37.

Exercise II.15. Let V be a d-dimensional K-vector space. A map $\beta\colon V^n \to K$ is called a *multilinear form* if it is linear in each component, i. e.,

$$\beta(v_1, \ldots, v_{i-1}, v_i + \lambda w, v_{i+1}, \ldots, v_n) = \beta(v_1, \ldots, v_n) + \lambda\beta(v_1, \ldots, w, \ldots, v_n)$$

for all $1 \leq i \leq n$, $\lambda \in K$ and $v_1, \ldots, v_n, w \in V$. Show:

(a) The multilinear forms form a subspace of $\mathrm{Map}(V^n, K)$ of dimension d^n.

(b) For every multilinear form β the following statements are equivalent:

(1) For $v_1, \ldots, v_n \in V$ we have $\beta(v_1, \ldots, v_n) = 0$ if there exists a $1 \leq i < n$ with $v_i = v_{i+1}$.

(2) For $v_1, \ldots, v_n \in V$ we have $\beta(v_1, \ldots, v_n) = 0$ if there exist $1 \leq i < j \leq n$ with $v_i = v_j$.

(3) For all linearly dependent $v_1, \ldots, v_n \in V$ we have $\beta(v_1, \ldots, v_n) = 0$.

If this holds, we call β *alternating*.

(c) For $f_1, \ldots, f_n \in V^*$, $\beta(v_1, \ldots, v_n) := \det\big((f_i(v_j))_{ij}\big)$ defines an alternating multilinear form.

(d) The alternating multilinear forms form a vector space of dimension $\binom{d}{n}$ (where $\binom{d}{n} = 0$ for $n > d$).

(e) For $V = K^n$ the determinant is the only alternating multilinear form β with $\beta(e_1, \ldots, e_n) = 1$.

Exercise II.16. 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 (or s) is the maximal dimension of a subspace $U \leq V$ such that the restriction of β to $U \times U$ is positive (or negative) definite.

Exercise II.17. Let $A \in \mathbb{R}^{n\times n}$ be 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.18. Let $A \in \mathbb{R}^{n\times n}$ be symmetric. How can one read off from χ_A whether A is positive semidefinite? Prove a criterion analogous to Theorem 12.44.

Exercise II.19. Let V be a unitary space and $f \in \mathrm{End}(V)$ with adjoint map f^*. Let $g \in \mathrm{End}(V^*)$ be the dual map of f (not denoted by f^* here). Show:

(a) The map $F\colon V \to V^*$, $v \mapsto F_v$ with $F_v(w) := [w, v]$ is a bijection (but not an isomorphism[3])

(b) We have $g = F \circ f^* \circ F^{-1}$.

Exercise II.20. Show for all $A \in \mathbb{C}^{n\times m}$:

(a) $\operatorname{Ker} A = \operatorname{Ker}(A^*A)$.

(b) $\operatorname{rk} A = \operatorname{rk}(A^*A)$.

Exercise II.21. Show that in the principal axis theorem and in the spectral theorem (for matrices) the transformation matrix S can be chosen with $\det S = 1$.

Exercise II.22 (Fuglede)**.** Let $A, B \in \mathbb{C}^{n\times n}$ commute. Show:

(a) If A is normal, then $A^*B = BA^*$.

(b) If A and B are normal, then so is AB.

Exercise II.23. A matrix $A \in \mathbb{C}^{n\times n}$ is called *positive (semi)definite* if $vAv^* > 0$ (or $vAv^* \geq 0$) for all $v \in \mathbb{C}^n \setminus \{0\}$. Show:

(a) Every positive (semi)definite matrix is Hermitian.
Remark: Real matrices with $vAv^{\mathrm{t}} \geq 0$ for all $v \in \mathbb{R}^n$ are not automatically symmetric.

(b) A Hermitian matrix A is positive (semi)definite if and only if all eigenvalues of A are positive (or nonnegative).

(c) For $k \in \mathbb{N}$ every positive (semi)definite matrix has exactly one positive (semi)definite k-th root.

Exercise II.24 (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 (nonreal) eigenvalues. In particular, $Q^{\mathrm{t}}AQ$ is an upper triangular matrix if A has only real eigenvalues.

Exercise II.25. 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 nontrivial! One may use Exercise I.16.

[3] This is called an *antilinear* map.

Exercise II.26. Let $A_1, \ldots, A_k \in K^{n\times n}$ be pairwise commuting triangularizable matrices. Show that $A_1, \ldots, A_k$ are simultaneously triangularizable.

Exercise II.27. Let $\lambda \in K$ and $A := J_n(\lambda)$ a Jordan block. Show that $C \in K^{n\times n}$ commutes with A if and only if

$$C = \begin{pmatrix} c_1 & & & 0 \\ c_2 & \ddots & & \\ \vdots & \ddots & \ddots & \\ c_n & \cdots & c_2 & c_1 \end{pmatrix}$$

for some $c_1, \ldots, c_n \in K$.

Exercise II.28. Let V be a K-vector space and $f \in \operatorname{End}(V)$. 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.29. Let V be a K-vector space and $f \in \operatorname{End}(V)$. Let $\lambda \in K$ be an eigenvalue of f that occurs with multiplicity k as a root of μ_f. Show

$$E_\lambda(f) \subsetneq \operatorname{Ker}\big((f - \lambda \operatorname{id})^2\big) \subsetneq \ldots \subsetneq \operatorname{Ker}\big((f - \lambda \operatorname{id})^k\big) = H_\lambda(f).$$

Hint: One may use the Weierstrass normal form.

Exercise II.30. Let $\alpha \in K[X] \setminus K$ be monic. Show $\chi_{B(\alpha)} = \alpha$ using the definition of the characteristic polynomial (and not via the minimal polynomial as in Lemma 15.20).

Exercise II.31. 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 $VB(\alpha)V^{-1} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Exercise II.32. Let $\alpha = X^n + a_{n-1}X^{n-1} + \ldots + a_0$ and

$$S = \begin{pmatrix} a_1 & a_2 & \cdots & a_{n-1} & 1 \\ a_2 & a_3 & \iddots & 1 & \\ \vdots & \iddots & \iddots & & \\ a_{n-1} & 1 & & & \\ 1 & & & & 0 \end{pmatrix} \in \operatorname{GL}(n, K)$$

(a) Show $S^{-1}B(\alpha)S = B(\alpha)^{\mathrm{t}}$.

(b) (Taussky) Conclude that every square matrix is the product of two symmetric matrices.

Exercise II.33. Prove Corollary 14.37 over an arbitrary field K instead of $\mathbb{C}$.

Exercise II.34.

(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

$$\operatorname{C}(\operatorname{diag}(A, B)) = \big\{\operatorname{diag}(C, D) : C \in \operatorname{C}(A),\ D \in \operatorname{C}(B)\big\} \cong \operatorname{C}(A) \times \operatorname{C}(B).$$

Hint: Bézout's lemma.

(b) Let A be semisimple with $\chi_A = \gamma_1^{a_1} \dots \gamma_k^{a_k}$ (prime factorization). Show

$$\dim \mathrm{C}(A) = \sum_{i=1}^{k} a_i^2 \deg(\gamma_i)$$

using (a). Compare with Remark 15.33.

Exercise II.35. Let V be a K-vector space and $F \subseteq \mathrm{End}(V)$. A subspace $U \leq V$ is called *F-invariant*, if $f(U) \subseteq U$ for all $f \in F$. Let

$$\mathrm{C}(F) := \bigcap_{f \in F} \mathrm{C}(f)$$

be the *centralizer* of F (similarly for matrices).

(a) Suppose that $\{0\}$ 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 nonzero element is invertible.
Remark: In contrast to Theorem 15.37, the multiplication in $\mathbb{H}$ is not commutative. We call $\mathbb{H}$ Hamilton's *quaternions*.

Exercise II.36. Let $K := \mathbb{F}_2(X)$ and $\gamma := Y^2 + X \in K[Y]$ as in Example 16.9. Show:

(a) $B(\gamma^2) \in K^{4\times 4}$ has no Jordan-Chevalley decomposition.

(b) $\mathrm{diag}(B(\gamma), B(\gamma))$ has several Jordan-Chevalley decompositions.

Exercise II.37. Complete the proof from Example 17.13 that $\overline{\mathbb{Q}}$ is a field.

Exercise II.38.

(a) Show for all $T \in K^{2\times 2\times 2}$ (Remark 17.3):

(1) There exist $A, B \in K^{2\times 2}$ with $T = A \otimes e_1 + B \otimes e_2$.

(2) There exists $C \in K^{2\times 2} \setminus \{0\}$ with $\mathrm{tr}(CA) = \mathrm{tr}(CB) = 0$.

(3) There exist $N_1, N_2, N_3 \in K^{2\times 2}$ of rank 1 with $A, B \in \langle N_1, N_2, N_3 \rangle$.
Hint: One may assume $C \in \{E_{11}, 1_2\}$.

(4) $\mathrm{rk}\, T \leq 3$.

(b) Let $T := 1_2 \otimes e_1 + B(X^2+1) \otimes e_2 \in \mathbb{R}^{2\times 2\times 2}$. Show $\mathrm{rk}\, T = 3$ and $\mathrm{rk}_{\mathbb{C}}\, T = 2$, where $\mathrm{rk}_{\mathbb{C}}$ is the rank of $T \in \mathbb{C}^{2\times 2\times 2}$.

Linear Algebra III

18 Numerical Methods

18.1 Efficient Arithmetic

Remark 18.1. In linear algebra I and II we mainly studied facts from a theoretical point of view. For example, eigenvalues were computed 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 (with respect to rounding errors) in practice. Eigenvalues are computed more or less via the definition (Theorem 18.72).

Example 18.2. On computers, the multiplication of two numbers is usually more expensive than addition (there are exceptions, however: multiplication by 2 corresponds to a shift of the binary sequence by one digit to the left.). In the following running time analyses we will therefore neglect additions. The multiplication of two n-digit decimal numbers with the "schoolbook method" requires n^2 digit multiplications (the multiplication table can be stored permanently 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 18.3 (Karatsuba algorithm)**.** *Let $x, y \in \mathbb{N}$ be decimal numbers with n digits.*

(1) Split 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 $xy = 10^{2m} z_2 + 10^m z_1 + z_0$.

This algorithm requires approximately $n^{\log_2(3)} \approx n^{1.58} < n^2$ digit multiplications.

Proof. We have $z_1 = x_1 y_0 + x_0 y_1 - x_1 y_1 - x_0 y_0 + z_0 + z_2 = x_1 y_0 + x_0 y_1$ 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 claim we argue by induction on n: For $n = 1$ one needs $1 = 1^{\log_2(3)}$ digit multiplication. The computation 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 18.4. 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 18.5. Before the introduction of electronic pocket 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 18.6. 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 map $\mathcal{F}_n \colon \mathbb{C}^n \to \mathbb{C}^n$, $x \mapsto xW =: \hat{x}$ is called the *discrete Fourier transform*.

Example 18.7. By Example 11.28 we have

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{pmatrix}.$$

Lemma 18.8. *For $x \in \mathbb{C}^n$ we have $\mathcal{F}_n^{-1}(x) = \frac{1}{n}\overline{\mathcal{F}_n(\overline{x})}$.*

Proof. Since $\zeta\overline{\zeta} = |\zeta|^2 = 1$, we have $\zeta^{-1} = \overline{\zeta}$. For $1 \le i, j \le n$, also $\sigma := \zeta^{i-j}$ is an n-th root of unity. In the case $i = j$ we have $\sigma = 1$, and otherwise $\sum_{k=0}^{n-1} \sigma^k = \frac{\sigma^n - 1}{\sigma - 1} = 0$ by 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 18.9 (Convolution theorem)**.** *For $\alpha = \sum a_k X^{k-1} \in \mathbb{C}[X]$ let $[\alpha] := (a_1, \dots, a_n)$. For all $\alpha, \beta \in \mathbb{C}[X]$ with $\deg(\alpha) + \deg(\beta) < n$ we have*

$$\mathcal{F}_n([\alpha\beta]) = (\hat{a}_1\hat{b}_1, \dots, \hat{a}_n\hat{b}_n).$$

Proof. Because of

$$\alpha\beta = \sum_{k=2}^{n+1}\Big(\sum_{i+j=k} a_i b_j\Big)X^{k-2} = \sum_{k=1}^{n}\Big(\sum_{i+j=k+1} a_i b_j\Big)X^{k-1}$$

we have $[\alpha\beta]_k = \sum_{i+j=k+1} a_i b_j$ 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 18.10.

(a) To multiply two numbers $a, b \in \mathbb{N}$ (with $ab < 10^n$), we regard them as polynomials in $X = 10$, $a = \sum a_k 10^{k-1}$, $b = \sum b_k 10^{k-1}$, and apply the convolution theorem:

$$ab = \mathcal{F}_n^{-1}(\hat{a}_1\hat{b}_1, \dots, \hat{a}_n\hat{b}_n) \cdot (1, 10, \dots, 10^{n-1})^{\mathrm{t}}.$$

Note that carries, as in schoolbook multiplication, are not resolved here. Carries can generally be avoided by representing a and b in *binary*, i. e., as a polynomial in $X = 2$. However, this way of computing does not yet yield any speedup.

(b) Let $n = 2m$ be even. With the notation $y_k := x_{2k-1}$ and $z_k := x_{2k}$ we have

$$\begin{aligned} \mathcal{F}_n(x)_i &= \sum_{k=1}^{n} x_k \zeta_n^{(i-1)(k-1)} = \sum_{k=1}^{m} y_k \zeta_m^{(i-1)(k-1)} + \zeta_n^{i-1} \sum_{k=1}^{m} z_k \zeta_m^{(i-1)(k-1)} \\ &= \mathcal{F}_m(y)_i + \zeta_n^{i-1} \mathcal{F}_m(z)_i \end{aligned} \tag{18.1}$$

for $i = 1, \dots, n$, where $\mathcal{F}_m(y)_i = \mathcal{F}_m(y)_{i-m}$ for $i > m$. With Lemma 18.8, 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 can always be achieved by appending zeros). As with the Karatsuba algorithm, we can reduce $\mathcal{F}_n$ to $\mathcal{F}_1(x) = x$ by repeated application of (18.1). For a fixed n, the n-th roots of unity can be computed once and stored. The addition of such numbers can be carried out cheaply with limited precision.

(c) The widely used Schönhage-Strassen algorithm performs the multiplication of n-digit numbers by means of FFT in an asymptotic running time of $n \log(n) \log\log(n)$. In 2019, Harvey-van der Hoeven improved the running time 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 the integral

$$\mathcal{F}(y) = \frac{1}{\sqrt{2\pi}^n} \int_{\mathbb{R}^n} f(x) e^{2\pi \mathrm{i}[x,y]} \mathrm{d}x$$

to an integrable function $f \colon \mathbb{R}^n \to \mathbb{R}$.

Example 18.11. We multiply $342 \cdot 87$ with the fast Fourier transform. By (18.1) we have

$$\mathcal{F}_4(2,4,3,0)^{\mathrm{t}} = \begin{pmatrix} \mathcal{F}_2(2,3)_1 + \mathcal{F}_2(4,0)_1 \\ \mathcal{F}_2(2,3)_2 + \mathrm{i}\mathcal{F}_2(4,0)_2 \\ \mathcal{F}_2(2,3)_1 - \mathcal{F}_2(4,0)_1 \\ \mathcal{F}_2(2,3)_2 - \mathrm{i}\mathcal{F}_2(4,0)_2 \end{pmatrix} = \begin{pmatrix} (2+3)+4 \\ (2-3)+4\mathrm{i} \\ (2+3)-4 \\ (2-4)-4\mathrm{i} \end{pmatrix} = \begin{pmatrix} 9 \\ -1+4\mathrm{i} \\ 1 \\ -1-4\mathrm{i} \end{pmatrix},$$

$$\mathcal{F}_4(7,8,0,0) = (15, 7+8\mathrm{i}, -1, 7-8\mathrm{i}),$$

$$\mathcal{F}_4^{-1}(135, -39+20\mathrm{i}, -1, -39-20\mathrm{i})^{\mathrm{t}} = \frac{1}{4}\begin{pmatrix} 135-1-2\cdot 39 \\ 135+1-\mathrm{i}40\mathrm{i} \\ 135-1+2\cdot 39 \\ 135+1+\mathrm{i}40\mathrm{i} \end{pmatrix} = \big(14, 44, 53, 24\big),$$

$$342 \cdot 87 = \mathcal{F}_4^{-1}(\dots)(1, 10, 100, 1000)^{\mathrm{t}} = 14 + 440 + 5{,}300 + 24{,}000 = 29{,}754.$$

Remark 18.12. The direct multiplication of two $n \times n$ matrices A, B requires n^3 multiplications of numbers $a_{ij}b_{jk}$ with $1 \leq i, j, k \leq n$, which in turn can be carried out with one of the above algorithms. This can also be done better.

Theorem 18.13 (STRASSEN algorithm)**.** *Let $A, B \in K^{n\times n}$.*

(1) By appending a zero row and column we may assume $n = 2m$.

(2) Split 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 $n^{\log_2(7)} \approx n^{2.81}$ multiplications of numbers.

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 computation of each C_i requires inductively approximately $m^{\log_2(7)} = \frac{1}{7}n^{\log_2(7)}$ multiplications of numbers. Therefore one needs approximately $n^{\log_2(7)}$ multiplications of numbers for AB. □

Remark 18.14.

(a) The Strassen algorithm has also been improved further. With the *large language model* AlphaEvolve, an algorithm with $n^{2.371177}$ multiplications of numbers was discovered in 2026.[1] The asymptotically best possible complexity is an open problem in computer science.

(b) Matrix powers can be computed efficiently by iterated squaring: $A^{11} = A(A^5)^2 = A(A(A^2)^2)^2$ needs only five multiplications instead of ten (Exercise III.1).

[1] see [Dupont et al., *Improving the matrix multiplication exponent with modern optimization and AlphaEvolve*, arXiv:2608.16884]

(c) The multiplication of arbitrary (not necessarily square) matrices $A \in K^{n\times m}$ and $B \in K^{m\times k}$ requires nmk multiplications of numbers. If more than two such matrices are multiplied, then the placement of parentheses has a considerable influence on the number of multiplications of numbers. This becomes particularly clear 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 18.15. *Let $A \in K^{n\times m}$, $B \in K^{m\times k}$ and $C \in K^{k\times l}$. The computation of $A(BC)$ requires fewer multiplications of numbers than $(AB)C$ if and only if $\frac{1}{n} + \frac{1}{k} < \frac{1}{m} + \frac{1}{l}$.*

Proof. The computation of $A(BC)$ and $(AB)C$ requires $mkl + nml = ml(n + k)$ and $nmk + nkl = nk(m + l)$ multiplications of numbers, respectively. We have

$$ml(n + k) < nk(m + l) \iff \frac{n + k}{nk} < \frac{m + l}{ml} \iff \frac{1}{n} + \frac{1}{k} < \frac{1}{m} + \frac{1}{l}. \qquad \square$$

Remark 18.16. For a concrete implementation, it is advisable to use proven libraries such as BLAS, LAPACK, Armadillo, NumPy and SciPy, programming languages such as Julia, or programs such as MATLAB, Scilab and Octave.

18.2 The Condition Number

Example 18.17. Even though linear maps are continuous in the sense of analysis, small changes in the arguments can nevertheless have large effects on the values (in the ϵ-δ definition of continuity, one has to choose $\delta \ll \epsilon$):

$$Ax = \begin{pmatrix} 9 & 8 \\ 8 & 7 \end{pmatrix} x = \begin{pmatrix} 17 \\ 15 \end{pmatrix} \Longrightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\begin{pmatrix} 9 & 8 \\ 8 & 7 \end{pmatrix} x = \begin{pmatrix} 17 \\ 15.\mathbf{1} \end{pmatrix} \Longrightarrow x = \begin{pmatrix} 1.8 \\ 0.1 \end{pmatrix}.$$

Definition 18.18. 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 18.19.

(a) We want to measure how small changes in 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

$$\frac{|x - \tilde{x}|}{|x|} = \frac{|x - \tilde{x}|}{|A(x - \tilde{x})|} \frac{|Ax|}{|x|} \frac{|b - \tilde{b}|}{|b|} \leq \frac{\max\{|Ax|/|x| : 0 \neq x \in \mathbb{C}^{m\times 1}\}}{\min\{|Ay|/|y| : 0 \neq y \in \mathbb{C}^{m\times 1}\}} \frac{|b - \tilde{b}|}{|b|} = \kappa(A) \frac{|b - \tilde{b}|}{|b|}.$$

Thus the condition number is the maximal factor by which the relative error can be amplified from b 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 18.20.

(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 $S \in \mathrm{U}(n, \mathbb{C})$ we know 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 $A, S \in \mathbb{C}^{n\times n}$ we have $\kappa(SA) \neq \kappa(A)$ in general. 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 classical 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 reduce the computation of $\kappa(A)$ to a diagonal matrix using the spectral theorem and (b). The following theorem allows this for arbitrary (in particular nonsquare) matrices.

Theorem 18.21 (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 nonnegative entries. The positive entries (on the main diagonal) are called the* singular values *of A. They are uniquely determined up to order.*

Proof. Existence: By Exercise II.23, the positive semidefinite matrix $B := A^*A \in \mathbb{C}^{m\times m}$ has only nonnegative eigenvalues

$$\lambda_1 \geq \ldots \geq \lambda_k > \lambda_{k+1} = \ldots = \lambda_m = 0,$$

[2] not necessarily square

where $k = \operatorname{rk} B = \operatorname{rk} A \leq \min\{n, m\}$ by Exercise II.20. 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

$$U_1 := AV_1 \operatorname{diag}\big(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}\big) \in \mathbb{C}^{n \times k}.$$

Since $U_1^*U_1 = 1_k$, we can extend U_1 to $U = (U_1, U_2) \in \mathrm{U}(n, \mathbb{C})$ with Gram-Schmidt. Now

$$\begin{aligned} U_2^*AV_1 &= U_2^*U_1 \operatorname{diag}\big(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\big) = 0, \\ U^*AV &= \begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix} A \begin{pmatrix} V_1 & V_2 \end{pmatrix} = \begin{pmatrix} U_1^*AV_1 & U_1^*AV_2 \\ U_2^*AV_1 & U_2^*AV_2 \end{pmatrix} = (\delta_{ij}\sqrt{\lambda_i}), \end{aligned}$$

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 nonnegative entries. Let $\mu_1, \ldots, \mu_l$ be its positive entries. Then $\mu_1^2, \ldots, \mu_l^2$ are the positive eigenvalues of

$$(SAT)^2 = (SAT)^*(SAT) = T^*BT.$$

With suitable numbering we therefore have $k = l$ 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 $i = 1, \ldots, k$. Hence the singular values are uniquely determined up to order. □

Remark 18.22. The proof shows: If $\lambda_1, \ldots, \lambda_k$ are the nonzero eigenvalues of the positive semidefinite matrix A^*A, then $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}$ are the singular values of A.

Corollary 18.23. *Let $A \in \mathbb{C}^{n \times m}$ have full rank and singular values $\sigma_1 \geq \ldots \geq \sigma_k$. Then $\kappa(A) = \sigma_1/\sigma_k$.*

Proof. Let $D := (\delta_{ij}\sigma_i) \in \mathbb{R}^{n \times m}$ be the matrix from the singular value decomposition of A. By Example 18.20 we have $\kappa(A) = \kappa(D) = \kappa(\operatorname{diag}(\sigma_1, \ldots, \sigma_k))$. 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 18.24. *For all $A \in \mathbb{C}^{n \times m}$ we have $\kappa(A) = \kappa(A^*)$.*

Proof. By Lemma 10.38, A^*A and AA^* have the same nonzero eigenvalues. Hence the claim follows from Remark 18.22. □

Theorem 18.25. *If $A \in \mathbb{C}^{n \times n}$ is normal, then the singular values of A are the absolute values of the nonzero eigenvalues.*

Proof. By the spectral theorem we may 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 18.26. For the matrix from Example 18.17 we have $\kappa(A) = \frac{8+\sqrt{65}}{8-\sqrt{65}} \approx 258$.

Theorem 18.27 (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

$$\tilde{\sigma}_i := \begin{cases} \sigma_i^{-1} & \text{if } i \le k, \\ 0 & \text{if } i > k \end{cases}$$

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 clearly 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 18.28.

(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 18.27 reduces the computation 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 $SAS = \operatorname{diag}(2, 0)$. Hence $A^+ = S\operatorname{diag}(1/2, 0)S = \frac{1}{4}A$.

(e) Clearly, $(A^{\mathrm{t}})^+ = (A^+)^{\mathrm{t}}$ and $\overline{A}^+ = \overline{A^+}$. In particular, A^+ is symmetric (or real, Hermitian) if A is symmetric (or 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 $A^+ = A^*(AA^*)^{-1}$, respectively. For $n \ge m$ one can compute 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 18.42.

[3] or *Moore-Penrose inverse*

18.3 Stable Variants of the Gaussian Algorithm

Remark 18.29. 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` (single precision) by the following layout: 1 bit for the sign, 23 bits for the *mantissa* (binary expansion) and 8 bits for the exponent. Thus the numbers of the form

$$\pm\Big(2^e + \sum_{i=1}^{23} a_i 2^{e-i}\Big) \qquad (a_i \in \{0,1\},\ -126 \le e \le 127)$$

are representable ($e = -127$ and $e = 128$ are reserved for ∞ and `NaN`). In the following we consider a simpler model with only three decimal digits, that is, 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 rarely occur, we do not bound 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 almost equal numbers leads to a loss of accuracy (*catastrophic cancellation*):[4]

$$1.23\text{e}0 - 1.22\text{e}0 = 1.??\text{e}{-3}.$$

The digits marked with a question mark contain no meaningful information. Moreover, addition of floating-point numbers is not associative:

$$(1.00\text{e}{-3} + 1.00\text{e}1) - 1.00\text{e}1 = 0.00\text{e}0 \neq 1.00\text{e}{-3} = 1.00\text{e}{-3} + (1.00\text{e}1 - 1.00\text{e}1).$$

Multiplication by large numbers is also problematic, because it amplifies existing rounding errors.

Example 18.30. 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 algorithm with floating-point numbers yields

$$\left(\begin{array}{cc|c} 1\text{e}{-4} & 1 & 1\text{e}4 \\ 1 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1\text{e}4 & 1\text{e}8 \\ 1 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1\text{e}4 & 1\text{e}8 \\ 0 & -1\text{e}4 & -1\text{e}8 \end{array}\right) \quad \Longrightarrow \quad x = (0, 1\text{e}4).$$

Thus the pure Gaussian algorithm is numerically *unstable*.

Remark 18.31. One approach to improve the stability of a system of equations is to compute the singular value decomposition of the coefficient matrix A. Here 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.

[4] see Exercise II.11

Theorem 18.32 (Pivoting)**.** *Let a system of linear equations $Ax = b$ with $A \in \mathbb{C}^{n\times m}$ and $b \in \mathbb{C}^{n\times 1}$ be given. The following modification of the Gaussian algorithm 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} of largest absolute value* (pivot) *with $i \geq z$ and $j \geq s$.*
- *If $a_{ij} = 0$, we are done. Otherwise swap the i-th row with the z-th row and the j-th column with the s-th column. Swap x_j with x_s in the solution vector.*
- *Divide the z-th row by a_{zs}.*
- *Eliminate $a_{ws} = 0$ for $w \neq z$ as usual.*
- *Increase z by 1.*

Proof. Applying the stated procedure as in Theorem 6.15, we obtain 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 solution set

$$L = \begin{pmatrix} c_1 \\ \vdots \\ c_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \left\langle \begin{pmatrix} s_{11} \\ \vdots \\ s_{k1} \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, \begin{pmatrix} s_{1,m-k} \\ \vdots \\ s_{k,m-k} \\ 0 \\ \vdots \\ 0 \\ -1 \end{pmatrix} \right\rangle.$$

We obtain the solution of the original system by permuting the coordinates according to the chosen swaps $j \leftrightarrow s$. The choice of the pivot guarantees that the algorithm never multiplies by large numbers, which could amplify possible rounding errors. □

Example 18.33. In Example 18.30 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 18.34. For arbitrary fields, the following theorem says that pivoting theoretically only has to be done once at the beginning.

Theorem 18.35 (LU decomposition)**.** *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 $A = PLU$.*

Proof. Induction on n: For $n = 1$ we 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 afterwards $a_{11} \neq 0$ holds. One achieves the elimination of a_{i1} for $i > 1$ by multiplying 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 we have

$$L_1 P_1 A = \begin{pmatrix} a_{11} & * \\ 0 & A_2 \end{pmatrix}$$

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 shape 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, we have $L_1^{-1} \hat{P}_2 = \hat{P}_2 L_1'$ for a lower triangular matrix L_1' with ones on the main diagonal. Altogether we have

$$A = P_1^{-1} L_1^{-1} \hat{P}_2 \hat{L}_2 \hat{U}_2 = P_1 \hat{P}_2 L_1' \hat{L}_2 \hat{U}_2 = PLU. \qquad \square$$

Example 18.36.

$$\begin{pmatrix} 0 & 0 & 2 \\ 2 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix} = P_{(1,2)} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 2 & 1 & -1 \end{pmatrix} = P_{(1,2)} \begin{pmatrix} 1 & . & . \\ . & 1 & . \\ 1 & . & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 1 & -2 \end{pmatrix}$$
$$= P_{(1,2)} L_1 P_{(2,3)} \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 2 \end{pmatrix} = P_{(1,2)} P_{(2,3)} \begin{pmatrix} 1 & . & . \\ 1 & 1 & . \\ . & . & 1 \end{pmatrix} U = P_{(1,2,3)} LU$$

Remark 18.37.

(a) Since there are usually many ways to pivot 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 yields a variant of the LU decomposition in which P lies 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 $i \geq 2$. Similarly, one obtains $a_{1,j_1} = 1$ and $a_{1j} = 0$ for $j > j_1$ by right multiplication with an upper triangular matrix. Now one chooses the minimal j_2 with $a_{2,j_2} \neq 0$ and proceeds 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 $k = 1, \dots, n$. Hence one has a decomposition of the form $A = LPU$ (cf. Exercise III.10).

(c) The next theorem yields a further approximation of the singular value decomposition.

Theorem 18.38 (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, nonnegative diagonal entries and $A = QR$.*

(a) If A has full rank, then R is uniquely determined.

(b) If A is real, 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 $s := \operatorname{rk} A$. 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 written as a linear combination of the a_{i_k} with $i_k < j$. We extend $a_{i_1}, \dots, a_{i_s}$ by 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 $\mu \in \mathbb{R}_{>0}$ and so on. Rearranging yields

$$a_{i_k} = \sum_{j=1}^{k} \lambda_j q_j \qquad (\lambda_k \in \mathbb{R}_{>0}).$$

By also expressing the remaining columns of A with respect to $q_1, \dots, q_n$, we obtain 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 $Q = (q_1, \dots, q_n) \in \mathrm{U}(n, \mathbb{C})$ we now have $A = QR$.

(a) W. l. o. g. let A itself be an upper triangular matrix with positive diagonal entries. For the first column we have $a_{11} e_1 = r_{11} q_1$. From $a_{11}, r_{11} > 0$ and $|q_1| = 1$ it follows that $q_1 = e_1$ and $a_{11} = r_{11}$. Inductively, suppose that $r_{ij} = a_{ij}$ and $q_j = e_j$ have already been proved for $i = 1, \dots, n$ and $j = 1, \dots, k-1$. For the k-th column of A we then have

$$\sum_{i=1}^{k} a_{ik} e_i = r_{kk} q_k + \sum_{i=1}^{k-1} r_{ik} e_i.$$

This shows $q_{ik} = 0$ for $i > k$. Since $[e_i, q_k] = [q_i, q_k] = 0$ for $i < k$, it follows that $q_k = e_k$ and $r_{ik} = a_{ik}$ for $i = 1, \dots, n$. Inductively we obtain $A = R$ (but not necessarily $Q = 1_n$ if $n > m$).

(b) If A is real, the Gram-Schmidt process can be carried out in $\mathbb{R}^n$ (note that we extended $a_{i_1}, \dots, a_{i_s}$ by real vectors to a basis). □

Remark 18.39.

(a) If $\operatorname{rk} A = n \leq m$, then the proof shows that Q and R are uniquely determined in the QR decomposition.

(b) In the case $n = m = 1$ one obtains the *polar decomposition* 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).

(c) We will see in Section 18.7 that the direct application of the Gram-Schmidt process is unstable.

Example 18.40.

$$\begin{pmatrix} 1 & \mathrm{i} \\ 1 & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \sqrt{2} & \mathrm{i} - 1 \\ \sqrt{2} & 1 - \mathrm{i} \end{pmatrix} \begin{pmatrix} \sqrt{2} & (1+\mathrm{i})/\sqrt{2} \\ 0 & 1 \end{pmatrix}$$

Remark 18.41. In practice, inaccurate measurements often give rise to overdetermined systems of equations $Ax = b$ without an exact solution. 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 18.42 (Method of least squares)**.** *Let $A \in \mathbb{C}^{n \times m}$ have full rank $m \leq n$ and let $b \in \mathbb{C}^{n \times 1}$. Then the system of equations $A^* A x = A^* b$ has a unique solution $\tilde{x}$, and for all $x \neq \tilde{x}$ we have $|A\tilde{x} - b| < |Ax - b|$.*

Proof. Since A has full rank $m \leq n$, the matrix A^*A is positive definite, in particular invertible. This shows existence and uniqueness of $\tilde{x}$. For $y \neq 0$ we have

$$\begin{aligned}[Ay, A\tilde{x} - b] &= y^*A^*(A\tilde{x} - b) = 0 = [A\tilde{x} - b, Ay], \\ |A(\tilde{x} + y) - b|^2 &= [(A\tilde{x} - b) + Ay, (A\tilde{x} - b) + Ay] = |A\tilde{x} - b|^2 + |Ay|^2 > |A\tilde{x} - b|^2.\end{aligned}$$ □

Remark 18.43.

(a) On New Year's night of 1801, the astronomer Piazzi discovered the dwarf planet *Ceres*. Its position was recorded for 40 days, until Ceres disappeared from view. Gauss used the method of least squares to extrapolate the orbit of Ceres from the available measurement data and was thus able to predict its position successfully. This made him internationally famous.

(b) By Example 18.28 we have $\tilde{x} = A^+b$ in the situation of Theorem 18.42.

(c) The method of least squares leads to the so-called *system of normal equations* $A^*Ax = A^*b$ with positive definite coefficient matrix. There are special methods for this. The following decomposition refines Lemma 12.40.

Theorem 18.44 (Cholesky decomposition)**.** *A matrix $A \in \mathbb{C}^{n\times n}$ 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 $A = R^*R$. In this case, R is uniquely determined.*

Proof. For $R \in \mathrm{GL}(n, \mathbb{C})$ the matrix R^*R is well known to be positive definite. Conversely, let A be positive definite and $\sqrt{A} = QR$ the (unique) QR decomposition of the square root of A (Exercise II.23). Then we have

$$A = \sqrt{A}^2 = \sqrt{A}^*\sqrt{A} = R^*Q^*QR = R^*R.$$

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, hence a diagonal matrix with positive main diagonal entries. Since

$$(PR^{-1})^*PR^{-1} = R^{-*}AR^{-1} = 1_n$$

the matrix PR^{-1} is unitary. It follows that $P = R$. □

Remark 18.45.

(a) The matrix $R = (r_{ij})$ can be computed iteratively without the detour via the QR decomposition: Since $a_{11} = e_1Ae_1^{\mathrm{t}} > 0$, 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 $A = \left(\begin{smallmatrix} A_{nn} & a \\ a^* & a_{nn} \end{smallmatrix}\right)$ und $R = \left(\begin{smallmatrix} R_1 & v \\ 0 & r_{nn} \end{smallmatrix}\right)$ we obtain $R_1^*v = a$ and $r_{nn} = \sqrt{a_{nn} - v^*v}$.

(b) The system $Ax = b$ with $A = R^*R$ can conveniently be 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.23

and then $Rx = y$ by *back 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 18.46. 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 reads $\left(\begin{smallmatrix} 2 & 3 \\ 3 & 6 \end{smallmatrix}\right) x = -\left(\begin{smallmatrix} 1 \\ 4 \end{smallmatrix}\right)$. The matrix R is obtained via $r_{11} = \sqrt{2}$, $r_{12} = \frac{3}{\sqrt{2}}$ 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 we obtain the approximation $Ax = \frac{1}{3}(1, -4, 5)^{\mathrm{t}}$. In practice, the square root expressions are approximated by floating-point numbers.

18.4 Iterative Methods

Remark 18.47.

(a) For large matrices, factorization methods such as the QR decomposition are slow and memory-intensive. To solve systems of linear equations in these cases, one can resort to faster iterative methods. For convergence proofs we need some tools from analysis.

(b) So far, norms have always arisen from inner products (Definition 11.2 and Definition 13.2). However, they can also be defined directly via the properties from Lemma 13.5.

Definition 18.48. 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 $\lambda \in \mathbb{C}$ we have:

(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 18.49.

(a) If $[.,.]$ is an inner product on V, then $\|v\| := \sqrt{[v, v]}$ defines a norm by 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 an inner product, since the parallelogram law

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

from Remark 11.3 is violated 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).[6] For $p = 1$ we obtain $\|x\|_1 = |x_1| + \ldots + |x_n|$. For $p = 2$ we obtain 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 18.50. 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 18.51. *Any two norms on a finite-dimensional $\mathbb{C}$-vector space are equivalent.*

Proof. W. l. o. g. let $V = \mathbb{C}^n$. It suffices to show that every norm $\|.\|$ is equivalent to $\|.\|_1$. For $\lambda := \max\{\|e_1\|, \ldots, \|e_n\|\} > 0$ we have

$$\|x\| = \Big\|\sum_{i=1}^n x_i e_i\Big\| \leq \sum_{i=1}^n |x_i|\|e_i\| \leq \lambda\|x\|_1$$

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 we have $\|x\|_1 = [x,y] \leq |x||y| = \sqrt{n}|x|$. 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 with respect to the Euclidean norm. Hence f attains its minimum $\mu > 0$ on the compact set $\{x \in \mathbb{R}^n : \|x\|_1 = 1\} \neq \varnothing$. For $x \neq 0$ we now have

$$\|x\| = \|x\|_1 f(\|x\|_1^{-1}x) \geq \|x\|_1\mu. \qquad \square$$

Remark 18.52. As is well known, the Euclidean norm on $\mathbb{C}^n$ is complete (i. e., every Cauchy sequence converges). By Lemma 18.51, every norm on a finite-dimensional $\mathbb{C}$-vector space V is therefore complete, i. e., V is a *Banach space*.

Theorem 18.53 (Banach's fixed-point theorem)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be a* contraction *with respect to a norm $\|.\|$, i. e., there is a constant $c < 1$ with $\|f(x) - f(y)\| \leq c\|x-y\|$ for all $x, y \in \mathbb{R}^n$. Then for all $x_0 \in \mathbb{R}^n$ the sequence $x_{k+1} := f(x_k)$ converges to the unique fixed point of f.*

[6] In analysis, one defines $x^p := \exp(p \ln x)$ for $x > 0$.

Proof. For $k, l \in \mathbb{N}$ we have

$$\|x_{k+l} - x_k\| = \Big\|\sum_{i=0}^{l-1}(x_{k+i+1} - x_{k+i})\Big\| \le \sum_{i=0}^{l-1}\|x_{k+i+1} - x_{k+i}\| = \sum_{i=0}^{l-1}\|f^{k+i}(x_1) - f^{k+i}(x_0)\|$$
$$\le \|x_1 - x_0\|\sum_{i=0}^{l-1} c^{k+i} = \|x_1 - x_0\|c^k\frac{1-c^l}{1-c} \xrightarrow{k\to\infty} 0,$$

i. e., $(x_k)_k$ is a Cauchy sequence. By Remark 18.52, the limit $\tilde{x} := \lim_{k\to\infty} x_k$ exists. Since f, being a contraction, is (uniformly) continuous (set $\delta := \frac{\epsilon}{c}$), we have $f(\tilde{x}) = \lim_{k\to\infty} f(x_k) = \tilde{x}$. If $y \neq \tilde{x}$ were also a fixed point of f, we would obtain the contradiction

$$\|\tilde{x} - y\| = \|f(\tilde{x}) - f(y)\| \le c\|\tilde{x} - y\| < \|\tilde{x} - y\|. \qquad \square$$

Remark 18.54.

(a) A linear map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is a contraction if and only if there exists $c < 1$ with $\|f(x)\| \le c\|x\|$ 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 18.49.

Lemma 18.55. *Let $A \in \mathbb{C}^{n\times n}$ be normal with eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$. Let*

$$\lambda_{\min} := \min\{|\lambda_i| : i = 1, \dots, n\}, \qquad \lambda_{\max} := \max\{|\lambda_i| : i = 1, \dots, n\}.$$

For $x \in \mathbb{C}^n$ we have $\lambda_{\min}|x| \le |Ax| \le \lambda_{\max}|x|$ and $|x^*Ax| \le \lambda_{\max}|x|^2$. If A is positive definite, then $|x^*Ax| \ge \lambda_{\min}|x|^2$.

Proof. Let $S \in \mathrm{U}(n, \mathbb{C})$ with $S^*AS = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ (spectral theorem). Since $|x| = |Sx|$ and $|Ax| = |S^*Ax|$, we may assume $A = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Then we have

$$|Ax|^2 = (Ax)^*Ax = \sum_{i=1}^n |\lambda_i|^2|x_i|^2 \le \lambda_{\max}^2\sum_{i=1}^n |x_i|^2 = \lambda_{\max}^2|x|^2$$

and similarly $|Ax| \ge \lambda_{\min}|x|$. Likewise,

$$|x^*Ax| = \Big|\sum_{i=1}^n \lambda_i|x_i|^2\Big| \le \sum_{i=1}^n |\lambda_i||x_i|^2 \le \lambda_{\max}|x|^2.$$

If A is positive definite, then $\lambda_1, \dots, \lambda_n$ are real and positive. Then as before we obtain $|x^{\mathrm{t}}Ax| \ge \lambda_{\min}|x|^2$. $\square$

Theorem 18.56 (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 strictly upper triangular matrix. Let $b \in \mathbb{C}^{n\times 1}$. Then for all initial values $x_0 \in \mathbb{C}^{n\times 1}$ the sequence*

$$x_{k+1} := L^{-1}b - L^{-1}Rx_k$$

converges to the solution of $Ax = b$.

Proof. Since A has real, positive diagonal entries (Remark 12.39) and R is a strictly upper triangular matrix, L must be invertible. For $A\tilde{x} = b$ we have

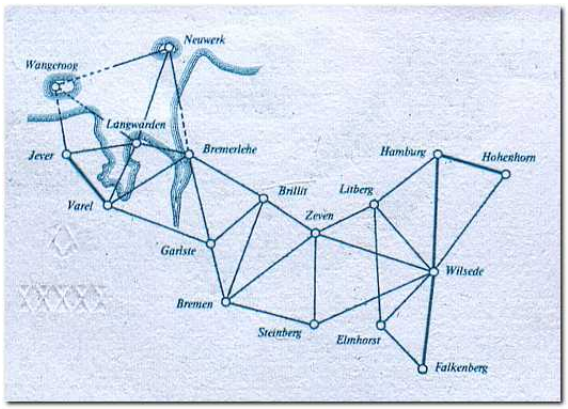
$$\tilde{x} = L^{-1}L\tilde{x} = L^{-1}(b - R\tilde{x}) = L^{-1}b - L^{-1}R\tilde{x},$$

i. e., $\tilde{x}$ is a fixed point of the map $f(x) := L^{-1}b - L^{-1}Rx$. 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 with respect to $\|.\|$. Since $L^{-1}b$ is constant, we can consider the linear map $g(x) := L^{-1}Rx$ instead of f. By Remark 18.54 and Lemma 18.55 we have to show that every eigenvalue $\lambda \in \mathbb{C}$ of $L^{-1}R$ has absolute value less than 1. 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. Hence we have

$$\begin{aligned} |\lambda|\|x\| = \|g(x)\|^2 = \|x - y\|^2 &= \|x\|^2 - x^*A^*y - y^*Ax + y^*Ay \\ &= \|x\|^2 - y^*L^*y - y^*Ly + y^*(L + R)y = \|x\|^2 - y^*Dy < \|x\|. \end{aligned}$$

This shows $|\lambda| < 1$. □

Remark 18.57. Gauss used this method to survey the Kingdom of Hanover, as attested by a drawing on the 10 DM banknote:



He wrote in a letter:

> "I recommend this method to you for imitation. You will hardly ever again eliminate directly, at least not if you have more than 2 unknowns. The indirect procedure can be carried out half asleep, or one can think of other things while doing it."

Example 18.58. 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:

$$\begin{array}{rcl} x_0 & := & (\ 0, \quad 0, \quad 0, \quad 0\ ) \\ x_{10} & \approx & (\ -1.90, \quad -2.97, \quad 4.45, \quad 8.06\ ) \\ x_{100} & \approx & (\ 2.84, \quad -14.20, \quad -5.17, \quad 27.94\ ) \\ x_{1000} & \approx & (\ -1.10, \quad 28.86, \quad -106.96, \quad 93.31\ ) \\ x_{10000} & \approx & (\ -4.000, \quad 59.995, \quad -179.988, \quad 139.992\ ) \end{array}$$

18.5 Matrix Norms

Remark 18.59. As is well known, the set of $n \times m$ matrices forms a vector space of dimension nm. It is therefore natural to introduce the norms defined in Definition 18.48 for matrices as well.

Definition 18.60. A map $\mathbb{C}^{n\times m} \to \mathbb{R}$, $A \mapsto \|A\|$ is called a *matrix norm* if for $A, B \in \mathbb{C}^{n\times m}$ and $\lambda \in \mathbb{C}$ we have

- $\|A\| \geq 0$ 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 18.61.

(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 $A, B \in \mathbb{C}^{n\times n}$ the Cauchy-Schwarz inequality yields

$$|AB|^2 = \sum_{i,j}\Big|\sum_{k=1}^{n} a_{ik}b_{kj}\Big|^2 \leq \sum_{i,j}\Big(\sum_{k=1}^{n} |a_{ik}||b_{kj}|\Big)^2 \leq \sum_{i,j}\Big(\sum_{k=1}^{n} |a_{ik}|^2\Big)\Big(\sum_{s=1}^{n} |b_{sj}|^2\Big) = |A|^2|B|^2,$$

i. e., $|.|$ is submultiplicative. If $\sigma_1, \ldots, \sigma_k$ are the singular values of A, then

$$|A| = \sqrt{\operatorname{tr}(A^*A)} = \sqrt{\sigma_1^2 + \ldots + \sigma_k^2}$$

by Remark 18.22.

(b) Every "natural"[7] vector norm $\|.\|$ *induces* a matrix norm by[8]

$$\|A\| = \max_{0\neq x\in\mathbb{C}^{m\times 1}} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

(Exercise III.14). In the case $m = 1$ the norms coincide. For all $x \in \mathbb{C}^{m\times 1}$ we have $\|Ax\| \leq \|A\|\|x\|$. For $A, B \in \mathbb{C}^{n\times n}$ we obtain

$$\|AB\| = \max_{\|x\|=1} \|ABx\| \leq \max_{\|x\|=1} \|A\|\|Bx\| = \|A\|\|B\|,$$

i. e., $\|.\|$ is submultiplicative.

(c) By (the proof of) Corollary 18.23, $\|A\|_2$ 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 $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ we have $\|A^2\|_{\max} = 2 > 1 = \|A\|_{\max}^2$.

[7] defined for arbitrary dimension

[8] As in Remark 18.19, the maximum is in fact attained.

Lemma 18.62. *For $A \in \mathbb{C}^{n\times m}$ we have*

(a) $$\|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^{n} |a_{ij}| \qquad \textit{(column-sum norm)}.$$

(b) $$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}| = \|A^{\mathrm{t}}\|_1 \qquad \textit{(row-sum norm)}.$$

(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 we have

$$\|A\|_1 = \max_{\|x\|_1=1} \sum_{i=1}^{n} \Big|\sum_{j=1}^{m} a_{ij}x_j\Big| \le \max_{\sum|x_i|=1} \sum_{j=1}^{m} |x_j| \sum_{i=1}^{n} |a_{ij}| \le \max_{1\le j\le m} \sum_{i=1}^{n} |a_{ij}|.$$

If the maximum on the right is attained for j, then $x = e_j$ yields equality.

(b) We have

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \max_{1\le i\le n} \Big|\sum_{j=1}^{m} a_{ij}x_j\Big| \le \max_{\max_i|x_i|=1} \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}||x_j| \le \max_{1\le i\le n} \sum_{j=1}^{m} |a_{ij}|.$$

If the maximum on the right is attained for i, then $x_j = \overline{a_{ij}}/|a_{ij}|$ for $a_{ij} \neq 0$ and $x_j = 0$ otherwise yields equality.

(c) If $\sigma_1 \ge \ldots \ge \sigma_k$ are the singular values of A, then we have

$$\|A\|_2 = \sigma_1 \le \sqrt{\sigma_1^2 + \ldots + \sigma_k^2} = |A| \le \sqrt{k}\sigma_1 \le \sqrt{\min\{n,m\}}\|A\|_2.$$

(d) For $x \in \mathbb{C}^n$ we have

$$\|x\|_\infty = \max_{1\le i\le n} |x_i| \le \sqrt{|x_1|^2 + \ldots + |x_n|^2} = |x| \le \sqrt{n}\|x\|_\infty.$$

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) The singular value decomposition yields $\|A^{\mathrm{t}}\|_2 = \|A\|_2$. Hence the claim follows from (b) and (d). □

Lemma 18.63. *The condition number of $A \in \mathrm{GL}(n,\mathbb{C})$ is $\kappa(A) = \|A\|_2\|A^{-1}\|_2$.*

Proof. By definition we have

$$\|A\|_2 = \max_{x \in \mathbb{C}^{n\times 1}\setminus\{0\}} |Ax|/|x|,$$
$$\|A^{-1}\|_2 = \max_x |A^{-1}x|/|x| = \max_y |y|/|Ay| = (\min_y |Ay|/|y|)^{-1}.$$

The claim follows from the definition of the condition number. □

Definition 18.64. 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 by Lemma 18.51, this definition does not depend on the choice of the norm. In this case, we write $A = \lim_{k\to\infty} A_k$

Remark 18.65.

(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) Clearly, the maps $A \to A^{\mathrm{t}}$, $A \to \overline{A}$ and tr are continuous (with respect to any matrix norm). As is well known, addition and multiplication of complex numbers are continuous maps. Hence matrix multiplication is also continuous. For $A = \lim_{k\to\infty} A_k$ and $B = \lim_{k\to\infty} B_k$ we therefore have $\lim_{k\to\infty} A_k B_k = AB$. The Leibniz formula shows that det is continuous as well. The representation $A^{-1} = \det(A)^{-1}\widetilde{A}$ (Theorem 9.22) shows that the inversion map $A \mapsto A^{-1}$ 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 18.66. *The set of* (*invertible*) *diagonalizable matrices is* dense *in* $\mathbb{C}^{n\times n}$*, i. e., for every* $A \in \mathbb{C}^{n\times n}$ *there exists a sequence of diagonalizable matrices* $(A_k)_k$ *with* $\lim_{k\to\infty} A_k = A$.

Proof. Let $S \in \mathrm{GL}(n, \mathbb{C})$ such that $D := S^{-1}AS$ is a triangular matrix (e. g., Jordan normal form). Then

$$B_k := D + \frac{1}{k}\operatorname{diag}(1, 2, \ldots, n) \in \mathbb{C}^{n\times n}$$

is also a triangular matrix for $k \in \mathbb{N}$. If k is sufficiently large, the main diagonal of B_k consists of pairwise distinct eigenvalues (nonzero). By Corollary 8.12, B_k and hence also

$$A_k := SB_kS^{-1} = A + \frac{1}{k}S\operatorname{diag}(1, \ldots, n)S^{-1}$$

is diagonalizable (and invertible). Clearly, $\lim_k A_k = A$. □

Example 18.67. Theorem 18.66 allows us to transfer statements about diagonalizable (or invertible) matrices to all matrices:

(a) The Cayley-Hamilton theorem, $\chi_A(A) = 0$, is easily verified for diagonal matrices A (see the proof of Theorem 10.52). Because of $S\chi_A(A)S^{-1} = \chi_{SAS^{-1}}(SAS^{-1})$ for $S \in \mathrm{GL}(n, \mathbb{C})$, it holds for diagonalizable matrices. The map $A \mapsto \chi_A(A)$ is continuous, since the coefficients of χ_A are polynomials in the coefficients of A (Remark 18.65). By Theorem 18.66, $\chi_A(A) = 0$ therefore holds for all $A \in \mathbb{C}^{n\times n}$.

(b) The equation $\widetilde{AB} = \widetilde{B}\widetilde{A}$ from Theorem 9.25 follows for invertible matrices directly from Theorem 9.22. By Remark 18.65, the map $A \mapsto \widetilde{A}$ is continuous. Hence $\widetilde{AB} = \widetilde{B}\widetilde{A}$ holds for all $A, B \in \mathbb{C}^{n\times n}$.

Theorem 18.68. *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 we have $|\lambda| < 1$ or $\lambda = 1$.

(b) If 1 is an eigenvalue, then the algebraic multiplicity equals the geometric multiplicity.

Proof. W. l. o. g. 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 $|\lambda_i| < 1$. By Lemma 14.41, the absolute values of the entries of $J_{n_i}(\lambda_i)^k$ have 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 $J_{n_i}(\lambda_i)_{21} = k \to \infty$. Hence $n_i = 1$ must hold for all i with $\lambda_i = 1$. W. l. o. g. let $\lambda_1 = \ldots = \lambda_t = 1 > |\lambda_j|$ for $j > t$. Clearly, (A^k) then converges to $\operatorname{diag}(1_t, 0_{n-t})$. □

Corollary 18.69. *Suppose that the conditions of Theorem 18.68 hold 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 $\lim_{k\to\infty} A^k = vw^*$.*

Proof. Since $\chi_{A^*} = \overline{\chi_A}$, 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 $S^{-1}AS = \operatorname{diag}(1, B)$ with $\lim_{k\to\infty} B^k = 0$. Let u be the first row of S^{-1}. Then we have

$$C := \lim_{k\to\infty} A^k = \lim_{k\to\infty} S\operatorname{diag}(1, B^k)S^{-1} = S\operatorname{diag}(1, 0_{n-1})S^{-1} = vu.$$

On the other hand, $w = C^*w = u^*v^*w = u^*[v, w] = u^*$. □

18.6 Computing Eigenvalues

Remark 18.70.

(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, the polynomial χ_A is ill-conditioned as well. The conditioning of μ_A is even worse, since here smaller changes can even change $\deg \mu_A$.

(b) The eigenvalues as roots of χ_A (or μ_A) are ill-conditioned as well:

$$\begin{aligned} &X^2 - 21X + 110 && \text{roots: } 10, 11 \\ &X^2 - 21.\mathbf{1}X + 110 && \text{roots: } \approx 9.41, 11.69 \\ &X^2 - 20.\mathbf{9}X + 110 && \text{roots: } \approx 10.45 \pm 0.89\mathrm{i} \end{aligned}$$

Perturbations of A thus affect the computation of eigenvalues "doubly". Moreover, 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 by the following theorem.

Theorem 18.71 (Bauer-Fike)**.** *Let $A \in \mathbb{C}^{n\times n}$ be diagonalizable with $S^{-1}AS = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Let $X \in \mathbb{C}^{n\times n}$ and let $\mu \in \mathbb{C}$ 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. W. l. o. g. we may assume $\mu \notin \{\lambda_1, \ldots, \lambda_n\}$. Let $D := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Then $D - \mu 1_n$ is invertible. By hypothesis, $A + X - \mu 1_n$ is not invertible. Hence

$$(D - \mu 1_n)^{-1} S^{-1}(A - \mu 1_n + X)S = 1_n + (D - \mu 1_n)^{-1} S^{-1} X S$$

is not invertible either. From Theorem 19.8 and Lemma 18.63 we obtain

$$1 \leq \rho\big((D - \mu 1_n)^{-1} S^{-1} X S\big) \leq \|(D - \mu 1_n)^{-1} S^{-1} X S\|_2 \leq \|(D - \mu 1_n)^{-1}\|_2 \kappa(S)\|X\|_2.$$

By Example 18.61, $\|(D - \mu 1_n)^{-1}\|_2 = \max_{i=1,\ldots,n}(|\lambda_i - \mu|)^{-1}$. 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, we can choose $S \in \mathrm{U}(n, \mathbb{C})$ by the spectral theorem. Then $\kappa(S) = 1$. □

Theorem 18.72 (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 $i \geq 2$. Then the sequence*

$$x_{k+1} := \frac{Ax_k}{|Ax_k|}$$

converges to an eigenvector for λ for "almost all" starting vectors $x_0 \in \mathbb{C}^{n\times 1}$. In this case, $\lim_{k\to\infty} x_k^* A x_k = \lambda$.

Proof. By hypothesis, 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 we order $\lambda_2, \ldots, \lambda_n$ suitably. Let $b_1, \ldots, b_n \in \mathbb{C}^{n\times 1}$ be a corresponding basis with $Ab_1 = \lambda b_1$. For $S := (b_1, \ldots, b_n) \in \mathrm{GL}(n, \mathbb{C})$ we have $S^{-1}AS = J$. 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 $i \geq 2$, by 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 $|\lambda_j| < |\lambda|$. Together with $A^k x_0$, also x_k tends to a multiple of b_1. The second claim follows from $|x_k| = 1$ for all $k \geq 1$. □

Remark 18.73. If one has (approximately) determined an eigenvector b_1 for the eigenvalue λ of A of 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. We have $S^{-1}AS = \left(\begin{smallmatrix} \lambda & * \\ 0 & A_1 \end{smallmatrix}\right)$, 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 18.72 to A_1 and iterate. We will see that this can be accomplished with just one iteration.

Lemma 18.74. *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. Since $|Q_k| = \sqrt{\operatorname{tr}(Q_k^* Q_k)} = \sqrt{\operatorname{tr} 1_n} = \sqrt{n}$, the sequence $(Q_k)_k$ is bounded. By Bolzano-Weierstrass, there exists a convergent subsequence $(Q_{k_i})_i$ with $Q := \lim_{i\to\infty} Q_{k_i}$. Since

$$Q^*Q = \lim_{i\to\infty} Q_{k_i}^* Q_{k_i} = 1_n$$

the matrix Q is unitary. From the continuity of matrix multiplication we obtain

$$\lim_{i\to\infty} R_{k_i} = \lim_{i\to\infty} Q_{k_i}^* (Q_{k_i} R_{k_i}) = Q^* \lim_{i\to\infty} A_{k_i} = Q^*.$$

Since the set of upper triangular matrices with nonnegative main diagonal entries is closed, Q^* must also belong to this set. Clearly, this is only possible if $Q = 1_n$. In particular, every convergent subsequence of $(Q_k)_k$ converges to the same limit. Hence (Q_k) must also converge to 1_n, and consequently so must (R_k). □

Definition 18.75. 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 18.76. In the situation of Definition 18.75, $D_k = \operatorname{diag}(d_1, \ldots, d_n)$ with $|d_i| = 1$ for $i = 1, \ldots, n$. The main diagonals of A_k and $D_k^* A_k D_k$ are identical for all k. Hence the main diagonal of A_k converges to the main diagonal of A. The entries of A_k off the main diagonal, however, can "oscillate" by a factor of absolute value 1 in every iteration.

Theorem 18.77 (FRANCIS algorithm[9])**.** *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.*

[9] also called the *QR algorithm*

Proof (WILKINSON). Let $\mathcal{Q}_k := Q_1 \dots Q_k$ and $\mathcal{R}_k := R_k \dots R_1$. 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 $k \geq 2$ we have 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 $\lambda_1, \dots, \lambda_n$ of A we have $|\lambda_1| > \dots > |\lambda_n| > 0$ by hypothesis. By Corollary 8.12 there exists $S \in \mathrm{GL}(n, \mathbb{C})$ with $S^{-1} A S = \mathrm{diag}(\lambda_1, \dots, \lambda_n) =: D$. Let $S^{-1} = L_{S'} P U_{S'}$ be the modified LU decomposition of S^{-1} from Remark 18.37. 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 $x_{ij} = \delta_{ij}$ for $i \leq j$ and $|\lambda_i| < |\lambda_j|$ for $i > j$. Hence also $M_k := R_S P^{\mathrm{t}} D^k L_{S'} D^{-k} P R_S^{-1}$ tends to 1_n. For the QR decomposition $M_k = Q_{M_k} R_{M_k}$ we have $\lim_{k \to \infty} Q_{M_k} = 1_n$ by Lemma 18.74. Altogether,

$$\begin{aligned} \mathcal{Q}_k \mathcal{R}_k = A^k = S D^k S^{-1} &= S P P^{\mathrm{t}} D^k L_{S'} P U_{S'} = Q_S R_S P^{\mathrm{t}} D^k L_{S'} D^{-k} D^k P U_{S'} \\ &= Q_S M_k R_S P^{\mathrm{t}} D^k P U_{S'} = Q_S Q_{M_k} R_{M_k} P^{\mathrm{t}} D^k P U_{S'}. \end{aligned}$$

Together with D^k, also $P^{\mathrm{t}} D^k P$ is 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

$$U_k := \mathrm{diag}\big(\overline{t_{11}}/|t_{11}|, \dots, \overline{t_{nn}}/|t_{nn}|\big) \in \mathrm{U}(n, \mathbb{C})$$

the matrix $U_k T_k$ 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 we obtain $\mathcal{Q}_k = Q_S Q_{M_k} U_k^*$. 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 main diagonal contains the eigenvalues of D and A, respectively. □

Remark 18.78.

(a) Clearly, (A_k) quasi-converges to a Schur decomposition (Theorem 13.23) of A. Let A be real. Then the hypothesis implies that the eigenvalues are real as well, since $|\lambda| = |\overline{\lambda}|$ for $\lambda \in \mathbb{C}$. If this hypothesis is not satisfied, one can arrange that (A_k) converges to a real block diagonal matrix with 2×2 blocks (Exercise II.24).

(b) In its pure form, the Francis algorithm converges only slowly. In practice, one accelerates the method 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 every iteration. This algorithm is counted (alongside the FFT and the Simplex algorithm) among the most important algorithms of the 20th century.[10]

Example 18.79.

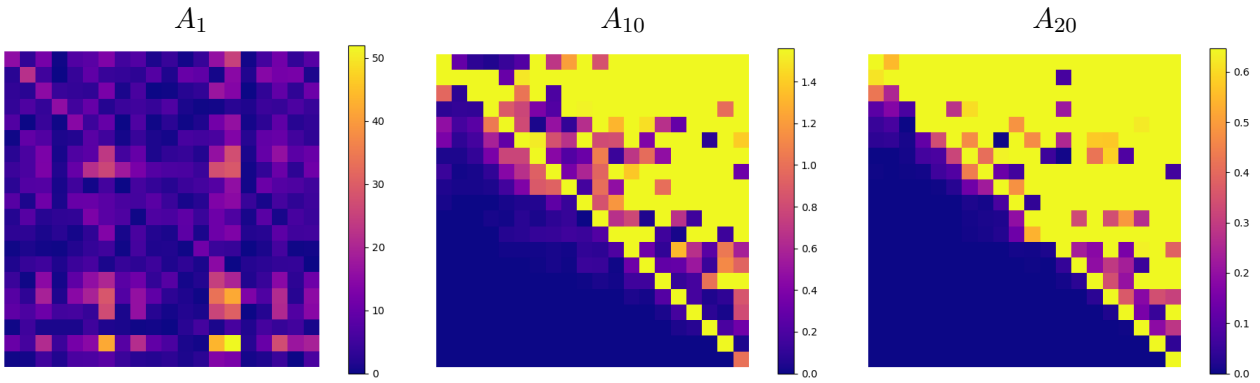
(a) For

$$A = \begin{pmatrix} 2 & -1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} = Q_1 R_1$$

we have $A_2 = R_1 Q_1 = \left(\begin{smallmatrix} 2 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ and inductively $A_k = \left(\begin{smallmatrix} 2 & (-1)^k \\ 0 & -1 \end{smallmatrix}\right)$. This shows the phenomenon of quasi-convergence.

[10] see [Dongarra-Sullivan, *Top Ten Algorithms of the Century*, Comput. Sci. Eng. 2 (2000), 22–23]

(b) Let $A = S^{-1}\operatorname{diag}(1,\dots,20)S \in \mathbb{R}^{20\times 20}$ for a randomly chosen matrix $S \in \mathrm{GL}(20,\mathbb{R})$. The following graphic illustrates the Francis algorithm for A. The color scale represents the absolute values of the entries below the main diagonal (the structure of A was explained in Remark 9.24).



Remark 18.80. For positive definite matrices, one can replace the QR decomposition in the Francis algorithm by the Cholesky decomposition.

Theorem 18.81 (Cholesky algorithm)**.** *Let $A = A_1 \in \mathbb{C}^{n\times n}$ be positive definite. For $k = 1, 2, \dots$ let $A_k = R_k^* R_k$ be the Cholesky decomposition and $A_{k+1} := R_k R_k^*$. Then $(A_k)_k$ converges to a diagonal matrix with the eigenvalues of A on the main diagonal.*

Proof (Schatzman). Since $A_{k+1} = R_k A_k R_k^{-1}$, all A_k have the same eigenvalues. Let $A_k = (a_{ij}^{(k)})$ and $R_k = (r_{ij}^{(k)})$. Since

$$\delta_k(s) := \sum_{i=1}^{s} a_{ii}^{(k)} = \sum_{i=1}^{s}\sum_{j=1}^{i} |r_{ji}^{(k)}|^2 = \sum_{i=1}^{s}\sum_{j=i}^{s} |r_{ij}^{(k)}|^2 \le \sum_{i=1}^{s}\sum_{j=i}^{n} |r_{ij}^{(k)}|^2 = \delta_{k+1}(s)$$

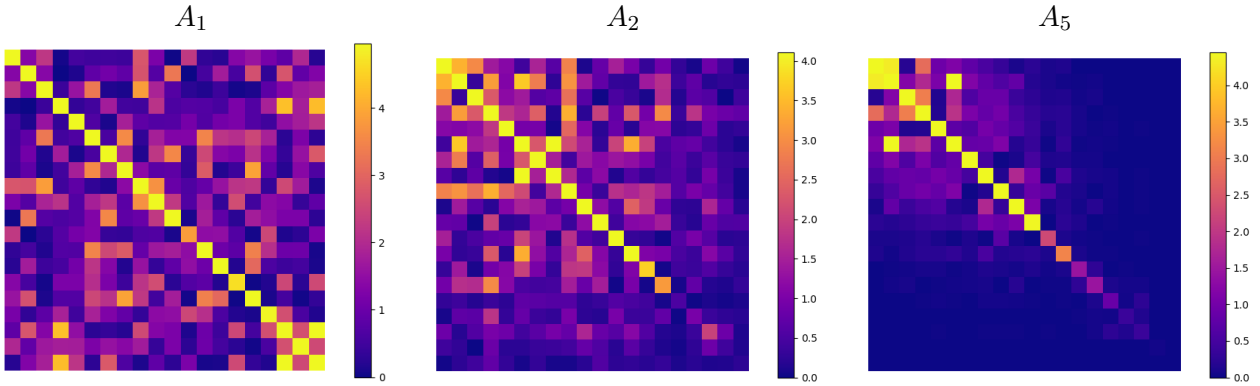
the sequences $(\delta_k(s))_k$ for $s = 0,\dots,n$ are monotonically increasing. Since $\delta_k(s) \le \delta_k(n) = \operatorname{tr} A_k = \operatorname{tr} A$, the sequences are bounded. Hence $\delta(s) := \lim_{k\to\infty} \delta_k(s)$ exists for $s = 0,\dots,n$. It follows that $\lim_{k\to\infty} a_{ss}^{(k)} = \delta(s) - \delta(s-1)$ 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 grows, we must have $\lim_{k\to\infty} r_{ij}^{(k)} = 0$ for all $i \ne j$. Together with R_k, also A_k must converge to a diagonal matrix. □

Example 18.82. 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 algorithm with A. The color

scale represents the absolute values of all entries off the main diagonal.



Clearly, the method works from the bottom right to the top left.

Remark 18.83. The approximate location of the eigenvalues of a matrix can be enclosed without any computational effort by the following theorem.

Theorem 18.84 (Gershgorin). *For every eigenvalue $\lambda \in \mathbb{C}$ of $(a_{ij}) \in \mathbb{C}^{n\times n}$ there exists $i \in \{1,\dots,n\}$ with $|\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|$.*

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 may assume $x_i = 1$ and $|x_j| \leq 1$ for $j \neq i$. By the triangle inequality we then have

$$|\lambda - a_{ii}| = |(Ax)_i - a_{ii}x_i| = \left|\sum_{j=1}^{n} a_{ij}x_j - a_{ii}x_i\right| = \left|\sum_{j\neq i} a_{ij}x_j\right| \leq \sum_{j\neq i} |a_{ij}|. \qquad \square$$

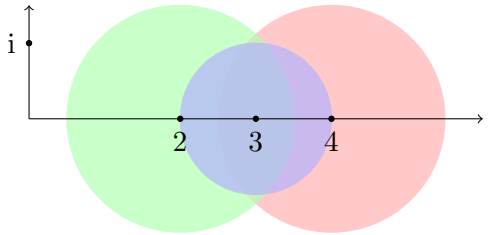
Corollary 18.85. *Every matrix $A \in \mathbb{C}^{n\times n}$ with $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for $i = 1,\dots,n$ is invertible.*[11]

Proof. By Gershgorin, 0 is not an eigenvalue of A. $\square$

Example 18.86.

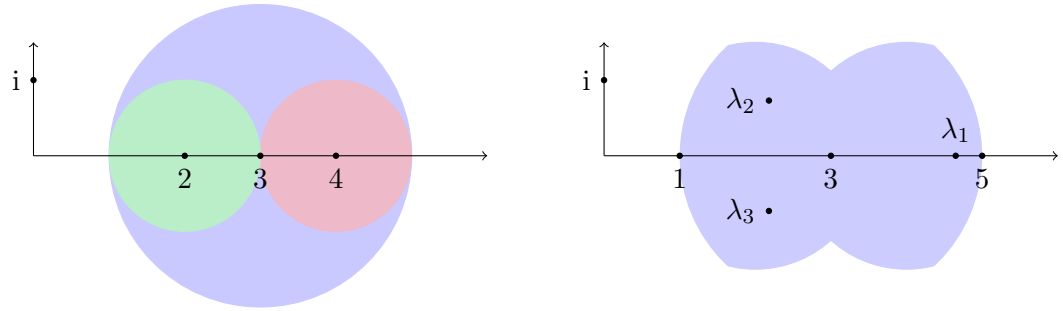
(a) Every eigenvalue lies in one of the circles shown:

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 0.5 & 4 & 1 \\ 1.5 & 0 & 2 \end{pmatrix}$$



[11] This property is called *diagonal dominance*.

Since A and A^{t} have the same eigenvalues, one can also use the columns and then intersect both sets:



(b) If we apply Theorem 18.84 to the companion matrix of $\alpha = X^n + a_1 X^{n-1} + \ldots + a_n \in \mathbb{C}[X]$, we obtain

$$|x| \leq \max\bigl\{|a_n|, |a_{n-1}| + 1, \ldots, |a_1| + 1\bigr\}$$

for every root $x \in \mathbb{C}$ of α (note that $|x + a_1| \leq 1 \Longrightarrow |x| \leq |a_1| + 1$).

18.7 Orthonormalization

Remark 18.87. In our formulation of the Gram-Schmidt process (11.10) we first constructed an orthogonal basis and then obtained an orthonormal basis by normalization. However, this can produce very large or very small factors, which lead to numerical instabilities. Hence it is better to normalize every vector immediately after it has been found:

$$b'_s := v_s - \sum_{i=1}^{s-1} [v_s, b_i] b_i, \qquad b_s := \frac{b'_s}{|b'_s|}$$

Rounding errors can additionally accumulate in the summation. Hence an iterated procedure is more stable.

Theorem 18.88 (Modified Gram-Schmidt process)**.** *Let V be a unitary space and let $v_1, \ldots, v_k \in V$ be 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 $b_1, \ldots, b_k$ 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 suppose the claim has already been shown for $b_1, \ldots, b_{s-1}$. Then $b_{s,1} = v_s - [v_s, b_1] b_1$. Inductively, suppose we have already shown

$$b_{s,i} = v_s - \sum_{j=1}^{i} [v_s, b_j] b_j$$

for some $i < s - 1$. Since $[b_j, b_{i+1}] = 0$ 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 computed in Theorem 11.10 (before normalization). □

Remark 18.89. When computing the QR decomposition of an invertible matrix $A \in \mathrm{GL}(n,\mathbb{C})$, besides the orthonormalized columns of A (as the columns of Q) one also needs 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 carry out 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 $Q_1 \in \mathrm{O}(n,\mathbb{R})$ in the hyperplane $\langle b\rangle^\perp$ can be computed according to Exercise II.13 as a HOUSEHOLDER *transformation*:

$$Q_1 = 1_n - \frac{2}{|b|^2} bb^{\mathrm{t}}$$

(for $b = 0$ let $Q_1 = 1_n$). Since

$$[a_1 + |a_1|e_1, b] = |a_1|^2 - |a_1|[a_1, e_1] + |a_1|[e_1, a_1] - |a_1|^2 = 0$$

we have $a_1 + |a_1|e_1 \in \langle b\rangle^\perp$ 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 $Q_1 A = \left(\begin{smallmatrix} |a_1| & * \\ 0 & * \end{smallmatrix}\right)$. For the columns $2,\dots,n$ one proceeds analogously and obtains corresponding matrices $Q_2,\dots,Q_n \in \mathrm{O}(n,\mathbb{R})$ (in fact, it suffices to consider the truncated 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):

$$D_{st}(\varphi) := \begin{pmatrix} 1_{s-1} & & & & \\ & \cos\varphi & & -\sin\varphi & \\ & & 1_{t-s-1} & & \\ & \sin\varphi & & \cos\varphi & \\ & & & & 1_{n-t} \end{pmatrix} \in \mathrm{O}(n,\mathbb{R}).$$

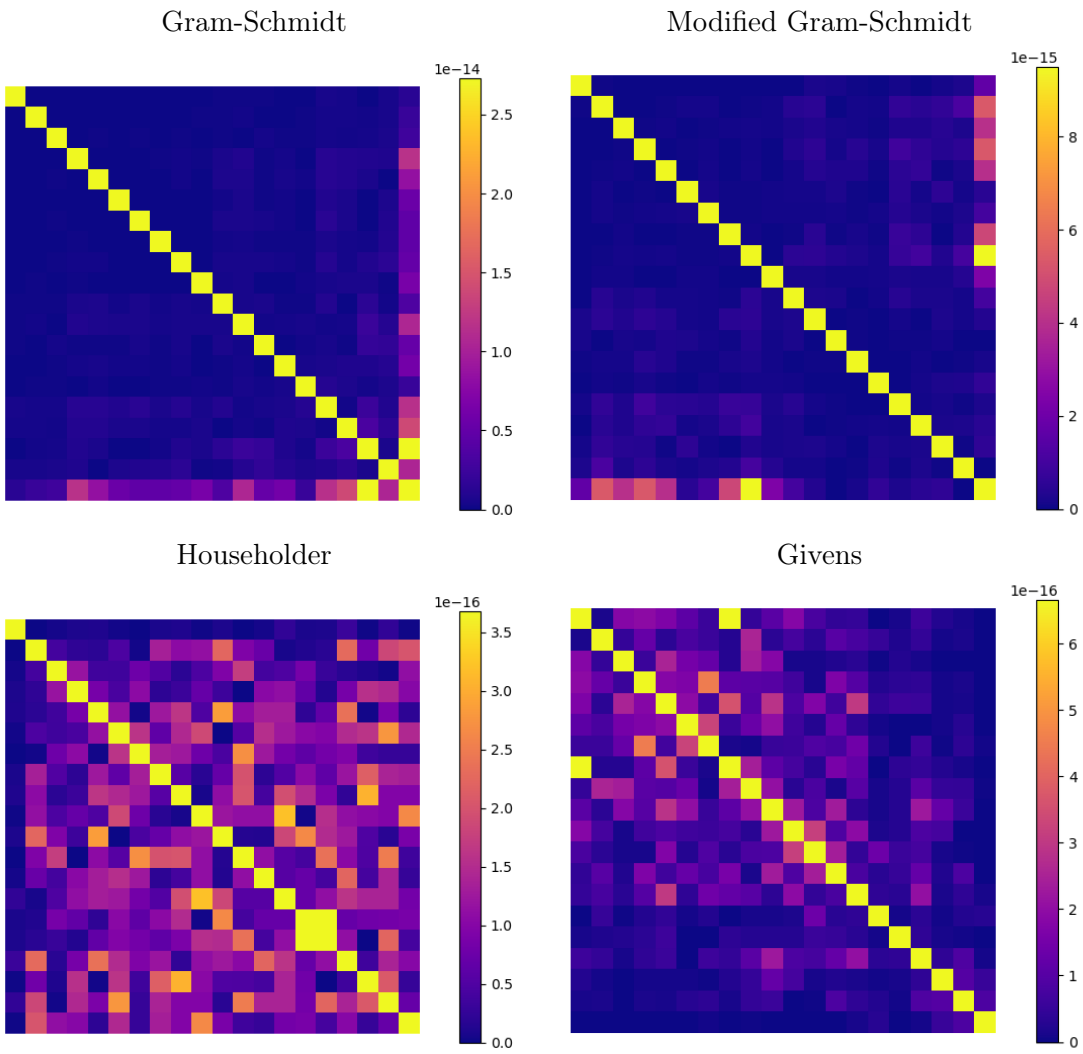
We proceed as in the Gaussian algorithm. Suppose the first $k-1$ columns of A are already in upper triangular form. After a row interchange we may assume $a_{kk} \neq 0$ (A is invertible). To eliminate the entry a_{ik} for $i > k$, let φ be the oriented angle from (a_{kk}, a_{ik}) to e_1 (see Example 11.22). For $c := |(a_{kk}, a_{ik})| = \sqrt{|a_{kk}|^2 + |a_{ik}|^2}$ we have

$$D_{ki}(\varphi) = \begin{pmatrix} 1_{k-1} & & & & \\ & a_{kk}/c & & a_{ik}/c & \\ & & 1_{i-k-1} & & \\ & -a_{ik}/c & & a_{kk}/c & \\ & & & & 1_{n-i} \end{pmatrix}.$$

By replacing A by $D_{ki}(\varphi)A$, we achieve $a_{ik} = 0$. In this way we obtain a sequence of Givens rotations $Q_1,\dots,Q_l \in \mathrm{O}(n,\mathbb{R})$ such that $R := Q_l \dots Q_1 A$ is an upper triangular matrix. For $Q := Q_1^{\mathrm{t}} \dots Q_l^{\mathrm{t}} \in \mathrm{O}(n,\mathbb{R})$ we have $A = QR$. This does require significantly more matrices than (a), but the Q_i are sparse, so the multiplications cost little.

Example 18.90. We compute the QR decomposition $A = QR$ of a random (invertible) matrix $A \in \mathbb{R}^{20\times 20}$ with the four methods presented. The quality of the results shows in a comparison of $Q^{\mathrm{t}}Q$

with 1_{20}. The color scale refers to the absolute values of the entries off the main diagonal (the smaller, the more accurate):



19 Analytic Aspects

19.1 Eigenvalue Estimates

Remark 19.1. By Corollary 13.20, a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ has 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 18.55.

Theorem 19.2 (COURANT-FISCHER)**.** *For every Hermitian matrix $A \in \mathbb{C}^{n\times n}$ we have*

$$\lambda_k(A) = \max_{\substack{V\leq\mathbb{C}^n \\ \dim V=k}} \min_{0\neq v\in V} \frac{vAv^*}{vv^*} = \min_{\substack{V\leq\mathbb{C}^n \\ \dim V=n-k+1}} \max_{0\neq v\in V} \frac{vAv^*}{vv^*}.$$

Proof. Let $\mu_k(A)$ be the middle part of the formula. By the spectral theorem there exists $S \in \mathrm{U}(n,\mathbb{C})$ 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 $\{vS : v \in V\}$, we may assume $A = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Moreover, we may restrict to normalized $v \in V$ by replacing v with $v/|v|$. For $V := \langle e_1,\ldots,e_k\rangle$ we have

$$\mu_k(A) \geq \min_{\substack{v\in V \\ |v|=1}} vAv^* = \min_{\substack{v\in V \\ |v|=1}} \sum_{i=1}^{k} \lambda_k|v_k|^2 \geq \min_{|v|=1} \lambda_k|v|^2 = \lambda_k.$$

Conversely, let $V \leq \mathbb{C}^n$ with $\dim V = k$ be arbitrary. By the dimension formula there exists a normalized vector $w \in V \cap \langle e_k,\ldots,e_n\rangle$. It follows

$$\min_{\substack{v\in V \\ |v|=1}} vAv^* \leq wAw^* = \sum_{i=k}^{n} \lambda_i|w_i|^2 \leq \lambda_k.$$

Since V is arbitrary, we have $\mu_k(A) \leq \lambda_k = \lambda_k(A)$. For the second equality, we consider analogously $V := \langle e_k,\ldots,e_n\rangle$ and $V \cap \langle e_1,\ldots,e_k\rangle$. □

Theorem 19.3 (CAUCHY's interlacing 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. Together with A, the matrix $B := P^*AP \in \mathbb{C}^{k\times k}$ is Hermitian as well. For $U \leq \mathbb{C}^k$ let $P(U) := \{uP^* : u \in U\} \leq \mathbb{C}^n$. Since P has full rank, we have $\dim U = \dim P(U)$. Moreover, $|uP^*| = |u|$ for all $u \in U$. By Courant-Fischer it follows

$$\lambda_i(B) = \max_{\substack{U\leq\mathbb{C}^k \\ \dim U=i}} \min_{\substack{v\in P(U) \\ |v|=1}} vAv^* \leq \max_{\substack{V\leq\mathbb{C}^n \\ \dim V=i}} \min_{\substack{v\in V \\ |v|=1}} vAv^* = \lambda_i(A).$$

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 19.2. □

Example 19.4. For $I \subseteq \{1, \ldots, n\}$, the matrix $P = (e_i : i \in I)$ has orthonormal columns. Cauchy's interlacing theorem then holds with the submatrix $B := P^*AP = (a_{ij} : i, j \in I)$. In the case $|I| = n - 1$ we obtain

$$\lambda_1(A) \geq \lambda_1(B) \geq \lambda_2(A) \geq \lambda_2(B) \geq \ldots \geq \lambda_{n-1}(B) \geq \lambda_n(A).$$

19.2 The Spectral Radius

Definition 19.5. For $A \in \mathbb{C}^{n \times n}$ the number

$$\rho(A) := \max\{|\lambda| : \lambda \in \mathbb{C} \text{ eigenvalue of } A\}$$

is called the *spectral radius* of A.[1]

Example 19.6.

(a) For unitary matrices A we have $\rho(A) = 1$.

(b) A matrix $A \in \mathbb{C}^{n \times n}$ 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 19.7. *For all $A \in \mathbb{C}^{n \times n}$ we have:*

(a) $\|A\| \geq \rho(A)$ for every submultiplicative matrix norm.

(b) For every $\epsilon > 0$ 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 $B \in \mathbb{C}^{n \times n}$ have first column v and zeros otherwise. Then

$$\rho(A)\|B\| = \|\lambda B\| = \|AB\| \leq \|A\|\|B\|$$

and thus $\rho(A) \leq \|A\|$.

(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 $\|B\| := \|DS^{-1}BSD^{-1}\|_1$ we now have $\|A\| := \|N\|_1 \leq \rho(A) + \epsilon$ by Lemma 18.62. □

Theorem 19.8 (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 $q := \rho(A) + \epsilon < 1$. By Theorem 19.7 there exists a submultiplicative matrix norm with $\|A\| \leq q$. It follows $\|A^k\| \leq \|A\|^k \leq q^k \to 0$ as $k \to \infty$.

(2) ⇒ (1): Let λ be an eigenvalue of A with absolute value $\rho(A)$ and eigenvector v. Since $\lambda^k v = A^k v \to 0$, we have $\rho(A) = |\lambda| < 1$.

(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 $B := \lim_{k\to\infty} B_{m_k}$ we have

$$(1_n - A)B = \lim_{k\to\infty} (1_n - A) \sum_{i=0}^{m_k} A^i = \lim_{k\to\infty} (1_n - A^{m_k+1}) \overset{(2)}{=} 1_n,$$

i. e., $B = (1_n - A)^{-1}$. Since all convergent subsequences of (B_m) have the same limit, the whole sequence must converge to B as well.

(3) ⇒ (2): Since the partial sums B_m converge, their differences $B_m - B_{m-1} = A^m$ must form a null sequence. □

Corollary 19.9. *Let $A \in \mathbb{C}^{n\times n}$ with $\|A\| < 1$ for a submultiplicative matrix norm. Then $\sum_{k=0}^{n} A^k = (1_n - A)^{-1}$.*

Proof. By Theorem 19.7 we have $\rho(A) < 1$. □

Example 19.10. Submultiplicativity in Corollary 19.9 is indispensable: For $A = (3/4)_{i,j=1}^{2}$ we have $\|A\|_{\max} = 3/4 < 1$ with the matrix norm from Example 18.61, but $A^k = \left(\frac{3}{2}\right)^{k-1} A$.

Theorem 19.11. *For every matrix norm $\|.\|$ and all $A \in \mathbb{C}^{n\times n}$ we have*

$$\boxed{\rho(A) = \lim_{k\to\infty} \sqrt[k]{\|A^k\|}.}$$

Proof. For an arbitrary (not necessarily submultiplicative) matrix norm $\|.\|'$ there exist constants $\lambda, \mu > 0$ by Lemma 18.51 with $\lambda\|A^k\| \leq \|A^k\|' \leq \mu\|A^k\|$ for all k. It follows

$$\sqrt[k]{\lambda}\sqrt[k]{\|A^k\|} \leq \sqrt[k]{\|A^k\|'} \leq \sqrt[k]{\mu}\sqrt[k]{\|A^k\|}.$$

Since $\lim_{k\to\infty} \sqrt[k]{\lambda} = 1 = \lim_{k\to\infty} \sqrt[k]{\mu}$, it suffices to prove the statement for one particular 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 18.61. By Lemma 14.41, every nonzero entry of A^k has the form $\binom{k}{l}\lambda_i^{k-l}$ with $1 \le i \le s$ and $0 \le l \le n-1$. For every $\epsilon > 0$ we have $k < (1+\epsilon)^k$ if k is large enough. It follows

$$\lim_{k\to\infty} \sqrt[k]{\binom{k}{l}|\lambda_i|^{k-l}} = |\lambda_i| \lim_{k\to\infty} \sqrt[k]{k(k-1)\dots(k-l+1)} \lim_{k\to\infty} |\lambda_i|^{-l/k} = |\lambda_i|.$$

This shows $\lim_{k\to\infty} \sqrt[k]{\|A^k\|} = \rho(A)$. □

19.3 The Exponential Function of a Matrix

Definition 19.12. 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 19.13.

(a) Since $\left|\frac{A^k}{k!}\right| \le \frac{|A|^k}{k!} \to 0$, $\exp A$ is well defined by Theorem 19.8.

(b) Via $\exp(\lambda 1_n) = e^\lambda 1_n$ 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 $S \in \mathrm{GL}(n,\mathbb{C})$. By the continuity of matrix multiplication we have

$$S^{-1}\exp(A)S = \exp J = \mathrm{diag}\big(\exp J_{n_1}(\lambda_1), \dots, \exp J_{n_s}(\lambda_s)\big).$$

By Lemma 14.41, the nonzero entries of $\exp(J_{n_i}(\lambda_i))$ have the form

$$\sum_{k=l}^{\infty} \frac{\binom{k}{l}\lambda_i^{k-l}}{k!} = \sum_{k=l}^{\infty} \frac{\lambda_i^{k-l}}{l!(k-l)!} = \frac{e^{\lambda_i}}{l!}.$$

Hence

$$L_m(\lambda) := \exp(J_m(\lambda)) = e^\lambda \begin{pmatrix} 1 & & & \\ \frac{1}{1!} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \frac{1}{(m-1)!} & \cdots & \frac{1}{1!} & 1 \end{pmatrix} \in \mathbb{C}^{m\times m}$$

and

$$\exp A = S\,\mathrm{diag}\big(L_{n_1}(\lambda_1), \dots, L_{n_s}(\lambda_s)\big)S^{-1}.$$

Example 19.14.

(a) For the permutation matrix $P = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ we have

$$\exp P = \sum_{k=0}^{\infty} \frac{1}{(2k)!} 1_n + \sum_{k=0}^{\infty} \frac{1}{(2k+1)!} P = \begin{pmatrix} \cosh(1) & \sinh(1) \\ \sinh(1) & \cosh(1) \end{pmatrix} = \begin{pmatrix} 1.54\dots & 1.17\dots \\ 1.17\dots & 1.54\dots \end{pmatrix}$$

with the hyperbolic 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) By Remark 19.13 we have $\exp A = 1_n$ if and only if A is diagonalizable and the eigenvalues are integer multiples of $2\pi\mathrm{i}$.

Theorem 19.15 (Jacobi)**.** *For $A \in \mathbb{C}^{n\times n}$ we have* $\boxed{\det(\exp A) = e^{\operatorname{tr} A}.}$

Proof. W. l. o. g. 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 19.16. If A is real, then $\det(\exp A) > 0$.

Theorem 19.17 (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 16.21.[3] Since all matrices involved are polynomials in A or B, respectively, they commute pairwise. By Lemma 14.11, D_A and D_B can be diagonalized simultaneously. Hence W. l. o. g. 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, we have

$$\exp N = \sum_{k=0}^{\infty} \frac{1}{k!}(N_A+N_B)^k = \sum_{k=0}^{2n}\sum_{l=0}^{k} \frac{1}{k!}\binom{k}{l} N_A^l N_B^{k-l} = \sum_{k=0}^{2n}\sum_{l=0}^{k} \frac{N_A^l}{l!}\frac{N_B^{k-l}}{(k-l)!} = \exp(N_A)\exp(N_B).$$

Similarly,

$$\exp(D+N) = \sum_{k=0}^{\infty}\sum_{l=0}^{k} \frac{D^{k-l}}{(k-l)!}\frac{N^l}{l!} = \sum_{l=0}^{n} \frac{N^l}{l!}\sum_{k=l}^{\infty} \frac{D^{k-l}}{(k-l)!} = \exp(N)\exp(D).$$

The corresponding equations also hold for $D_A + N_A$ and $D_B + N_B$. Altogether we obtain

$$\begin{aligned} \exp(A+B) = \exp(D+N) &= \exp(D)\exp(N) = \exp(D_A)\exp(D_B)\exp(N_A)\exp(N_B) \\ &= \exp(D_A)\exp(N_A)\exp(D_B)\exp(N_B) = \exp(A)\exp(B). \end{aligned}$$

The second claim follows from $\exp(A)\exp(-A) = \exp(A-A) = \exp 0_n = 1_n$. □

[3] This can be avoided by using the Cauchy product for absolutely convergent matrix series instead.

Remark 19.18.

(a) The commutativity of A and B in the functional equation cannot be dropped, since otherwise $\exp(A)\exp(B) = \exp(A+B) = \exp(B+A) = \exp(B)\exp(A)$ for all A, B, contradicting Theorem 19.19. 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 written uniquely 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 19.19. *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 the existence we may assume $A = J_n(\mu)$ for some $\mu \neq 0$ by Remark 19.13. Let $\lambda = \log(\mu)$ and $B := J_n(\lambda)$. By Remark 19.13, μ is an eigenvalue of $\exp B = L_n(\lambda)$ with algebraic multiplicity n and geometric multiplicity 1. From Theorem 14.28 it follows $\exp B \approx A$. Hence there exists $S \in \mathrm{GL}(n,\mathbb{C})$ 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 stated properties. After a change of basis we may assume $B = \mathrm{diag}(J_{n_1}(\lambda_1), \ldots, J_{n_s}(\lambda_s))$. By Remark 19.13 we have

$$A = \exp B = \mathrm{diag}(L_{n_1}(\lambda_1), \ldots, L_{n_s}(\lambda_s)).$$

Because of the special choice of the eigenvalues via 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 $S \in \mathrm{C}(A)$. We have to show $S \in \mathrm{C}(B)$. Clearly, $\chi_A = (X - e^{\lambda_1})^{n_1} \ldots (X - e^{\lambda_s})^{n_s}$. W. l. o. g. let $\lambda_1, \ldots, \lambda_s$ be sorted such that equal λ_i are adjacent. By Exercise II.34, $C = \mathrm{diag}(C_1, \ldots, C_t)$, where each C_k belongs to a block $\mathrm{diag}(L_{n_i}(\lambda_i), \ldots, L_{n_j}(\lambda_j))$ with $\lambda_i = \ldots = \lambda_j$. Hence we may 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, $\mathrm{C}(J) \subseteq \mathrm{C}(A)$. By Frobenius (Theorem 15.32) we have

$$\dim \mathrm{C}(A) = \sum_{i=1}^{s} (2i-1)n_i = \dim \mathrm{C}(J).$$

It follows $\mathrm{C}(J) = \mathrm{C}(A)$. 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 19.20. By Corollary 13.20, 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 $\varphi_1, \ldots, \varphi_n \in \mathbb{R}$. By the spectral theorem there exists $S \in \mathrm{U}(n,\mathbb{C})$ with $U = S\,\mathrm{diag}(e^{\varphi_1 \mathrm{i}}, \ldots, e^{\varphi_n \mathrm{i}})S^*$. For the Hermitian matrix $H := S\,\mathrm{diag}(\varphi_1, \ldots, \varphi_n)S^*$ we have

$$U = \exp(H\mathrm{i}).$$

This relationship plays a role in quantum mechanics. If one drops the special choice of the eigenvalues in Theorem 19.19, one can choose $\varphi_1, \ldots, \varphi_n > 0$. Then H is even positive definite.

Remark 19.21.

(a) In analysis, one can compute $\ln(1-x)$ for $x \in \mathbb{R}$ with $|x| < 1$ by the *Mercator series*

$$\ln(1-x) = -\sum_{k=1}^{\infty} \frac{1}{k} x^k$$

By Theorem 19.11, for $\rho(A) < 1$ the corresponding series with A in place of x also converges. If A is nilpotent (i. e., $\rho(A) = 0$), the series terminates after $n-1$ summands. In that case, $B := \log(1-A)$ is the unique matrix from Theorem 19.19 (Exercise III.19).

(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 given A all solutions f have the form $f(t) = \exp(At)c$ with $c \in \mathbb{R}^n$ (note: $t \in \mathbb{R}$). This is a generalization of the well-known differentiation rule $(e^{at})' = ae^{at}$.

19.4 Nonnegative Matrices

Remark 19.22. In probability theory and other applied areas, real matrices with only nonnegative entries occur. We show that the eigenvalues and eigenvectors of such matrices have a special structure. This is the theoretical foundation of important applications such as the Google search algorithm.

Definition 19.23. We call $A = (a_{ij}) \in \mathbb{R}^{n\times m}$

- *positive* (resp. *nonnegative*), if $a_{ij} > 0$ (resp. $a_{ij} \geq 0$) holds for all i, j.
- *reducible*, if $n = m$ and there exists $\varnothing \neq I \subsetneq \{1, \ldots, n\}$ with $a_{ij} = 0$ for all $i \in I$ and $j \notin I$.
- *irreducible*, if $n = m$ and A is not reducible.

We write $A < B$ (resp. $A \leq B$) if $B - A$ is positive (resp. nonnegative). Moreover, let $A_+ := (|a_{ij}|)_{ij}$.

Remark 19.24. One easily sees that $\leq$ defines an order relation on $\mathbb{R}^{n\times m}$. We also use the notation $v \geq w$ and v_+ for vectors v, w (regarded 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 19.25.

(a) In stochastics, one studies the long-term development of random events by means of *Markov chains*. In the simplest case, a Markov chain consists of states $Z_1, \ldots, Z_n$. Let w_{ij} be the probability that a transition from Z_i to Z_j takes place. We call $W = (w_{ij})$ 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. Hence one is interested in $\lim_{k\to\infty} W^k$. Clearly, W is nonnegative and every row sum is 1. Matrices with this property are called (row) *stochastic*. Clearly, $v = (1, \ldots, 1)^{\mathrm{t}}$ is an eigenvector of W for the eigenvalue 1. By Exercise III.20, W^k is also stochastic for $k \geq 0$. From Lemma 18.62 and Theorem 19.11 it follows that

$$\rho(W) = \lim_{k\to\infty} \sqrt[k]{\|W^k\|_\infty} = 1.$$

(b) Clearly, 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. Indeed, if $I \subseteq \{1, \dots, n\}$ 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 exist 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) Together with A, also A^{t} is irreducible.

Theorem 19.26 (Perron)**.** *For every nonnegative matrix $A \in \mathbb{R}^{n\times n}$, $\rho(A)$ is an eigenvalue with a nonnegative 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 $\rho(A)^{-1}A \geq 0$, we may assume $\rho(A) = 1$. For $0 < t < 1$ and $m \in \mathbb{N}$ we have $\rho(tA) = t < 1$ and

$$(1 - tA)^{-1} \overset{19.8}{=} \sum_{k=0}^{\infty} (tA)^k \geq 1_n + tA + \ldots + (tA)^m.$$

If 1 is not an eigenvalue of A, then $1 - A$ is invertible and we have

$$(1 - A)^{-1} = (\lim_{t\to 1^-}(1 - tA))^{-1} = \lim_{t\to 1^-}(1 - tA)^{-1} \geq \lim_{t\to 1^-}(1_n + tA + \ldots + (tA)^m) = 1_n + A + \ldots + A^m$$

for all $m \in \mathbb{N}$. In particular, $\lim_{m\to\infty} A^m = 0$, contradicting Theorem 19.8.

Step 2: Every positive matrix A has a positive eigenvector for the eigenvalue $\rho(A)$.[4]
As above, we may assume $\rho(A) = 1$. Let $v = (v_1, \dots, v_n)$ be an eigenvector of A for the eigenvalue 1. Then

$$v_+ = (Av)_+ \leq A_+v_+ = Av_+,$$

so $w := (A - 1_n)v_+ \geq 0$. In the case $w = 0$, $v_+ = Av_+ > 0$ is a positive eigenvector of A. Now let $w \neq 0$. Then $Aw > 0$ because of $A > 0$. Hence there exists an $\epsilon > 0$ with $Aw \geq \epsilon Av_+$. For $z := Av_+ > 0$ we have

$$(A - 1_n)z = A(A - 1_n)v_+ = Aw \geq \epsilon z,$$

so $Az \geq (1 + \epsilon)z$. 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 $\lim_{k\to\infty} B^k = 0$ by Theorem 19.8. Contradiction.

Step 3: Every nonnegative matrix A has a nonnegative eigenvector for the eigenvalue $\rho(A)$.
For $k \in \mathbb{N}$ let $A_k := A + \frac{1}{k}J > 0$, where $J = (1)_{i,j=1}^n$. Clearly, $A_1 > A_2 > \ldots > A$ and $A_1^m > A_2^m > \ldots > A^m$ for all $m \in \mathbb{N}$. From Theorem 19.11 (applied, for example, with the Euclidean norm) it follows that $\rho(A_1) \geq \rho(A_2) \geq \ldots \geq \rho(A)$. In particular,

$$\mu := \lim_{k\to\infty} \rho(A_k) \geq \rho(A).$$

exists. By Step 2, there exist positive eigenvectors v_k of A_k for the eigenvalue $\rho(A_k)$. After normalization, $|v_k| = 1$. By Bolzano-Weierstrass, $(v_k)_k$ has a convergent subsequence. W. l. o. g. let $v := \lim_{k\to\infty} v_k \geq 0$. Since $|v| = 1$, we have $v \neq 0$. Moreover,

$$Av = \lim_{k\to\infty} A_k \lim_{k\to\infty} v_k = \lim_{k\to\infty} A_k v_k = \lim_{k\to\infty} \rho(A_k)v_k = \mu v.$$

Because $\mu \geq \rho(A)$, we obtain $\mu = \rho(A)$. □

[4] This is what Perron originally proved.

Theorem 19.27 (Perron-Frobenius)**.** *For every nonnegative irreducible matrix $A \in \mathbb{R}^{n\times n}$ we have:*

(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 nonnegative eigenvector of A.

Proof.

(a,b) By Perron's theorem, there exists an eigenvector $v \geq 0$ for the eigenvalue $\rho(A)$. Let $I := \{1 \leq i \leq n : v_i = 0\}$. For $i \in I$ we have

$$0 = \rho(A)v_i = (Av)_i = \sum_{j=1}^{n} a_{ij}v_j = \sum_{j\notin I} a_{ij}v_j.$$

This is only possible if $a_{ij} = 0$ for all $i \in I$ and $j \notin I$. Since A is irreducible, we must have $I = \varnothing$, i.e., $v > 0$.

Now suppose 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 in at least one coordinate. Since we have already seen that every nonnegative eigenvector for the eigenvalue $\rho(A)$ is positive, it follows that $w = \lambda v$. Hence at least the geometric multiplicity of $\rho(A)$ is equal to 1. The Jordan normal form of A therefore has only one block for $\rho(A)$. So it suffices 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

$$\lambda[w,u] = [Aw,u] = w^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[w,u]$$

we have $\lambda = \rho(A)$. Hence v is, up to scaling, the only nonnegative eigenvector of A. □

Remark 19.28. The next theorem generalizes the estimate $\rho(A) \leq \|A\|_\infty$ from Theorem 19.7 (put $x = (1,\dots,1)$ and use Lemma 18.62).

Theorem 19.29 (Collatz-Wielandt)**.** *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}_{\geq 0}$ be irreducible and $x > 0$. Then*

$$\min\Big\{\sum_{j=1}^{n} a_{ij}x_j/x_i : i = 1,\dots,n\Big\} \leq \rho(A) \leq \max\Big\{\sum_{j=1}^{n} a_{ij}x_j/x_i : i = 1,\dots,n\Big\}.$$

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

$$\sum_{j=1}^n (z_j - \rho(A))x_j u_j = \sum_{j=1}^n y_j u_j - \sum_{j=1}^n x_j \sum_{k=1}^n a_{kj}u_k = \sum_{j=1}^n y_j u_j - \sum_{k=1}^n u_k \sum_{j=1}^n a_{kj}x_j = 0.$$

Because $x, u > 0$, there exist $1 \le s, t \le n$ with $z_s \le \rho(A)$ and $z_t \ge \rho(A)$. Hence $\min_i z_i \le \rho(A) \le \max_i z_i$. □

Theorem 19.30 (VON MISES)**.** *Let $A \in \mathbb{R}^{n\times n}$ be nonnegative and irreducible. Then the sequence*

$$x_{k+1} := \frac{Ax_k}{|Ax_k|} \qquad (k = 1, 2, \ldots)$$

converges to a positive eigenvector for $\rho(A)$ for every positive starting vector $x_1 \in \mathbb{R}^{n\times 1}$. In particular, $\rho(A) = \lim_{k\to\infty} |Ax_k|$.

Proof. By Perron-Frobenius, the hypotheses of the power method (Theorem 18.72) are satisfied. However, we have to check 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 $x \in u^\perp$ we have

$$[Ax, u] = x^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[x, u] = 0.$$

This shows that $u^\perp$ is A-invariant. If we extend v by 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 $[x_1, u] > 0$, we have $x_1 \notin u^\perp$. The proof of Theorem 18.72 now shows that the power method converges for x_1. □

Lemma 19.31. *Let $A \ge 0$ be irreducible and λ 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. W. l. o. g. let $\rho(A) > 0$. Let $w \in \mathbb{C}^{n\times 1}$ be an eigenvector of A for the eigenvalue λ. For $z := w_+ \ge 0$ we have

$$\rho(A)z = (\lambda w)_+ = (Aw)_+ \le Az.$$

Let $u > 0$ with $A^{\mathrm{t}}u = \rho(A)u$. From

$$0 < \rho(A)[z, u] \le [Az, u] = z^{\mathrm{t}}A^{\mathrm{t}}u = \rho(A)[z, u]$$

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 $B := (b_{ij}) = \rho(A)\lambda^{-1}U^*AU \in \mathbb{C}^{n\times n}$ we have

$$Bz = \frac{\rho(A)}{\lambda}U^*Aw = \rho(A)U^*w = \rho(A)z = Az.$$

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

$$(Az)_i = (Bz)_i = \sum_{j=1}^n b_{ij}z_j \overset{\in\mathbb{R}}{=} \sum_{j=1}^n \mathrm{Re}(b_{ij})z_j \le \sum_{j=1}^n |\mathrm{Re}(b_{ij})|z_j \le \sum_{j=1}^n |b_{ij}|z_j = (Az)_i.$$

This is only possible if $B = B_+ = A$. □

Remark 19.32. Note that A and UAU^* in Lemma 19.31 have the same main diagonal. Hence 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 19.33. *For every positive matrix A, $\rho(A)$ is the only eigenvalue of absolute value $\rho(A)$.*

Proof. Follows from Remark 19.32. □

Example 19.34. We consider a Markov chain with transition matrix

$$W = \frac{1}{10}\begin{pmatrix} 5 & 1 & 4 \\ 2 & 4 & 4 \\ 0 & 9 & 1 \end{pmatrix}.$$

Clearly, W is irreducible. By Remark 19.32, $\rho(W) = 1$ is the only eigenvalue of absolute value 1. By Theorem 18.68, $W_\infty := \lim_{k\to\infty} W^k$ exists. To compute W_∞, by Corollary 18.69 we have to determine eigenvectors of W and W^{t}. Since W is stochastic, $v := (1,1,1)^{\mathrm{t}}$ is an eigenvector of W for the eigenvalue 1. One checks that $w := \frac{1}{91}(18,45,28)^{\mathrm{t}}$ is an eigenvector of W^{t} with $[v,w] = 1$. Now we have

$$W_\infty = vw^{\mathrm{t}} = \frac{1}{91}\begin{pmatrix} 18 & 45 & 28 \\ 18 & 45 & 28 \\ 18 & 45 & 28 \end{pmatrix}.$$

In the long run, we are therefore in state Z_1, Z_2, Z_3 with probability $\frac{18}{91} \approx 20\%$, 49%, 31%, respectively.

Theorem 19.35 (Wielandt)**.** *Let $A \in \mathbb{R}^{n\times n}$ be nonnegative and irreducible with exactly k eigenvalues of absolute value $\rho(A)$. Then:*

(a) The eigenvalues of absolute value $\rho(A)$ are $e^{2\pi \mathrm{i} j/k}\rho(A)$ for $j = 1,\ldots,k$, i. e., they are distributed evenly on a circle in the complex plane.

(b) If μ is an arbitrary 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 absolute value $\rho(A)$, i. e., $|\lambda_1| = \ldots = |\lambda_k| = 1$. By Lemma 19.31, for every i there exists a $U_i \in \mathrm{GL}(n,\mathbb{C})$ with $\lambda_i A = U_i^{-1}AU_i \approx A$ (unitarity is not needed). Comparing 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 $i = 1,\ldots,k$. Hence $\lambda_1,\ldots,\lambda_k$ are exactly the k-th roots of unity defined in Example 11.28 and Definition 18.6.

(b) W. l. o. g. 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 19.36. *Let $A \in \mathbb{R}^{n\times n}$ be nonnegative. Then for every eigenvalue $\lambda \in \mathbb{C}$ of A of absolute value $\rho(A)$ there exists an $m \leq n$ with $\lambda^m = \rho(A)^m$.*

Proof. By Example 19.25, there exists a permutation matrix P such that

$$P^{\mathrm{t}}AP = \begin{pmatrix} A_1 & & * \\ & \ddots & \\ 0 & & A_k \end{pmatrix}$$

with square nonnegative irreducible matrices $A_1, \ldots, A_k$. Every eigenvalue of A_i is also an eigenvalue of A (extend a corresponding eigenvector by 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 with Theorem 19.35. □

Example 19.37.

(a) Let P_σ be a permutation matrix and $\sigma = \sigma_1 \ldots \sigma_k$ the decomposition into pairwise disjoint cycles (including 1-cycles). Let l_i be the length of σ_i. By Example 15.19, P has the eigenvalues

$$\zeta_{l_1}, \zeta_{l_1}^2, \ldots, \zeta_{l_1}^{l_1}, \zeta_{l_2}, \ldots, \zeta_{l_2}^{l_2}, \ldots, \zeta_{l_k}^{l_k}$$

with $\zeta_m = e^{2\pi \mathrm{i}/m}$ for $m \in \mathbb{N}$.

(b) Now let $\sigma = (1, 3, 2)$ and

$$A = a1_3 + bP_\sigma + cP_{\sigma^2} = \begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

with $a, b, c \geq 0$.[5] Let $\zeta = \frac{-1+\sqrt{3}\mathrm{i}}{2} \in \mathbb{C}$ be a 3rd root of unity. Then $(1, \zeta, \zeta^2)^{\mathrm{t}}$ is an eigenvector of A for the eigenvalue

$$a + \zeta b + \zeta^2 c = a - \frac{1}{2}(b + c) + \frac{\sqrt{3}\mathrm{i}}{2}(b - c).$$

By a suitable choice of a, b, c, one can in this way realize every complex number as an eigenvalue of a nonnegative matrix.

Remark 19.38. By Schur-Horn, for given $\lambda_1, \ldots, \lambda_n \geq 0$ there always exists a positive semidefinite real matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. The determination of all possible sets of eigenvalues of (symmetric) nonnegative matrices, however, is an open problem.

19.5 PageRank

Remark 19.39. To sort results, the search engine Google assigns to every web page w a weight that measures how many other web pages link to w. We write $w_i \to w_j$ if the web page w_i contains a hyperlink to w_j. Let l_i be the number of (indexed) hyperlinks on w_i. We "define" the PageRank[6] of w_i by

$$p_i := \frac{1-d}{n} + d \sum_{\substack{1 \leq j \leq n \\ w_j \to w_i}} \frac{p_j}{l_j}, \tag{19.1}$$

[5] Matrices of this form are called *circulant*.

[6] Named after one of the Google founders, Larry Page (not after "web page").

where n is the number of all indexed pages and $0 < d < 1$ is a constant damping factor.[7] Thus the PageRank of w_i is "large" if many pages w_j (with high p_j and low l_j) link to w_i. Because of $d < 1$, isolated web pages (that are not linked to) cannot receive PageRank 0. The p_i cannot be computed directly from (19.1). It is not even clear whether they are uniquely determined by (19.1).

Definition 19.40. With the above notation, we call $G = (g_{ij})$ with

$$g_{ij} = \begin{cases} \frac{d}{l_i} + \frac{1-d}{n} & \text{if } w_i \to w_j, \\ \frac{1-d}{n} & \text{otherwise} \end{cases}$$

the *Google matrix*.

Theorem 19.41. *The Google matrix is stochastic. Moreover, $(p_1, \ldots, p_n)$ is the only 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 we have

$$\sum_{i=1}^{n} p_i = 1 - d + d\sum_{i=1}^{n} \sum_{\substack{1 \le j \le n \\ w_j \to w_i}} \frac{p_j}{l_j} = 1 - d + d\sum_{j=1}^{n} p_j \sum_{\substack{1 \le i \le n \\ w_j \to w_i}} \frac{1}{l_j} = 1 - d + d\sum_{j=1}^{n} p_j.$$

It follows that $\sum_{i=1}^{n} p_i = 1$. For $v := (p_1, \ldots, p_n)$ we have

$$(G^{\mathrm{t}}v)_i = \sum_{j=1}^{n} g_{ji} p_j = \frac{1-d}{n}\sum_{j=1}^{n} p_j + \sum_{\substack{1 \le j \le n \\ w_j \to w_i}} \frac{dp_j}{l_j} = p_i,$$

i. e., v is an eigenvector of G^{t} for the eigenvalue $\rho(G^{\mathrm{t}}) = \rho(G) = 1$. By Perron-Frobenius, v is uniquely determined by $\sum p_i = 1$. □

Remark 19.42.

(a) In practice, p_i is determined (approximately) with the von Mises theorem. One can show that $|\lambda| \le d$ for all eigenvalues $\lambda \ne 1$ of G. Hence the power method converges at least with the factor $\frac{1}{d}$ (see the proof of Theorem 18.72).

(b) One can also regard 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 visits a random other web page with probability $1 - d$. Once the p_i have been determined, one can use Corollary 18.69 to find the long-term probabilities of the surfer's whereabouts (see Example 19.34).

[7] Google uses $d = 0.85$. The exact value of n is unknown, but should be larger than 10^{10}.

20 Linear Optimization

20.1 Linear Programs

Remark 20.1.

(a) In linear optimization one maximizes (or minimizes) an *objective function* (for example the profit or the costs of a company) subject to constraints (for example demand or capacity).

(b) For the sake of uniformity, we will only use column vectors. In contrast to earlier chapters, let $\mathbb{R}^n$ be the space of real *column* vectors.

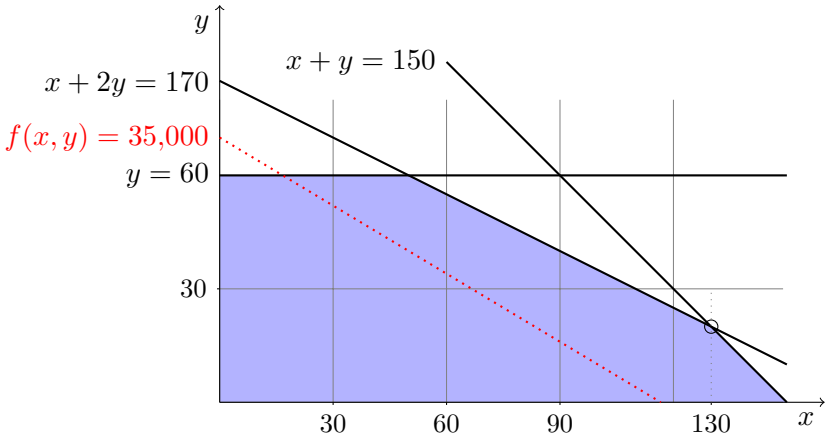
Example 20.2. A company produces two products P and Q using three machines A, B, C. In the production period, the machines have a maximum operating time of 170 hours (A), 150 hours (B) and 180 hours (C), respectively. The production of P and Q requires the following resources:

Product	Time A	Time B	Time C	Revenue
P	$1h$	$1h$	0	300 €
Q	$2h$	$1h$	$3h$	500 €

In which ratio should one produce units of P and Q in order to maximize the total revenue? Let x and y be the numbers 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 subject to the constraints

$$x, y \geq 0, \qquad x + 2y \leq 170, \qquad x + y \leq 150, \qquad 3y \leq 180$$

These inequalities describe the region in $\mathbb{R}^2$ marked in blue:



The parameters for a revenue of $f(x, y) = 35{,}000$ are marked by the red dotted line. Since the slope $-3/5$ of this line does not depend on the revenue, one easily sees that f is maximal for $(x, y) = (130, 20)$. The revenue is then 49,000 €. We are lucky that the solution (x, y) is unique and integral here. For problems with significantly more parameters, this graphical approach is unsuitable.

Remark 20.3.

(a) 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, we may assume $a = 0$. Because of $\max f = -\min(-f)$, we can restrict ourselves to the search for maxima.

(b) 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 $c^{\mathrm{t}}x \geq b \iff (-c)^{\mathrm{t}}x \leq -b$, we can do without the third variant. By adding a so-called *slack variable* y, we can replace $c^{\mathrm{t}}x \leq b$ by $c^{\mathrm{t}}x + y = b$ and $y \geq 0$. Slack variables do not occur in the objective function. In general, we can assume $x_i \geq 0$ for $i = 1, \ldots, m$ by replacing x_i with two variables $x_i^+, x_i^- \geq 0$ such that $x = x_+ - x_-$. All constraints can now be summarized in matrix form $Ax = b$ with $b \in \mathbb{R}^n$ and $x \geq 0$. With the Gaussian algorithm, we can transform A into reduced row echelon form and delete zero rows. In particular, we can assume that A has full rank. In the case $n \geq m$, $Ax = b$ can have at most one solution by 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, we achieve $b \geq 0$.

(c) Alternatively, constraints of the form $c^{\mathrm{t}}x = b$ can be split into $c^{\mathrm{t}}x \leq b$ and $-c^{\mathrm{t}}x \leq -b$. This yields $Ax \leq b$ without adding slack variables and without restrictions on x. We will come back to this in Definition 20.14.

Definition 20.4. A *linear program* (in standard form) $L = (A, b, c)$ consists of $A \in \mathbb{R}^{n\times m}$ of rank $n < m$, $b \in \mathbb{R}^n$ with $b \geq 0$ and $c \in \mathbb{R}^m$. A nonnegative vector $x \in \mathbb{R}^m$ is called *feasible* if x satisfies the constraint $Ax = b$. The goal is to find 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 $x \in \mathbb{R}^m$. Moreover, 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 20.5. Let L be a linear program with objective function f. Suppose that 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, we have $f(x_{\max}) = s = \max_{x\in M} f(x)$, i. e., L is solvable. This shows that every linear program is either solvable or infeasible or unbounded.

Example 20.6. The problem from Example 20.2 can be written as a linear program with three slack variables:

$$A = \begin{pmatrix} 1 & 2 & 1 & . & . \\ 1 & 1 & . & 1 & . \\ . & 3 & . & . & 1 \end{pmatrix}, \qquad b = (170, 150, 180)^{\mathrm{t}}, \qquad c = (300, 500, 0, 0, 0)^{\mathrm{t}}.$$

20.2 Convex Sets

Definition 20.7.

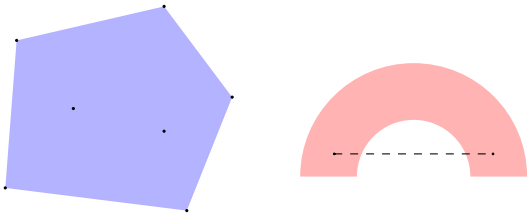
- A subset $\Delta \subseteq \mathbb{R}^n$ is called *convex*, if $\lambda x + (1-\lambda)y \in \Delta$ for all $x, y \in \Delta$ and $0 \leq \lambda \leq 1$. Intuitively, this means that for any two points in Δ, the line segment connecting them also lies in Δ.

- For $v_1, \ldots, v_k \in \mathbb{R}^n$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ with $\lambda_1 + \ldots + \lambda_k = 1$, we call $\lambda_1 v_1 + \ldots + \lambda_k v_k$ a *convex combination* of $v_1, \ldots, v_k$.
- The set of all convex combinations of elements of Δ is called the *convex hull* of Δ and is denoted by $\mathrm{con}\Delta$ (or $\mathrm{con}(x_1, \ldots, x_k)$ if $\Delta = \{x_1, \ldots, x_k\}$). Clearly, $\mathrm{con}\Delta$ is the "smallest" convex subset containing Δ, i. e.,

$$\mathrm{con}\Delta = \bigcap_{\substack{\Delta \subseteq \Gamma \subseteq \mathbb{R}^n \\ \Gamma \text{ convex}}} \Gamma.$$

Example 20.8.

(a) Convex sets in $\mathbb{R}$ are (open or (half-)closed) intervals. In $\mathbb{R}^2$, convex sets have neither holes nor inward-pointing "kinks". Here is the convex hull of seven points (blue) and a nonconvex 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 *cross-polytope* $\mathrm{con}(\pm e_1, \ldots, \pm e_n)$ (for $n = 3$ this is an octahedron).

Theorem 20.9 (Carathéodory)**.** *Let $\Delta \subseteq \mathbb{R}^n$ and $x \in \mathrm{con}\Delta$. Then x is a convex combination of at most $n+1$ elements of 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 that $k > n+1$. We form the matrix $A \in \mathbb{R}^{(n+1)\times k}$ with columns $v_1, \ldots, v_k$ and the additional row $(1, \ldots, 1)$. Since $\mathrm{rk}\, A \leq n+1 < k$, there exists a $y \in \mathbb{R}^k \setminus \{0\}$ with $Ay = 0$. Hence $y_1 v_1 + \ldots + y_k v_k = 0$ 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 $\lambda_i' := \lambda_i\big(1 - \frac{y_i \lambda_s}{\lambda_i y_s}\big) \geq 0$ we have

$$\sum_{i \neq s} \lambda_i' v_i = \sum_{i=1}^k \lambda_i v_i - \frac{\lambda_s}{y_s} \sum_{i=1}^k y_i v_i = x,$$

$$\sum_{i \neq s} \lambda_i' = \sum_{i=1}^k \lambda_i - \frac{\lambda_s}{y_s} \sum_{i=1}^k y_i = 1$$

in contradiction to the choice of k. □

Example 20.10. In contrast to subspaces, $\mathrm{con}\Delta \subseteq \mathbb{R}^n$ cannot always be "spanned" (as a convex combination) by finitely (or countably) many elements of Δ. To see this, consider the convex unit disk $\Delta := \{x \in \mathbb{R}^2 : |x| \leq 1\}$. Suppose that there exist countably many $v_1, v_2, \ldots \in \Delta$ such that every element of 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 may assume $x \neq \pm v_i$ for $i \in \mathbb{N}$. Then w. l. o. g. we may assume $x = \lambda_1 v_1 + \ldots + \lambda_k v_k$ with $k \geq 2$, $\lambda_i > 0$ and $\lambda_1 + \ldots + \lambda_k = 1$. By the triangle inequality and the Cauchy-Schwarz inequality, we have

$$1 = [x, x] \leq \sum_{i=1}^k \lambda_i |[x, v_i]| \leq \sum_{i=1}^k \lambda_i |x||v_i| \leq \sum_{i=1}^k \lambda_i = 1$$

only if x is linearly dependent on every v_i and $|v_1| = \ldots = |v_k| = 1$. But then we would have $x = \pm v_1$. Hence one needs uncountably many elements of Δ in order to represent every element as a convex combination.

Lemma 20.11 (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) hold simultaneously with u and v, respectively, then we would have $0 = [u, v] = u_1 v_1 + \ldots + u_n v_n \geq u_1 v_1 > 0$. We argue by induction on n. For $n = 1$ we have $U = \mathbb{R}$ or $U^\perp = \mathbb{R}$. So let $n \geq 2$ and

$$\widetilde{U} := \big\{x \in \mathbb{R}^{n-1} : (x, 0) \in U\big\} \leq \mathbb{R}^{n-1},$$
$$\widehat{U} := \big\{x \in \mathbb{R}^{n-1} : \exists y \in \mathbb{R} : (x, y) \in U\big\} \leq \mathbb{R}^{n-1}.$$

For dimension reasons, there exists a $u_0 \in U$ with $U = (\widetilde{U} \times \{0\}) + \langle u_0 \rangle$. From Exercise II.7 it follows that

$$U^\perp = (\widetilde{U}^\perp \times \mathbb{R}) \cap u_0^\perp. \tag{20.1}$$

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 $u = (\hat{u}, u_n) \in U$ we have $[u, v] = [\hat{u}, \hat{v}] = 0$, 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. So let $y < 0$. By (20.1), there exists a $z \in \mathbb{R}$ with $v = (\tilde{v}, z) \in u_0^\perp \subseteq U^\perp$. Since $\hat{u}, \tilde{v} \geq 0$, we have
$$0 = [u, v] = [\hat{u}, \tilde{v}] + yz \geq yz$$
and $z \geq 0$. Hence (2) holds for v. □

Example 20.12. 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 20.13. *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 $b^{\mathrm{t}} y < 0$ and $A^{\mathrm{t}} y \geq 0$.

Proof. If x and y are solutions of (1) and (2), respectively, then we would have $0 \leq (A^{\mathrm{t}} y)^{\mathrm{t}} x = y^{\mathrm{t}}(Ax) = y^{\mathrm{t}} b < 0$. So 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 $U := \operatorname{Ker} S \leq \mathbb{R}^{m+1}$. Suppose that there exists an $\tilde{x} = (x_1, x)^{\mathrm{t}} \in U$ with $x \geq 0$ and $x_1 > 0$. After scaling, we may assume $x_1 = 1$. Then x^{t} is a solution of (1).

By Farkas' lemma, we may now assume that there exists a $\tilde{z} = (z_1, z)^{\mathrm{t}} \in U^\perp$ with $z \geq 0$ and $z_1 > 0$. By Example 11.16, we have $U^\perp = \{S^{\mathrm{t}} y : y \in \mathbb{R}^n\}$. So let $y \in \mathbb{R}^n$ with $S^{\mathrm{t}} y = \tilde{z}$. Then $b^{\mathrm{t}} y = -z_1 < 0$ and $A^{\mathrm{t}} y = z \geq 0$. Hence (2) holds for y. □

Definition 20.14. Let $L = (A, b, c)$ be a linear program. The *dual* linear program L^* of L minimizes the objective function $f^*\colon \mathbb{R}^n \to \mathbb{R}$, $y \mapsto b^{\mathrm{t}}y$ subject to the constraint $A^{\mathrm{t}}y \geq c$. Analogously to Definition 20.4, one defines solvability, feasibility and boundedness of L^*.

Example 20.15. The dual program of $L = (A, b, c)$ from Example 20.6 is

$$\min_{y\in\mathbb{R}^3}(170, 150, 180)y \qquad \begin{pmatrix} 1 & 1 & . \\ 2 & 1 & 3 \\ 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{pmatrix} y \geq \begin{pmatrix} 300 \\ 500 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Since $y, c \geq 0$, L^* is bounded. Clearly, 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

$$f^*(y) = c^{\mathrm{t}}y = 170y_1 + 150(300 - y_1) + 60(200 - y_1) = 57{,}000 - 40y_1.$$

The minimum $f^*(y) = 49{,}000$ is attained 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 20.16 (Duality theorem)**.** *For every linear program $L = (A, b, c)$ we have:*

(a) L is solvable if and only if L^* is solvable. For solutions $x_{\max}$ and $y_{\min}$ we have $f(x_{\max}) = f^*(y_{\min})$.

(b) If L (or L^*) is unbounded, then L^* (or 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

$$f(x) = c^{\mathrm{t}}x \leq y^{\mathrm{t}}Ax = y^{\mathrm{t}}b = f^*(y). \tag{20.2}$$

If L (or L^*) is unbounded, then L^* (or L) must therefore be infeasible.

(a) Let $x_{\max} \in \mathbb{R}^m$ be a solution of L. Let

$$A_1 := \begin{pmatrix} 0 & A \\ -1 & c^{\mathrm{t}} \end{pmatrix} \in \mathbb{R}^{(n+1)\times(m+1)}$$

and $U := \operatorname{Ker} A_1$. If there exists a $\binom{t}{w} \in \operatorname{Ker} A_1$ with $t > 0$ and $w \in \mathbb{R}^m_{\geq 0}$, then we would have $x_{\max} + w \geq 0$, $A(x_{\max} + w) = b$ 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 $v := -\frac{1}{s}u$ we finally have $A^{\mathrm{t}}v \geq c$, 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 that there exist $x \in \mathbb{R}^m$, $u \in \mathbb{R}^n$ and $s \in \mathbb{R}$ with $(x^t, u^t, s)A_2 \geq 0$ and positive first coordinate. Then

$$c^t x > u^t b, \qquad Ax \leq sb, \qquad Ax \geq sb, \qquad x, s \geq 0, \qquad A^t u \geq sc.$$

Let $s > 0$. By passing to $\frac{1}{s}(x^t, u^t, s)$, we may assume $s = 1$. Then x is feasible for L and u for L^*, but (20.2) is violated. Hence $s = 0$ 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 $c^t x = 0$) sufficiently large, (20.2) becomes invalid again. Therefore (x^t, u^t, s) cannot exist.

With Farkas' lemma, we find a nonnegative vector $(1, u_+, u_-, z, x, s) \in \operatorname{Ker} A_2$. This means

$$Ax = b, \qquad A^t(u_+ - u_-) = c + z \geq c, \qquad b^t(u_+ - u_-) = c^t x - s \leq c^t x.$$

Hence x is feasible for L and $u := u_+ - u_-$ is feasible for L^*. Because of $f(x) = c^t x \geq b^t u = f^*(u)$, (20.2) implies $s = 0$. Moreover, u must be a solution of L^* and $f(x_{\max}) = c^t x = b^t u = f^*(u)$.

Finally, let $y_{\min}$ be a solution of L^*. If (2) in Corollary 20.13 holds for y, then we would have $A^t(y_{\min} + y) \geq c$ and $f^*(y_{\min} + y) = b^t(y_{\min} + y) < b^t y_{\min} = f^*(y_{\min})$. Hence (1) must hold, i. e., L is solvable. □

Example 20.17. It can happen that both L and L^* are infeasible. For example for $A = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, $b = c = (1, 1)^t$.

20.3 The Simplex Algorithm

Definition 20.18.

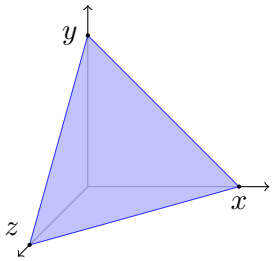
- The feasible set $P := \{x \in \mathbb{R}^m_{\geq 0} : Ax = b\}$ of a linear program $L = (A, b, c)$ describes a *polyhedron*. Clearly, P is convex. We call $x \in P$ a *vertex* of P or L if $P \setminus \{x\}$ is also convex.
- For $x \in P$ we call $\operatorname{supp} x := \{i : x_i > 0\} \subseteq \{1, \ldots, m\}$ 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 $I' := \{1, \ldots, m\} \setminus I$. We call I a *basis* of L if A_I is invertible. If in addition $A_I^{-1} b \geq 0$, we call I *feasible*.

Remark 20.19.

(a) In geometry one defines polyhedra 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 such as the feasible region in Example 20.2. 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" corners.

(b) We saw in Remark 20.3 how one can transform polyhedra of the form $\{x : Ax \leq b\}$ into polyhedra of the form $\{x \geq 0 : Ax = b\}$ by introducing slack variables. Hence the geometric picture of the polyhedra defined in Definition 20.18 refers to a proper subspace. For example,

$P = \{(x, y, z) \geq 0 : x + y + z = 1\}$ describes an equilateral triangle in a plane in $\mathbb{R}^3$ with vertices e_1, e_2 and e_3.



Lemma 20.20. *For every feasible point x of a linear program $L = (A, b, c)$ the following are equivalent:*

(1) x is a vertex.

(2) For $I := \operatorname{supp} x$ we have $\operatorname{rk} A_I = |I|$.

Proof.

(1) $\Rightarrow$ (2): Suppose that the columns of A_I are linearly dependent. Then there exists $y \in \mathbb{R}^m$ with $\operatorname{supp} y \subseteq I$ and $Ay = 0$. Moreover, there exists $\epsilon > 0$ with $x \pm \epsilon y \geq 0$. Now $x \pm \epsilon y$ lie in the polyhedron P of L and $x = \frac{1}{2}(x - \epsilon y) + \frac{1}{2}(x + \epsilon y)$. But then $P \setminus \{x\}$ would not be convex. Contradiction.

(2) $\Rightarrow$ (1): Suppose that $P \setminus \{x\}$ is not convex. Then there exist $y, z \in P \setminus \{x\}$ and $0 < \lambda < 1$ with $x = \lambda y + (1 - \lambda)z$. Since $y, z \geq 0$ we have $\operatorname{supp} y, \operatorname{supp} z \subseteq I$. 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, we would now have $y = z = x$. □

Example 20.21.

(a) In the case $b = 0$, $x = 0$ is a vertex with $I = \varnothing$.

(b) If I is a feasible basis 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 bases.

Corollary 20.22. *Every linear program in m variables has at most $\sum_{k=0}^{n} \binom{m}{k} \leq 2^m$ vertices.*

Proof. Every vertex x is uniquely determined by $I := \operatorname{supp} x$, since $A_I x_I = b$ has (at most) one solution by Lemma 20.20. The number of subsets $I \subseteq \{1, \ldots, m\}$ with $|I| \leq n$ is well 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 20.23. The estimate in Corollary 20.22 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. So one only needs to consider so-called *antichains* $\{I_1, \ldots, I_k\}$ of $\{1, \ldots, m\}$, i. e., $I_s \nsubseteq I_t$ for all $s \neq t$. A theorem of Sperner yields the sharper estimate $k \leq \binom{m}{\lfloor m/2 \rfloor}$.[1]

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 (every I contains exactly one element of $\{i, i + n\}$ for $i = 1, \ldots, n$).

[1] See lecture notes on logic and set theory

Theorem 20.24 (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 may assume that $x_{\max}$ is not a vertex. As in the proof of Lemma 20.20 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

$$f(x_{\max}) \pm \epsilon f(y) = f(x_\pm) \leq f(x_{\max})$$

we have $f(y) = 0$. We can choose ϵ such that $\operatorname{supp}(x_+) \subsetneq I$ or $\operatorname{supp}(x_-) \subsetneq I$ holds. In this case, x_+ or x_-, respectively, is a solution of L with 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 20.25. By Theorem 20.24 and Corollary 20.22, to solve a linear program one only has to consider finitely many (but possibly exponentially many) feasible vertices. Nevertheless, the maximum may be attained at infinitely many points, for example in the trivial case $c = 0$. The following theorem allows us to check whether a vertex that has been found is a solution.

Theorem 20.26 (Simplex criterion)**.** *Let $L = (A, b, c)$ be a linear program with feasible basis I. If*

$$\gamma(I) := c_{I'}^{\mathrm{t}} - c_I^{\mathrm{t}} A_I^{-1} A_{I'} \leq 0,$$

then the vertex belonging to I is a solution of L.

Proof. W. l. o. g. let $I = \{1, \ldots, n\}$ and $A = (A_I, A_{I'})$. For every feasible vector x we have $A_I x_I + A_{I'} x_{I'} = b$. 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 20.27. In practice, it is expensive to enumerate all vertices. The idea of the simplex method is to move from a vertex x to an "adjacent" vertex y with $f(y) > f(x)$. Intuitively, adjacent means that the line 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 20.28. *Let $L = (A, b, c)$ be a linear program with feasible basis I. Let x be the corresponding vertex (Example 20.21), $y := x_I$ and $M := (m_{ij}) = A_I^{-1} A_{I'}$. Let $j \in I'$ with $\gamma(I)_j > 0$.*[2] *Then:*

(a) If the j-th column of M is nonpositive (i. e., ≤ 0), then L is unbounded.

(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 and for the corresponding vertex x' we have $f(x') \geq f(x)$.

(c) If $y > 0$, then $f(x') > f(x)$ holds in (b).

[2] We index the components of $\gamma(I)$ and the columns of M with I' instead of $1, \ldots, n - m$.

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

$$f(z) = c_I^{\mathrm{t}} z_I + \mu c_{I'}^{\mathrm{t}} e_j = c^{\mathrm{t}} x - \mu c_I^{\mathrm{t}} M_j + \mu c_{I'}^{\mathrm{t}} e_j = c^{\mathrm{t}} x + \mu \gamma(I)_j \to \infty \qquad (\mu \to \infty).$$

Hence L is unbounded.

(b,c) Clearly, $S := A_I^{-1} A_J = (e_1, \dots, e_{i-1}, M_j, e_{i+1}, \dots, e_n)$. By expansion along the i-th row we get $\det S = m_{ij} > 0$. In particular, S is invertible. Hence A_J must be invertible as well, i. e., J is a basis. 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 we have $y_k \geq \frac{y_i}{m_{ij}} m_{kj}$ for $k \in I \setminus \{i\}$. This shows $x' \geq 0$. Moreover,

$$\begin{aligned} A_I^{-1} A x' = A_I^{-1} A_J x'_J &= \sum_{k \neq i} \Big(y_k - \frac{y_i}{m_{ij}} m_{kj} \Big) e_k + \frac{y_i}{m_{ij}} M_j \\ &= \sum_{k \in I} \Big(y_k - \frac{y_i}{m_{ij}} m_{kj} \Big) e_k + \frac{y_i}{m_{ij}} M_j = y = x_I \end{aligned}$$

and $Ax' = A_I x_I = Ax = b$. Hence x' is a feasible vertex for J. Finally we have

$$f(x') = \sum_{k \in I} c_k \Big(y_k - \frac{y_i}{m_{ij}} m_{kj} \Big) + \frac{y_i}{m_{ij}} c_j = f(x) - \frac{y_i}{m_{ij}} c_I^{\mathrm{t}} M_j + \frac{y_i}{m_{ij}} c_j = f(x) + \frac{y_i}{m_{ij}} \gamma(I)_j \geq f(x)$$

with equality if and only if $y_i = 0$. □

Remark 20.29 (Simplex algorithm)**.** Let a linear program $L = (A, b, c)$ be given.

(1) One determines a feasible basis I_1 and the corresponding vertex x_1. If necessary, this can be achieved by appending b as a column of A and extending $\{m+1\}$ to a basis.

(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 we can apply Lemma 20.28. If Lemma 20.28(a) holds, then L is unbounded and we can stop.

   (c) Otherwise Lemma 20.28(b) yields a vertex x_{k+1} with $f(x_{k+1}) \geq f(x_k)$.

(3) In practice, it can happen that the loop gets into an infinite cycle of vertices $x_k, x_{k+1}, \dots, x_l, x_k, \dots$. In this case one can start again with a new initial vertex x_1 or build in artificial perturbations.

One can construct examples in which the simplex algorithm runs through all (exponentially many) vertices, but in practice this occurs only very rarely.

Example 20.30. We again consider the linear program from Example 20.6:

$$A = \begin{pmatrix} 1 & 2 & 1 & . & . \\ 1 & 1 & . & 1 & . \\ . & 3 & . & . & 1 \end{pmatrix}, \qquad b = (170, 150, 180), \qquad c = (300, 500, 0, 0, 0).$$

Clearly, $I_1 = \{3, 4, 5\}$ is a feasible basis with vertex $x_1 = (0, 0, 170, 150, 180)^{\mathrm{t}}$, $f(x_1) = 0$ and $A_{I_1} = 1_3$. We have $\gamma(I) = (300, 500)$ and

$$M_1 = A_{I_1'} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 0 & 3 \end{pmatrix}.$$

We can choose $j = 1$ in Lemma 20.28. Then $i = 4$ because of $170 > 150$. Now $I_2 = \{1, 3, 5\}$ and $x_2 = (150, 0, 20, 0, 180)^{\mathrm{t}}$ with $f(x_2) = 45{,}000$. We compute

$$M_2 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 1 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 3 & 0 \end{pmatrix}$$

and $\gamma(I_2) = (500, 0) - (300, 0, 0)M = (200, -300)$. Here one must choose $j = 2$ and $i = 3$. So $I_3 = \{1, 2, 5\}$ and $x_3 = (130, 20, 0, 0, 120)^{\mathrm{t}}$ with $f(x_3) = 49{,}000$. Next,

$$M_3 = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 & 0 \\ 1 & -1 & 0 \\ -3 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \\ -3 & 3 \end{pmatrix}$$

and $\gamma(I_3) = (0, 0) - (300, 500, 0)M = (-200, -100)$. By the simplex criterion, x_3 is a solution. Except for the slack variable x_5, this is the solution found in Example 20.2.

21 Lattices and Quadratic Forms

21.1 Lattices

Remark 21.1. The vector space axioms do not use that the field K has multiplicative inverses. Hence one can replace K by an arbitrary ring R and then speaks of *R-modules* (more precisely: *left modules*). In general, the theory of modules is arbitrarily complicated and has little to do with linear algebra anymore. Even $\mathbb{Z}$-modules usually do not have a basis (for example, $\mathbb{F}_2$ is naturally a $\mathbb{Z}$-module in which every element is linearly dependent since $2 \cdot x = 0$ for $x \in \mathbb{F}_2$). In this section we study a family of $\mathbb{Z}$-modules inside $\mathbb{R}^n$ which have a basis by definition.

Definition 21.2. 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. In this case, B is called a *basis* of Δ and $\operatorname{rk} \Delta := m$ the *rank* of Δ. In the case $m = n$ we say that 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. Moreover,

$$_B[\Delta]_B := AA^{\mathrm{t}} = ([b_i, b_j])_{ij} \in \mathbb{R}^{m \times m}$$

is called the *Gram matrix* of Δ with respect to B.

Remark 21.3.

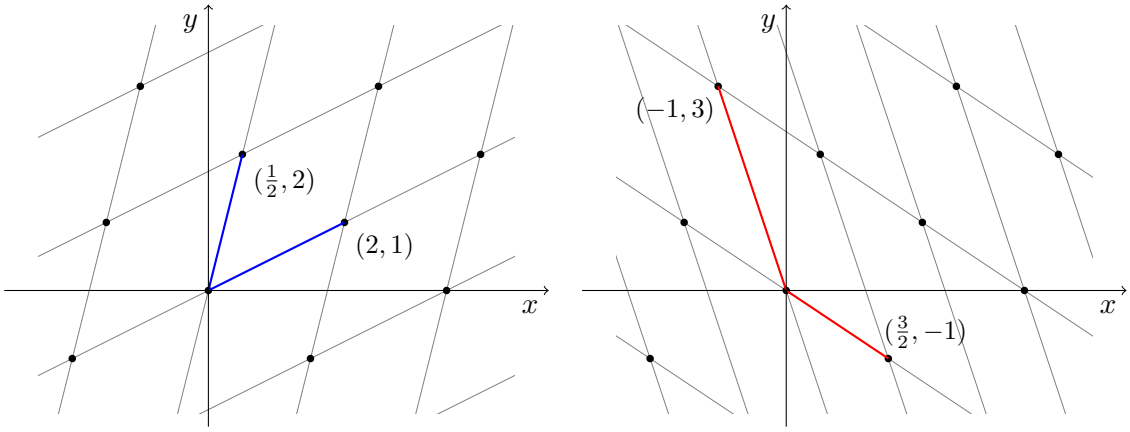
(a) For every basis B of a lattice Δ we have $|B| = \dim\langle B \rangle = \dim\langle \Delta \rangle$. Hence $\operatorname{rk} \Delta$ does not depend on the choice of the basis. By replacing $\mathbb{R}^n$ with $\langle \Delta \rangle$, we can often assume that Δ has full rank.

(b) Since the generator matrix A has full rank, the Gram matrix AA^{t} is positive definite (Lemma 12.40).

Example 21.4.

(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 with respect to B are 1_n.

(b) A portion 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 $x = yA$. We have

$$x \in \Delta^* \iff \forall b \in B : [x, b] \in \mathbb{Z} \iff xA^{\mathrm{t}} \in \mathbb{Z}^m \iff z := yAA^{\mathrm{t}} \in \mathbb{Z}^m.$$

For $S := (AA^{\mathrm{t}})^{-1}A = {}_B[\Delta]_B^{-1}A \in \mathbb{R}^{m\times n}$ we therefore have $\Delta^* = \{zS : z \in \mathbb{Z}^m\}$.[1] This shows that Δ^* is a lattice with generator matrix S. We call Δ^* the *dual* lattice of Δ. The Gram matrix of Δ^* with respect to 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 clearly *self-dual*, i. e., $\Delta^* = \Delta$. In general, $(\Delta^*)^* = \Delta$, because

$$(SS^{\mathrm{t}})^{-1}S = {}_B[\Delta]_B S = A.$$

Definition 21.5. 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 21.6. *For $A \in \mathbb{Z}^{n\times n}$ we have $A \in \mathrm{GL}(n, \mathbb{Z})$ if and only if $\det A = \pm 1$.*

Proof. For $A \in \mathrm{GL}(n, \mathbb{Z})$ we have $\det A \in \mathbb{Z}$ and $\det(A)^{-1} = \det(A^{-1}) \in \mathbb{Z}$, so $\det A = \pm 1$. The converse inclusion follows from Remark 9.24. □

Lemma 21.7. *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 $S\,{}_B[\Delta]_B S^{\mathrm{t}} = {}_C[\Delta]_C$.*

[1] Incidentally, $S^{\mathrm{t}} = A^+$ is the pseudoinverse of A.

Proof. Let $G_B, G_C \in \mathbb{R}^{m\times n}$ be the generator matrices of Δ with respect to 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 $\det({}_C[\Delta]_C) = \det(S)^2 \det({}_B[\Delta]_B) \geq \det({}_B[\Delta]_B)$. By symmetry we also have $\det({}_B[\Delta]_B) \geq \det({}_C[\Delta]_C)$. This shows $\det S = \pm 1$ and $S \in \mathrm{GL}(m,\mathbb{Z})$ by Lemma 21.6. □

Definition 21.8.

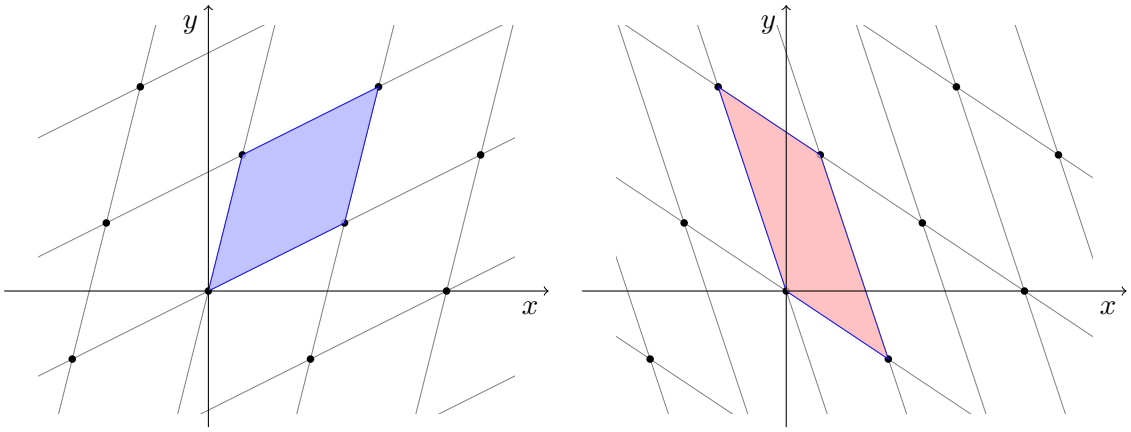
- A lattice Δ is called *integral* if ${}_B[\Delta]_B$ is an integer matrix for some basis B. By Lemma 21.7, this property does not depend on the choice of B.
- We call $\operatorname{disc}\Delta := \sqrt{\det({}_B[\Delta]_B)}$ the *discriminant* of Δ. This is also independent of B.

Example 21.9.

(a) If Δ has full rank with generator matrix A, then $\operatorname{disc}\Delta = |\det A|$. Geometrically, in this case $\operatorname{disc}\Delta$ is the volume of the so-called *fundamental parallelepiped*

$$ \{\lambda_1 b_1 + \ldots + \lambda_n b_n : 0 \leq \lambda_1, \ldots, \lambda_n \leq 1\} $$

for a basis $b_1, \ldots, b_n$ of Δ (cf. Example 9.2). In particular, $\operatorname{disc}\Delta = 7/2$ for the lattice from Example 21.4(b).



(b) A lattice Δ is integral if and only if $\Delta \subseteq \Delta^*$. In general, $\operatorname{disc}(\Delta^*) = \operatorname{disc}(\Delta)^{-1}$.

21.2 The Minimal Norm

Definition 21.10. 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 21.11.

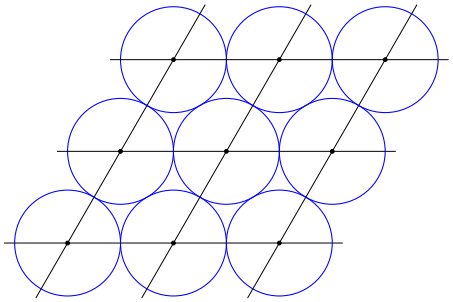
(a) Since lattices are groups, we have $\min\Delta = \min\big\{|x-y| : x, y \in \Delta, x \neq y\big\}$.

(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. By Lemma 18.55 we have $|x| \geq \sqrt{\lambda}|s|$. In particular, $\min \Delta \geq \sqrt{\lambda}$. Equality will usually not occur, since there need not be a normalized integer eigenvector of G. However, there are only finitely many vectors $x \in \Delta$ with $|x| = \min \Delta$. They are called *shortest* vectors. Both determining the shortest vectors and determining $\min \Delta$ are difficult algorithmic problems for $n > 10$. Modern encryption schemes such as LWE[2] are based on this; in contrast to classical schemes such as RSA, they cannot be broken by quantum computers so far.

(c) Harriot investigated in 1591 how to arrange (cannon)balls 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 parallelepiped $\operatorname{disc} \Delta$: $\rho(\Delta) = \frac{\tau_n}{\operatorname{disc} \Delta}$.

(1) For $n = 1$, $\Delta = \mathbb{Z}2$ is optimal with $\rho(\Delta) = 1$.

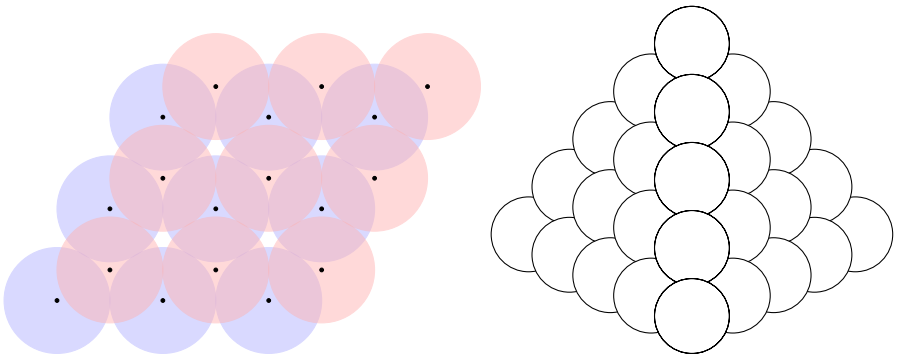
(2) For $n = 2$ one obtains an optimal arrangement from the hexagonal lattice $\Delta = \mathbb{Z}(2, 0) + \mathbb{Z}(1, \sqrt{3})$



with $\min \Delta = 2$ and $\rho(\Delta) = \frac{\pi}{2\sqrt{3}} \approx 0.91$. The square lattice $\mathbb{Z}(2, 0) + \mathbb{Z}(0, 2)$ only has density $\pi/4 \approx 0.79$.

(3) For $n = 3$ one takes the optimal arrangement in the plane and stacks these layers on top of each other, each shifted:

$$\Delta = \mathbb{Z}(2, 0, 0) + \mathbb{Z}(1, \sqrt{3}, 0) + \mathbb{Z}(1, 1/\sqrt{3}, \sqrt{8/3})$$



We have $\min \Delta = 2$ and $\rho(\Delta) = \frac{\pi}{3\sqrt{2}} \approx 0.74$. Gauss proved that there is no lattice with higher density (Theorem A.14, cf. Exercise III.28). Kepler conjectured in 1611 that it is

[2] Learning with errors

impossible in general to arrange balls even more densely (even irregularly).[3] This was only proved in 1998 by HALES with the help of a computer.[4]

(4) For $4 \le n \le 7$ the densest sphere packing is not known. For $n = 8$ there exists the so-called E_8 lattice with generator and Gram matrix

$$\begin{pmatrix} 2 & . & . & . & . & . & . & . \\ -1 & 1 & . & . & . & . & . & . \\ . & -1 & 1 & . & . & . & . & . \\ . & . & -1 & 1 & . & . & . & . \\ . & . & . & -1 & 1 & . & . & . \\ . & . & . & . & -1 & 1 & . & . \\ . & . & . & . & . & -1 & 1 & . \\ 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 & 1/2 \end{pmatrix}, \quad \begin{pmatrix} 4 & -2 & . & . & . & . & . & 1 \\ -2 & 2 & -1 & . & . & . & . & . \\ . & -1 & 2 & -1 & . & . & . & . \\ . & . & -1 & 2 & -1 & . & . & . \\ . & . & . & -1 & 2 & -1 & . & . \\ . & . & . & . & -1 & 2 & -1 & . \\ . & . & . & . & . & -1 & 2 & . \\ 1 & . & . & . & . & . & . & 2 \end{pmatrix}.$$

One sees that E_8 consists of the vectors x and $x + \frac{1}{2}(1, \dots, 1)$ for $x \in \mathbb{Z}^8$ with $2 \mid x_1 + \ldots + x_8$. In particular, $\min E_8 = \sqrt{2}$. For the correspondingly scaled lattice we have $\rho(\sqrt{2}E_8) = \frac{\pi^4}{384} \approx 0.25$. VIAZOVSKA proved in 2016 that $\sqrt{2}E_8$ yields the densest arrangement. One year later, the case $n = 24$ was solved as well. For these achievements, Viazovska received the Fields Medal in 2022.

Example 21.12. For the lattice $\Delta = \mathbb{Z}(2, 1) + \mathbb{Z}(1/2, 2)$ from Example 21.4(b), the shortest vectors are $\pm(3/2, -1)$ with $\min \Delta = \sqrt{13}/2$.

Theorem 21.13. *A subgroup $U \le (\mathbb{R}^n, +)$ is a lattice if and only if for every $d \in \mathbb{N}$ there are only finitely many $x \in U$ with $|x| \le d$.*

Proof. For every lattice, the stated bound holds by Remark 21.11. Conversely, let $U \le (\mathbb{R}^n, +)$ such that the condition holds. Let $u_1, \dots, u_k \in U$ be a largest possible set of linearly independent vectors. Then $U \subseteq \langle u_1, \dots, u_k \rangle =: V$. In the case $k = 0$, $U = \{0\} = \langle \varnothing \rangle$ is a lattice of rank 0. Now suppose the claim has already been proved for $k - 1$. Since $W := U \cap \langle u_1, \dots, u_{k-1} \rangle$ also satisfies the hypothesis of the theorem, W is a lattice. Let $b_1, \dots, b_{k-1}$ be a basis of W. We consider

$$S := \{\lambda_1 b_1 + \ldots + \lambda_{k-1} b_{k-1} + \lambda u_k : 0 \le \lambda_1, \dots, \lambda_{k-1} < 1,\ 0 \le \lambda \le 1\} \cap U.$$

By the triangle inequality, S is bounded and hence finite by hypothesis. Since $u_k \in S$, there exists a $b_k = \lambda_1 b_1 + \ldots + \lambda_{k-1} b_{k-1} + \lambda u_k \in S$ with $\lambda > 0$ minimal. Clearly, $b_1, \dots, b_k \in U$ are linearly independent. Every element $u \in U$ has the form $u = \mu_1 b_1 + \ldots + \mu_k b_k$ with $\mu_1, \dots, \mu_k \in \mathbb{R}$. Let $x_k \in \mathbb{Z}$ with $0 \le \mu_k - x_k < 1$. Then there exist $x_1, \dots, x_{k-1} \in \mathbb{Z}$ 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 $0 \le \lambda'_1, \dots, \lambda'_{k-1} < 1$ and $0 \le \lambda'_k < \lambda$. By the choice of b_k it follows that $\lambda'_k = 0$ and $\lambda'_1 b_1 + \ldots + \lambda'_{k-1} b_{k-1} \in W$. Since $b_1, \dots, b_{k-1}$ is a basis of W, we must have $\lambda'_1 = \ldots = \lambda'_{k-1} = 0$. This shows $u = x_1 b_1 + \ldots + x_k b_k$. Hence U is a lattice with basis $b_1, \dots, b_k$. □

Corollary 21.14. *If $\Delta, \Lambda \subseteq \mathbb{R}^n$ are lattices, then so is $\Delta \cap \Lambda$.*

[3] There are infinitely many ways to stack the planes shifted with the same density, but these are formally not lattices. For example, magnesium atoms form such an arrangement, which repeats periodically after two layers.

[4] Since 2017 there has been a computer-verified "formal" proof.

Proof. As is well known, $\Delta \cap \Lambda$ is a subgroup of $(\mathbb{R}^n, +)$. Like Δ, the intersection $\Delta \cap \Lambda$ can only contain finitely many elements of bounded norm. □

21.3 Integer Matrices

Remark 21.15. The study of integral lattices leads to matrices with integer entries. In Section 10.3 we already constructed matrices with polynomial entries and noted that the Gaussian algorithm is not applicable, since it would require division. With a little number theory[5] we develop a substitute for the Gaussian algorithm in $\mathbb{Z}^{n\times m}$, which allows us to solve systems of equations over $\mathbb{Z}$.

Definition 21.16. For $a, b \in \mathbb{Z}$ we write $a \mid b$ (a *divides* b) if there exists a $c \in \mathbb{Z}$ with $b = ac$. 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 $\operatorname{cd}(a_1, \ldots, a_n)$ be the set of common divisors of $a_1, \ldots, a_n$. We call $d \in \operatorname{cd}(a_1, \ldots, a_n)$ the *greatest common divisor* and write $\gcd(a_1, \ldots, a_n) := d$ if $d \geq 0$ and $e \mid d$ for all $e \in \operatorname{cd}(a_1, \ldots, a_n)$. In the case $d = 1$ we call $a_1, \ldots, a_n$ *coprime* (as for polynomials).

Remark 21.17.

(a) If g and g' are greatest common divisors of $a_1, \ldots, a_n$, then $g \mid g' \mid g$ and $g = \pm g'$. Since $g, g' \geq 0$, we obtain $g = g'$, i. e., there exists at most one greatest common divisor of $a_1, \ldots, a_n$ (this justifies the notation gcd).

(b) Despite its name, the gcd is not necessarily the *greatest* common divisor. For example, $\operatorname{cd}(0, 0) = \mathbb{Z}$, 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) Compare the following lemma with Lemma 15.6.

Lemma 21.18 (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. In the case $a_1 = \ldots = a_n = 0$ we have $\gcd(a_1, \ldots, a_n) = 0$, and $b_1 = \ldots = b_n = 1$ satisfies the claim. So let at least one $a_i \neq 0$. We consider the set

$$I := \mathbb{Z}a_1 + \ldots + \mathbb{Z}a_n = \{a_1b_1 + \ldots + a_nb_n : b_1, \ldots, b_n \in \mathbb{Z}\}.$$

Since $a_i^2 \in I$, the set I contains a positive element. Let

$$d := \min\{e \in I : e > 0\}$$

and let $b_1, \ldots, b_n \in \mathbb{Z}$ with $d = a_1b_1 + \ldots + a_nb_n$. For $e \in \operatorname{cd}(a_1, \ldots, a_n)$ we have $e \mid d$, hence also $\gcd(a_1, \ldots, a_n) \mid d$. Division with remainder yields for every i numbers $q_i \in \mathbb{Z}$ and $0 \leq r_i < d$ with $a_i = q_id + r_i$. Since $r_i = a_i - q_id \in I$, the minimality of d implies $r_i = 0$. Hence $d \in \operatorname{cd}(a_1, \ldots, a_n)$ and $d \mid \gcd(a_1, \ldots, a_n)$. This shows $\gcd(a_1, \ldots, a_n) = d = a_1b_1 + \ldots + a_nb_n$. Since $d \mid a_i$, the numbers $\frac{a_i}{d}$ are integers. From

$$\frac{a_1}{d}b_1 + \ldots + \frac{a_n}{d}b_n = 1$$

we obtain $\gcd(b_1, \ldots, b_n) = 1$. □

[5] More details can be found in my number theory lecture notes

Lemma 21.19. *For $a_1, \ldots, a_n \in \mathbb{Z}$ there exists an $A \in \mathbb{Z}^{n\times n}$ with first row (or column) $(a_1, \ldots, a_n)$ and $|\det A| = \gcd(a_1, \ldots, a_n)$.*

Proof. In the case $a_1 = \ldots = a_n = 0$ the matrix 0_n satisfies the claim. So let w. l. o. g. $a_1 \neq 0$ (swap columns/rows if necessary). Since the determinant is linear in every row (and column), we may 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 $|\det A_1| = \gcd(a_1, \ldots, a_{n-1}) > 0$. Let $d := \det A_1$. By Bézout there exist $s, t \in \mathbb{Z}$ with $ds + a_n t = 1$. Let $b_i := -a_i t/d \in \mathbb{Z}$ for $i = 1, \ldots, n-1$. We construct $A_2 \in \mathbb{Z}^{(n-1)\times(n-1)}$ from A_1 by deleting the first row of A_1 and appending $(b_1, \ldots, b_{n-1}) = -\frac{t}{d}(a_1, \ldots, a_{n-1})$ as the last row. Then $\det A_2 = (-1)^{n+1}t$. Let

$$A := \begin{pmatrix} & A_1 & & a_n \\ & & & 0 \\ b_1 & \cdots & b_{n-1} & s \end{pmatrix}.$$

By expansion along the last column we see

$$\det A = (-1)^{n+1} a_n \det(A_2) + s \det(A_1) = a_n t + ds = 1.$$

The analogous statement for columns follows by transposition. □

Remark 21.20. The next theorem provides an integer version of the reduced row echelon form (cf. Theorem 6.10).

Theorem 21.21 (Hermite normal form)**.** *For every matrix $A \in \mathbb{Z}^{n\times m}$ there exists exactly one nonnegative matrix $H = (h_{ij}) \in \mathbb{Z}^{n\times m}$ with the following properties:*

(a) $H = SA$ for some $S \in \mathrm{GL}(n, \mathbb{Z})$.

(b) Zero rows are located at the bottom of H.

(c) For $\tau_i := \min\{1 \leq j \leq m : h_{ij} \neq 0\}$ we have $\tau_1 < \tau_2 < \ldots$ (if defined) and $h_{j,\tau_i} < h_{i,\tau_i}$ 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 go through 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 21.19 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 $a_1 = d \geq 0$. Here $a_2, \ldots, a_n$ are replaced by integer linear combinations of $a_1, \ldots, a_n$. Hence $a_1 \in \mathrm{cd}(a_2, \ldots, a_n)$. 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 21.19 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, we have $a_i = 0$. In this way we achieve $a_2 = \ldots = a_n = 0$.

In the second column of A we can achieve $a_{22} \geq 0 = a_{32} = \ldots = a_{n2}$ analogously. In the case $a_{22} = 0$ we move on to the third column. In the case $a_{12} < 0$ or $a_{12} > a_{22}$ we add or subtract a_{22} from a_{12}, respectively, by multiplying with an elementary matrix from the left (this does not change the first column of A). By repeating this step sufficiently often, one obtains $0 \leq a_{12} < a_{22}$. Once all columns have been processed in this way, A (or H) has the desired properties.

Uniqueness: For the uniqueness of H we follow the proof of Theorem 6.10. We may assume w. l. o. g. 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 $a_i \in \langle e_1, \ldots, e_k \rangle$. We show $a_i = h_i$ and $s_k = e_k$ by induction on k. In the case $k = 0$ we have $a_i = 0$ and $h_i = Sa_i = 0$. Now suppose the claim has already been proved up to $k-1$. The first column (from the left) of A that does not lie in $\langle e_1, \ldots, e_{k-1} \rangle$ has the form $a_i = (a_{1i}, \ldots, a_{ki}, 0, \ldots, 0)^{\mathrm{t}}$ with $a_{ki} > 0$. By the induction hypothesis and $\tau_k < \tau_{k+1}$ we have $s_{jk} a_{ki} = h_{ji} = 0$ and $s_{jk} = 0$ for $j > k$. Since S is invertible, it follows that $s_{kk} \neq 0$ and $h_{ki} = s_{kk} a_{ki} \neq 0$. Hence we must have $i = \tau_k$ and $h_{ki} > 0$. For $j < k$ we have

$$0 \leq h_{ji} = a_{ji} + s_{jk} a_{ki} < h_{ki} = a_{ki}.$$

From $0 \leq a_{ji} < a_{ki}$ we obtain $s_{jk} = 0$. Since $\det S = \pm 1$, we have $s_{kk} = 1$ and $h_{ki} = a_{ki}$. This shows $s_k = e_k$ and $a_i = h_i$. □

Example 21.22. If one has a lattice Δ with 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 21.23.

(a) The properties of the Hermite normal form do not make use of integrality. Hence one can also transform rational matrices into Hermite normal form by temporarily multiplying by the common denominator of all entries and applying the algorithm to the resulting integer matrix.

(b) In the Hermite normal form one operates only on the rows (that is, from the left) of a given matrix. If one also allows column operations, one arrives at a uniquely determined diagonal matrix.

Theorem 21.24. *For $A \in \mathbb{Z}^{n \times m}$ there exist uniquely determined nonnegative 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 $T \in \mathrm{GL}(m, \mathbb{Z})$. We call $d_1, \ldots, d_{\min\{n,m\}}$ the elementary divisors *of A.*

Proof. Existence: For the zero matrix the claim is clear (note $0 \mid 0$). As in the proof of Theorem 21.21, we may assume $a_{11} > 0 = a_{21} = \ldots = a_{n1}$. The same procedure with the columns (that is, by multiplying 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 over 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. Here the divisibility by d is preserved in every 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. Hence

$$D_k(A) \mid D_k(AT) \mid D_k(ATT^{-1}) = D_k(A).$$

Similarly, $D_k(A) = D_k(SA)$ and therefore $D_k(A) = D_k((d_i \delta_{ij})) = d_1 \ldots d_k$. Hence the elementary divisors are uniquely determined by A. □

Remark 21.25.

(a) If A is square, then $|\det A|$ is the product of the elementary divisors of A. The proof above shows how one can characterize the elementary divisors by minors.

(b) In contrast to the Gaussian algorithm, the procedure for the Smith normal form is considerably more laborious.

(c) The Smith normal form allows us to solve integer linear systems of equations $Ax = b$. To this end, let $SAT = D = (d_i\delta_{ij})$ as in Theorem 21.24 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 given by $y = (b_1/d_1, \ldots, b_k/d_k, *, \ldots, *)$, where $*$ stands for an arbitrary integer. The solutions of $Ax = b$ are obtained by $x = Ty$.

Example 21.26. We look for all $x \in \mathbb{Z}^3$ with

$$Ax := \begin{pmatrix} 21 & -5 & 26 \\ 3 & -1 & 4 \\ -8 & 2 & -10 \end{pmatrix} x = \begin{pmatrix} 47 \\ 7 \\ -18 \end{pmatrix} =: b. \tag{21.1}$$

The algorithm in the proof of Theorem 21.24 (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

$$S := \begin{pmatrix} 1 & -2 & 2 \\ -1 & 1 & -2 \\ 1 & 1 & 3 \end{pmatrix} \in \mathrm{GL}(3, \mathbb{Z}), \qquad T := \begin{pmatrix} 2 & -1 & 1 \\ -1 & -1 & -1 \\ -2 & 0 & -1 \end{pmatrix} \in \mathrm{GL}(3, \mathbb{Z})$$

we have $SAT = D$. Hence (21.1) is equivalent to

$$Dy = Sb = \begin{pmatrix} -3 \\ -4 \\ 0 \end{pmatrix}$$

with $y := T^{-1}x$. This shows $y = (-3, -2, a)$ with $a \in \mathbb{Z}$. Hence $x = Ty = (a - 4, 5 - a, 6 - a)$ for $a \in \mathbb{Z}$.

Theorem 21.27. *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 $d_1 b_1, \ldots, d_l b_l$ is a basis of Λ.*

Proof. Existence: Let $a_1, \ldots, a_k$ and $c_1, \ldots, c_l$ be arbitrary bases of Δ and Λ, respectively, for now. Here $l = \operatorname{rk}\Lambda \leq \operatorname{rk}\Delta = k$. 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 $A := (x_{ij}) \in \mathbb{Z}^{l \times k}$. By Theorem 21.24 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. Since

$$c_i = \sum_{j=1}^{k} x_{ij} a_j = \sum_{j=1}^{k} \sum_{m=1}^{l} s_{im} d_m t_{mj} a_j = \sum_{m=1}^{l} s_{im} d_m b_m$$

we see that $\{d_1 b_1, \ldots, d_l b_l\}$ 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 $b'_i = \sum_{j=1}^{k} s_{ij} b_j$ for $1 \le i \le k$ and $d'_i b'_i = \sum_{j=1}^{l} t_{ij} d_j b_j$ for $1 \le i \le l$. Comparing coefficients shows $d'_i s_{ij} = t_{ij} d_j$ for $1 \le i, j \le l$ and $s_{ij} = 0$ for $l < j \le k$. For the $l \times k$ matrices we have

$$(\delta_{ij} d'_i) S = T(\delta_{ij} d_i).$$

From the uniqueness of the Smith normal form we obtain $d'_i = d_i$ for $i = 1, \ldots, l$. □

21.4 Theorems of Hermite and Minkowski

Lemma 21.28. *Every lattice $\Delta \subseteq \mathbb{R}^n$ has 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 $\lambda_1, \ldots, \lambda_k \in \mathbb{Z}$. By Lemma 21.19 there exists an $S \in \mathrm{GL}(k, \mathbb{Z})$ with first column $\lambda_1, \ldots, \lambda_k$. By a change of basis (Lemma 21.7) we may therefore assume $|b_1| = |v| = \min \Delta$. □

Theorem 21.29 (Hermite)**.** *For every lattice $\Delta \subseteq \mathbb{R}^n$ of rank k we have*

$$\boxed{\min \Delta \le \left(\frac{4}{3}\right)^{(k-1)/4} \sqrt[k]{\operatorname{disc} \Delta}.}$$

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. By Lemma 21.28 we may assume $\min \Delta = |b_1| = \sqrt{g_{11}}$. For $k = 1$ we have $\sqrt{g_{11}} = \operatorname{disc} \Delta$ as claimed. Let $k \ge 2$ and suppose the claim has already been proved 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 $2 \le i \le k$ are orthogonal to b_1. We consider the lattices Λ with basis $c_2, \ldots, c_k$ and Λ_1 with basis $b_1, c_2, \ldots, c_k$. For the Gram matrices we have $G_{\Lambda_1} = S^{\mathrm{t}} G S = \operatorname{diag}(g_{11}, G_\Lambda)$, 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 $\operatorname{disc} \Delta = \operatorname{disc} \Lambda_1 = \sqrt{g_{11}} \operatorname{disc}(\Lambda)$. 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

$$|y| = \min \Lambda \le \left(\frac{4}{3}\right)^{(k-2)/4} \left(\frac{\operatorname{disc} \Delta}{\sqrt{g_{11}}}\right)^{\frac{1}{k-1}}.$$

Moreover, there exists an $x_1 \in \mathbb{Z}$ with

$$\mu := \left| x_1 + \frac{g_{12}}{g_{11}} x_2 + \ldots + \frac{g_{1k}}{g_{11}} x_k \right| \le \frac{1}{2}.$$

One easily checks $S^{-1} = 2 \cdot 1_k - S$ (the entries off the main diagonal change sign). For $v := x_1 b_1 + \ldots + x_k b_k \in \Delta$ we therefore have

$$\begin{aligned} g_{11} = (\min \Delta)^2 \le |v|^2 &= x G x^{\mathrm{t}} = x S^{-\mathrm{t}} G_{\Lambda_1} S^{-1} x^{\mathrm{t}} = \mu^2 g_{11} + (x_2, \ldots, x_k) G_\Lambda (x_2, \ldots, x_k)^{\mathrm{t}} \\ &\le \frac{g_{11}}{4} + |y|^2 \le \frac{g_{11}}{4} + \Big(\frac{4}{3}\Big)^{(k-2)/2} \Big(\frac{\operatorname{disc}(\Delta)^2}{g_{11}}\Big)^{\frac{1}{k-1}}. \end{aligned}$$

Rearranging yields

$$g_{11} \le \Big(\frac{4}{3}\Big)^{k/2} \Big(\frac{\operatorname{disc}(\Delta)^2}{g_{11}}\Big)^{\frac{1}{k-1}}, \qquad g_{11}^k \le \Big(\frac{4}{3}\Big)^{k(k-1)/2} \operatorname{disc}(\Delta)^2. \qquad \square$$

Example 21.30. A lattice with Gram matrix $\left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ attains equality in the Hermite bound. We will replace the exponential factor $(4/3)^{(k-1)/4}$ by the polynomial factor $\sqrt{k}$.

Theorem 21.31 (Minkowski's linear forms theorem)**.** *Let $\Delta \subseteq \mathbb{R}^n$ be a lattice of full rank. For all $d_1, \ldots, d_n \in \mathbb{R}_{>0}$ with $d_1 \ldots d_n \ge \operatorname{disc} \Delta$ there exists an $x \in \Delta \setminus \{0\}$ with $|x_i| \le d_i$ 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: $\Delta \subseteq \mathbb{Z}^n$.
By the Hermite normal form we may assume that A is an upper triangular matrix. Then

$$d_1 \ldots d_n \ge \operatorname{disc} \Delta = |\det A| = |a_{11} \ldots a_{nn}|.$$

For integers $0 \le \alpha_i < |a_{ii}|$ and $0 \le \delta_i \le d_i$ we consider the system of equations $xA = \alpha + \delta$. We show by induction on n that for 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 $n \ge 2$ and $B := A_{nn}$. Inductively, there exist unique $\alpha_1, \ldots, \alpha_{n-1}$ 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

$$\alpha_n = x_1 a_{1n} + \ldots + x_n a_{nn} - \delta_n \equiv a_{1n} x_1 + \ldots + x_{n-1} a_{n-1,n} - \delta_n \pmod{|a_{nn}|}$$

(since $a_{ni} = 0$ for $i < n$, the unknown x_n occurs only in this equation).

Clearly there are $(\lfloor d_1 \rfloor + 1) \ldots (\lfloor d_n \rfloor + 1) > d_1 \ldots d_n$ possible vectors δ, but only $|a_{11} \ldots a_{nn}| \le d_1 \ldots d_n$ possible α. Hence for at least one α there must exist $\delta \ne \delta'$ such that $yA = \alpha + \delta$ and $zA = \alpha + \delta'$ have integer solutions $y \ne z$. For $x := (y - z)A \in \Delta \setminus \{0\}$ we have $|x_i| = |\delta_i - \delta'_i| \le d_i$ 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 $|x_i| \le m d_i$ for $i = 1, \ldots, n$. Clearly $\frac{1}{m} x \in \Delta$ satisfies the claim.

Case 3: $\Delta \subseteq \mathbb{R}^n$.
Let $A_1, A_2, \ldots \in \mathbb{Q}^{n \times n}$ with $\lim_{k \to \infty} A_k = A$ and $|\det A_k| \le |\det A|$ for $k \in \mathbb{N}$ (this is easily achieved by scaling a row). By Case 2 there exist $x_1, x_2, \ldots \in \mathbb{Z}^n \setminus \{0\}$ with $|(x_k A_k)_i| \le d_i$ for $k \in \mathbb{N}$ and $i = 1, \ldots, n$. Since $|x_k| = |(x_k A_k) A_k^{-1}| \le |d| |A_k^{-1}|$, the sequence $(x_k)_k$ is bounded. Hence it must contain a constant subsequence $x := x_{k_1} = x_{k_2} = \ldots$. Now we have

$$|(xA)_i| = \lim_{j \to \infty} |(x A_{k_j})_i| \le d_i$$

for $i = 1, \ldots, n$. $\square$

Remark 21.32. 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.29.

Corollary 21.33. *For every lattice Δ of rank k we have* $\boxed{\min \Delta \leq \sqrt{k}\sqrt[k]{\operatorname{disc} \Delta}.}$

Proof. Since the statement depends only on the Gram matrix of Δ, we may assume $\Delta \subseteq \mathbb{R}^k$ by Example 21.4. 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 that

$$\min \Delta \leq |x| \leq \sqrt{k \operatorname{disc}(\Delta)^{2/k}} = \sqrt{k}\sqrt[k]{\operatorname{disc} \Delta}. \qquad \square$$

Remark 21.34. A computation shows that the estimate from Corollary 21.33 is sharper than the Hermite 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 21.50). 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 21.59 and Theorem A.13. For $n = 8$ the lattice E_8 from Remark 21.11 yields the optimal value $\gamma_8 = 2$.

21.5 The LLL Algorithm

Remark 21.35. Many problems for lattices Δ can be solved algorithmically by constructing a basis of vectors that are as "short" and "nearly orthogonal" as possible. Over $K \in \{\mathbb{R}, \mathbb{C}\}$ one can always construct an orthonormal basis with the Gram-Schmidt process. Over $\mathbb{Z}$ the situation is more complicated. As a first step, we bring the Gram matrix into block diagonal form with blocks as small as possible.

Definition 21.36. 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 $\Delta_1 \subseteq \Delta_2^\perp$ (Definition 11.13). In this case, one calls $\Delta = \Delta_1 \perp \Delta_2$ an *orthogonal decomposition*. If $\Delta \neq \{0\}$ is not decomposable, then Δ is called *indecomposable*.

Remark 21.37. 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 21.38 (Eichler)**.** *Every lattice Δ has an orthogonal decomposition $\Delta = \Delta_1 \perp \ldots \perp \Delta_k$ into indecomposable lattices $\Delta_1, \ldots, \Delta_k$ which are uniquely determined up to the order.*

Proof. Since the rank can only decrease finitely many times, Δ can always be decomposed orthogonally into indecomposable lattices $\Delta_1, \ldots, \Delta_k$. We call $x \in \Delta \setminus \{0\}$ *minimal* if x is not the sum of shorter vectors, i. e., for $y, z \in \Delta$ with $x = y + z$ we have $|y| \geq |x|$ or $|z| \geq |x|$. Since only finitely many vectors in Δ have a given norm (Theorem 21.13), every $x \in \Delta$ can be written as a sum of minimal elements. Because of $|x_1 + \ldots + x_k|^2 = |x_1|^2 + \ldots + |x_k|^2$ for $x_i \in \Delta_i$, every minimal element lies in some 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 $y_j, z_j \in \Delta_j$ for $j = 1, \ldots, k$. Then $x = y_i + z_i$ and

$$|y|^2 = |y_1|^2 + \ldots + |y_k|^2 \geq |y_i|^2 \geq |x|^2$$

or $|z| \geq |x|$. Hence 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 hence linearly independent. Therefore, there can only be finitely many equivalence classes $M_1, \ldots, M_l \subseteq M$. Every M_i lies in some Δ_j and generates as a group a lattice $\Lambda_i \subseteq \Delta_j$ (the set of all integral linear combinations of elements of 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 we can decompose x into minimal elements (within Δ_j), at least one of which does not lie in M_i. Then Δ_j would have to contain further equivalence classes of M, and we would have an orthogonal decomposition $\Delta_j = \Lambda_i \perp \Gamma$. This contradicts the indecomposability of Δ_j. Hence $\Delta_j = \Lambda_i$. Similarly, $k = l$ and $\Delta_i = \Lambda_i$ for $i = 1, \ldots, k$ follow after renumbering. □

Theorem 21.39. *For every indecomposable integral lattice Δ of rank $m \geq 2$ we have* $\min \Delta \geq \sqrt{2}$.

Proof. Let $A = (a_{ij}) \in \mathbb{Z}^{m \times m}$ be the Gram matrix of Δ with respect to some basis. Suppose that there exists an $x \in \mathbb{Z}^m$ with $xAx^{\mathrm{t}} = 1$. Then $\gcd(x_1, \ldots, x_m) = 1$. By Lemma 21.19 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 $A_1 \in \mathbb{Z}^{(m-1)\times(m-1)}$. This contradicts Remark 21.37. □

Lemma 21.40. *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 obtained by applying the Gram-Schmidt process to $v_1, \ldots, v_k$ (without normalization). Then* $\operatorname{disc} \Delta = |b_1| \ldots |b_k|$.

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 $b_1, \ldots, b_k$. By the Gram-Schmidt process there exists a lower triangular matrix $R \in \mathbb{R}^{k \times k}$ with ones on the main diagonal and $RA = B$. Since $\det R = 1$, we have

$$\operatorname{disc} \Delta = \sqrt{\det(AA^{\mathrm{t}})} = \sqrt{\det(BB^{\mathrm{t}})} = \sqrt{\det(\mathrm{diag}(|b_1|^2, \ldots, |b_k|^2))} = |b_1| \ldots |b_k|. \qquad \square$$

Remark 21.41.

(a) In the following, let $\lfloor x \rceil \in \mathbb{Z}$ denote the rounded value of $x \in \mathbb{R}$, i. e., $|x - \lfloor x \rceil| \leq \frac{1}{2}$ (it does not matter whether $*.5$ is rounded up or down).

(b) Let $v_1, \dots, v_k \in \mathbb{R}^n$ be linearly independent and let $b_1, \dots, b_k$ be the corresponding Gram-Schmidt orthogonal basis. Then there exists a $w_k \in \langle b_1, \dots, b_{k-1}\rangle = \langle v_1, \dots, v_{k-1}\rangle$ with $v_k = b_k + w_k$. Hence b_k is the image of the projection of v_k onto $\langle b_1, \dots, b_{k-1}\rangle^\perp$. In particular, b_k does not change if we replace v_k by an element of $v_k + \langle v_1, \dots, v_{k-1}\rangle$.

Definition 21.42. Let $\frac{1}{4} < \delta < 1$. Let $v_1, \dots, v_k$ be a basis of a lattice $\Delta \subseteq \mathbb{R}^n$ and let $b_1, \dots, b_k$ be 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 21.43 (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) While $i \leq k$ repeat:

(a) For $j = i-1, i-2, \dots, 1$:

- *Replace v_i by $v_i - \lfloor \mu_{ij} \rceil v_j$.*
- *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 that v_i violates the Lovász condition, i. e., $|b_i|^2 < (\delta - \mu_{i,i-1}^2)|b_{i-1}|^2$. By the swap $v_i \leftrightarrow v_{i-1}$, the numbers $d_1, \dots, d_{i-2}$ remain untouched, while b_{i-1} is replaced by

$$b'_{i-1} = v_i - \sum_{j<i-1} \mu_{ij} b_j = b_i + \mu_{i,i-1} b_{i-1}$$

Because of $[b_{i-1}, b_i] = 0$, we have

$$|b'_{i-1}|^2 = |b_i|^2 + \mu_{i,i-1}^2 |b_{i-1}|^2 < \delta |b_{i-1}|^2.$$

Hence the new value of d_{i-1} is

$$|b_1| \dots |b_{i-2}||b'_{i-1}| < \sqrt{\delta} d_{i-1}$$

[6]Pronounced: triple-L algorithm. Named after A. Lenstra, H. Lenstra and L. Lovász.

In particular, this value is smaller than before. For $j \geq i$, clearly $v_1, \ldots, v_j$ and $v_1, \ldots, v_i, v_{i-1}, \ldots, v_j$ generate the same lattice Δ_j. Hence $d_i, \ldots, d_k$ do not change by Lemma 21.40. Altogether, D is therefore decreased at least by the factor $\sqrt{\delta} < 1$. By Lemma 21.40 and the Hermite bound, we have

$$d_j = \operatorname{disc} \Delta_j \geq \Big(\frac{3}{4}\Big)^{j(j-1)/4} \big(\min \Delta_j\big)^j \geq \Big(\frac{3}{4}\Big)^{j(j-1)/4} \big(\min \Delta\big)^j.$$

Hence D is bounded from below by a positive constant which depends only on Δ. This shows that the Lovász condition can be violated only finitely many times. Thus $i = k + 1$ is reached after finitely many steps and the algorithm terminates.

At the end of the algorithm, the Lovász condition holds for $v_1, \ldots, v_k$. In step (4), v_i is modified. By Remark 21.41, $b_1, \ldots, b_k$ remain unaffected by this. For $j = i - 1$ we replace 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 that $v_j \in b_j + \langle b_1, \ldots, b_{j-1} \rangle$). Afterwards, $|\mu_{i,i-1}| \leq 1/2$. For $j = i - 2$ we obtain analogously $|\mu_{i,i-2}| \leq 1/2$, while $\mu_{i,i-1}$ is not changed anymore, since j runs through the values in descending order. In the end, $|\mu_{ij}| \leq 1/2$ for all $j < i$, i. e., the length condition is satisfied. Altogether, $v_1, \ldots, v_k$ is δ-reduced. □

Remark 21.44. One can show that the LLL algorithm has polynomial running time in k. Larger values of δ (i. e., close to 1) yield shorter basis vectors at the cost of a longer running time. 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 running time (without proof).

Theorem 21.45. *Let $v_1, \ldots, v_k$ be a δ-reduced basis of a lattice $\Delta \subseteq \mathbb{R}^n$. Let $\rho := \frac{2}{\sqrt{4\delta - 1}}$. Then*

$$|v_1| \leq \rho^{k-1} \min \Delta, \qquad |v_1| \ldots |v_k| \leq \rho^{k(k-1)/2} \operatorname{disc}(\Delta).$$

Proof. Let $b_1, \ldots, b_k$ be the Gram-Schmidt orthogonal basis for $v_1, \ldots, v_k$. We have $\sigma := \rho^2 = \frac{1}{\delta - 1/4} > \frac{4}{3}$. From the δ-reducedness we obtain

$$|b_i|^2 \geq (\delta - \mu_{i,i-1}^2)|b_{i-1}|^2 \geq \sigma^{-1}|b_{i-1}|^2 \geq \ldots \geq \sigma^{j-i}|b_j|^2$$

for $j < i$. Let $v = \lambda_1 v_1 + \ldots + \lambda_i v_i = \eta_1 b_1 + \ldots + \eta_i b_i \in \Delta$ be a shortest vector with $\lambda_1, \ldots, \lambda_i \in \mathbb{Z}$, $\eta_1, \ldots, \eta_i \in \mathbb{R}$ and $\lambda_i \neq 0$. By the construction of $b_1, \ldots, b_k$ we have $\eta_i = \lambda_i$. In particular, $|\eta_i| \geq 1$. Since $b_1, \ldots, b_k$ are pairwise orthogonal, it follows that

$$(\min \Delta)^2 = |v|^2 = \eta_1^2 |b_1|^2 + \ldots + \eta_i^2 |b_i|^2 \geq |b_i|^2 \geq \sigma^{1-i}|b_1|^2 = \sigma^{1-i}|v_1|^2 \geq \sigma^{1-k}|v_1|^2$$

and $|v_1| \leq \rho^{k-1} \min \Delta$. We show by induction on i that

$$1 + \frac{1}{4} \sum_{j=1}^{i-1} \sigma^j \leq \sigma^{i-1}$$

For $i = 1$ we have $1 = \sigma^0$. Let $i \geq 2$ and suppose that the claim has already been proved for $i - 1$. 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, we therefore have

$$|v_i|^2 = \Big|b_i + \sum_{j=1}^{i-1} \mu_{ij} b_j\Big|^2 = |b_i|^2 + \sum_{j=1}^{i-1} \mu_{ij}^2 |b_j|^2 \le |b_i|^2 + \frac{1}{4}\sum_{j=1}^{i-1} \sigma^{i-j}|b_i|^2 \le \sigma^{i-1}|b_i|^2.$$

It follows that

$$|v_1| \dots |v_k| \le \rho^{k(k-1)/2} |b_1| \dots |b_k| \overset{21.40}{=} \rho^{k(k-1)/2} \operatorname{disc}(\Delta). \qquad \square$$

Remark 21.46. For $\delta = 3/4$ we obtain $\rho = \sqrt{2} \approx 1.41$. As $\delta \to 1$, ρ approaches the factor $2/\sqrt{3} \approx 1.15$ from the Hermite bound. We will come back to this in Remark 21.67. 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 21.47. 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 $\operatorname{disc}\Delta = 169$. The Hermite bound gives $\min\Delta \le \frac{4}{3}\sqrt[5]{169} \approx 3.72$. The LLL algorithm with $\delta = 3/4$ and $\delta = 9/10$, respectively, 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 $\min\Delta = \sqrt{7} \approx 2.65$.

21.6 Quadratic Forms

Remark 21.48. We change our perspective on lattices by neglecting the generator matrix and considering only Gram matrices instead. We then still know the rank, but no longer the ambient Euclidean space. The situation here is similar to that of 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 21.49.

- 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.

- We call q_C
  - *nondegenerate*, if $C \in \mathrm{GL}(n, \mathbb{R})$.
  - *positive* (respectively *negative*), if C is positive (respectively negative) definite, i. e., $q(x) > 0$ (respectively $q(x) < 0$) for all $x \neq 0$. If q is positive, we call
$$\min q := \min\bigl\{q(x) : x \in \mathbb{Z}^n \setminus \{0\}\bigr\}$$
the *minimum* of q.
  - *integral*, if $C \in \mathbb{Z}^{n\times n}$.
- We call $\det q := \det C$ 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 21.50.

(a) The symmetry of C is inessential, since for an arbitrary matrix $A \in \mathbb{R}^{n\times n}$, $C := \frac{1}{2}(A + A^{\mathrm{t}})$ is symmetric with $q_A = q_C$. If A is an integer matrix, however, C need not be an integer matrix. 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, then by passing to an equivalent form one can eliminate a variable and reduce the rank. We are mainly interested in positive quadratic forms (cf. Exercise III.31). 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 $q_G(x_1, \ldots, x_k) = |x|^2$ for all $x \in \Delta$, where $x_1, \ldots, x_k$ are the coefficients with respect to the chosen basis. By Lemma 21.7, a change of basis of Δ corresponds to passing to an equivalent quadratic form. Conversely, in Example 21.4 we saw 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 lattices with the same Gram matrix in different dimensions. Note that $\min q_G = (\min \Delta)^2$ and $\det q = \mathrm{disc}(\Delta)^2$. The Hermite bound therefore has the form
$$\min q \leq \Bigl(\frac{4}{3}\Bigr)^{(k-1)/2} \sqrt[k]{\det q}.$$

(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 by an equivalent form that is as "simple" as possible.

(e) As for lattices, for a given $d > 0$ there are only finitely many $x \in \mathbb{Z}^n$ with $q(x) \leq d$ (Remark 21.11). Moreover, by Lemma 21.28 there exists an equivalent form q_C with $c_{11} = \min q$.

Theorem 21.51. *Every integral quadratic form q is equivalent to a form q_C with Gram matrix*
$$C = \begin{pmatrix} * & * & & 0 \\ * & \ddots & \ddots & \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix} \geq 0$$
of tridiagonal shape.

Proof. The proof proceeds similarly to Theorem 21.24. 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

$$d := \gcd(a_2, \ldots, a_n) = a_2 b_2 + \ldots + a_n b_n.$$

By Lemma 21.19 there exists a $U \in \mathrm{GL}(n-1, \mathbb{Z})$ with first column $(b_2, \ldots, b_n)$. Let $S := \mathrm{diag}(1, U) \in \mathrm{GL}(n, \mathbb{Z})$. 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 $i = 3, \ldots, n$. Multiplying AS by S^{t} from the left does not change the first row. Since $S^{\mathrm{t}}AS$ is symmetric, the first column of A now has the same shape. One proceeds analogously with the rows $2, \ldots, n-1$. In the end, A has been transformed into the desired matrix C. □

Example 21.52.

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

Remark 21.53.

(a) Theorem 21.51 is not suitable for counting equivalence classes of quadratic forms, since infinitely many such *tridiagonal matrices* are equivalent to each other. For *binary* quadratic forms (i. e., $n = 2$), Theorem 21.51 merely yields a sign reduction. We treat this case in Theorem 21.59.

(b) An integral positive quadratic form q is called *unimodular* if $\det q = 1$. Let q_C be unimodular with $n \leq 5$. Since $(4/3)^2 < 2$, we have $\min q = 1$ by Hermite. As in Theorem 21.39, one constructs an equivalent form with $C = \mathrm{diag}(1, C_1)$ and $\det C_1 = 1$. One can now repeat the same argument with q_{C_1}. Eventually q is equivalent to q_{1_n}. In particular, up to equivalence, q is the only unimodular quadratic form of rank n. By a theorem of Mordell, this also holds for $n \in \{6, 7\}$. For $n = 8$, the E_8 lattice introduced in Remark 21.11 determines a unimodular form q with $\min q = 2$. In particular, q is not equivalent to q_{1_8}.

Theorem 21.54. *For all $n \in \mathbb{N}$ and $d \in \mathbb{R}$ there exist, up to equivalence, only finitely many integral positive quadratic forms q of rank n with $\det q \leq d$.*

Proof. Let q be a quadratic form of rank n with $\det q \leq d$. Let Δ be a corresponding lattice (Remark 21.50). Let $v_1, \ldots, v_n$ be a δ-reduced basis of Δ (for example for $\delta = 3/4$). By passing to an equivalent quadratic form, we may assume q_C with $c_{ii} = |v_i|^2$ for $i = 1, \ldots, n$. By Theorem 21.45, $c_{11} \ldots c_{nn}$ is bounded by a function of d. Since C is positive definite, we also have $c_{ii}c_{jj} - c_{ij}^2 > 0$, i. e., $|c_{ij}| < \sqrt{c_{ii}c_{jj}}$ for $i \neq j$. Hence there are only finitely many possibilities for C. □

Definition 21.55. A positive quadratic form q with Gram matrix $C = (c_{ij}) \in \mathbb{R}^{n \times n}$ is called (*Minkowski*)*-reduced*, if for $i = 1, \ldots, n$ we have:

- $c_{ii} \leq q(x)$ 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 21.56 (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 $1 \le i \le n$. Among the vectors $x \in \mathbb{Z}^n$ with $\gcd(x_i, \ldots, x_n) = 1$ we can choose, by Remark 21.50, one with minimal $q(x)$. By Lemma 21.19 there exists a $T \in \mathrm{GL}(n-i+1, \mathbb{Z})$ with first column $(x_i, \ldots, x_n)$. Now let

$$S := \begin{pmatrix} 1 & & & x_1 & 0 \\ & \ddots & & \vdots & \vdots \\ & & 1 & x_{i-1} & 0 \\ & & & & \\ & & & T & \end{pmatrix} \in \mathrm{GL}(n, \mathbb{Z}).$$

Replacing C by $S^{\mathrm{t}}CS$, the entries $c_{11}, \ldots, c_{i-1,i-1}$ remain unchanged, while c_{ii} now satisfies the first Minkowski condition. In this way one achieves that the condition holds for $c_{11}, \ldots, c_{nn}$. The signs of $c_{i,i+1}$ can be changed by $S = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_n)$ with $\epsilon_1, \ldots, \epsilon_n \in \{\pm 1\}$ (while $c_{11}, \ldots, c_{nn}$ remain unchanged). □

Example 21.57. Let q be the quadratic form belonging to the lattice $\Delta \subseteq \mathbb{R}^5$ from Example 21.47. The Gram matrix

$$\begin{pmatrix} 7 & 0 & -1 & 3 & 0 \\ 0 & 9 & 0 & 4 & 3 \\ -1 & 0 & 10 & 1 & 5 \\ 3 & 4 & 1 & 10 & 0 \\ 0 & 3 & 5 & 0 & 11 \end{pmatrix}$$

defines an equivalent reduced form. Compared to the δ-reduction, the diagonal entries have become even smaller.

Remark 21.58.

(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 $c_{11} \le q(y_1, \ldots, y_n) = g^{-2} q(x) \le q(x)$. This shows $c_{11} = \min q$. Moreover, we have $c_{11} \le q(e_2) = c_{22} \le q(e_3) = c_{33} \le \ldots \le c_{nn}$. For $i < j$ we have

$$c_{jj} \le q(e_i \pm e_j) = c_{ii} \pm 2c_{ij} + c_{jj},$$

i. e., $2|c_{ij}| \le c_{ii}$.

(b) By Remark 21.50, only finitely many x can violate the inequality $c_{ii} \le q(x)$. In fact, the number of Minkowski conditions to be checked depends only on n and not on q (without proof). However, this number grows exponentially with n. For $n \ge 10$ it is very costly to reduce a given quadratic form. The KORKINE-ZOLOTAREV reduction provides a compromise between the efficiency of the LLL algorithm and the quality of the Minkowski reduction. We do not go into detail here.

(c) The diagonal entries c_{ii} of a reduced quadratic form q_C are uniquely determined by the Minkowski condition. Moreover, there are only finitely many vectors $x \in \mathbb{Z}^n$ with $q_C(x) = c_{ii}$. Hence there are 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 fixing an order on the matrix entries (e.g., among all equivalent forms let c_{12} be minimal, etc.). For $n \le 2$ there is in general only one reduced form per equivalence class.

Theorem 21.59 (GAUSS)**.** *A positive quadratic form $q = q_C$ is reduced if and only if $0 \le 2c_{12} \le c_{11} \le c_{22}$. In this case, $c_{11}c_{22} \le \frac{4}{3}\det(q)$. 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 $0 \le 2b \le a \le c$ by Remark 21.58. Conversely, let $0 \le 2b \le a \le c$ be given. For $(x, y) \in \mathbb{Z}^2 \setminus \{0\}$ we have

$$\begin{aligned} q(x,y) &= ax^2 + 2bxy + cy^2 \ge a(x^2+y^2) + 2bxy \\ &\ge a(x^2+y^2) - a|xy| = a(|x|-|y|)^2 + a|xy| \ge a \end{aligned} \tag{21.2}$$

with equality if $(x, y) = (1, 0)$. Hence $a = \min q \le q(x_1, x_2)$ for $\gcd(x_1, x_2) = 1$. Moreover, $c = q(0, 1) \le q(x_1, x_2)$ for $\gcd(x_2) = 1$. So q is reduced.

For the uniqueness, let $S = \left(\begin{smallmatrix} x & u \\ y & v \end{smallmatrix}\right) \in \mathrm{GL}(2, \mathbb{Z})$ with $S^{\mathrm{t}} C S = \left(\begin{smallmatrix} a & b' \\ b' & c' \end{smallmatrix}\right)$ and $0 \le 2b' \le a \le c'$ (note that $a = \min q$). W. l. o. g. let $c' \le c$. For (x, y) equality holds in (21.2). There are two cases:

Case 1: $y \ne 0$.
From (21.2) we get $c = a \le c' \le c$. So $c = c'$. From $a^2 - b^2 = \det C = a^2 - (b')^2$ and $b' \ge 0$ we obtain $b = b'$.

Case 2: $(x, y) = (\pm 1, 0)$.
By replacing S with $-S$, we may assume $x = 1$. Then $v = \det S = \pm 1$. From

$$ua - \frac{a}{2} \le ua \pm b = b' \le \frac{a}{2}$$

we obtain $u \le 1$. In the case $u < 0$ we would have $b' < 0$. For $u = 0$ we have $S = \mathrm{diag}(1, \pm 1)$ and $C = D$. So let $u = 1$. Then $b = a/2 = b'$. From $ac - b^2 = \det C = ac' - b^2$ we get $c = c'$. □

Remark 21.60.

(a) The bound $c_{11}c_{22} \le \frac{4}{3}\det(q)$ corresponds to the estimate from Theorem 21.45 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 we have

$$\frac{|c_{12}|}{c_{11}} = \frac{|[v_2, v_1]|}{|v_1|^2} = |\mu_{21}| \le \frac{1}{2},$$

i. e., $2|c_{12}| \le c_{11}$. The (only) Lovász condition reads

$$(\delta - \mu_{21}^2)|v_1|^2 \le |b_2|^2 = |v_2 - \mu_{21}v_1|^2 = |v_2|^2 - 2\mu_{21}[v_1, v_2] + \mu_{21}^2|v_1|^2 = |v_2|^2 - \mu_{21}^2|v_1|^2,$$

i. e., $\delta c_{11} \le c_{22}$. 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, $2|c_{12}| \le c_{11}$ holds.

(3) If $c_{11} > c_{22}$, go to (1).

(4) Replace c_{12} by $|c_{12}|$.

(d) The equivalence classes of positive quadratic forms q_C of rank 2 and $\det q \le 10$ 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 $n \le 5$ on the internet (cf. Exercise III.33).[7]

[7] `https://www.math.rwth-aachen.de/homes/Gabriele.Nebe/LATTICES/`

(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 $k \in \{0, 1\}$. By Example 21.52, q_0 and q_1 are equivalent. By Example 12.35, q_0 is positive and $\min q_0 = 2$. Since $c_{ii} = 2 = \min q_k$ for $i = 1, 2, 3$ and $c_{12} = c_{23} = 1$, both q_0 and q_1 are reduced.

21.7 Successive Minima

Definition 21.61. Let $q\colon \mathbb{Z}^n \to \mathbb{R}$ be a positive quadratic form. Following Courant-Fischer, we define the *successive minima*

$$\mu_i(q) := \min_{\substack{s_1,\ldots,s_i \in \mathbb{Z}^n \\ \text{linearly independent}}} \max\{q(s_1), \ldots, q(s_i)\}.$$

Remark 21.62. Clearly, $\min q = \mu_1(q) \leq \ldots \leq \mu_n(q)$. If q_C is reduced, then we have

$$\mu_i(q) \leq \max\{q(e_1), \ldots, q(e_i)\} = c_{ii}$$

for $i = 1, \ldots, n$. In general, however, equality does not necessarily hold.

Example 21.63. Let

$$C = \begin{pmatrix} 1 & . & . & . & 1/2 \\ . & 1 & . & . & 1/2 \\ . & . & 1 & . & 1/2 \\ . & . & . & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1/2 & 5/4 \end{pmatrix}.$$

Since

$$q(x) = x_1^2 + \ldots + x_4^2 + \frac{5}{4}x_5^2 + x_1x_5 + \ldots + x_4x_5 = \sum_{i=1}^{4}\Big(x_i + \frac{1}{2}x_5\Big)^2 + \frac{1}{4}x_5^2 \geq 1$$

for $x \in \mathbb{Z}^n \setminus \{0\}$, the form 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 $\mu_5(q) = 1 < \frac{5}{4} = c_{55}$.

Lemma 21.64. *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. Clearly there exists $s_1 \in \mathbb{Z}^n$ with $q(s_1) = \mu_1(q)$. Suppose that $s_1, \ldots, s_k$ 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

$$\mu_i(q) \leq \max\big\{q(s_1), \ldots, q(s_{i-1}), q(t_1)\big\} \leq \mu_{k+1}(q)$$

we obtain $q(t_1) = \mu_{k+1}(q)$. Hence we can set $s_{k+1} := t_1$. The claim now follows by induction. □

Remark 21.65. The following theorem improves the Hermite bound (Remark 21.50).

Theorem 21.66 (Minkowski)**.** *For every positive quadratic form q of rank n we have*

$$\boxed{\mu_1(q)\dots\mu_n(q) \le \left(\frac{4}{3}\right)^{n(n-1)/2} \det(q).}$$

Proof. Let $S \in \mathbb{Z}^{n\times n}$ with columns $s_1, \dots, s_n$ as in Lemma 21.64 (note that S does not necessarily lie in $\mathrm{GL}(n,\mathbb{Z})$). Let $q = q_C$ and let $S^{\mathrm{t}}CS = R^{\mathrm{t}}R$ 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 we have

$$\det q_1 = \det(D)^{-1} \det(S^{-\mathrm{t}}R^{\mathrm{t}}RS^{-1}) = \frac{\det q}{\mu_1(q)\dots\mu_n(q)}.$$

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, we also have $z_{k+1} = \dots = z_n = 0$. Hence

$$q_1(x) = y^{\mathrm{t}}R^{\mathrm{t}}D^{-1}Ry = z^{\mathrm{t}}D^{-1}z = \frac{z_1^2}{\mu_1(q)} + \dots + \frac{z_k^2}{\mu_k(q)} \ge \frac{1}{\mu_k(q)}|z|^2 = \frac{1}{\mu_k(q)}y^{\mathrm{t}}R^{\mathrm{t}}Ry = \frac{1}{\mu_k(q)}q(x) \ge 1.$$

In particular, q_1 is positive and

$$1 \le (\min q_1)^n \le \left(\frac{4}{3}\right)^{n(n-1)/2} \det(q_1) = \left(\frac{4}{3}\right)^{n(n-1)/2} \frac{\det q}{\mu_1(q)\dots\mu_n(q)}$$

by Hermite (Remark 21.50). □

Remark 21.67.

(a) The proof shows that one may replace $(4/3)^{(n-1)/2}$ in the Minkowski bound by the n-th Hermite constant γ_n (Example 21.30).

(b) For a δ-reduced lattice basis $v_1, \dots, v_n$, by Theorem 21.45 with $\delta \to 1$ the same bound holds asymptotically: $|v_1|^2 \dots |v_n|^2 \le (4/3)^{n(n-1)/2} \det(q)$.

(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 lecture notes on number theory

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 computation of A^k by repeated squaring.

(b) Show that this way one needs at most $2\lfloor \log_2(k) \rfloor$ matrix multiplications (here $\lfloor \log_2(k) \rfloor$ 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 computations 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 by the data type `float` (Remark 18.29).

Exercise III.9. Let $A \in K^{n\times n}$ with leading principal minors $\det A_k \neq 0$ for $k = 1, \ldots, n-1$ (see Remark 12.43). Show that a unique LU decomposition of the form $A = LU$ exists (see Theorem 18.35).

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 matrices $U \in \mathrm{U}(n, \mathbb{C})$ and $P \in \mathbb{C}^{n\times n}$ with $A = UP$, where P is positive semidefinite. In the case $A \in \mathrm{GL}(n, \mathbb{C})$, U and P are uniquely determined.

Exercise III.12 (HADAMARD inequality)**.** Let $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ with columns $s_1, \dots, s_n$. Show:

(a) $|\det A| \le |s_1| \dots |s_n|$ with equality if and only if $s_1, \dots, s_n$ are pairwise orthogonal or $s_i = 0$ for some i.
Hint: QR decomposition.

(b) If A is positive semidefinite, then $\det A \le a_{11} \dots a_{nn}$.

Exercise III.13. Let $p \in \mathbb{R}$ with $p \ge 1$. For $x, y \in \mathbb{C}^n$, prove the MINKOWSKI *inequality*

$$\sqrt[p]{\sum_{i=1}^{n} |x_i + y_i|^p} \le \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} + \sqrt[p]{\sum_{i=1}^{n} |y_i|^p}.$$

Deduce that

$$\|x\|_p := \sqrt[p]{\sum_{i=1}^{n} |x_i|^p} \qquad (x \in \mathbb{C}^n)$$

defines a norm on $\mathbb{C}^n$.
Hint: One can assume $\|x\|_p + \|y\|_p = 1$ and use the convexity of the function $\mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$, $r \mapsto r^p$ (Definition A.3).

Exercise III.14. Let $\|.\|$ be a norm that is defined on $\mathbb{C}^{n\times 1}$ for all n. Show that

$$\|A\| = \max_{0\ne x\in\mathbb{C}^{m\times 1}} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

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 18.89 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. Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{m\times m}$. Show the following relation between the Kronecker sum and the Kronecker product:

$$\exp(A \oplus B) = \exp(A) \otimes \exp(B).$$

Exercise III.19. Show:

(a) The number of permutations in S_n with exactly k cycles (including 1-cycles) is

$$\begin{bmatrix} n \\ k \end{bmatrix} := \frac{n!}{k!} \sum_{\substack{1\le l_1,\dots,l_k\le n \\ l_1+\dots+l_k=n}} \frac{1}{l_1 \dots l_k}.$$

Hint: One calls $\begin{bmatrix} n \\ k \end{bmatrix}$ a *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., $\sum_{k=1}^{n} (-1)^k \left[{n \atop k} \right] = 0$.

(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 $\exp A = 1_n - N$.

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 $A := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$. Compute $\lim_{k \to \infty} A^k$.

Exercise III.23. Let $A \in \mathbb{R}^{n \times n}$ be nonnegative and irreducible. Show $(1_n + A)^{n-1} > 0$.

Exercise III.24. Let V be a K-vector space of dimension n. A subset $A \subseteq V$ is called *affinely dependent*, if there exist 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 $\lambda_1 + \ldots + \lambda_k = 0$. 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 of $\Delta \subseteq K^n$ is called the *affine hull* $\operatorname{aff}\Delta$.

(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., $\operatorname{aff}\Delta = \Delta$) 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 20.13:

(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 $(1_n, A, -A)$.

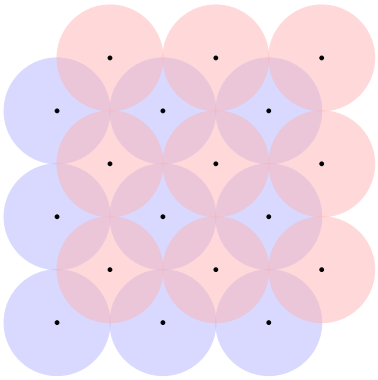
(b) There exists an $x \geq 0$ with $Ax \leq b$ if and only if $y^{\mathrm{t}} b \geq 0$ for all $y \geq 0$ with $y^{\mathrm{t}} A \geq 0$.

Exercise III.27. Show that a subgroup $U \leq (\mathbb{R}^n, +)$ is already a lattice if there exist only finitely many $x \in U$ with $|x| < 1$.
Hint: Bolzano-Weierstrass.

[9] In contrast to convex combinations, the λ_i are not restricted.

Exercise III.28.

(a) Determine the lattice $\Delta \subseteq \mathbb{R}^3$ corresponding to the square arrangement of unit spheres (the spheres in each plane touch the spheres in the plane below, see Remark 21.11):



(b) Compute $\min \Delta$ and the density $\rho(\Delta)$.

(c) Explain why this arrangement is not a counterexample to Kepler's conjecture.

Exercise III.29. Let $\Delta \subseteq \mathbb{R}^n$ be a lattice of full rank and $\delta := \frac{1}{\sqrt{n}} \min \Delta$. Let $F \subseteq \mathbb{R}^n$ be the fundamental parallelepiped of Δ (Example 21.9). 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 $x, y \in \Delta$ we have $(x + W) \cap (y + W) = \varnothing$.

(b) There exist $x_1, \ldots, x_k \in \Delta$ with $W \subseteq \bigcup_{i=1}^{k} (x_i + F)$.

(c) For the volume we have

$$\delta^n = \operatorname{vol} W = \operatorname{vol}\big(F \cap ((W - x_1) \cup \ldots \cup (W - x_k))\big) \leq \operatorname{vol} F = \operatorname{disc} \Delta,$$

i. e., $\min \Delta \leq \sqrt{n} \sqrt[n]{\operatorname{disc} \Delta}$.

(d) For the dual lattice Δ^* we have $(\min \Delta)(\min \Delta^*) \leq n$.

(e) For every positive quadratic form q of rank n we have $\min q \leq n \sqrt[n]{\det q}$.

Exercise III.30. Let q_C be a reduced positive quadratic form of rank 2. Show $\mu_2(q) = c_{22}$ for the successive minimum.

Exercise III.31. 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. 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.32. Show that a positive quadratic form q_C of rank 3 is reduced if and only if the following holds:

$$0 \le 2c_{12} \le c_{11} \le c_{22}, \qquad 2|c_{13}| \le c_{11},$$
$$0 \le 2c_{23} \le c_{22} \le c_{33}, \qquad 2(c_{12} + c_{23} - c_{13}) \le c_{11} + c_{22}.$$

Exercise III.33. Determine all reduced integral quadratic forms of rank 3 with determinant ≤ 5. Which of them are equivalent?

Exercise III.34. Let $A, B \in \mathbb{R}^{n\times n}$ be positive definite. Show $\min q_{A\otimes B} \le (\min q_A)(\min q_B)$.
Remark: KITAOKA proved that equality holds for quadratic forms of rank ≤ 43. For rank 292 there exist examples of strict inequality.

Appendix

Convex Optimization

Remark A.1. In this section we study optimization problems with a convex (or concave) objective function subject to linear constraints. Since the simplex algorithm does not work in this setting, there is no reason to bring the constraints to 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 20.3). As usual, $x \in M$ is called a *vertex*, if $M \setminus \{x\}$ is convex. For $I \subseteq \{1, \dots, n\}$ let $A_I := (a_{ij} : i \in I, j = 1, \dots, m)$. We call I a *basis*, if A_I is invertible. In this case, $|I| = m$.

Theorem A.2. *Let $M := \{x \in \mathbb{R}^m : Ax \leq b\}$ with $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. Then $x \in M$ is a vertex if and only if there exists a basis I with $A_I x = b_I$.*

Proof. Let I be a basis with $A_I x = b_I$. Let $y, z \in M$ and $0 < \lambda < 1$ with $x = \lambda y + (1-\lambda)z$. From

$$b_I = \lambda A_I y + (1-\lambda) A_I z \leq \lambda b_I + (1-\lambda) b_I = b_I$$

we obtain $A_I y = b_I$ and $A_I z = b_I$. Since A_I is invertible, we have $y = x = z$. Hence x is a vertex.

Conversely, let x be a vertex. Let I be the set of indices i with $\sum_{j=1}^m a_{ij} x_j = b_i$. Suppose that $\operatorname{rk} A_I < m$. Then there exists a $y \in \mathbb{R}^m \setminus \{0\}$ with $A_I y = 0$. For $i \notin I$ we have $\sum_{j=1}^m a_{ij} x_j < b_i$. Hence 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. Hence $\operatorname{rk} A_I = m$. Therefore there exists a basis $J \subseteq I$ with $A_J x = b_J$. □

Definition A.3. Let $M \subseteq \mathbb{R}^n$ be convex. A function $f\colon M \to \mathbb{R}$ is called

- *convex*, if
$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$
for all $x, y \in M$ and $0 < \lambda < 1$. If strict inequality holds for $x \neq y$, then f is called *strictly* convex.
- (*strictly*) *concave*, if $-f$ is (strictly) convex, i. e.,
$$f\big(\lambda x + (1-\lambda)y\big) \geq \lambda f(x) + (1-\lambda) f(y).$$

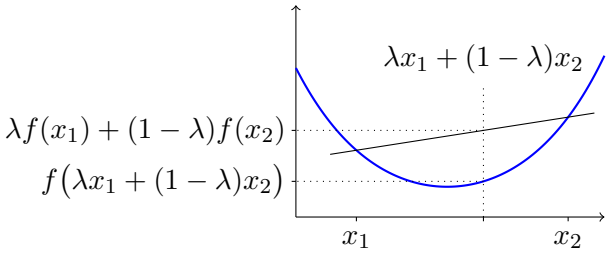
Remark A.4.

(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 $0 \leq \lambda \leq 1$ we have

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda y_1 + (1-\lambda) y_2.$$

(c) In analysis one shows that f is convex if and only if the Hessian matrix $(\frac{\partial^2 f}{\partial x_i \partial x_j})$ is positive semidefinite for all $x \in M$ (cf. Example 12.35). 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.5. *Let $M = \{x \in \mathbb{R}^m : Ax \leq b\} \neq \varnothing$ be convex and $f\colon M \to \mathbb{R}$ convex (or concave). If f has a maximum (or minimum) on M, then it is attained at a vertex.*

Proof. We argue as in the proof of Theorem 20.24. Let $x \in M$ be a maximum of f. Let $I(x) = I$ be the set of indices i with $\sum_{j=1}^m a_{ij}x_j = b_i$. Suppose that $\operatorname{rk} A_I < m$. As in the proof of Theorem A.2, there exist $y \in \mathbb{R}^m$ and $\epsilon > 0$ with $x_\pm := x \pm \epsilon y \in M$. Since f is convex, we have

$$f(x) = f\Big(\frac{1}{2}x_+ + \frac{1}{2}x_-\Big) \leq \frac{1}{2}f(x_+) + \frac{1}{2}f(x_-) \leq \max\big\{f(x_+), f(x_-)\big\} \leq f(x)$$

and $f(x_+) = f(x) = f(x_-)$. Since we always assume that A has full rank, we have $Ay \neq 0$. Hence we can choose ϵ such that $I(x_+) \supsetneq I$ or $I(x_-) \supsetneq I$ holds. Then we replace x by x_+ (or 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 is analogous. □

Remark A.6. If f is *strictly* convex (or concave), then the maximum (or *minimum*) can be attained *only* at vertices (for this one does not need Theorem A.2).

Theorem A.7 (Inequality of the harmonic, geometric, arithmetic and quadratic means)**.** *For all positive $x_1, \ldots, x_n \in \mathbb{R}$ we have*

$$\frac{n}{\frac{1}{x_1} + \ldots + \frac{1}{x_n}} \leq \sqrt[n]{x_1 \ldots x_n} \leq \frac{x_1 + \ldots + x_n}{n} \leq \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}$$

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 $1/x_i$. The third inequality (with the case of equality) follows from the Cauchy-Schwarz inequality applied to the vectors $(x_1, \ldots, x_n)$ and $(1, \ldots, 1)$. We prove 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 suppose that the statement holds for n. Then we have

$$x_1 \ldots x_{2n} \leq \Big(\sum_{i=1}^n \frac{x_i}{n} \sum_{i=n+1}^{2n} \frac{x_i}{n}\Big)^n \leq \Big(\sum_{i=1}^{2n} \frac{x_i}{2n}\Big)^{2n}$$

with equality if and only if $x_1 = \ldots = x_{2n}$. Hence the claim holds for $2n$. Now let $A := \sum_{i=1}^{n-1} \frac{x_i}{n-1}$. Then we have

$$x_1 \ldots x_{n-1} A \leq \Big(\sum_{i=1}^{n-1} \frac{x_i}{n} + \frac{A}{n}\Big)^n = \Big(\frac{A(n-1)}{n} + \frac{A}{n}\Big)^n = A^n$$

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 holds for $n-1$. Combining both induction steps, one obtains the statement for $2n-1$, $2n-2, \ldots, n+1$. □

Theorem A.8. *For every positive semidefinite matrix $A \in \mathbb{C}^{n\times n}$ we have $\sqrt[n]{\det A} \leq \operatorname{tr}(A)/n$ with equality if and only if A is a scalar matrix.*

Proof. By Exercise II.23, A has nonnegative eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. By Remark 10.35 we have

$$\sqrt[n]{\det A} = \sqrt[n]{\lambda_1 \ldots \lambda_n} \leq \frac{1}{n}(\lambda_1 + \ldots + \lambda_n) = \frac{1}{n}\operatorname{tr}(A)$$

with equality if and only if $\lambda_1 = \ldots = \lambda_n$. Since A is diagonalizable by the spectral theorem, in this case A must be a scalar matrix. □

Lemma A.9. *For $x_1, \ldots, x_n, y_1, \ldots, y_n \geq 0$ we have*

$$\sqrt[n]{(x_1 + y_1) \ldots (x_n + y_n)} \geq \sqrt[n]{x_1 \ldots x_n} + \sqrt[n]{y_1 \ldots y_n}$$

with equality if and only if $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are linearly dependent or $x_i = 0 = y_i$ for some i.

Proof. W. l. o. g. let $x_i + y_i > 0$ for $i = 1, \ldots, n$ (otherwise both sides are 0). By the inequality between the arithmetic and geometric means we have

$$1 = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i}{x_i + y_i} + \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i + y_i} \geq \sqrt[n]{\prod_{i=1}^{n} \frac{x_i}{x_i + y_i}} + \sqrt[n]{\prod_{i=1}^{n} \frac{y_i}{x_i + y_i}} = \frac{\sqrt[n]{x_1 \ldots x_n} + \sqrt[n]{y_1 \ldots y_n}}{\sqrt[n]{(x_1 + y_1) \ldots (x_n + y_n)}}$$

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, w. l. o. g. we may assume $x_1 \neq 0$, and then $y_i = \frac{y_1}{x_1} x_i$ for $i = 1, \ldots, n$. Hence x and y are linearly dependent. □

Lemma A.10 (Simultaneous diagonalization)**.** *Let $A \in \mathbb{C}^{n\times n}$ be positive definite and $B \in \mathbb{C}^{n\times n}$ 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. By Exercise II.23 there exists a Hermitian square root $\sqrt{A}$ of A. Clearly, $C := \sqrt{A}^{-1} B \sqrt{A}^{-1}$ is Hermitian as well. Let $U \in \mathrm{U}(n, \mathbb{C})$ with $U^* C U = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ (spectral theorem). Now the claim holds for $S := \sqrt{A}^{-1} U$. □

Theorem A.11 (Minkowski inequality)**.** *Let $A \in \mathbb{C}^{n\times n}$ be positive definite and $B \in \mathbb{C}^{n\times n}$ positive semidefinite. Then we have*

$$\sqrt[n]{\det(A + B)} \geq \sqrt[n]{\det A} + \sqrt[n]{\det B}$$

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. By Lemma A.10 there exists an $S \in \mathrm{GL}(n, \mathbb{C})$ with $S^* A S = 1_n$ and

$$D := S^* B S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

By multiplying from the left and right by $\sqrt[n]{\det(S^*)}$ and $\sqrt[n]{\det S}$, respectively, we may assume $A = 1_n$ and $B = D$. By assumption, $\lambda_1, \ldots, \lambda_n \geq 0$. By Lemma A.9 we have

$$\sqrt[n]{\det(1_n + D)} = \sqrt[n]{(1+\lambda_1)\ldots(1+\lambda_n)} \geq 1 + \sqrt[n]{\lambda_1 \ldots \lambda_n} = \sqrt[n]{\det 1_n} + \sqrt[n]{\det D}$$

with equality if and only if $\lambda := \lambda_1 = \ldots = \lambda_n$. For the original matrix this means $B = S^{-*} D S^{-1} = \lambda A$.

The second claim follows from

$$\begin{aligned} \det(A+B) &= \sqrt[n]{\det(A+B)}^n \geq \left(\sqrt[n]{\det A} + \sqrt[n]{\det B} \right)^n \\ &= \sum_{k=0}^{n} \binom{n}{k} \sqrt[n]{\det A}^k \sqrt[n]{\det B}^{n-k} \geq \det(A) + \det(B). \end{aligned}$$

□

Theorem A.12. *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 $0 < \lambda < 1$. Clearly, $C := \lambda A + (1-\lambda) B$ has main diagonal v. For $x \in \mathbb{R}^n \setminus \{0\}$ we have

$$x C x^{\mathrm{t}} = \lambda x A x^{\mathrm{t}} + (1-\lambda) x B x^{\mathrm{t}} > 0.$$

This shows $C \in \mathcal{M}$. By the Minkowski inequality we have

$$\sqrt[n]{\det C} \geq \sqrt[n]{\det(\lambda A)} + \sqrt[n]{\det((1-\lambda)B)} = \lambda \sqrt[n]{\det A} + (1-\lambda) \sqrt[n]{\det B}$$

with equality if and only if $\lambda \mu A = (1-\lambda) B$ for some $\mu > 0$. Since A and B have the same main diagonal, it follows that $\lambda \mu = (1-\lambda)$ and $A = B$. Hence $A \mapsto \sqrt[n]{\det A}$ is strictly concave. □

Theorem A.13 (Oppenheim)**.** *Let q_C be a reduced positive quadratic form of rank 3. Then we have*

$$c_{11} c_{22} c_{33} \leq c_{11} c_{22} c_{33} + \frac{1}{2} c_{11} c_{22} (c_{33} - c_{22}) + \frac{1}{2} c_{11} c_{33} (c_{22} - c_{11}) \leq 2 \det(q).$$

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

$$C := \begin{pmatrix} a & x & z \\ x & b & y \\ z & y & c \end{pmatrix} \in \mathcal{M}$$

we have $0 \le x \le a/2$, $0 \le y \le b/2$, $z \le a/2$, $-z \le a/2$ and $x + y - z \le (a+b)/2$ by Exercise III.32. Therefore, we may identify $\mathcal{M}$ with the convex set $M \subseteq \mathbb{R}^3$ of all (x, y, z) such that

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \le \frac{1}{2} \begin{pmatrix} a \\ 0 \\ b \\ 0 \\ a \\ a \\ a+b \end{pmatrix} =: v$$

holds. By the rule of Sarrus we have

$$\det C = abc + 2xyz - ay^2 - bz^2 - cx^2.$$

By Theorem A.12 and Remark A.6, the minimum of $C \mapsto \sqrt[3]{\det C}$ (and hence also the minimum of det) on M is attained only at vertices. By Theorem A.2, the vertices are obtained by choosing three linearly independent rows of A and solving the corresponding system of equations. There are eight ways to choose three linearly independent rows among the first six rows of A. However, the choice $(x, y, z) = (a, b, -a)$ does not satisfy the last inequality. If we choose the seventh row as an 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)	$abc - \frac{1}{4}(ab^2 + a^2c)$	$\{1,3,7\}$	$(a,b,0)$	$abc - \frac{1}{4}(ab^2 + a^2c)$
$\{1,4,5\}$	$(a,0,a)$	$abc - \frac{1}{4}a^2(b+c)$	$\{1,4,7\}$	$(a,0,-b)$	as $I = \{1,4,6\}$ with $a = b$
$\{1,4,6\}$	$(a,0,-a)$	$abc - \frac{1}{4}a^2(b+c)$	$\{1,6,7\}$	$(a,b-a,-a)$	$abc - \frac{1}{4}(ab^2 + a^2c)$
$\{2,3,5\}$	$(0,b,a)$	$abc - \frac{1}{4}(ab^2 + a^2b)$	$\{2,3,7\}$	$(0,b,-a)$	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 $a = b$

Clearly, $\det C \ge abc - \frac{1}{4}(ab^2 + a^2c)$ 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 $2(x, y, z) \in \{(a,a,a), (a,a,0), (a,0,\pm a), (0,a,\pm a)\}$. By Example 21.52 we have $(a,a,a) \sim (a,a,0)$. Permuting x, y, z and changing signs shows that $(a,a,0) \sim (a,0,\pm a) \sim (0,a,\pm a)$. □

Theorem A.14 (Gauss). *For every lattice $\Delta \subseteq \mathbb{R}^3$ we have $\rho(\Delta) \le \frac{\pi}{3\sqrt{2}}$. If equality holds, then the Gram matrix of Δ has the form*

$$c \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

with respect to a suitable choice of basis.

Proof. Let C be the Gram matrix of Δ. By a change of basis we may assume that q_C is reduced. As in Remark 21.11, we arrange unit balls with centers in Δ. Since the balls do not overlap, we have $c_{11} = \min(\Delta)^2 \ge 4$. By Oppenheim's theorem, $\det C \ge 2^5$. 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 by Oppenheim's theorem we can construct a basis of Δ such that C has the stated form. □

Theorem A.15 (Fischer inequality)**.** *Let* $M = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right) \in \mathbb{C}^{n\times n}$ *be positive definite. Then* $\det M \leq \det(A)\det(C)$ *with equality if and only if* $B = 0$*.*

Proof. By Sylvester's criterion, A, C as well as A^{-1}, C^{-1} are positive definite. By Exercise II.23 we have

$$D := \operatorname{diag}(\sqrt{A}, \sqrt{C})^{-1} M \operatorname{diag}(\sqrt{A}, \sqrt{C})^{-1} = \begin{pmatrix} 1 & \sqrt{A}^{-1} B \sqrt{C}^{-1} \\ \sqrt{C}^{-1} B^* \sqrt{A}^{-1} & 1 \end{pmatrix}.$$

By Theorem A.8 we have

$$\det(A)^{-1} \det(C)^{-1} \det(M) = \det D \leq (\operatorname{tr}(D)/n)^n = 1,$$

hence $\det M \leq \det(A)\det(C)$. Equality holds if and only if D is a scalar matrix, i. e., $\sqrt{A}^{-1} B \sqrt{C}^{-1} = 0 = B$. □

Corollary A.16. *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. By the formula for the adjugate $\widetilde{A}$ we have $b_{ii} = \det(A_{ii})/\det(A)$. By swapping the first and i-th row/column, A is similar to $\left(\begin{smallmatrix} a_{ii} & B \\ B^* & A_{ii} \end{smallmatrix}\right)$. The Fischer inequality yields $\det A \leq a_{ii} \det(A_{ii}) = a_{ii} b_{ii} \det(A)$. □

Index

A

absolute value, 99
  of a complex number, 108
addition formulas, 106
adjacent transposition, 83
affine hull, 257
affine space, 28
affinely (in)dependent, 257
AlphaEvolve, 184
angle, 101
  oriented, 106
annihilator, 60
associative law, 21, 23
axiom, 13
axiom of choice, 13, 17

B

back substitution, 194
Banach fixed-point theorem, 195
Banach space, 195
Banach-Tarski paradox, 13
Barnes, 264
basis
  dual, 59
  feasible, 228
  of a lattice, 233
  of a linear program, 228, 261
  of a vector space, 30
basis extension theorem, 31
Bauer-Fike, 202
Bézout, 147
bijection, 19
bilinear form, 113
  alternating, 113
  antisymmetric, 113
  degenerate, 113
  indefinite, 120
  index, 118
  negative (semi)definite, 120
  positive (semi)definite, 120
  symmetric, 113
binary exponentiation, 255
Binet formula, 94
Blichfeldt, 244
block diagonal matrix, 39
Bruhat decomposition, 255
bubble sort, 83

C

Cantor, 13, 21
Carathéodory, 225
cardinal number, 20
Cartan-Dieudonné, 112
Cartesian product, 17
catastrophic cancellation, 189
Cauchy, 262
Cauchy's interlacing theorem, 210
Cauchy-Binet formula, 71
Cauchy-Schwarz inequality, 100, 126
Cayley-Hamilton, 96
centralizer, *154*, 179
Ceres, 193
Chan-Li, 132
change of basis, 57
change-of-basis matrix, 54
characteristic, 114
characteristic polynomial
  of a linear map, 92
  of a matrix, 92
Chinese remainder theorem, 158
Cholesky algorithm, 205
Cholesky decomposition, 193
circular reasoning, 31
classical adjoint, 73
codomain, 19
coefficient, 86
  leading, 86
coefficient matrix, 41
  augmented, 41
Collatz-Wielandt, 218
column operation, 43
column vector, 35
commutative law, 23
companion matrix, 150
complement, 32
  orthogonal, 103, 116
complex conjugation, 109
composite, 19
composition, 19
condition number, 185, 199
congruence
  of matrices, 115
  of polynomials, 158
constant term, 86
continuum hypothesis, 13, 20
contraction, 195
contraposition, 12
converse, 12
convex combination, 225
convex hull, 225
convolution theorem, 182
Conway, 254

coordinate vector, 31
correspondence theorem, 81
cosine, 101
Courant-Fischer, 210
Cramer's rule, 74
cross product, 104
cross-polytope, 225
cycle, 75
    disjoint, 76

D

De Morgan's laws, 12, 15
decimal expansion, 14
Dedekind identity, 81
degree measure, 101
derivative, 159
determinant
    multiplicativity, 70
    of a linear map, 71
    of a matrix, 68
    of a quadratic form, 249
diagonal argument, 20
diagonal matrix, 36
diagonalizability, 64
    simultaneous, 135
        of bilinear forms, 263
differential equation, 216
dimension, 33
dimension formula, 34
direct product, 24
discriminant, 235
distributive law, 12, 15, 24
division with remainder, 88
divisor, 88
    common, 238
        greatest, 238
domain, 19
double dual space, 59
dual space, 59
duality theorem, 227

E

Eichler, 244
eigenspace, 63
eigenvalue, 63
eigenvector, 63
elementary divisor, 240
elementary matrix, 43
endomorphism, 63
equivalence class, 18
equivalence relation, 18
Euclidean algorithm, 146
Euclidean space, 99
Euler, 111
exchange lemma, 32
exponential function, 22, 213

F

Farkas' lemma, 226
FFT, *see* Fourier transform
Fibonacci numbers, 94
field, 24
    of algebraic numbers, 170
    of complex numbers, 108
    of rational functions, 160
    ordered, 108
    with four elements, 157
    with three elements, 81
    with two elements, 24
Fillmore, 58
Fischer inequality, 266
Fitting, 136
Flanders-Wimmer, 171
floating-point number, 189
forward substitution, 193
Fourier matrix, 182
Fourier transform
    continuous, 183
    discrete, 182
    fast, 183
Francis algorithm, 203
Frobenius, 155
Frobenius inequality, 82
Frobenius norm, 198
Frobenius normal form, 150
Fuglede, 177
function, *see* map
functional, 59
functional analysis, 33
functional equation, 214
fundamental parallelepiped, 235
fundamental theorem
    of algebra, 109
    of linear programming, 230

G

Galois connection, 60
Gauss, 251, 265
Gauss-Seidel method, 196
Gaussian algorithm, 44
    with pivoting, 190
generalized eigenspace, 136
generalized eigenspace decomposition, 137
generator matrix, 233
geometric series, 211
Gershgorin, 206
Givens rotation, 208
Gödel's incompleteness theorems, 13
Goldbach's conjecture, 11
golden ratio, 94
Golden-Thompson inequality, 215
Google matrix, 222
Gram matrix
    of a bilinear form, 114
    of a lattice, 233
    of a quadratic form, 248
Gram-Schmidt process, 102
    modified, 207
group, 23
    abelian, 23
    affine, 83
    alternating, 77

general linear, 39
orthogonal, 104
special linear, 71
special orthogonal, 106
special unitary, 129
symmetric, 75
unitary, 127, 129

H
Hadamard inequality, 256
Hadamard product, 172
Hales, 237
Hamilton's quaternions, 179
Harriot, 236
Harvey-van der Hoeven, 183
Håstad, 167
Hermite, 242
Hermite constant, 244
Hermite normal form, 239
Heron's method, 255
Hessenberg matrix, 204
Hessian matrix, 121
Hilbert matrix, 186
HM-GM-AM-QM inequality, 262
homeomorphism, 50
homogeneity, 100, 125, 194
homomorphism, 50
homomorphism theorem, 53
Horner scheme, 88
Householder transformation, 176, 208
hypercube, 67
hyperplane, 33

I
identity map, 19
identity matrix, 35
image, 19
imaginary part, 108
inclusion map, 19
index, 118
inner product, 99, 125
interpolation, 89
intersection, 13
invariant, 67
inverse element, 23
inverse function, 21
isomorphism, 50
isomorphism theorem
first, 53
second, 53

J
Jacob, 151
Jacobi, 214
Jordan block, 138
generalized, 162
Jordan normal form, 141
Jordan-Chevalley decomposition, 162

K
Karatsuba algorithm, 181
Kepler, 236
kernel, 51
Kitaoka, 259
Korkine-Zolotarev, 251
Kronecker delta, 35
Kronecker product, 166
Kronecker sum, 169
Kronecker-Capelli, 41

L
Lagrange polynomial, 90
Laplace expansion, 72
generalized, 83
large language model, 184
lattice, 233
basis
δ-reduced, 246
dual, 234
E_8, 237
(in)decomposable, 244
integral, 235
self-dual, 234
law of cosines, 175
law of excluded middle, 12
law of noncontradiction, 12
law of sines, 175
lead coefficient, 86
leading principal minor, 123
Leibniz formula, 77
length
of a cycle, 75
Lights Out, 84
linear combination, 25
affine, 257
convex, 225
linear factor, 91
linear program, 224
dual, 227
infeasible, 224
solvable, 224
unbounded, 224
linearly (in)dependent, 30
LLL algorithm, 246
logarithm, 22
logarithm table, 182
Lovász condition, 246
LU decomposition, 190
LWE, 236

M
main diagonal, 36
mantissa, 189
map, 19
adjoint, 127
affine, 51
antilinear, 177
bijective, 19
concave, 261
strictly, 261
convex, 261
strictly, 261

diagonalizable, 64
simultaneously, 135
dual, 62
Hermitian, 127
injective, 19
linear, 50
nilpotent, 138
normal, 127
orthogonal, 104
semisimple, 160
separable, 160
surjective, 19
symmetric, 104
triangularizable, 134
simultaneously, 135
unitary, 127
Markov chain, 216
mathematical induction, 16
matrix, 35
adjoint, 128
adjugate, 73
antisymmetric, 114
block, 38
circulant, 221
commuting, 38
congruent, 115
convergence, 200
diagonalizable, 64
simultaneously, 177
diagonally dominant, 206
equivalent, 48
Hermitian, 128
positive (semi)definite, 177
inverse, 38
invertible, 38
(ir)reducible, 216
nilpotent, 138
nonnegative, 216
nonsingular, 38
normal, 128
of a linear map, 54
orthogonal, 106
positive, 216
quasi-convergence, 203
row equivalent, 43
similar, 57
singular, 38
skew-symmetric, 114
sparse, 72
square, 35
stochastic, 216
symmetric, 36
transpose, 36
triangularizable, 130, *134*
simultaneously, 178
unitary, 128
well-conditioned/ill-conditioned, 185
matrix inversion, 47
matrix norm, 198
induced, 198
submultiplicative, 198
matrix unit, 36
Mazur-Ulam, 105
Mercator series, 216
method of least squares, 192
Millennium Prize Problem, 12
min-max theorem, 210
minimal norm, 235
minimal polynomial, 95
Minkowski, 29, 250, 254
Minkowski inequality, 256, 263
Minkowski space, 119
Minkowski's lattice point theorem, 244
Minkowski's linear forms theorem, 243
Mirsky, 94
von Mises, 219
mnemonic, 37, 53, 54, 56, 65, 89, 111
module, 233
modus ponens, 12
Moore-Penrose, 188
Mordell, 250
multilinear form, 113, *176*
alternating, 176
multiplicity
algebraic, 91
geometric, 63

N

Neumann series, 211
neutral element, 23
Newton's method, 109
norm, 99, 125, *194*
equivalent, 195
normal equations, 193
normal form
Frobenius, 150
Hermite, 239
Jordan, 141
Smith, 240
Weierstrass, 151
numbers
complex, 108
integer, 14
natural, 14
rational, 14
real, 14

O

one-way function, 22
Oppenheim, 264
Oppenheim inequality, 173
order relation, 18
lexicographic, 141
total, 18
orthogonal basis, 116
orthogonal decomposition, 244
orthonormal basis, 102
with respect to a bilinear form, 116

P

PageRank, 221
paradox, 11

parallelogram law, 100
partition
of a set, 14
of an integer, 138
permutation, 75
permutation matrix, 76
Perron, 217
Perron-Frobenius, 218
Piazzi, 193
pivot, 190
polar decomposition, 192, 255
polarization, 114
polyhedron, 228
polynomial, 86
companion matrix, 150
congruent, 158
constant, 86
coprime, 146
degree, 86
derivative, 159
irreducible, 146
monic, 86
separable, 159
polytope, 228
power method, 202
power set, 15
predicate, 11
preimage, 19
primary decomposition, 148
prime factorization in $K[X]$, 147
principal axis theorem, 110
product rule, 160
projection, 51
pseudoinverse, 188
Pythagoras, 100
trigonometric, 175

Q

QR algorithm, 203
QR decomposition, 191
quadratic form, 114, *248*
binary, 250
equivalent, 249
indefinite, 258
integral, 249
minimum, 249
nondegenerate, 249
positive, 249
reduced, 250
unimodular, 250
universal, 254
quantum mechanics, 215
quotient space, 28

R

Rado, 243
rank
of a lattice, 233
of a linear map, 51
of a matrix, 40
Rayleigh quotient, 210
real part, 108
reduced row echelon form, 44
reflection
in $\mathbb{R}^2$, 106
in $\mathbb{R}^n$, 112
relation, 17
antisymmetric, 17
asymmetric, 17
reflexive, 17
symmetric, 17
transitive, 18
trivial, 18
remainder, 88
restriction, 19
right-hand rule, 104
ring, 37
root, 89
double, 91
simple/multiple, 91
root of unity, 108
rotary reflection, 112
rotation, 106
Roth, 171, 172
row operation, 43
row vector, 35
RSA, 236
Ruffini's rule, 88
rule of Sarrus, 78
Russell's paradox, 13

S

SAT problem, 12
scalar, 25
scalar matrix, 36
Schatzman, 205
Schönhage-Strassen algorithm, 183
Schur
product theorem, 172
Schur decomposition, 130
real, 177
Schur product, 172
Schur's lemma, 156
Schur-Horn, 132
selection sort, 77
sesquilinear form, 125
set
convex, 224
countable, 20
disjoint, 14
empty, 13
equinumerous, 19
(in)finite, 13
uncountable, 20
set difference, 14
sign, 76
signum, 76, 84
simplex algorithm, 231
simplex criterion, 230
sine, 102
singular value decomposition, 186
slack variable, 224

Smith normal form, 240
solution set, 41
span, 29
spanning set, 29
spectral radius, 211
spectral theorem, 129
spectrum, 129
splitting field, 157
standard basis, 30
standard inner product, 100, 125
statement, 11
    equivalent, 11
Steinitz, 32
Stirling number, 256
Strassen algorithm, 184
subgroup, 26
    proper, 26
subset, 14
    proper, 14
subspace, 26
    cyclic, 149
    F-invariant, 179
    f-invariant, 134
    proper, 27
successive minima, 253
sum of subspaces, 29, 64
    direct, 29
sum rule, 160
support, 228
Sylvester equation, 170
Sylvester inequality, 82
Sylvester's criterion, 122
Sylvester's determinant theorem, 174
Sylvester's law of inertia
    for Hermitian matrices, 131
    for symmetric bilinear forms, 118
symplectic space, 120
system of linear equations, 41
    (in)homogeneous, 41
    overdetermined, 42
    solvable, 41
    underdetermined, 42
system of representatives, 18

T
Taussky, 178
tensor product, 166
trace
    of a linear map, 58
    of a matrix, 57
transition matrix, 216
transitivity, 12
translation, 83
transposition, 76
triangle inequality, 100, 126, 194
triangular matrix, 66
    strictly, 66
tridiagonal matrix, 250
triple, 17
tuple, 17

U
union, 13
    disjoint, 14
unitary space, 125

V
Vandermonde matrix, 72
variable, 86
vector, 25
    normalized, 99, 125
    orthogonal, 99, 125
    shortest, 236
vector product, 104
vector space, 25
    Euclidean, 99
    finite-dimensional, 33
    finitely generated, 29
    isomorphic, 50
    unitary, 125
vectorization, 168
Venn diagram, 14
vertex, 228, 261
Viazovska, 237

W
Weierstrass normal form, 151
Wielandt, 220
Wilkinson, 204

Z
Zassenhaus algorithm, 48
Zermelo-Fraenkel axioms, 13
zero matrix, 35
zero polynomial, 86
zero space, 25
zero vector, 25
Zorn's lemma, 33