Graduate Texts in Mathematics

Graduate Texts in Mathematics 216

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

This page intentionally left blank

Denis Serre

Matrices

Theory and Applications

Denis Serre
Ecole Normale Supérieure de Lyon
UMPA
Lyon Cedex 07, F-69364
France
Denis.SERRE@umpa.ens-lyon.fr

Mathematics Subject Classification (2000): 15-01

Library of Congress Cataloging-in-Publication Data
Serre, D. (Denis)
[Matrices. English.]
Matrices : theory and applications / Denis Serre.
p. cm.—(Graduate texts in mathematics ; 216)
Includes bibliographical references and index.
ISBN 0-387-95460-0 (alk. paper)
1. Matrices I. Title. II. Series.
QA188 .S4713 2002
512.9′434—dc21 2002022926

ISBN 0-387-95460-0 Printed on acid-free paper.

Translated from *Les Matrices: Théorie et pratique*, published by Dunod (Paris), 2001.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10869456

Typesetting: Pages created by the author in LaTeX2e.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To Pascale and Joachim

This page intentionally left blank

Preface

The study of matrices occupies a singular place within mathematics. It is still an area of active research, and it is used by every mathematician and by many scientists working in various specialities. Several examples illustrate its versatility:

- Scientific computing libraries began growing around matrix calculus. As a matter of fact, the discretization of partial differential operators is an endless source of linear finite-dimensional problems.
- At a discrete level, the maximum principle is related to nonnegative matrices.
- Control theory and stabilization of systems with finitely many degrees of freedom involve spectral analysis of matrices.
- The discrete Fourier transform, including the fast Fourier transform, makes use of Toeplitz matrices.
- Statistics is widely based on correlation matrices.
- The generalized inverse is involved in least-squares approximation.
- Symmetric matrices are inertia, deformation, or viscous tensors in continuum mechanics.
- Markov processes involve stochastic or bistochastic matrices.
- Graphs can be described in a useful way by square matrices.

- Quantum chemistry is intimately related to matrix groups and their representations.
- The case of quantum mechanics is especially interesting: Observables are Hermitian operators, their eigenvalues are energy levels. In the early years, quantum mechanics was called "mechanics of matrices," and it has now given rise to the development of the theory of large random matrices. See [23] for a thorough account of this fashionable topic.

This text was conceived during the years 1998–2001, on the occasion of a course that I taught at the École Normale Supérieure de Lyon. As such, every result is accompanied by a detailed proof. During this course I tried to investigate all the principal mathematical aspects of matrices: algebraic, geometric, and analytic.

In some sense, this is not a specialized book. For instance, it is not as detailed as [19] concerning numerics, or as [35] on eigenvalue problems, or as [21] about Weyl-type inequalities. But it covers, at a slightly higher than basic level, all these aspects, and is therefore well suited for a graduate program. Students attracted by more advanced material will find one or two deeper results in each chapter but the first one, given with full proofs. They will also find further information in about the half of the 170 exercises. The solutions for exercises are available on the author's site `http://www.umpa.ens-lyon.fr/~serre/exercises.pdf`.

This book is organized into ten chapters. The first three contain the basics of matrix theory and should be known by almost every graduate student in any mathematical field. The other parts can be read more or less independently of each other. However, exercises in a given chapter sometimes refer to the material introduced in another one.

This text was first published in French by Masson (Paris) in 2000, under the title *Les Matrices: théorie et pratique*. I have taken the opportunity during the translation process to correct typos and errors, to index a list of symbols, to rewrite some unclear paragraphs, and to add a modest amount of material and exercises. In particular, I added three sections, concerning alternate matrices, the singular value decomposition, and the Moore–Penrose generalized inverse. Therefore, this edition differs from the French one by about 10 percent of the contents.

Acknowledgments. Many thanks to the Ecole Normale Supérieure de Lyon and to my colleagues who have had to put up with my talking to them so often about matrices. Special thanks to Sylvie Benzoni for her constant interest and useful comments.

Lyon, France — Denis Serre
December 2001

Contents

This page intentionally left blank

List of Symbols

I_n, 5

J, 151
$J(a;r)$, 110
J_{ik}, 100
J_2, 132
J_3, 132
J_4, 132

$K(A)$, 162
$\overline{K}$, 4
$\ker M$, 7
$\ker u$, 7
K^I, 2
$K(M)$, 6
K_n, 57
$K[X]$, 15
$k[X,Y]$, 99

$\lambda_k(A)$, 57
$\mathcal{L}(E,F)$, 7
$\mathcal{L}_\omega$, 152

$\operatorname{adj} M$, 17
$\bar{M}$, 40
$\hat{M}$, 17
$M\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$, 17
M^k, 6
M^{-1}, 20
M^{-k}, 20
M^{-T}, 20
$[M,N]$, 6
$\mathbf{M}_n(K)$, 5
$\mathbf{M}_{n\times m}(K)$, 5
M^*, 40
M^{-*}, 40
M^T, 10

$\| A \|$, 64
$\| A \|_p$, 65
$\| x \|_p$, 61
$\| x \|_A$, 154
$\| x \|_\infty$, 61
$\| \cdot \|'$, 64
$|||\cdot|||$, 65

ω_J, 158
0_n, 5

$\mathbf{O}_n(K)$, 20
0_{nm}, 5
$\mathbf{O}_n^-$, 123
$\mathbf{O}(p,q)$, 120
$\perp_A$, 160

Pf, 22
P_G, 156
π_0, 125
P_J, 156
P_M, 24
P_ω, 156
p', 62
$\mathbf{PSL}_2(\mathbb{R})$, 56

$R_A(F)$, 57
$\operatorname{rk} M$, 5
$\Re$, real part, 63
$R(h;A)$, 92
$\rho(A)$, 61
$R(M)$, 8
$r(x)$, 70, 160

$\langle x,y\rangle$, 11, 41
$\mathbf{S}\Delta_n$, 90
σ_r, 188
$s_j(A)$, 75
$s_k(a)$, 52
$\mathbf{SL}_n(A)$, 20
s_m, 189
$\boldsymbol{S}_n$, 15
$\mathbf{SO}_n(K)$, 20
S^1, 86
$\operatorname{Sp}(M)$, 24
$\operatorname{Sp}_K(M)$, 24
$\mathbf{SPD}_n$, 42
$\mathbf{Sp}_m$, 120
$\mathbf{Sp}_m$, 120
S^2, 56, 126
$\mathbf{SU}_n$, 41
$\mathbf{Sym}_n(K)$, 10

τ, 151
τ_{CG}, 164
T_k, 162
$\operatorname{Tr} M$, 25

$\mathbf{U}_n$, 41
$\boldsymbol{\mathcal{U}}_p$, 85

$\mathbf{U}(p,q)$, 120
u^*, 42
u^T, 11

$V(a)$, 173

$|x|$, 80
$x \leq y$, 80
$x > 0$, 80
$x \geq 0$, 80

This page intentionally left blank

1

Elementary Theory

1.1 Basics

1.1.1 Vectors and Scalars

Fields. Let $(K, +, \cdot)$ be a field. It could be $\mathbb{R}$, the field of real numbers, $\mathbb{C}$ (complex numbers), or, more rarely, $\mathbb{Q}$ (rational numbers). Other choices are possible, of course. The elements of K are called *scalars*.

Given a field k, one may build larger fields containing k: algebraic extensions $k(\alpha_1, \ldots, \alpha_n)$, fields of rational fractions $k(X_1, \ldots, X_n)$, fields of formal power series $k[[X_1, \ldots, X_n]]$. Since they are rarely used in this book, we do not define them and let the reader consult his or her favorite textbook on abstract algebra.

The digits 0 and 1 have the usual meaning in a field K, with $0 + x = 1 \cdot x = x$. Let us consider the subring $\mathbb{Z}1$, composed of all sums (possibly empty) of the form $\pm(1 + \cdots + 1)$. Then $\mathbb{Z}1$ is isomorphic to either $\mathbb{Z}$ or to a field $\mathbb{Z}/p\mathbb{Z}$. In the latter case, p is a prime number, and we call it the *characteristic* of K. In the former case, K is said to have characteristic 0.

Vector spaces. Let $(E, +)$ be a commutative group. Since E is usually not a subset of K, it is an abuse of notation that we use $+$ for the additive laws of both E and K. Finally, let

$$\begin{array}{rcl} (a, x) & \mapsto & ax, \\ K \times E & \to & E, \end{array}$$

be a map such that

$$(a+b)x = ax + bx, \quad a(x+y) = ax + ay.$$

One says that E is a *vector space* over K (one often speaks of a K-vector space) if moreover,

$$a(bx) = (ab)x, \quad 1x = x,$$

hold for all $a, b \in K$ and $x \in E$. The elements of E are called *vectors*. In a vector space one always has $0x = 0$ (more precisely, $0_K x = 0_E$).

When $P, Q \subset K$ and $F, G \subset E$, one denotes by PQ (respectively $P+Q, F+G, PF$) the set of products pq as (p,q) ranges over $P \times Q$ (respectively $p+q$, $f+g$, pf as p, q, f, g range over P, Q, F, G). A subgroup $(F,+)$ of $(E,+)$ that is stable under multiplication by scalars, i.e., such that $KF \subset F$, is again a K-vector space. One says that it is a *linear subspace* of E, or just a subspace. Observe that F, as a subgroup, is nonempty, since it contains 0_E. The intersection of any family of linear subspaces is a linear subspace. The sum $F+G$ of two linear subspaces is again a linear subspace. The trivial formula $(F+G)+H = F+(G+H)$ allows us to define unambiguously $F+G+H$ and, by induction, the sum of any finite family of subsets of E. When these subsets are linear subspaces, their sum is also a linear subspace.

Let I be a set. One denotes by K^I the set of maps $a = (a_i)_{i \in I} : I \to K$ where only finitely many of the a_i's are nonzero. This set is naturally endowed with a K-vector space structure, by the addition and product laws

$$(a+b)_i := a_i + b_i, \quad (\lambda a)_i := \lambda a_i.$$

Let E be a vector space and let $i \mapsto f_i$ be a map from I to E. A *linear combination* of $(f_i)_{i \in I}$ is a sum

$$\sum_{i \in I} a_i f_i,$$

where the a_i's are scalars, only finitely many of which are nonzero (in other words, $(a_i)_{i \in I} \in K^I$). This sum involves only finitely many terms. It is a vector of E. The family $(f_i)_{i \in I}$ is *free* if every linear combination but the trivial one (when all coefficients are zero) is nonzero. It is a *generating* family if every vector of E is a linear combination of its elements. In other words, $(f_i)_{i \in I}$ is free (respectively generating) if the map

$$\begin{array}{ccc} K^I & \to & E, \\ (a_i)_{i \in I} & \mapsto & \displaystyle\sum_{i \in I} a_i f_i, \end{array}$$

is injective (respectively onto). Last, one says that $(f_i)_{i \in I}$ is a *basis* of E if it is free and generating. In that case, the above map is bijective, and it is actually an isomorphism between vector spaces.

If $\mathcal{G} \subset E$, one often identifies $\mathcal{G}$ and the associated family $(g)_{g\in\mathcal{G}}$. The set G of linear combinations of elements of $\mathcal{G}$ is a linear subspace E, called the linear subspace *spanned* by $\mathcal{G}$. It is the smallest linear subspace E containing $\mathcal{G}$, equal to the intersection of all linear subspaces containing $\mathcal{G}$. The subset $\mathcal{G}$ is generating when $G = E$.

One can prove that every K-vector space admits at least one basis. In the most general setting, this is a consequence of the axiom of choice. All the bases of E have the same cardinality, which is therefore called the *dimension* of E, denoted by $\dim E$. The dimension is an upper (respectively a lower) bound for the cardinality of free (respectively generating) families. In this book we shall only use finite-dimensional vector spaces. If F, G are two linear subspaces of E, the following formula holds:

$$\dim F + \dim G = \dim F \cap G + \dim(F + G).$$

If $F \cap G = \{0\}$, one writes $F \oplus G$ instead of $F + G$, and one says that F and G are in *direct sum*. One has then

$$\dim F \oplus G = \dim F + \dim G.$$

Given a set I, the family $(\mathbf{e}^i)_{i\in I}$, defined by

$$(\mathbf{e}^i)_j = \begin{cases} 0, & j \neq i, \\ 1, & j = i, \end{cases}$$

is a basis of K^I, called the *canonical basis*. The dimension of K^I is therefore equal to the cardinality of I.

In a vector space, every generating family contains at least one basis of E. Similarly, given a free family, it is contained in at least one basis of E. This is the *incomplete basis theorem*.

Let L be a field and K a subfield of L. If F is an L-vector space, then F is also a K-vector space. As a matter of fact, L is itself a K-vector space, and one has

$$\dim_K F = \dim_L F \cdot \dim_K L.$$

The most common example (the only one that we shall consider) is $K = \mathbb{R}$, $L = \mathbb{C}$, for which we have

$$\dim_{\mathbb{R}} F = 2 \dim_{\mathbb{C}} F.$$

Conversely, if G is an $\mathbb{R}$-vector space, one builds its *complexification* $G^{\mathbb{C}}$ as follows:

$$G^{\mathbb{C}} = G \times G,$$

with the induced structure of an additive group. An element (x, y) of $G^{\mathbb{C}}$ is also denoted $x + iy$. One defines multiplication by a complex number by

$$(\lambda = a + ib, z = x + iy) \mapsto \lambda z := (ax - by, ay + bx).$$

One verifies easily that $G^{\mathbb{C}}$ is a $\mathbb{C}$-vector space, with

$$\dim_{\mathbb{C}} G^{\mathbb{C}} = \dim_{\mathbb{R}} G.$$

Furthermore, G may be identified with an $\mathbb{R}$-linear subspace of $G^{\mathbb{C}}$ by

$$x \mapsto (x, 0).$$

Under this identification, one has $G^{\mathbb{C}} = G + iG$. In a more general setting, one may consider two fields K and L with $K \subset L$, instead of $\mathbb{R}$ and $\mathbb{C}$, but the construction of G^L is more delicate and involves the notion of tensor product. We shall not use it in this book.

One says that a polynomial $P \in L[X]$ *splits* over L if it can be written as a product of the form

$$a \prod_{i=1}^{r} (X - a_i)^{n_i}, \quad a, a_i \in L, \quad r \in \mathbb{N}, \, n_i \in \mathbb{N}^*.$$

Such a factorization is unique, up to the order of the factors. A field L in which every nonconstant polynomial $P \in L[X]$ admits a root, or equivalently in which every polynomial $P \in L[X]$ splits, is *algebraically closed*. If the field K' contains the field K and if every polynomial $P \in K[X]$ admits a root in K', then the set of roots in K' of polynomials in $K[X]$ is an algebraically closed field that contains K, and it is the smallest such field. One calls K' the *algebraic closure* of K. Every field K admits an algebraic closure, unique up to isomorphism, denoted by $\overline{K}$. The fundamental theorem of algebra asserts that $\overline{\mathbb{R}} = \mathbb{C}$. The algebraic closure of $\mathbb{Q}$, for instance, is the set of *algebraic* complex numbers, meaning that they are roots of polynomials $P \in \mathbb{Z}[X]$.

1.1.2 Matrices

Let K be a field. If $n, m \geq 1$, a matrix of size $n \times m$ with entries in K is a map from $\{1, \ldots, n\} \times \{1, \ldots, m\}$ with values in K. One represents it as an array with n rows and m columns, an element of K (an *entry*) at each point of intersection of a row an a column. In general, if M is the name of the matrix, one denotes by m_{ij} the element at the intersection of the ith row and the jth column. One has therefore

$$M = \begin{pmatrix} m_{11} & \ldots & m_{1m} \\ \vdots & \ddots & \vdots \\ m_{n1} & \ldots & m_{nm} \end{pmatrix},$$

which one also writes

$$M = (m_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}.$$

In particular circumstances (extraction of matrices or minors, for example) the rows and the columns can be numbered in a different way, using non-

consecutive numbers. One needs only two finite sets, one for indexing the rows, the other for indexing the columns.

The set of matrices of size $n \times m$ with entries in K is denoted by $\mathbf{M}_{n\times m}(K)$. It is an additive group, where $M + M'$ denotes the matrix M'' whose entries are given by $m''_{ij} = m_{ij} + m'_{ij}$. One defines likewise multiplication by a scalar $a \in K$. The matrix $M' := aM$ is defined by $m'_{ij} = am_{ij}$. One has the formulas $a(bM) = (ab)M$, $a(M + M') = (aM) + (aM')$, and $(a + b)M = (aM) + (bM)$, which endow $\mathbf{M}_{n\times m}(K)$ with a K-vector space structure. The zero matrix is denoted by 0, or 0_{nm} when one needs to avoid ambiguity.

When $m = n$, one writes simply $\mathbf{M}_n(K)$ instead of $\mathbf{M}_{n\times n}(K)$, and 0_n instead of 0_{nn}. The matrices of sizes $n \times n$ are called *square* matrices. One writes I_n for the *identity* matrix, defined by

$$m_{ij} = \delta_i^j = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

In other words,

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The identity matrix is a special case of a *permutation matrix*, which are square matrices having exactly one nonzero entry in each row and each column, that entry being a 1. In other words, a permutation matrix M reads

$$m_{ij} = \delta_i^{\sigma(j)}$$

for some permutation $\sigma \in \boldsymbol{S}_n$.

A square matrix for which $i < j$ implies $m_{ij} = 0$ is called a *lower triangular* matrix. It is *upper* triangular if $i > j$ implies $m_{ij} = 0$. It is *strictly* upper triangular if $i \geq j$ implies $m_{ij} = 0$. Last, it is *diagonal* if m_{ij} vanishes for every pair (i, j) such that $i \neq j$. In particular, given n scalars $d_1, \ldots, d_n \in K$, one denotes by $\mathrm{diag}(d_1, \ldots, d_n)$ the diagonal matrix whose diagonal term m_{ii} equals d_i for every index i.

When $m = 1$, a matrix M of size $n \times 1$ is called a *column vector*. One identifies it with the vector of K^n whose ith coordinate in the canonical basis is m_{i1}. This identification is an isomorphism between $\mathbf{M}_{n\times 1}(K)$ and K^n. Likewise, the matrices of size $1 \times m$ are called *row vectors*.

A matrix $M \in \mathbf{M}_{n\times m}(K)$ may be viewed as the ordered list of its columns $M^{(j)}$ $(1 \leq j \leq m)$. The dimension of the linear subspace spanned by the $M^{(j)}$ in K^n is called the *rank* of M and denoted by $\mathrm{rk}\ M$.

1.1.3 Product of Matrices

Let $n, m, p \geq 1$ be three positive integers. We define a (noncommutative) multiplication law

$$\begin{aligned} \mathbf{M}_{n\times m}(K) \times \mathbf{M}_{m\times p}(K) &\to \mathbf{M}_{n\times p}(K), \\ (M, M') &\mapsto MM', \end{aligned}$$

which we call the *product* of M and M'. The matrix $M'' = MM'$ is given by the formula

$$m''_{ij} = \sum_{k=1}^{m} m_{ik} m'_{kj}, \quad 1 \leq i \leq n,\, 1 \leq j \leq p.$$

We check easily that this law is associative: if M, M', and M'' have respective sizes $n \times m$, $m \times p$, $p \times q$, one has

$$(MM')M'' = M(M'M'').$$

The product is distributive with respect to addition:

$$M(M' + M'') = MM' + MM'', \quad (M + M')M'' = MM'' + M'M''.$$

It also satisfies

$$a(MM') = (aM)M' = M(aM'), \quad \forall a \in K.$$

Last, if $m = n$, then $I_n M' = M'$. Similarly, if $m = p$, then $MI_m = M$.

The product is an internal composition law in $\mathbf{M}_n(K)$, which endows this space with a structure of a unitary K-algebra. It is noncommutative in general. For this reason, we define the *commutator* of M and N by $[M, N] := MN - NM$. For a square matrix $M \in \mathbf{M}_n(K)$, one defines $M^2 = MM$, $M^3 = MM^2 = M^2M$ (from associativity), ..., $M^{k+1} = M^k M$. One completes this notation by $M^1 = M$ and $M^0 = I_n$. One has $M^j M^k = M^{j+k}$ for all $j, k \in \mathbb{N}$. If $M^k = 0$ for some integer $k \in \mathbb{N}$, one says that M is *nilpotent*. One says that M is *idempotent* if $I_n - M$ is nilpotent.

One says that two matrices $M, N \in \mathbf{M}_n(K)$ *commute* with each other if $MN = NM$. The powers of a square matrix M commute pairwise. In particular , the set $K(M)$ formed by polynomials in M, which cinsists of matrices of the form

$$a_0 I_n + a_1 M + \cdots + a_r M^r, \quad a_0, \ldots, a_r \in K, \quad r \in \mathbb{N},$$

is a commutative algebra.

One also has the formula (see Exercise 2)

$$\mathrm{rk}(MM') \leq \min\{\mathrm{rk}\ M,\ \mathrm{rk}\ M'\}.$$

1.1.4 Matrices as Linear Maps

Let E, F be two K-vector spaces. A map $u : E \to F$ is *linear* (one also speaks of a *homomorphism*) if $u(x + y) = u(x) + u(y)$ and $u(ax) = au(x)$

for every $x, y \in E$ and $a \in K$. One then has $u(0) = 0$. The preimage $u^{-1}(0)$, denoted by $\ker u$, is the *kernel* of u. It is a linear subspace of E. The *range* $u(E)$ is also a linear subspace of F. The set of homomorphisms of E into F is a K-vector space, denoted by $\mathcal{L}(E, F)$. If $F = E$, one defines $\mathrm{End}(E) := \mathcal{L}(E, F)$; its elements are the *endomorphisms* of E.

The identification of $\mathbf{M}_{n\times 1}(K)$ with K^n allows us to consider the matrices of size $n \times m$ as linear maps from K^m to K^n. If $M \in \mathbf{M}_{n\times m}(K)$, one proceeds as in the following diagram:

$$\begin{array}{ccccccc} K^m & \to & \mathbf{M}_{m\times 1}(K) & \to & \mathbf{M}_{n\times 1}(K) & \to & K^n, \\ x & \mapsto & X & \mapsto & Y = MX & \mapsto & y. \end{array}$$

Namely, the image of the vector x with coordinates $x_1, \ldots, x_m$ is the vector y with coordinates $y_1, \ldots, y_n$ given by

$$y_i = \sum_{j=1}^{m} m_{ij} x_j. \tag{1.1}$$

One thus obtains an isomorphism between $\mathbf{M}_{n\times m}(K)$ and $\mathcal{L}(K^m; K^n)$, which we shall use frequently in studying matrix properties.

More generally, if E, F are K-vector spaces of respective dimensions m and n, in which one chooses bases $\beta = \{e_1, \ldots, e_m\}$ and $\gamma = \{f_1, \ldots, f_n\}$, one may construct the linear map $u : E \to F$ by

$$u(x_1 e_1 + \cdots + x_m e_m) = y_1 f_1 + \cdots + y_n f_n,$$

via the formulas (1.1). One says that M is the matrix of u in the bases β, γ.

Let E, F, G be three K-vector spaces of dimensions p, m, n. Let us choose respective bases α, β, γ. Given two matrices M, M' of sizes $n \times m$ and $m \times p$, corresponding to linear maps $u : F \mapsto G$ and $u' : E \mapsto F$, the product MM' is the matrix of the linear map $u \circ u' : E \mapsto G$. Here lies the origin of the definition of the product of matrices. The associativity of the product expresses that of the composition of maps. One will note, however, that the isomorphism between $\mathbf{M}_{n\times m}(K)$ and $\mathcal{L}(E, F)$ is by no means canonical, since the correspondence $M \mapsto u$ always depends on an arbitrary choice of two bases. One thus cannot reduce the entire theory of matrices to that of linear maps, and vice versa.

When $E = F$ is a K-vector space of dimension n, it is often worth choosing a single basis ($\gamma = \beta$ with the previous notation). One then has an algebra isomorphism $M \mapsto u$ between $\mathbf{M}_n(K)$ and $\mathrm{End}(E)$, the algebra of endomorphisms of E. Again, this isomorphism depends on an arbitrary choice of basis.

If M is the matrix of $u \in \mathcal{L}(E, F)$ in the bases α, β, the linear subspace $u(E)$ is spanned by the vectors of F whose representations in the basis β are the columns $M^{(j)}$ of M. Its dimension thus equals $\mathrm{rk}M$.

If $M \in \mathbf{M}_{n\times m}(K)$, one defines the *kernel* of M to be the set $\ker M$ of those $X \in \mathbf{M}_{m\times 1}(K)$ such that $MX = 0_n$. The image of K^m under M is

called the *range* of M, sometimes denoted by $R(M)$. The kernel and the range of M are linear subspaces of K^m and K^n, respectively. The range is spanned by the columns of M and therefore has dimension rk M.

Proposition 1.1.1 *Let K be a field. If $M \in \mathbf{M}_{n\times m}(K)$, then*

$$m = \dim \ker M + rk\ M.$$

Proof
Let $\{f_1, \dots, f_r\}$ be a basis of $R(M)$. By construction, there exist vectors $\{e_1, \dots, e_r\}$ of K^m such that $Me_j = f_j$. Let E be the linear subspace spanned by the e_j. If $e = \sum_j a_j e_j \in \ker M$, then $\sum_j a_j f_j = 0$, and thus the a_j vanish. It follows that the restriction $M : E \to R(M)$ is an isomorphism, so that $\dim E = \text{rk } M$.

If $e \in K^m$, then $Me \in R(M)$, and there exists $e' \in E$ such that $Me' = Me$. Therefore, $e = e' + (e - e') \in E + \ker M$, so that $K^m = E + \ker M$. Since $E \cap \ker M = \{0\}$, one has $m = \dim E + \dim \ker M$.

■

1.2 Change of Basis

Let E be a K-vector space, in which one chooses a basis $\beta = \{e_1, \dots, e_n\}$. Let $P \in \mathbf{M}_n(K)$ be an invertible matrix.[1] The set $\beta' = \{e'_1, \dots, e'_n\}$ defined by

$$e'_i = \sum_{j=1}^{n} p_{ji} e_j$$

is a basis of E. One says that P is the matrix of the change of basis $\beta \mapsto \beta'$, or the change-of-basis matrix. If $x \in E$ has coordinates $(x_1, \dots, x_n)$ in the basis β and $(x'_1, \dots, x'_n)$ in the basis β', one then has the formulas

$$x_j = \sum_{i=1}^{n} p_{ji} x'_i.$$

If $u : E \to F$ is a linear map, one may compare the matrices of u for different choices of the bases of E and F. Let β, β' be bases of E and let γ, γ' be bases of F. Let us denote by P, Q the change-of-basis matrices of $\beta \mapsto \beta'$ and $\gamma \mapsto \gamma'$. Finally, let M, M' be the matrices of u in the bases β, γ and β', γ', respectively. Then

$$MP = QM',$$

or $M' = Q^{-1}MP$, where Q^{-1} denotes the inverse of Q. One says that M and M' are *equivalent*. Two equivalent matrices have same rank.

---

[1] See Section 2.2 for the meaning of this notion.

If $E = F$ and $u \in \mathrm{End}(E)$, one may compare the matrices M, M' of u in two different bases β, β' (here $\gamma = \beta$ and $\gamma' = \beta'$). The above formula becomes

$$M' = P^{-1}MP.$$

One says that M and M' are *similar*, or that they are *conjugate* (the latter term comes from group theory). One also says that M' is the conjugate of M by P.

The equivalence and the similarity of matrices are two equivalence relations. They will be studied in Chapter 6.

1.2.1 Block Decomposition

Considering matrices with entries in a ring A does not cause difficulties, as long as one limits oneself to addition and multiplication. However, when A is not commutative, it is important to choose the formula

$$\sum_{j=1}^{m} M_{ij}M'_{jk}$$

when computing $(MM')_{ik}$, since this one corresponds to the composition law when one identifies matrices with A-linear maps from A^m to A^n.

When $m = n$, the product is a composition law in $\mathbf{M}_n(K)$. This space is thus a K-algebra. In particular, it is a ring, and one may consider the matrices with entries in $B = \mathbf{M}_n(K)$. Let $M \in \mathbf{M}_{p\times q}(B)$ have entries M_{ij} (one chooses uppercase letters in order to keep in mind that the entries are themselves matrices). One naturally identifies M with the matrix $M' \in \mathbf{M}_{pn\times qn}(K)$, whose entry of indices $((i-1)n+k, (j-1)n+l)$, for $i \leq p$, $j \leq q$, and $k, l \leq n$, is nothing but

$$(M_{ij})_{kl}.$$

One verifies easily that this identification is an isomorphism between $\mathbf{M}_{p\times q}(B)$ and $\mathbf{M}_{pn\times qn}(K)$ as K-vector spaces.

More generally, choosing decompositions $n = n_1 + \cdots + n_r$, $m = m_1 + \cdots + m_s$ with $n_k, m_l \geq 1$, one may associate to every matrix $M \in \mathbf{M}_{n\times m}(K)$ an array $\tilde{M}$ with r rows and s columns whose element of index (k, l) is a matrix $\tilde{M}_{kl} \in \mathbf{M}_{n_k\times m_l}(K)$. Defining

$$\nu_k = \sum_{t<k} n_t, \quad \mu_l = \sum_{t<l} m_t \quad (\nu_1 = \mu_1 = 0),$$

one has by definition

$$(\tilde{M}_{kl})_{ij} = m_{\nu_k+i,\mu_l+j}, \quad 1 \leq i \leq n_k, 1 \leq j \leq m_l.$$

This procedure, which depends on the choice of n_k, m_l, is called *block decomposition*.

Though $\tilde{M}$ is not strictly speaking a matrix (except in the case studied previously where the n_k, m_l are all equal to each other), one still may define the sum and the product of such objects. Concerning the product of $\tilde{M}$ and $\tilde{M}'$, we must of course be able to compute the products $\tilde{M}_{jk}\tilde{M}'_{kl}$, and thus the sizes of blocks must be compatible. One verifies easily that the block decomposition behaves well with respect to the addition and the product. For instance, if $n = n_1 + n_2$, $m = m_1 + m_2$ and $p = p_1 + p_2$, two matrices M, M' of sizes $n \times m$ and $m \times p$, with block decomposition M_{ij}, M'_{kl}, have a product $M'' = MM' \in \mathbf{M}_{n\times p}(K)$, whose block decomposition M''_{ij} is given by

$$M''_{ij} = M_{i1}M'_{1j} + M_{i2}M'_{2j}.$$

A square matrix M, whose block decomposition is the same according to rows and columns (that is $m_k = n_k$, in particular the diagonal blocks are square matrices) is said *lower block-triangular* if the blocks M_{kl} with $k < l$ are null blocks. One defines similarly the *upper block*-triangular matrices or the *block-diagonal* matrices.

1.2.2 Transposition

If $M \in \mathbf{M}_{n\times m}(K)$, one defines the *transposed* matrix of M (or simply the *transpose* of M) by

$$M^T = (m_{ji})_{1\leq i\leq m, 1\leq j\leq n}.$$

The transposed matrix has size $m \times n$, and its entries $\hat{m}_{ij}$ are given by $\hat{m}_{ij} = m_{ji}$. When the product MM' makes sense, one has $(MM')^T = (M')^T M^T$ (note that the orders in the two products are reversed). For two matrices of the same size, $(M + M')^T = M^T + (M')^T$. Finally, if $a \in K$, then $(aM)^T = a(M^T)$. The map $M \mapsto M'$ defined on $\mathbf{M}_n(K)$ is thus linear, but it is not an algebra endomorphism.

A matrix and its transpose have the same rank. A proof of this fact is given at the end of this section.

For every matrix $M \in \mathbf{M}_{n\times m}(K)$, the products $M^T M$ and MM^T always make sense. These products are square matrices of sizes $m \times m$ and $n \times n$, respectively.

A square matrix is said to be *symmetric* if $M^T = M$, and *skew-symmetric* if $M^T = -M$ (notice that these two notions coincide when K has characteristic 2). When $M \in \mathbf{M}_{n\times m}(K)$, the matrices $M^T M$ and MM^T are symmetric. We denote by $\mathbf{Sym}_n(K)$ the subset of symmetric matrices in $\mathbf{M}_n(K)$. It is a linear subspace of $\mathbf{M}_n(K)$. The product of two symmetric matrices need not be symmetric.

A square matrix is called *orthogonal* if $M^T M = I_n$. We shall see in Section 2.2 that this condition is equivalent to $MM^T = I_n$.

If $M \in \mathbf{M}_{n\times m}(K)$, $y \in K^m$, and $x \in K^n$, then the product $x^T My$ belongs to $\mathbf{M}_1(K)$ and is therefore a scalar, equal to $y^T M^T x$. Saying that

$M = 0$ amounts to writing $x^T My = 0$ for every x and y. If $m = n$ and $x^T Mx = 0$ for every x, one says that M is *alternate*. An alternate matrix is skew-symmetric, since

$$x^T(M+M^T)y = x^T My+y^T Mx = (x+y)^T M(x+y)-x^T Mx-y^T My = 0.$$

The converse holds whenever the characteristic of K is not 2, since

$$2x^T Mx = x^T(M + M^T)x = 0.$$

However, in characteristic 2 there exist matrices that are skew-symmetric but not alternate. As a matter of fact, the diagonal of an alternate matrix must vanish, though this need not be the case for a skew-symmetric matrix in characteristic 2.

The interpretation of transposition in terms of linear maps is the following. One provides K^n with the bilinear form

$$\langle x, y\rangle := x^T y = y^T x = x_1y_1 + \cdots + x_ny_n,$$

called the canonical *scalar product*; one proceeds similarly in K^m. If $M \in \mathbf{M}_{n\times m}(K)$, there exists a unique matrix $N \in \mathbf{M}_{m\times n}(K)$ satisfying

$$\langle Mx, y\rangle = \langle x, Ny\rangle,$$

for all $x \in K^m$ and $y \in K^n$ (notice that the scalar products are defined on distinct vector spaces). One checks easily that $N = M^T$. More generally, if E, F are K-vector spaces endowed with nondegenerate symmetric bilinear forms, and if $u \in \mathcal{L}(E, F)$, then one can define a unique $u^T \in \mathcal{L}(F, E)$ from the identity

$$\langle u(x), y\rangle_F = \langle x, u^T(y)\rangle_E, \quad \forall x \in E, y \in F.$$

When $E = K^m$ and $F = K^n$ are endowed with their canonical bases and canonical scalar products, the matrix associated to u^T is the transpose of the matrix associated to u.

Let K be a field. Let us endow K^m with its canonical scalar product. If F is a linear subspace of K^m, one defines the *orthogonal subspace* of F by

$$F^\perp := \{x \in K^m;\ \langle x, y\rangle = 0,\ \forall y \in F\}.$$

It is a linear subspace of K^m. We observe that for a general field, the intersection $F \cap F^\perp$ can be nontrivial, and K^m may differ from $F + F^\perp$. One has nevertheless

$$\dim F + \dim F^\perp = m.$$

Actually, $F^\perp$ is the kernel of the linear map $T : K^m \to \mathcal{L}(F; K) =: F^*$, defined by $T(x)(y) = \langle x, y\rangle$ for $x \in K^m$, $y \in F$. Let us show that T is onto. If $\{f_1, \dots, f_r\}$ is a basis of F, then every linear form l on F is a map

$$f = \sum_j z_j f_j \mapsto l(f) = \sum_j l(f_j)z_j.$$

Completing the basis of F as a basis of K^m, one sees that l is the restriction of a linear form L on K^m. Let us define the vector $x \in K^m$ by its coordinates in the canonical basis: $x_j = L(\mathbf{e}^j)$. One has $L(y) = \langle x, y \rangle$ for every $y \in K^m$; that is, $l = T(x)$. Finally, we obtain

$$m = \dim \ker T + \text{rk } T = \dim F^\perp + \dim F^*.$$

The dual formulas between kernels and ranges are frequently used. If $M \in \mathbf{M}_{n\times m}(K)$, one has

$$K^m = \ker M \oplus^\perp R(M^T), \quad K^n = \ker(M^T) \oplus^\perp R(M),$$

where $\oplus^\perp$ means a direct sum of orthogonal subspaces. We conclude that

$$\text{rk } M^T = \dim R(M^T) = m - \dim R(M^T)^\perp = m - \dim \ker M,$$

and finally, that

$$\text{rk } M^T = \text{rk } M.$$

1.2.3 Matrices and Bilinear Forms

Let E, F be two K-vector spaces. One chooses two respective bases $\beta = \{e_1, \dots, e_n\}$ and $\gamma = \{f_1, \dots, f_m\}$. If $B : E \times F \to K$ is a bilinear form, then

$$B(x, y) = \sum_{i,j} B(e_i, f_j) x_i y_j,$$

where the x_i, y_j are the coordinates of x, y. One can define a matrix $M \in \mathbf{M}_{n\times m}(K)$ by $m_{ij} = B(e_i, f_j)$. Conversely, if $M \in \mathbf{M}_{n\times m}(K)$ is given, one can construct a bilinear form on $E \times F$ by the formula

$$B(x, y) := x^T M y = \sum_{i,j} m_{ij} x_i y_j.$$

Therefore, there is an isomorphism between $\mathbf{M}_{n\times m}(K)$ and the set of bilinear forms on $E \times F$. One says that M is the matrix of B with respect to the bases β, γ. This isomorphism depends on the choice of the bases. A particular case arises when $E = K^n$ and $F = K^m$ are endowed with canonical bases.

If M is associated to B, it is clear that M^T is associated to the bilinear form defined on $F \times E$ by

$$(y, x) \mapsto B(x, y).$$

When M is a square matrix, one may take $F = E$ and $\gamma = \beta$. In that case, M is symmetric if and only if B is symmetric: $B(x, y) = B(y, x)$. Likewise, one says that B is *alternate* if $B(x, x) \equiv 0$, that is if M itself is an alternate matrix.

If $B : E \times F \to K$ is bilinear, one can compare the matrices M and M' of B with respect to the bases β, γ and β', γ'. Denoting by P, Q the change-of-basis matrices of $\beta \mapsto \beta'$ and $\gamma \mapsto \gamma'$, one has

$$m'_{ij} = B(e'_i, f'_j) = \sum_{k,l} p_{ki} q_{lj} B(e_k, f_l) = \sum_{k,l} p_{ki} q_{lj} m_{kl}.$$

Therefore,

$$M' = P^T M Q.$$

When $F = E$ and $\gamma = \beta$, $\gamma' = \beta'$, the change of basis has the effect of replacing M by $M' = P^T M P$. In general, M' is not similar to M, though it is so if P is orthogonal. If M is symmetric, then M' is too. This was expected, since one expresses the symmetry of the underlying bilinear form B.

If the characteristic of K is distinct from 2, there is an isomorphism between $\mathbf{Sym}_n(K)$ and the set of quadratic forms on K^n. This isomorphism is given by the formula

$$Q(e_i + e_j) - Q(e_i) - Q(e_j) = 2m_{ij}.$$

In particular, $Q(e_i) = m_{ii}$.

1.3 Exercises

1. Let G be an $\mathbb{R}$-vector space. Verify that its complexification $G^{\mathbb{C}}$ is a $\mathbb{C}$-vector space and that $\dim_{\mathbb{C}} G^{\mathbb{C}} = \dim_{\mathbb{R}} G$.

2. Let $M \in \mathbf{M}_{n\times m}(K)$ and $M' \in \mathbf{M}_{m\times p}(K)$ be given. Show that

$$\mathrm{rk}(MM') \leq \min\{\mathrm{rk}\ M,\ \mathrm{rk}\ M'\}.$$

   First show that $\mathrm{rk}(MM') \leq \mathrm{rk}\ M$, and then apply this result to the transpose matrix.

3. Let K be a field and let A, B, C be matrices with entries in K, of respective sizes $n \times m$, $m \times p$, and $p \times q$.

   (a) Show that $\mathrm{rk}\, A + \mathrm{rk}\, B \leq m + \mathrm{rk}\, AB$. It is sufficient to consider the case where B is onto, by considering the restriction of A to the range of B.

   (b) Show that $\mathrm{rk}\, AB + \mathrm{rk}\, BC \leq \mathrm{rk}\, B + \mathrm{rk}\, ABC$. One may use the vector spaces $K^p / \ker B$ and $R(B)$, and construct three homomorphisms u, v, w, with v being onto.

4. (a) Let $n, n', m, m' \in \mathbb{N}^*$ and let K be a field. If $B \in \mathbf{M}_{n\times m}(K)$ and $C \in \mathbf{M}_{n'\times m'}(K)$, one defines a matrix $B \otimes C \in \mathbf{M}_{nn'\times mm'}(K)$,

the tensor product, whose block form is

$$B \otimes C = \begin{pmatrix} b_{11}C & \cdots & b_{1m}C \\ \vdots & & \vdots \\ b_{n1}C & \cdots & b_{nm}C \end{pmatrix}.$$

Show that $(B, C) \mapsto B \otimes C$ is a bilinear map and that its range spans $\mathbf{M}_{nn' \times mm'}(K)$. Is this map onto?

(b) If $p, p' \in \mathbb{N}^*$ and $D \in \mathbf{M}_{m \times p}(K)$, $E \in \mathbf{M}_{m' \times p'}(K)$, then compute $(B \otimes C)(D \otimes E)$.

(c) Show that for every bilinear form $\phi : \mathbf{M}_{n \times m}(K) \times \mathbf{M}_{n' \times m'}(K) \to K$, there exists one and only one linear form

$$L : \mathbf{M}_{nn' \times mm'}(K) \to K$$

such that $L(B \otimes C) = \phi(B, C)$.

2

Square Matrices

The essential ingredient for the study of square matrices is the determinant. For reasons that will be given in Section 2.5, as well as in Chapter 6, it is useful to consider matrices with entries in a ring. This allows us to consider matrices with entries in $\mathbb{Z}$ (rational integers) as well as in $K[X]$ (polynomials with coefficients in K). We shall assume that the ring A of scalars is a commutative (meaning that the multiplication is commutative) integral domain (meaning that it does not have zero divisors: $ab = 0$ implies either $a = 0$ or $b = 0$), with a unit denoted by 1, that is, an element satisfying $1x = x1 = x$ for every $x \in A$. Observe that the ring $\mathbf{M}_n(A)$ is not commutative if $n \geq 2$. For instance,

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

An element a of A is *invertible* if there exists $b \in A$ such that $ab = 1$. The element b is unique (because A is an integral domain), and one calls it the *inverse* of a, with the notation $b = a^{-1}$. The set of invertible elements of A is a multiplicative group, denoted by A^*. One has

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

2.1 Determinants and Minors

We recall that $\boldsymbol{S}_n$, the *symmetric group*, denotes the group of permutations over the set $\{1, \ldots, n\}$.

Let $M \in \mathbf{M}_n(A)$ be a square matrix. Its determinant is defined by

$$\det M := \sum_{\sigma \in \boldsymbol{S}_n} \epsilon(\sigma) m_{1\sigma(1)} \cdots m_{n\sigma(n)},$$

where the sum ranges over all the permutations of the integers $1, \ldots, n$. We denote by $\epsilon(\sigma) = \pm 1$ the signature of σ, equal to $+1$ if σ is the product an even number of transpositions, and -1 otherwise. Recall that $\epsilon(\sigma\sigma') = \epsilon(\sigma)\epsilon(\sigma')$.

If M is triangular, then all the products vanish other than the one associated with the identity (that is, $\sigma(j) = j$). The determinant of a triangular M is thus equal to the product of diagonal entries m_{ii}. In particular, $\det I_n = 1$ and $\det 0_n = 0$. An analogous calculation shows that the determinant of a block triangular matrix is equal to the product of the determinants of the diagonal blocks M_{jj}.

Since $\epsilon(\sigma^{-1}) = \epsilon(\sigma)$, one has

$$\det M^T = \det M.$$

Looking at M as a row matrix with entries in A^n, one may view the determinant as a multilinear form of the n columns of M:

$$\det M = \det \left(M^{(1)}, \ldots, M^{(n)} \right).$$

This form is *alternate*: If two columns are equal, the determinant vanishes. As a matter of fact, if the ith and the jth columns are equal, one groups the permutations pairwise $(\sigma, \tau\sigma)$, where τ is the transposition (i, j). For each pair, both products are equal, up to the signatures, which are opposite; their sum is thus zero. Likewise, if two rows are equal, the determinant is zero.

More generally, if the columns of M satisfy a non trivial linear relation ($a_1, \ldots, a_n$ not all zero) of linear dependence

$$a_1 M_1 + \cdots + a_n M_n = 0$$

(that is, if $\mathrm{rk}\ M < n$), then $\det M$ is zero. Let us assume, for instance, that a_1 is nonzero. For $j \geq 2$, one has

$$\det \left(M^{(j)}, M^{(2)}, \ldots, M^{(n)} \right) = 0.$$

Using the multilinearity, one has thus

$$\begin{aligned} a_1 \det M &= \det \left(a_1 M^{(1)} + \cdots + a_n M^{(n)}, M^{(2)}, \ldots, M^{(n)} \right) \\ &= \det \left(0, M^{(2)}, \ldots \right) = 0. \end{aligned}$$

Since A is an integral domain, we conclude that $\det M = 0$.

For a matrix $M \in \mathbf{M}_{n \times m}(A)$, not necessarily square, and $p \geq 1$ an integer with $p \leq m, n$, one may extract a $p \times p$ matrix $M' \in \mathbf{M}_p(A)$ by retaining only p rows and p columns of M. The determinant of such a matrix M' is

called a *minor of order p*. Once the choice of the row indices $i_1 < \cdots < i_p$ and column indices $j_1 < \cdots < j_p$ has been made, one denotes by

$$M \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$$

the corresponding minor. A *principal minor* is a minor with equal row and column indices, that is, of the form

$$M \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ i_1 & i_2 & \cdots & i_p \end{pmatrix}.$$

In particular, the *leading* principal minor of order p is

$$M \begin{pmatrix} 1 & 2 & \cdots & p \\ 1 & 2 & \cdots & p \end{pmatrix}.$$

Given a matrix $M \in \mathbf{M}_n(A)$, one associates the matrix $\hat{M}$ of *cofactors*, defined as follows: its (i,j)-th entry $\hat{m}_{ij}$ is the minor of order $n-1$ obtained by removing the ith row and the jth column multiplied by $(-1)^{i+j}$. It is also the factor of m_{ij} in the formula for the determinant of M. Finally, we define the *adjoint* matrix $\operatorname{adj} M$ by

$$\operatorname{adj} M := \hat{M}^T.$$

Proposition 2.1.1 *If $M \in \mathbf{M}_n(A)$, one has*

$$M(\operatorname{adj} M) = (\operatorname{adj} M)M = \det M \cdot I_n. \tag{2.1}$$

Proof

The identity is clear as far as diagonal terms are concerned; it amounts to the definition of the determinant (see also below). The off-diagonal terms m'_{ij} of $M(\operatorname{adj} M)$ are sums involving on the one hand an index, and on the other hand a permutation $\sigma \in \boldsymbol{S}_n$. One groups the terms pairwise, corresponding to permutations σ and $\sigma\tau$, where τ is the tranposition (i,j). The sum of two such terms is zero, so that $m'_{ij} = 0$.

■

Proposition 2.1.1 contains the well-known and important expansion formula for the determinant with respect to either a row or a column. The expansion with respect to the ith row is written

$$\det M = (-1)^{i+1} m_{i1}\hat{m}_{i1} + \cdots + (-1)^{i+n} m_{in}\hat{m}_{in},$$

while the expansion with respect to the ith column is

$$\det M = (-1)^{i+1} m_{1i}\hat{m}_{1i} + \cdots + (-1)^{i+n} m_{ni}\hat{m}_{ni}.$$

2.1.1 Irreducibility of the Determinant

By definition, the determinant is a polynomial function, in the sense that $\det M$ is the value taken by a polynomial $\operatorname{Det}_A \in A[x_{11}, \ldots, x_{nn}]$ when the

x_{ij}'s are replaced by the scalars m_{ij}. We observe that Det_A does not really depend on the ring A, in the sense that it is the image of $\mathrm{Det}_{\mathbb{Z}}$ through the canonical ring homomorphism $\mathbb{Z} \to A$. For this reason, we shall simply write Det. The polynomial Det may be viewed as the determinant of the matrix $X = (x_{ij})_{1 \leq i,j \leq n} \in \mathbf{M}_n(A[x_{11}, \ldots, x_{nn}])$.

Theorem 2.1.1 *The polynomial* Det *is irreducible in* $A[x_{11}, \ldots, x_{nn}]$.

Proof

We shall proceed by induction on the size n. If $n = 1$, there is nothing to prove. Thus let us assume that $n \geq 2$. We denote by D the ring of polynomials in the x_{ij} with $(i,j) \neq (1,1)$, so that $A[x_{11}, \ldots, x_{nn}] = D[x_{11}]$. From the expansion with respect to the first row, we see that $\mathrm{Det} = x_{11}P + Q$, with $P, Q \in D$. Since Det is of degree one as a polynomial in x_{11}, any factorization must be of the form $(x_{11}R + S)T$, with $R, S, T \in D$. In particular, $RT = P$.

By induction, and since P is the polynomial Det of $(n-1) \times (n-1)$ matrices, it is irreducible in E, the ring of polynomials in the x_{ij}'s with $i, j > 1$. Therefore, it is also irreducible in D, since D is the polynomial ring $E[x_{12}, \ldots, x_{1n}, x_{21}, \ldots, x_{n1}]$. Therefore, we may assume that either R or T equals 1.

If the factorization is nontrivial, then $R = 1$ and $T = P$. It follows that P divides Det. An expansion with respect to various rows shows similarly that every minor of size $n-1$, considered as an element of $A[x_{11}, \ldots, x_{nn}]$, divides Det. However, each such minor is irreducible, and they are pairwise distinct, since they do not depend on the same set of x_{ij}'s. We conclude that the product of all minors of size $n-1$ divides Det. In particular, the degree n of Det is greater than or equal to the degree $n^2(n-1)$ of this product, an obvious contradiction.

■

2.1.2 The Cauchy–Binet Formula

In the sequel, we shall use also the following result.

Proposition 2.1.2 *Let* $B \in \mathbf{M}_{n \times m}(A)$, $C \in \mathbf{M}_{m \times l}(A)$, *and an integer* $p \leq n, l$ *be given. Let* $1 \leq i_1 < \cdots < i_p \leq n$ *and* $1 \leq k_1 < \cdots < k_p \leq l$ *be indices. Then the minor*

$$(BC)\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ k_1 & k_2 & \cdots & k_p \end{pmatrix}$$

is given by the formula

$$\sum_{1 \leq j_1 < j_2 < \cdots < j_p \leq m} B\begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix} \cdot C\begin{pmatrix} j_1 & j_2 & \cdots & j_p \\ k_1 & k_2 & \cdots & k_p \end{pmatrix}.$$

Corollary 2.1.1 *Let $b, c \in A$. If b divides every minor of order p of B and if c divides every minor of order p of C, then bc divides every minor of order p of BC.*

The particular case $l = m = n$ is fundamental:

Theorem 2.1.2 *If $B, C \in \mathbf{M}_n(A)$, then $\det(BC) = \det B \cdot \det C$.*

In other words, the determinant is a multiplicative homomorphism from $\mathbf{M}_n(A)$ to A.

Proof

The corollaries are trivial. We only prove the Cauchy–Binet formula. Since the calculation of the ith row (respectively the jth column) of BC involves only the ith row of B (respectively the jth column of C), one may assume that $p = n = l$. The minor to be evaluated is then $\det BC$. If $m < n$, there is nothing to prove, since on the one hand the rank of BC is less than or equal to m, thus $\det BC$ is zero, and on the other hand the left-hand side sum in the formula is empty.

There remains the case $m \geq n$. Let us write the determinant of a matrix P as that of its columns P_j and let us use the multilinearity of the determinant:

$$\begin{aligned}
\det BC &= \det\left(\sum_{j_1=1}^{n} c_{j_1 1} B_{j_1}, (BC)_2, \dots, (BC)_n\right) \\
&= \sum_{j_1=1}^{n} c_{j_1 1} \det\left(B_{j_1}, \sum_{j_2=1}^{n} c_{j_2 2} B_{j_2}, (BC)_3, \dots, (BC)_n\right) \\
&= \cdots = \sum_{1 \leq j_1, \dots, j_n \leq n} c_{j_1 1} \cdots c_{j_n n} \det(B_{j_1}, \dots, B_{j_n}).
\end{aligned}$$

In the sum the determinant is zero as soon as $f \mapsto j_f$ is not injective, since then there are two identical columns. If on the contrary j is injective, this determinant is a minor of B, up to the sign. This sign is that of the permutation that puts $j_1, \dots, j_p$ in increasing order. Grouping in the sum the terms corresponding to the same minor, we find that $\det BC$ equals

$$\sum_{1 \leq k_1 < \cdots < k_n \leq m,} \sum_{\sigma \in \boldsymbol{S}_n} \epsilon(\sigma) c_{k_1 \sigma(1)} \cdots c_{k_n \sigma(n)} B \begin{pmatrix} 1 & 2 & \cdots & n \\ k_1 & k_2 & \cdots & k_n \end{pmatrix},$$

which is the required formula.

■

2.2 Invertibility

Since $\mathbf{M}_n(A)$ is not an integral domain, the notion of invertible elements of $\mathbf{M}_n(A)$ needs an auxiliary result, presented below.

Proposition 2.2.1 *Given $M \in \mathbf{M}_n(A)$, the following assertions are equivalent:*

1. *There exists $N \in \mathbf{M}_n(A)$ such that $MN = I_n$.*
2. *There exists $N' \in \mathbf{M}_n(A)$ such that $N'M = I_n$.*
3. $\det M$ *is invertible.*

If M satisfies one of these equivalent conditions, then the matrices N, N' are unique and one has $N = N'$.

Definition 2.2.1 *One then says that M is* invertible. *One also says sometimes that M is* nonsingular, *or* regular. *One calls the matrix $N = N'$ the* inverse *of M, and one denotes it by M^{-1}. If M is not invertible, one says that M is* singular.

Proof

Let us show that (1) is equivalent to (3). If $MN = I_n$, then $\det M \cdot \det N = 1$; hence $\det M \in A^*$. Conversely, if $\det M$ is invertible, $(\det M)^{-1}\hat{M}^T$ is an inverse of M by (2.1). Analogously, (2) is equivalent to (3). The three assertions are thus equivalent.

If $MN = N'M = I_n$, one has $N = (N'M)N = N'(MN) = N'$. This equality between the left and right inverses shows that these are unique. ■

The set of the invertible elements of $\mathbf{M}_n(A)$ is denoted by $\mathbf{GL}_n(A)$ (for "general linear group"). It is a multiplicative group, and one has

$$(MN)^{-1} = N^{-1}M^{-1}, \quad (M^k)^{-1} = (M^{-1})^k, \quad (M^T)^{-1} = (M^{-1})^T.$$

The matrix $(M^T)^{-1}$ is also written M^{-T}. If $k \in \mathbb{N}$, one writes $M^{-k} = (M^k)^{-1}$ and one has $M^j M^k = M^{j+k}$ for every $j, k \in \mathbb{Z}$.

The set of the matrices of determinant one is a normal subgroup of $\mathbf{GL}_n(A)$, since it is the kernel of the homomorphism $M \mapsto \det M$. It is called the *special linear group* and is denoted by $\mathbf{SL}_n(A)$.

The orthogonal matrices are invertible, and they satisfy the relation $M^{-1} = M^T$. In particular, orthogonality is equivalent to $MM^T = I_n$. The set of orthogonal matrices with entries in a field K is obviously a multiplicative group, and is denoted by $\mathbf{O}_n(K)$. It is called the *orthogonal group*. The determinant of an orthogonal matrix equals ± 1, since

$$1 = \det M \cdot \det M^T = (\det M)^2.$$

The set $\mathbf{SO}_n(K)$ of orthogonal matrices with determinant equal to 1 is obviously a normal subgroup of the orthogonal group. It is called the *special orthogonal group*. It is simply the intersection of $\mathbf{O}_n(K)$ with $\mathbf{SL}_n(K)$.

A triangular matrix is invertible if and only if its diagonal entries are invertible; its inverse is then triangular of the same type, upper or lower. The proposition below is an immediate application of Theorem 2.1.2.

Proposition 2.2.2 *If $M, M' \in \mathbf{M}_n(A)$ are similar (that is, $M' = P^{-1}MP$ with $P \in \mathbf{GL}_n(A)$), then*

$$\det M' = \det M.$$

2.3 Alternate Matrices and the Pfaffian

The very simple structure of alternate forms is described in the following statement.

Proposition 2.3.1 *Let B be an alternate bilinear form on a vector space E, of dimension n. Then there exists a basis*

$$\{x_1, y_1, \dots, x_k, y_k, z_1, \dots, z_{n-2k}\}$$

such that the matrix of B in this basis is block-diagonal, equal to $\mathrm{diag}(J, \dots, J, 0, \dots, 0)$, *with k blocks J defined by*

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

Proof
We proceed by induction on the dimension n. If $B = 0$, there is nothing to prove. If B is nonzero, there exist two vectors x_1, y_1 such that $B(x_1, y_1) \neq 0$. Multiplying one of them by $B(x_1, y_1)^{-1}$, one may assume that $B(x_1, y_1) = 1$. Since B is alternate, $\{x_1, y_1\}$ is free. Let N be the plane spanned by x_1, y_1. The set of vectors x satisfying $B(x, v) = 0$ (or equivalently $B(v, x) = 0$, since B must be skew-symmetric) for every v in N is denoted by $N^\perp$. The formulas

$$B(ax_1 + by_1, x_1) = -b, \quad B(ax_1 + by_1, y_1) = a$$

show that $N \cap N^\perp = \{0\}$. Additionally, every vector $x \in E$ can be written as $x = y + n$, where $n \in N$ and $y \in N^\perp$ are given by

$$n = B(x, y_1)x_1 - B(x, x_1)y_1, \quad y := x - n.$$

Therefore, $E = N \oplus N^\perp$. We now consider the restriction of B to the subspace $N^\perp$ and apply the induction hypothesis. There exists a basis $\{x_2, y_2, \dots, x_k, y_k, z_1, \dots, z_{n-2k}\}$ such that the matrix of the restriction of B in this basis is block-diagonal, equal to $\mathrm{diag}(J, \dots, J, 0, \dots, 0)$, with $k-1$ blocks J, which means that $B(x_j, y_j) = 1 = -B(y_j, x_j)$ and $B(u, v) = 0$ for every other choice of u, v in the basis. Obviously, this property extends to the form B itself and the basis $\{x_1, y_1, \dots, x_k, y_k, z_1, \dots, z_{n-2k}\}$.

■

We now choose an alternate matrix $M \in M_n(K)$ and apply Proposition 2.3.1 to the form defined by M. In view of Section 1.2.3, we have the following.

Corollary 2.3.1 *Given an alternate matrix $M \in M_n(K)$, there exists a matrix $Q \in \mathbf{GL}_n(K)$ such that*

$$M = Q^T \operatorname{diag}(J, \dots, J, 0, \dots, 0)Q. \tag{2.2}$$

Obviously, the rank of M, being the same as that of the block-diagonal matrix, equals twice the number of J blocks. Finally, since $\det J = 1$, we have $\det M = \epsilon(\det Q)^2$, where $\epsilon = 0$ if there is a zero diagonal block in the decomposition, and $\epsilon = 1$ otherwise. Thus we have proved the following result.

Proposition 2.3.2 *The rank of an alternate matrix M is even. The number of J blocks in the identity (2.2) is the half of that rank. In particular, it does not depend on the decomposition. Finally, the determinant of an alternate matrix is a square in K.*

A very important application of Proposition 2.3.2 concerns the *Pfaffian*, whose crude definition is a polynomial whose square is the determinant of the general alternate matrix. First of all, since the rank of an alternate matrix is even, $\det M = 0$ whenever n is odd. Therefore, we restrict our attention from now on to the even-dimensional case $n = 2m$. Let us consider the field $F = \mathbb{Q}(x_{ij})$ of rational functions with rational coefficients, in $n(n-1)/2$ indeterminates x_{ij}, $i < j$. We apply the proposition to the alternate matrix X whose (i,i)-entry is 0 and (i,j)-entry (respectively (j,i)-entry) is x_{ij} (respectively $-x_{ij}$). Its determinant, a polynomial in $\mathbb{Z}[x_{ij}]$, is the square of some irreducible rational function f/g, where f and g belong to $\mathbb{Z}[x_{ij}]$. From $g^2 \det X = f^2$, we see that g divides f in $\mathbb{Z}[x_{ij}]$. But since f and g are coprime, one finds that g is invertible; in other words $g = \pm 1$. Thus

$$\det X = f^2. \tag{2.3}$$

Now let k be a field and let $M \in M_n(k)$ be alternate. There exists a unique homomorphism from $\mathbb{Z}[x_{ij}]$ into k sending x_{ij} to m_{ij}. From equation (2.3) we obtain

$$\det M = (f(m_{12}, \dots, m_{n-1,n}))^2. \tag{2.4}$$

In particular, if $k = \mathbb{Q}$ and $M = \operatorname{diag}(J, \dots, J)$, one has $f^2 = 1$. Up to multiplication by ± 1, which leaves unchanged the identity (2.3), we may assume that $f = 1$ for this special case. This determination of the polynomial f is called the *Pfaffian* and is denoted by Pf. It may be viewed as a polynomial function on the vector space of alternate matrices with entries in a given field k. equation (2.4) now reads

$$\det M = (\operatorname{Pf}(M))^2. \tag{2.5}$$

Given an alternate matrix $M \in M_n(k)$ and a matrix $Q \in M_n(k)$, we consider the Pfaffian of the alternate matrix $Q^T M Q$. We first consider the case of the field of fractions $\mathbb{Q}(x_{ij}, y_{ij})$ in the $n^2 + n(n-1)/2$ indeterminates

x_{ij} $(1 \leq i < j \leq n)$ and y_{ij} $(1 \leq i, j \leq n)$. Let Y be the matrix whose (i,j)-entry is y_{ij}. Then, with X as above,

$$(\mathrm{Pf}(Y^T XY))^2 = \det Y^T XY = (\det Y)^2 \det X = (\mathrm{Pf}(X) \det Y)^2.$$

Since $\mathbb{Z}[x_{ij}, y_{ij}]$ is an integral domain, we have the polynomial identity

$$\mathrm{Pf}\left(Y^T XY\right) = \epsilon\, \mathrm{Pf}(X) \det Y, \quad \epsilon = \pm 1.$$

As above, one infers that $\mathrm{Pf}(Q^T MQ) = \pm\, \mathrm{Pf}(M) \det Q$ for every field k, matrix $Q \in M_n(k)$, and alternate matrix $M \in M_n(k)$. Inspection of the particular case $Q = I_n$ yields $\epsilon = 1$. We summarize these results now.

Theorem 2.3.1 *Let $n = 2m$ be an even integer. There exists a unique polynomial* Pf *in the indeterminates x_{ij} $(1 \leq i < j \leq n)$ with integer coefficients such that:*

- *For every field k and every alternate matrix $M \in M_n(k)$, one has $\det M = \mathrm{Pf}(M)^2$.*
- *If $M = \mathrm{diag}(J, \ldots, J)$, then $\mathrm{Pf}(M) = 1$.*

Moreover, if $Q \in M_n(k)$ is given, then $\mathrm{Pf}\left(Q^T MQ\right) = \mathrm{Pf}(M) \det Q$.

We warn the reader that if $m > 1$, there does not exist a matrix $Z \in \mathbb{Q}[x_{ij}]$ such that $X = Z^T \mathrm{diag}(J, \ldots, J)Z$. The factorization of the polynomial $\det X$ does not correspond to a similar factorization of X itself. In other words, the decomposition $X = Q^T \mathrm{diag}(J, \ldots, J)Q$ in $M_n(\mathbb{Q}(x_{ij}))$ cannot be written within $M_n(\mathbb{Q}[x_{ij}])$.

The Pfaffian is computed easily for small values of n. For instance, $\mathrm{Pf}(X) = x_{12}$ if $n = 2$, and $\mathrm{Pf} = x_{12}x_{34} - x_{13}x_{24} + x_{14}x_{23}$ if $n = 4$.

2.4 Eigenvalues and Eigenvectors

Let K be a field and E, F two vector spaces of finite dimension. Let us recall that if $u : E \mapsto F$ is a linear map, then

$$\dim E = \dim \ker u + \mathrm{rk}\, u,$$

where $\mathrm{rk}\, u$ denotes the dimension of $u(E)$ (the *rank* of u). In particular, if $u \in \mathrm{End}(E)$, then

$$u \text{ is bijective} \Longleftrightarrow u \text{ is injective} \Longleftrightarrow u \text{ is surjective.}$$

However, u is bijective, that is invertible, in $\mathrm{End}(E)$, if and only if its matrix M in some basis β is invertible, that is if its determinant is nonzero. As a matter of fact, the matrix of u^{-1} is M^{-1}; the existence of an inverse (either that of M or that of u) implies that of the other one. Finally, if $M \in \mathbf{M}_n(K)$, then $\det M \neq 0$ is equivalent to

$$\forall X \in K^n, \quad MX = 0 \Longrightarrow X = 0.$$

In other words,

$$\det M = 0 \Longleftrightarrow (\exists X \in K^n, X \neq 0,\ MX = 0).$$

More generally, since $MX = \lambda X$ ($\lambda \in K$) can also be written $(\lambda I_n - M)X = 0$, one sees that $\det(\lambda I_n - M)$ is zero if and only if there exists a nonzero vector in K^n such that $MX = \lambda X$. One then says that λ is an *eigenvalue* of M in K, and that X is an *eigenvector* associated to λ. An eigenvector is thus always a nonzero vector. The set of the eigenvalues of M in K is called the *spectrum* of M and is denoted by $\mathrm{Sp}_K(M)$.

A matrix in $\mathbf{M}_n(K)$ may have no eigenvalues in K, as the following example demonstrates, with $K = \mathbb{R}$:

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

In order to understand in detail in the structure of a square matrix $M \in \mathbf{M}_n(K)$, one is thus led to consider M as a matrix with entries in $\overline{K}$. One then writes $\mathrm{Sp}(M)$ instead of $\mathrm{Sp}_{\overline{K}}(M)$, and one has $\mathrm{Sp}_K(M) = K \cap \mathrm{Sp}(M)$, since the eigenvalues are characterized by $\det(\lambda I_n - M) = 0$, and this equality has the same meaning in $\overline{K}$ as in K when $\lambda \in K$.

2.5 The Characteristic Polynomial

The previous calculations show that the eigenvalues of $M \in \mathbf{M}_n(K)$ are the roots of the polynomial

$$P_M(X) := \det(XI_n - M).$$

Let us observe in passing that if X is an indeterminate, then $XI_n - M \in \mathbf{M}_n(K(X))$. Its determinant P_M is thus well-defined, since $K(X)$ is a commutative integral domain with a unit element. One calls P_M the *characteristic polynomial* of M. Substituting 0 for X, one sees that the constant term in P_M is simply $(-1)^n \det M$. Since the term corresponding to the permutation $\sigma = \mathrm{id}$ in the computation of the determinant is of degree n (it is $\prod_i (X - m_{ii})$) and since the products corresponding to the other permutations are of degree less than or equal to $n-2$, one sees that P_M is of degree n, with

$$P_M(X) = X^n - \left(\sum_{i=1}^{n} m_{ii}\right) X^{n-1} + \cdots + (-1)^n \det M.$$

The coefficient

$$\sum_{i=1}^{n} m_{ii}$$

is called the *trace* of M and is denoted by $\operatorname{Tr} M$. One has the trivial formula that if $N \in \mathbf{M}_{n\times m}(K)$ and $P \in \mathbf{M}_{m\times n}(K)$, then

$$\operatorname{Tr}(NP) = \operatorname{Tr}(PN).$$

For square matrices, this identity also becomes

$$\operatorname{Tr}[N, P] = 0.$$

Since P_M possesses n roots in $\overline{K}$, counting multiplicities, one sees that a square matrix has always at least one eigenvalue, which, however, does not necessarily belong to K. The multiplicity of λ as a root of P_M is called *algebraic multiplicity* of the eigenvalue λ. The *geometric multiplicity* of λ is the dimension of $\ker(\lambda I_n - M)$ in K^n. The sum of the algebraic multiplicities of the eigenvalues of M (considered in $\overline{K}$) is n, the size of the matrix. An eigenvalue of algebraic multiplicity one (that is, a simple root of P_M) is called *simple*. It is *geometrically simple* if its geometric multiplicity equals one.

The characteristic polynomial is a *similarity invariant*, in the following sense:

Proposition 2.5.1 *If M and M' are similar, then $P_M = P_{M'}$. In particular,* $\det M = \det M'$ *and* $\operatorname{Tr} M = \operatorname{Tr} M'$.

The proof is immediate. One deduces that the eigenvalues and their algebraic multiplicities are similarity invariants. This is also true for the geometric multiplicities, by a direct comparison of the kernel of $\lambda I_n - M$ and of $\lambda I_n - M'$. Furthermore, the expression obtained above for the characteristic polynomial provides the following result.

Proposition 2.5.2 *The product of the eigenvalues of M (considered in $\overline{K}$), counted with their algebraic multiplicities, is* $\det M$. *Their sum is* $\operatorname{Tr} M$.

Let μ be the geometric multiplicity of an eigenvalue λ of M. Let us choose a basis γ of $\ker(\lambda I_n - M)$, and then a basis of β of $\overline{K}^n$ that completes γ. Using the change-of-basis matrix from the canonical basis to β, one sees that M is similar to a matrix $M' = P^{-1}MP$, whose μ first columns have the form

$$\begin{pmatrix} \lambda I_\mu \\ 0_{n-\mu,\mu} \end{pmatrix}.$$

A direct calculation shows then that $(X-\lambda)^\mu$ divides $P_{M'}$, that is, P_M. The geometric multiplicity is thus less than or equal to the algebraic multiplicity.

The characteristic polynomials of M and M^T are equal. Thus, M and M^T have the same eigenvalues. We shall show in Chapter 6 a much deeper result, namely M and M^T are similar.

The main result concerning the characteristic polynomial is the Cayley–Hamilton theorem:

Theorem 2.5.1 *Let $M \in \mathbf{M}_n(K)$. Let*

$$P_M(X) = X^n + a_1 X^{n-1} + \cdots + a_n$$

be its characteristic polynomial. Then the matrix

$$M^n + a_1 M^{n-1} + \cdots + a_n I_n$$

equals 0_n.

One also writes $P_M(M) = 0$. Though this formula looks trivial (obviously, $\det(MI_n - M) = 0$), it is not. Actually, it must be understood in the following way. Let us consider the expression $XI_n - M$ as a matrix with entries in $K[X]$. When one substitutes a matrix N for the indeterminate X in $XI_n - M$, one obtains a matrix of $\mathbf{M}_n(A)$, where A is the subring of $\mathbf{M}_n(K)$ spanned by I_n and N (one denotes it by $K(N)$). The ring A is commutative (but is not an integral domain in general), since it is the set of the $q(N)$ for $q \in K[X]$. Therefore,

$$P_M(N) = \begin{pmatrix} N - m_{11}I_n & & & \\ & \ddots & -m_{ij}I_n & \\ & & \ddots & \\ & & & N - m_{nn}I_n \end{pmatrix}.$$

The Cayley–Hamilton theorem expresses that the determinant (which is an element of $\mathbf{M}_n(K)$, rather than of K) of this matrix is zero.

Proof

Let $R \in \mathbf{M}_n(K(X))$ be the matrix $XI_n - M$, and let S be the adjoint of R. Each s_{ij} is a polynomial of degree less than or equal to $n-1$, because the products arising in the calculation of the cofactors involve $n-1$ linear or constant terms. Thus we may write

$$S = S_0 X^{n-1} + \cdots + S_{n-1},$$

where $S_j \in \mathbf{M}_n(K)$. Let us now write $RS = (\det R)I_n = P_M(X)I_n$:

$$(XI_n - M)(S_0 X^{n-1} + \cdots + S_{n-1}) = (X^n + a_1 X^{n-1} + \cdots + a_n)I_n.$$

Identifying the powers of X, we obtain

$$\begin{aligned} S_0 &= I_n, \\ S_1 - MS_0 &= a_1 I_n, \\ &\vdots \\ S_j - MS_{j-1} &= a_j I_n, \\ &\vdots \\ S_{n-1} - MS_{n-2} &= a_{n-1} I_n, \\ -MS_{n-1} &= a_n I_n. \end{aligned}$$

Let us multiply these rows by the powers of M, beginning with M^n and ending with $M^0 = I_n$. Summing the obtained equalities, we obtain the expected formula.

■

For example, every 2×2 matrix satisfies the identity

$$M^2 - (\mathrm{Tr}\ M)M + (\det M)I_2 = 0.$$

2.5.1 The Minimal Polynomial

For a square matrix $M \in \mathbf{M}_n(K)$, let us denote by J_M the set of polynomials $Q \in K[X]$ such that $Q(M) = 0$. It is clearly an ideal of $K[X]$. Since $K[X]$ is Euclidean, hence principal (see Sections 6.1.1 and 6.1.2), there exists a polynomial Q_M such that $J_M = K[X]Q_M$. In other words, $Q(M) = 0$ and $Q \in K[X]$ imply $Q_M | Q$. Theorem 2.5.1 shows that the ideal J_M does not reduce to $\{0\}$, because it contains the characteristic polynomial. Hence, $Q_M \neq 0$ and one may choose it monic. This choice determines Q_M in a unique way, and one calls it the *minimal polynomial* of M. It divides the characteristic polynomial.

Contrary to the case of the characteristic polynomial, it is not immediate that the minimal polynomial is independent of the field in which one considers J_M (note that we consider only fields that contain the entries of M). We shall see in Section 6.3.2 that if L is a field containing K, then the minimal polynomials of M in $K[X]$ and $L[X]$ are the same. This explains the terminology.

Two similar matrices obviously have the same minimal polynomial, since

$$Q(P^{-1}MP) = P^{-1}Q(M)P.$$

If λ is an eigenvalue of M, associated to an eigenvector X, and if $q \in K[X]$, then $q(\lambda)X = q(M)X$. Applied to the minimal polynomial, this equality shows that the minimal polynomial is divisible by $X - \lambda$. Hence, if P_M splits over $\bar{K}$ in the form

$$\prod_{j=1}^{r} (X - \lambda_j)^{n_j},$$

the λ_j all being distinct, then the minimal polynomial can be written as

$$\prod_{j=1}^{r} (X - \lambda_j)^{m_j},$$

with $1 \leq m_j \leq n_j$. In particular, if every eigenvalue of M is simple, the minimal polynomial and the characteristic polynomial are equal.

An eigenvalue is called *semi-simple* if it is a simple root of the minimal polynomial.

2.6 Diagonalization

If $\lambda \in K$ is an eigenvalue of M, one calls the linear subspace $E_K(\lambda) = \ker(M - \lambda I_n)$ in K^n the *eigenspace* associated to λ. It is formed of eigenvectors associated to λ on the one hand, and of the zero vector on the other hand. Its dimension is nonzero. If L is a field containing K (an "extension" of K), then $\dim_K E_K(\lambda) = \dim_L E_L(\lambda)$. This equality is not obvious. It follows from the third canonical form with Jordan blocks, which we shall see in Section 6.3.3.

If $\lambda_1, \dots, \lambda_r$ are distinct eigenvalues, then the eigenspaces are in direct sum. That is,

$$(x_1 \in E_K(\lambda_1), \dots, x_r \in E_K(\lambda_r), x_1 + \cdots + x_r = 0) \Longrightarrow (x_1 = \cdots = x_r = 0).$$

As a matter of fact, if there existed a relation $x_1 + \cdots + x_s = 0$ where $x_1, \dots, x_s$ did not vanish simultaneously (we say that it has *length* s), one could choose such a relation of minimal length r. One then would have $r \geq 2$. Multiplying this relation by $M - \lambda_r I_n$, one would obtain

$$(\lambda_1 - \lambda_r)x_1 + \cdots + (\lambda_{r-1} - \lambda_r)x_{r-1} = 0,$$

which is a nontrivial relation of length $r - 1$ for the vectors $(\lambda_j - \lambda_r)x_j \in E_K(\lambda_j)$. This contradicts the minimality of r.

If all the eigenvalues of M are in K and if the algebraic and geometric multiplicities coincide for each eigenvalue of M, the sum of the dimensions of the eigenspaces equals n. Since these linear subspaces are in direct sum, one deduces that

$$K^n = E(\lambda_1) \oplus \cdots \oplus E(\lambda_r).$$

Thus one may choose a basis of K^n formed of eigenvectors. If P is the change-of-basis matrix from the canonical basis to the new one, then $M' = P^{-1}MP$ is diagonal, and its diagonal terms are the eigenvalues, repeated with their multiplicities. One says that M is *diagonalizable* in K. A particular case is that in which the eigenvalues of M are in K and are simple.

Conversely, if M is similar, in $\mathbf{M}_n(K)$, to a diagonal matrix $M' = P^{-1}MP$, then P is a change-of-basis matrix from the canonical basis to an *eigenbasis* (that is, a basis composed of eigenvectors) of M. Hence, M is diagonalizable if and only if the algebraic and geometric multiplicities of each eigenvalue coincide.

Two obstacles could prevent M from being diagonalizable in K. The first one is that an eigenvalue of M does not belong to K. One can always overcome this difficulty by moving towards $\mathbf{M}_n(\overline{K})$. The second one is more serious: In $\overline{K}$, the geometric multiplicity of an eigenvalue can be strictly less than its algebraic multiplicity. For instance, a triangular matrix whose diagonal vanishes has only one eigenvalue, zero, of algebraic multiplicity n. Such a matrix is nilpotent. However it is diagonalizable only if it is 0_n,

since $M = PM'P^{-1}$ and $M' = 0$ imply $M = 0$. Hence,

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

is not diagonalizable.

2.7 Trigonalization

Let us begin with an application of the Cayley–Hamilton theorem.

Proposition 2.7.1 *Let $M \in \mathbf{M}_n(K)$ and let P_M be its characteristic polynomial. If $P_M = QR$ with coprime factors $Q, R \in K[X]$, then $K^n = E \oplus F$, where E, F are the ranges of $Q(M)$ and $R(M)$, respectively. Moreover, one has $E = \ker R(M)$, $F = \ker Q(M)$.*

More generally, if $P_M = R_1 \cdots R_s$, where the R_s are coprime, one has $K^n = E_1 \oplus \cdots \oplus E_s$ with $E_j = \ker R_j(M)$.

Proof

It is sufficient to prove the first assertion. From Bézout's theorem, there exists $R_1, Q_1 \in K[X]$ such that $RR_1 + QQ_1 = 1$. Hence, every $x \in K^n$ can be written as a sum $y + z$ with $y = Q(M)(Q_1(M)x) \in E$, and similarly $z = R(M)(R_1(M)x) \in F$. Hence $K^n = E + F$.

Furthermore, for every $y \in E$, the Cayley–Hamilton theorem says that $R(M)y = 0$. Likewise, $z \in F$ implies $Q(M)z = 0$. If $x \in E \cap F$, one has thus $R(M)x = Q(M)x = 0$. Again using Bézout's theorem, one obtains $x = 0$. This proves $K^n = E \oplus F$.

Finally, $E \subset \ker R(M)$. Since these two vector spaces have the same dimension (namely $n - \dim F$), they are equal.

■

If K is algebraically closed, we can split P_M in the form

$$P_M(X) = \prod_{\lambda \in \mathrm{Sp}(M)} (X - \lambda)^{n_\lambda}.$$

From Proposition 2.7.1 one has $K^n = \oplus_\lambda E_\lambda$, where $E_\lambda = \ker(M - \lambda I)^{n_\lambda}$ is called a *generalized eigenspace*. Choosing a basis in each E_λ, we obtain a new basis $\mathcal{B}$ of K^n. If P is the matrix of the linear transformation from the canonical basis to $\mathcal{B}$, the matrix PMP^{-1} is block-diagonal, because each E_λ is stable under the action of M:

$$PMP^{-1} = \mathrm{diag}(\dots, M_\lambda, \dots).$$

The matrix M_λ is that of the restriction of M to E_λ. Since $E_\lambda = \ker(M - \lambda I)^{n_\lambda}$, one has $(M_\lambda - \lambda I)^{n_\lambda} = 0$, so that λ is the unique eigenvalue of M_λ.

Let us define $N_\lambda = M_\lambda - \lambda I_{n_\lambda}$, which is nilpotent. Let us also write

$$\begin{aligned} D' &= \mathrm{diag}(\dots, \lambda I_{n_\lambda}, \dots), \\ N' &= \mathrm{diag}(\dots, N_\lambda, \dots), \end{aligned}$$

and then $D = P^{-1}D'P$, $N = P^{-1}N'P$. The matrices D', N' are respectively diagonal and nilpotent. Moreover, they commute with each other: $D'N' = N'D'$. One deduces the following result.

Proposition 2.7.2 *If K is algebraically closed, every matrix $M \in \mathbf{M}_n(K)$ decomposes as a sum $M = D+N$, where D is diagonalizable, N is nilpotent, $DN = ND$, and $\mathrm{Sp}(D) = \mathrm{Sp}(M)$.*

Let us continue this analysis.

Lemma 2.7.1 *Every nilpotent matrix is similar to a strictly upper triangular matrix (and also to a strictly lower triangular one).*

Proof

Let us consider the nondecreasing sequence of linear subspaces $E_k = \ker N^k$. Since $E_0 = \{0\}$ and $E_r = K^n$ for a suitable r, one can find a basis $\{x^1, \dots, x^n\}$ of K^n such that $\{x^1, \dots, x^j\}$ is a basis of E_k if $j = \dim E_k$ (use the theorem that any linearly independent set can be enlarged to a basis). Since $N(E_{k+1}) = E_k$, $Nx^j \in E_k$. If P is the change-of-basis matrix from this basis to the canonical one, then PNP^{-1} is strictly upper triangular.

■

Let us return to the decomposition $PMP^{-1} = D' + N'$ above. Each N_λ can be written, from the lemma, in the form $R_\lambda^{-1}T_\lambda R_\lambda$, where T_λ is strictly upper triangular. Then $R_\lambda(D_\lambda + N_\lambda)R_\lambda^{-1} = D_\lambda + T_\lambda$ is triangular. Let us set

$$R = \mathrm{diag}(\dots, R_\lambda, \dots).$$

Then $(RP)M(RP)^{-1}$ is block-diagonal, with the diagonal blocks upper triangular, and hence this matrix is itself upper triangular.

Theorem 2.7.1 *If K is algebraically closed, then every square matrix is similar to a triangular matrix (one says that it is trigonalizable).*

More generally, if the characteristic polynomial of $M \in \mathbf{M}_n(K)$ splits as the product of linear factors, then M is trigonalizable.

A direct proof of this theorem that does not use the three previous statements is possible. Its strategy is used in the proof of Theorem 3.1.3

2.8 Irreducibility

A square matrix A is said *reducible* if there exists a nontrivial partition $\{1, \dots, n\} = I \cup J$ such that $(i, j) \in I \times J$ implies $a_{ij} = 0$. It is *irreducible*

otherwise. Saying that a matrix is reducible is equivalent to saying that there exists a permutation matrix P such that PAP^{-1} is of block-triangular form

$$\begin{pmatrix} B & C \\ 0_{p,n-p} & D \end{pmatrix},$$

with $1 \leq p \leq n-1$. As a matter of fact, P is the matrix of the transformation from a basis γ to the canonical one, γ being obtained by first writing the vectors $\mathbf{e}^j$ with $j \in J$, and then those with $j \in I$. Working in the new basis amounts to decomposing the linear system $Ax = b$ into two subsystems $Dz = d$ and $By = c - Cz$, which are to be solved successively. The spectrum of A is the union of those of B and D, so that many interesting questions concerning square matrices reduce to questions about irreducible matrices.

We shall see in the exercises a characterization of irreducible matrices in terms of graphs. Here is a useful consequence of irreducibility.

Proposition 2.8.1 *Let $M \in \mathbf{M}_n(K)$ be an irreducible matrix such that $i \geq j+2$ implies $m_{ij} = 0$. Then the eigenvalues of M are geometrically simple.*

Proof

The hypothesis implies that all entries $m_{i+1,i}$ are nonzero. If λ is an eigenvalue, let us consider the matrix $N \in \mathbf{M}_{n-1}(\bar{K})$, obtained from $M - \lambda I_n$ by deleting the first row and the last column. It is a triangular matrix, whose diagonal terms are nonzero. It is thus invertible, which implies $\mathrm{rk}(M - \lambda I_n) = n-1$. Hence $\ker(M - \lambda I_n)$ is of dimension one.

■

2.9 Exercises

1. Verify that the product of two triangular matrices of the same type (upper or lower) is triangular, of the same type.

2. Prove in full detail that the determinant of a triangular matrix (respectively a block-triangular one) equals the product of its diagonal terms (respectively the product of the determinants of its diagonal blocks).

3. Find matrices $M, N \in \mathbf{M}_2(K)$ such that $MN = 0_2$ and $NM \neq 0_2$. Such an example shows that MN and NM are not necessarily similar, though they would be in the case where M or N is invertible.

4. Characterize the square matrices that are simultaneously orthogonal and triangular.

5. One calls any square matrix M satisfying $M^2 = M$ a *projection matrix*, or *projector*.

   (a) Let $P \in \mathbf{M}_n(K)$ be a projector, and let $E = \ker P$, $F = \ker(I_n - P)$. Show that $K^n = E \oplus F$.

   (b) Let P, Q be two projectors. Show that $(P-Q)^2$ commute with P and with Q. Also, prove the identity

$$(P-Q)^2 + (I_n - P - Q)^2 = I_n.$$

6. Let M be a square matrix over a field K, which we write blockwise as

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

   The formula $\det M = \det(AD - BC)$ is meaningless in general, except when A, B, C, D have the same size. In that case the formula is false, with the exception of scalar blocks. Compare with Schur's formula (Proposition 8.1.2).

7. If $A, B, C, D \in \mathbf{M}_m(K)$ and if $AC = CA$, show that the determinant of

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

   equals $\det(AD - CB)$. Begin with the case where A is invertible, by computing the product

$$\begin{pmatrix} I_m & 0_m \\ -C & A \end{pmatrix} M.$$

   Then apply this intermediate result to the matrix $A - zI_n$, with $z \in \bar{K}$ a suitable scalar.
   Compare with the previous exercise.

8. Verify that the inverse of a triangular matrix, whenever it exists, is triangular of the same type.

9. Show that the eigenvalues of a triangular matrix are its diagonal entries. What are their algebraic multiplicities?

10. Let $A \in \mathbf{M}_n(K)$ be given. One says that a list $(a_{1\sigma(1)}, \dots, a_{n\sigma(n)})$ is a *diagonal* of A if σ is a permutation (in that case, the diagonal given by the identity is the *main* diagonal). Show the equivalence of the following properties.

    - Every diagonal of A contains a zero element.
    - There exists a null matrix extracted from A of size $k \times l$ with $k + l > n$.

11. Compute the number of elements in the group $\mathbf{GL}_2(\mathbb{Z}/2\mathbb{Z})$. Show that it is not commutative. Show that it is isomorphic to the symmetric group $\boldsymbol{S}_m$, for a suitable integer m.

12. If $(a_0, \dots, a_{n-1}) \in \mathbb{C}^n$ is given, one defines the *circulant matrix* $\mathrm{circ}(a_0, \dots, a_{n-1}) \in \mathbf{M}_n(\mathbb{C})$ by

$$\mathrm{circ}(a_0, \dots, a_{n-1}) := \begin{pmatrix} a_0 & a_1 & \dots & a_{n-1} \\ a_{n-1} & a_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_1 \\ a_1 & \dots & a_{n-1} & a_0 \end{pmatrix}.$$

We denote by $\mathcal{C}_n$ the set of circulant matrices. Obviously, the matrix $\mathrm{circ}(1, 0, 0, \dots, 0)$ is the identity. The matrix $\mathrm{circ}(0, 1, 0, \dots, 0)$ is denoted by π.

(a) Show that $\mathcal{C}_n$ is a subalgebra of $\mathbf{M}_n(\mathbb{C})$, equal to $\mathbb{C}[\pi]$. Deduce that it is isomorphic to the quotient ring $\mathbb{C}[X]/(X^n - 1)$.

(b) Let C be a circulant matrix. Show that C^*, as well as $P(C)$, is circulant for every polynomial P. If C is nonsingular, show that C^{-1} is circulant.

(c) Show that the elements of $\mathcal{C}_n$ are diagonalizable in a common eigenbasis.

(d) Replace $\mathbb{C}$ by any field K. If K contains a *primitive* nth root ω of unity (that is, $\omega^n = 1$, and $\omega^m = 1$ implies $m \in n\mathbb{Z}$), show that the elements of $\mathcal{C}_n$ are diagonalizable.
Note: A thorough presentation of circulant matrices and applications is given in Davis's book [12].

(e) One assumes that the characteristic of K divides n. Show that $\mathcal{C}_n$ contains matrices that are not diagonalizable.

13. Show that the Pfaffian is linear with respect to any row or column of an alternate matrix. Deduce that the Pfaffian is an irreducible polynomial in $\mathbb{Z}[x_{ij}]$.

14. (Schur's Lemma).
Let k be an algebraically closed field and S a subset of $\mathbf{M}_n(k)$. Assume that the only linear subspaces of k^n that are stable under every element of S are $\{0\}$ and k^n itself. Let $A \in \mathbf{M}_n(k)$ be a matrix that commutes with every element of S. Show that there exists $c \in k$ such that $A = cI_n$.

15. (a) Show that $A \in \mathbf{M}_n(K)$ is irreducible if and only if for every pair (j, k) with $1 \leq j, k \leq n$, there exists a finite sequence of indices $j = l_1, \dots, l_r = k$ such that $a_{l_p, l_{p+1}} \neq 0$.

(b) Show that a tridiagonal matrix $A \in \mathbf{M}_n(K)$, for which none of the $a_{j,j+1}$'s and $a_{j+1,j}$'s vanish, is irreducible.

16. Let $A \in \mathbf{M}_n(k)$ ($k = \mathbb{R}$ or $\mathbb{C}$) be given, with minimal polynomial q. If $x \in k^n$, the set

$$I_x := \{p \in k[X] \,|\, p(A)x = 0\}$$

is an ideal of $k[X]$, which is therefore principal.

(a) Show that $I_x \neq (0)$ and that its monic generator, denoted by p_x, divides q.
(b) One writes r_j instead of p_x when $x = \mathbf{e}^j$. Show that q is the least common multiple of $r_1, \ldots, r_n$.
(c) If $p \in k[X]$, show that the set

$$V_p := \{x \in k^n \,|\, p_x \in (p)\}$$

(the vectors x such that p divides p_x) is open.
(d) Let $x \in k^n$ be an element for which p_x is of maximal degree. Show that $p_x = q$. **Note**: In fact, the existence of an element x such that p_x equals the minimal polynomial holds true for every field k.

17. Let k be a field and $A \in \mathbf{M}_{n\times m}(k)$, $B \in \mathbf{M}_{m\times n}(k)$ be given.

(a) Let us define

$$M = \begin{pmatrix} XI_n & A \\ B & XI_m \end{pmatrix}.$$

Show that $X^m \det M = X^n \det(X^2 I_m - BA)$ (search for a lower triangular matrix M' such that $M'M$ is upper triangular).
(b) Find an analogous relation between $\det(X^2 I_n - AB)$ and $\det M$. Deduce that $X^n P_{BA}(X) = X^m P_{AB}(X)$.
(c) What do you deduce about the eigenvalues of A and of B?

18. Let k be a field and $\theta : \mathbf{M}_n(k) \to k$ a linear form satisfying $\theta(AB) = \theta(BA)$ for every $A, B \in \mathbf{M}_n(k)$.

(a) Show that there exists $\alpha \in k$ such that for all $X, Y \in k^n$, one has $\theta(XY^T) = \alpha \sum_j x_j y_j$.
(b) Deduce that $\theta = \alpha \operatorname{Tr}$.

19. Let A_n be the ring $K[X_1, \ldots, X_n]$ of polynomials in n variables. Consider the matrix $M \in \mathbf{M}_n(A_n)$ defined by

$$M = \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \\ X_1^2 & \cdots & X_n^2 \\ \vdots & & \vdots \\ X_1^{n-1} & \cdots & X_n^{n-1} \end{pmatrix}.$$

Let us denote by $\Delta(X_1, \ldots, X_n)$ the determinant of M.

(a) Show that for every $i \neq j$, the polynomial $X_j - X_i$ divides Δ.
(b) Deduce that

$$\Delta = a \prod_{i<j} (X_j - X_i),$$

where $a \in K$.
(c) Determine the value of a by considering the monomial

$$\prod_{j=1}^{n} X_j^j.$$

(d) Redo this analysis for the matrix

$$\begin{pmatrix} X_1^{p_1} & \cdots & X_n^{p_1} \\ \vdots & & \vdots \\ X_1^{p_n} & \cdots & X_n^{p_n} \end{pmatrix},$$

where $p_1, \ldots, p_n$ are nonnegative integers.

20. Deduce from the previous exercise that the determinant of the *Vandermonde* matrix

$$\begin{pmatrix} 1 & \cdots & 1 \\ a_1 & \cdots & a_n \\ a_1^2 & \cdots & a_n^2 \\ \vdots & & \vdots \\ a_1^{n-1} & \cdots & a_n^{n-1} \end{pmatrix}, \quad a_1, \ldots, a_n \in K,$$

is zero if and only if at least two of the a_j's coincide.

21. A matrix $A \in \mathbf{M}_n(\mathbb{R})$ is called a *totally positive* matrix when all minors

$$A \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$$

with $1 \leq p \leq n$, $1 \leq i_1 < \cdots < i_p \leq n$ and $1 \leq j_1 < \cdots < j_p \leq n$ are positive.

(a) Prove that the product of totally positive matrices is totally positive.
(b) Prove that a totally positive matrix admits an LU factorization (see Chapter 8), and that every "nontrivial" minor of L and U is positive. Here, "nontrivial" means

$$L \begin{pmatrix} i_1 & i_2 & \cdots & i_p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix}$$

with $1 \leq p \leq n$, $1 \leq i_1 < \cdots < i_p \leq n$, $1 \leq j_1 < \cdots < j_p \leq l$, and $i_s \geq j_s$ for every s. For U, read $i_s \leq j_s$ instead. **Note**: One says that L and U are *triangular totally positive*.

(c) Show that a Vandermonde matrix (see the previous exercise) is totally positive whenever $0 < a_1 < \cdots < a_n$.

22. Multiplying a Vandermonde matrix by its transpose, show that

$$\det \begin{pmatrix} n & s_1 & \cdots & s_{n-1} \\ s_1 & s_2 & \iddots & \vdots \\ \vdots & \iddots & & \vdots \\ s_{n-1} & \cdots & \cdots & s_{2n-2} \end{pmatrix} = \prod_{i<j} (a_j - a_i)^2,$$

where $s_q := a_1^q + \cdots + a_n^q$.

23. The *discriminant* of a matrix $A \in M_n(k)$ is the number

$$d(A) := \prod_{i<j} (\lambda_j - \lambda_i)^2,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, counted with multiplicity.

(a) Verify that the polynomial

$$\Delta(X_1, \ldots, X_n) := \prod_{i<j} (X_j - X_i)^2$$

is symmetric. Therefore, there exists a unique polynomial $Q \in \mathbb{Z}[Y_1, \ldots, Y_n]$ such that

$$\Delta = Q(\sigma_1, \ldots, \sigma_n),$$

where the σ_j's are the elementary symmetric polynomials

$$\sigma_1 = X_1 + \cdots + X_n, \ldots, \sigma_n = X_1 \cdots X_n.$$

(b) Deduce that there exits a polynomial $D \in \mathbb{Z}[x_{ij}]$ in the indeterminates x_{ij}, $1 \leq i, j \leq n$, such that for every k and every square matrix A,

$$d(A) = D(a_{11}, a_{12}, \ldots, a_{nn}).$$

(c) Consider the restriction D_S of the discriminant to symmetric matrices, where x_{ji} is replaced by x_{ij} whenever $i < j$. Prove that D_S takes only nonnegative values on $\mathbb{R}^{n(n+1)/2}$. Show, however, that D_S is not the square of a polynomial if $n \geq 2$ (consider first the case $n = 2$).

24. Let $P \in k[X]$ be a polynomial of degree n that splits completely in k. Let B_P be the companion matrix

$$B_P := \begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_n \\ 1 & \ddots & & \vdots & \vdots \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -a_1 \end{pmatrix}.$$

Find a matrix $H \in \mathbf{M}_n(k)$, whose transpose is of Vandermonde type, such that

$$HB_P = \operatorname{diag}(\lambda_1, \dots, \lambda_n)H.$$

This furnishes a direct proof of the fact that when the roots of P are simple, B_P is diagonalizable.

25. (E. Formanek [14])
Let k be a field of characteristic 0.

(a) Show that for every $A, B, C \in \mathbf{M}_2(k)$,

$$\left[[A, B]^2, C\right] = 0.$$

Hint: use the Cayley–Hamilton theorem.

(b) Show that for every $M, N \in \mathbf{M}_2(k)$,

$$\begin{aligned} MN + NM - \operatorname{Tr}(M)N - \operatorname{Tr}(N)M + \\ (\operatorname{Tr}(M)\operatorname{Tr}(N) - \operatorname{Tr}(MN))I_2 \quad = \quad 0. \end{aligned}$$

One may begin with the case $M = N$ and recognize a classical theorem, then "bilinearize" the formula.

(c) If $\pi \in \boldsymbol{S}_r$ ($\boldsymbol{S}_r$ is the symmetric group over $\{1, \dots, r\}$), one defines a map $T_\pi : \mathbf{M}_2(k)^r \to k$ in the following way. One decomposes π as a product of disjoint cycles, including the cycles of order one, which are the fixed points of π:

$$\pi = (a_1, \dots, a_{k_1})(b_1, \dots, b_{k_2}) \cdots .$$

One sets then

$$T_\pi(N_1, \dots, N_r) = \operatorname{Tr}(N_{a_1} \cdots N_{a_{k_1}}) \operatorname{Tr}(N_{b_1} \cdots N_{b_{k_2}}) \cdots$$

(note that the right-hand side depends neither on the order of the cycles in the product nor on the choice of the first index inside each cycle, because of the formula $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$). Show that for every $N_1, N_2, N_3 \in \mathbf{M}_2(k)$, one has

$$\sum_{\pi \in \boldsymbol{S}_3} \epsilon(\pi) T_\pi(N_1, N_2, N_3) = 0.$$

(d) Generalize this result to $\mathbf{M}_n(k)$: for every $N_1, \ldots, N_{n+1} \in \mathbf{M}_n(k)$, one has

$$\sum_{\pi \in \boldsymbol{S}_{n+1}} \epsilon(\pi) T_\pi(N_1, \ldots, N_{n+1}) = 0.$$

Note: Polynomial identities satisfied by every $n \times n$ matrix have been studied for decades. See [15] for a thorough account. One should at least mention the theorem of Amitsur and Levitzki:

Theorem 2.9.1 *Consider the free algebra $\mathbb{Z}[x_1, \ldots, x_r]$ (where $x_1, \ldots, x_r$ are noncommuting indeterminates) define the* standard polynomial $\mathcal{S}_r$ *by*

$$\mathcal{S}_r(x_1, \ldots, x_r) = \sum_{\pi \in \boldsymbol{S}_r} \epsilon(\pi) x_{\pi(1)} \cdots x_{\pi(r)}.$$

Then, given a commutative ring A, one has the polynomial identity

$$\mathcal{S}_{2n}(Q_1, \ldots, Q_{2n}) = 0_n, \quad \forall Q_1, \ldots, Q_{2n} \in \mathbf{M}_n(A).$$

26. Let k be a field and let $A \in \mathbf{M}_n(k)$ be given. For every set $J \subset \{1, \ldots, n\}$, denote by A_J the matrix extracted from A by keeping only the indices $i, j \in J$. Hence, $A_J \in \mathbf{M}_p(k)$ for $p = \operatorname{card} J$. Let $\lambda \in k$.

    (a) Assume that for every J whose cardinality is greater than or equal to $n - p$, λ is an eigenvalue of A_J. Show that λ is an eigenvalue of A, of algebraic multiplicity greater than or equal to $p+1$ (express the derivatives of the characteristic polynomial).
    (b) Conversely, let q be the geometric multiplicity of λ as an eigenvalue of A. Show that if $\operatorname{card} J > n - q$, then λ is an eigenvalue of A_J.

27. Let $A \in \mathbf{M}_n(k)$ and $l \in \mathbb{N}$ be given. Show that there exists a polynomial $q_l \in k[X]$, of degree at most $n - 1$, such that $A^l = q_l(A)$. If A is invertible, show that there exists $r_l \in k[X]$, of degree at most $n - 1$, such that $A^{-l} = r_l(A)$.

28. Let k be a field and $A, B \in \mathbf{M}_n(k)$. Assume that $\lambda \neq \mu$ for every $\lambda \in \operatorname{Sp} A$, $\mu \in \operatorname{Sp} B$. Show, using the Cayley–Hamilton theorem, that the linear map $M \mapsto AM - MB$ is an automorphism of $\mathbf{M}_n(k)$.

29. Let k be a field and $(M_{jk})_{1 \leq j,k \leq n}$ a set of matrices of $\mathbf{M}_n(k)$, at least one of which is nonzero, such that $M_{ij} M_{kl} = \delta^k_j M_{il}$ for all $1 \leq i, j, k, l \leq n$.

    (a) Show that none of the matrices M_{jk} vanishes.
    (b) Verify that each M_{ii} is a projector. Denote its range by E_i.

(c) Show that $E_1, \ldots, E_n$ are in direct sum. Deduce that each E_j is a line.

(d) Show that there exist generators e_j of each E_j such that $M_{jk}e_l = \delta^l_k e_j$.

(e) Deduce that every algebra automorphism of $\mathbf{M}_n(k)$ is interior: For every $\sigma \in \mathrm{Aut}\mathbf{M}_n(k)$, there exists $P \in \mathbf{GL}_n(k)$ such that $\sigma(M) = P^{-1}MP$ For every $M \in \mathbf{M}_n(k)$.

3

Matrices with Real or Complex Entries

Definitions

A square matrix $M \in \mathbf{M}_n(I\!R)$ is said to be *normal* if M and M^T commute: $M^T M = M M^T$. The real symmetric, skew-symmetric, and orthogonal matrices are normal.

In considering matrices with complex entries, a useful operation is complex conjugation $z \mapsto \bar{z}$. One denotes by $\bar{M}$ the matrix obtained from M by conjugating the entries. We then define the *Hermitian adjoint* matrix[1] M^* by

$$M^* := (\bar{M})^T = \overline{M^T}.$$

One therefore has $m^*_{ij} = \overline{m_{ji}}$ and $\det M^* = \overline{\det M}$. The map $M \mapsto M^*$ is an *anti-isomorphism*, which means that it is antilinear (meaning that $(\lambda M)^* = \bar{\lambda} M^*$) and satisfies, moreover, the product formula

$$(MN)^* = N^* M^*.$$

When a square matrix $M \in \mathbf{M}_n(\mathbb{C})$ is invertible, then $(M^*)^{-1} = (M^{-1})^*$. This matrix is sometimes denoted by M^{-*}.

One says that a square matrix $M \in \mathbf{M}_n(\mathbb{C})$ is *Hermitian* if $M^* = M$ and *skew-Hermitian* if $M^* = -M$. If $M \in \mathbf{M}_{n\times m}(\mathbb{C})$, the matrices MM^* and

---

[1] We warn the reader about the possible confusion between the *adjoint* and the *Hermitian adjoint* of a matrix. One may remark that the Hermitian adjoint is defined for every rectangular matrix with complex entries, while the adjoint is defined for every square matrix with entries in a commutative ring.

M^*M are Hermitian. We denote by $\mathbf{H}_n$ the set of Hermitian matrices in $\mathbf{M}_n(\mathbb{C})$. It is an $\mathbb{R}$-linear subspace of $\mathbf{M}_n(\mathbb{C})$, though it is not a $\mathbb{C}$-linear subspace, since iM is skew-Hermitian when M is Hermitian.

A square matrix $M \in \mathbf{M}_n(\mathbb{C})$ is said to be *unitary* if $M^*M = I_n$. Since this means that M is invertible, with inverse M^*, and since the left and the right inverses are equal, an equivalent criterion is $MM^* = I_n$. The set of unitary matrices in $\mathbf{M}_n(\mathbb{C})$ forms a multiplicative group, denoted by $\mathbf{U}_n$. Unitary matrices satisfy $|\det M| = 1$, since $\det M^*M = |\det M|^2$ for every matrix M. The set of unitary matrices whose determinant equals 1, denoted by $\mathbf{SU}_n$ is obviously a normal subgroup of $\mathbf{U}_n$. Finally, M is said to be *normal* if M and M^* commute: $MM^* = M^*M$. The Hermitian, skew-Hermitian, and unitary matrices are normal.

Observe that the *real* orthogonal (respectively symmetric, skew-symmetric) matrices are unitary (respectively Hermitian, skew-Hermitian). Conversely, if M is real and either unitary, symmetric, or skew-symmetric, then M is either orthogonal, Hermitian, or skew-Hermitian.

A *sesquilinear* form on a complex vector space is a map

$$(x, y) \mapsto \langle x, y\rangle,$$

linear in x and satisfying

$$\langle y, x\rangle = \overline{\langle x, y\rangle}.$$

It is thus antilinear in y:

$$\langle x, \lambda y\rangle = \bar{\lambda}\langle x, y\rangle.$$

When $y = x$, $\langle x, y\rangle = \langle x, x\rangle$ is a real number. The map $x \mapsto \langle x, x\rangle$ is called a *Hermitian* form. The correspondence between sesquilinear and Hermitian forms is one-to-one.

Given a matrix $M \in \mathbf{M}_n(\mathbb{C})$, the form

$$(x, y) \mapsto \sum_{j,k} m_{jk} x_j \overline{y_k},$$

defined on $\mathbb{C}^n \times \mathbb{C}^n$, is sesquilinear if and only if M is Hermitian. It follows that there is an isomorphism between the sets of Hermitian matrices, Hermitian, and sesquilinear forms on $\mathbb{C}^n$. As a matter of fact, a Hermitian form can be written in the form

$$x \mapsto \sum_{j,k} m_{jk} x_j \bar{x}_k.$$

The *kernel* of a Hermitian or a sesquilinear form is the set of vectors $x \in E$ such that $\langle x, y\rangle = 0$ for every $y \in E$. It equals the set of vectors $y \in E$ such that $\langle x, y\rangle = 0$ for every $x \in E$. If $E = \mathbb{C}^n$, it is also the kernel of M^T, where M is the (Hermitian) matrix associated to the Hermitian form. One says that the Hermitian form is *degenerate* if its kernel does not

reduce to $\{0\}$. When $E = \mathbb{C}^n$, this amounts to $\det M = 0$. One says that the form is *nondegenerate* otherwise.

If both E and F are endowed with nondegenerate sequilinear forms $\langle \cdot, \cdot \rangle_E$ and $\langle \cdot, \cdot \rangle_F$, respectively, and if $u \in \mathcal{L}(E, F)$, one defines u^* by the formula

$$\langle u^*(x), y \rangle_E = \langle x, u(y) \rangle_F, \quad \forall x \in F,\ y \in E.$$

The map $u \mapsto u^*$ is an $\mathbb{R}$-isomorphism from $\mathcal{L}(E, F)$ onto $\mathcal{L}(F, E)$, and one has $(\lambda u)^* = \bar{\lambda} u^*$, $(u^*)^* = u$. When $E = \mathbb{C}^n$ and $F = \mathbb{C}^m$ are endowed with the canonical sesquilinear forms $x_1 \overline{y_1} + \cdots$, the matrix associated to u^* is simply the Hermitian adjoint of the matrix associated to u. The canonical Hermitian form over $\mathbb{C}^n$ is positive definite: $\langle x, x \rangle > 0$ if $x \neq 0$. It allows us to define a norm by $\|x\| = \sqrt{\langle x, x \rangle}$. Identifying $\mathbb{C}^n$ with column vectors, one also defines $\|X\| = \sqrt{X^* X}$ if $X \in \mathbf{M}_{n \times 1}(\mathbb{C})$. This norm will be denoted by $\| \cdot \|_2$ in Chapter 4. A matrix is unitary if and only if it is associated with an *isometry* of $\mathbb{C}^n$:

$$\|u(x)\| = \|x\|, \quad \forall x \in \mathbb{C}^n.$$

More generally, let M be a Hermitian matrix and $\langle \cdot, \cdot \rangle$ the form that it defines on $\mathbb{C}^n$. One says that M is *positive definite* if $\langle x, x \rangle > 0$ for every $x \neq 0$. Again, $\sqrt{\langle x, x \rangle}$ is a norm on $\mathbb{C}^n$. We shall denote by $\mathbf{HPD}_n$ the set of the positive definite Hermitian matrices; it is an open cone in $\mathbf{H}_n$. Its closure consists of the Hermitian matrices M that define a positive semidefinite Hermitian form over $\mathbb{C}^n$ ($\langle x, x \rangle \geq 0$ for every x). They are called positive semidefinite Hermitian matrices. One defines similarly, among the real symmetric matrices, those that are positive definite, respectively positive semidefinite. The positive definite real symmetric matrices form an open cone in $\mathbf{Sym}_n(\mathbb{R})$, denoted by $\mathbf{SPD}_n$.

The natural ordering on Hermitian forms induces an ordering on Hermitian matrices. One writes $H \geq 0_n$ when the Hermitian form associated to H takes nonnegative values. More generally, one writes $H \geq h$ if $H - h \geq 0_n$. We likewise define an ordering on real-valued symmetric matrices, referring to the ordering on real-valued quadratic forms.[2]

If U is unitary, the matrix $U^* M U$ is similar to M. If M is Hermitian, skew-Hermitian, normal, or unitary and if U is unitary, then $U^* M U$ is still Hermitian, skew-Hermitian, normal, or unitary.

[2]We warn the reader that another, completely different, order still denoted by the symbol $\geq$ will be defined in Chapter 5. This one will concern real-valued matrices that are neither symmetric nor even square. One expects that the context is never ambiguous.

3.1 Eigenvalues of Real- and Complex-Valued Matrices

Since $\mathbb{C}$ is algebraically closed, every complex-valued square matrix, and every endomorphism of a $\mathbb{C}$-vector space of dimension $n \geq 1$, possesses eigenvalues. As a matter of fact, the characteristic polynomial has roots. A real-valued square matrix may not have eigenvalues in $\mathbb{R}$, but it has at least one in $\mathbb{C}$. If n is odd, $M \in \mathbf{M}_n(\mathbb{R})$ has at least a real eigenvalue, because P_M is real of odd degree.

Proposition 3.1.1 *The eigenvalues of Hermitian matrices, as well as those of real symmetric matrices, are real.*

Proof

Let $M \in \mathbf{M}_n(\mathbb{C})$ be a Hermitian matrix and let λ be one of its eigenvalues. Let us choose an eigenvector X: $MX = \lambda X$. Taking the Hermitian adjoint, we obtain $X^*M = \bar{\lambda}X$. Hence,

$$\lambda X^*X = X^*(MX) = (X^*M)X = \bar{\lambda}X^*X,$$

or

$$(\lambda - \bar{\lambda})X^*X = 0.$$

However $X^*X = \sum_j |x_j|^2 > 0$. Therefore, we are left with $\bar{\lambda} - \lambda = 0$. Hence λ is real.

■

We leave it to the reader to show, as an exercise, that the eigenvalues of skew-Hermitian matrices are purely imaginary.

Proposition 3.1.2 *The eigenvalues of the unitary matrices, as well as those of real orthogonal matrices, are complex numbers of modulus one.*

Proof

As before, if X is an eigenvector associated to λ, one has

$$|\lambda|^2\|X\|^2 = (\lambda X)^*(\lambda X) = (MX)^*MX = X^*M^*MX = X^*X = \|X\|^2,$$

and therefore $|\lambda|^2 = 1$.

■

3.1.1 Continuity of Eigenvalues

One of the more delicate statements in the elementary theory of matrices concerns the continuity of the eigenvalues. Though a proof might be provided througth explicit bounds, it is easier to use Rouché's theorem about holomorphic functions. We begin with a statement concerning polynomials, that is a bit less precise than Rouché's theorem.

Theorem 3.1.1 *Let $n \in \mathbb{N}$ and let $P \in \mathbb{C}[X]$ be a polynomial of degree n,*

$$P(X) = p_0 + p_1 X + \cdots + p_n X^n.$$

Let x be a root of P, with multiplicity μ, and let d be the distance from x to the other roots of P. Let D be an open disk, $D = D(x;\rho)$, with $0 < \rho < d$. Then there exists a number $\epsilon > 0$ such that if $Q \in \mathbb{C}[X]$ has degree n,

$$Q(X) = q_0 + q_1 X + \cdots + q_n X^n,$$

and if

$$\max_j |q_j - p_j| < \epsilon,$$

then D contains exactly μ roots of Q, counting multiplicities.

Let us apply this result to the characteristic polynomial of a given matrix. Since the coefficients of the characteristic polynomial p_M are polynomial functions of the entries of M, the map $M \mapsto p_M$ is continuous from $\mathbf{M}_n(\mathbb{C})$ to the set of polynomials of degree n. From Rouché's theorem, we have the following result.

Theorem 3.1.2 *Let $M \in \mathbf{M}_n(\mathbb{C})$, and let λ be one of its eigenvalues, with multiplicity μ, and let d be the distance from λ to the other eigenvalues of M. Let D be an open disk, $D = D(\lambda;\rho)$, with $0 < \rho < d$. Let us fix a norm on $\mathbf{M}_n(\mathbb{C})$.*

There exists an $\epsilon > 0$ such that if $A \in \mathbf{M}_n(\mathbb{C})$ and $\|A\| < \epsilon$, the sum of algebraic multiplicities of the eigenvalues of $M + A$ in D equals μ.

Let us remark that this statement becomes false if one considers the geometric multiplicities.

One often invokes this theorem by saying that *the eigenvalues of a matrix are continuous functions of its entries.* Here is an interpretation. One adapts the Hausdorff distance between compact sets so as to take into account the multiplicity of the eigenvalues. If $M, N \in \mathbf{M}_n(\mathbb{C})$, let us denote by $(\lambda_1, \ldots, \lambda_n)$ and $(\theta_1, \ldots, \theta_n)$ their eigenvalues, repeated according to their multiplicities. One then defines

$$d(\mathrm{Sp}\, M, \mathrm{Sp}\, N) := \inf_{\sigma \in \boldsymbol{S}_n} \max_j |\lambda_j - \theta_{\sigma(j)}|,$$

where $\boldsymbol{S}_n$ is the group of permutations of the indices $\{1, \ldots, n\}$. This number is called the distance between the spectra of M and N. With this notation, one may rewrite Theorem 3.1.2 in the following form.

Proposition 3.1.3 *If $M \in \mathbf{M}_n(\mathbb{C})$ and $\alpha > 0$, there exists $\epsilon > 0$ such that $\|N - M\| < \epsilon$ implies $d(\mathrm{Sp}\, M, \mathrm{Sp}\, N) < \alpha$.*

A useful consequence of Theorem 3.1.2 is the following.

Corollary 3.1.1 *In* $\mathbf{M}_n(k)$ *(*$k = \mathbb{R}$ *or* $\mathbb{C}$*) the set of diagonalizable matrices is an open subset.*

3.1.2 Trigonalization in an Orthonormal Basis

From now on we say that two matrices are *unitarily similar* if they are similar through a unitary transformation. Two real matrices are unitarily similar if they are similar through an orthogonal transformation.

If $K = \mathbb{C}$, one may sharpen Theorem 2.7.1:

Theorem 3.1.3 (Schur) *If* $M \in \mathbf{M}_n(\mathbb{C})$*, there exists a unitary matrix* U *such that* U^*MU *is upper triangular.*

One also says that every matrix with complex entries is *unitarily trigonalizable.*

Proof

We proceed by induction on the size n of the matrices. The statement is trivial if $n = 1$. Let us assume that it is true in $\mathbf{M}_{n-1}(\mathbb{C})$, with $n \geq 2$. Let $M \in \mathbf{M}_n(\mathbb{C})$ be a matrix. Since $\mathbb{C}$ is algebraically closed, M has at least one eigenvalue λ. Let X be an eigenvector associated to λ. By dividing X by $\|X\|$, one can assume that X is a unit vector. One can then find an orthonormal basis $\{X^1, X^2, \ldots, X^n\}$ of $\mathbb{C}^n$ whose first element is X. Let us consider the matrix $V := (X^1 = X, X^2, \ldots, X^n)$, which is unitary, and let us form the matrix $M' := V^*MV$. Since

$$VM'\mathbf{e}^1 = MV\mathbf{e}^1 = MX = \lambda X = \lambda V\mathbf{e}^1,$$

one obtains $M'\mathbf{e}^1 = \lambda\mathbf{e}^1$. In other words, M' has the block-triangular form:

$$M' = \begin{pmatrix} \lambda & \cdots \\ 0_{n-1} & N \end{pmatrix},$$

where $N \in \mathbf{M}_{n-1}(\mathbb{C})$. Applying the induction hypothesis, there exists $W \in \mathbf{U}_{n-1}$ such that W^*NW is upper triangular. Let us denote by $\hat{W}$ the (block-diagonal) matrix $\mathrm{diag}(1, W) \in \mathbf{U}_n$. Then $\hat{W}^*M'\hat{W}$ is upper triangular. Hence, $U = V\hat{W}$ satisfies the conditions of the theorem.

■

3.2 Spectral Decomposition of Normal Matrices

We recall that a matrix M is *normal* if M^* commutes with M. For real matrices, this amounts to saying that M^T commutes with M. Since it is equivalent for a Hermitian matrix H to be zero or to satisfy $x^*Hx = 0$ for every vector x, we see that M is normal if and only if $\|Ax\|_2 = \|A^*x\|_2$ for every vector, where $\|x\|_2$ denotes the standard Hermitian (Euclidean) norm (take $H = AA^* - A^*A$).

Theorem 3.2.1 *If $K = \mathbb{C}$, the normal matrices are diagonalizable, using unitary matrices:*

$$(M^*M = MM^*) \Longrightarrow (\exists U \in \mathbf{U}_n; \quad M = U^{-1}\,\mathrm{diag}(d_1, \ldots, d_n)U).$$

Again, one says that normal matrices are *unitarily diagonalizable*. This theorem contains the following properties.

Corollary 3.2.1 *Unitary, Hermitian, and skew-Hermitian matrices are unitarily diagonalizable.*

Observe that among normal matrices one distinguishes each of the above families by the nature of their eigenvalues. Those of unitary matrices have modulus one, while those of Hermitian matrices are real. Finally, those of skew-Hermitian matrices are purely imaginary.

Proof

We proceed by induction on the size n of the matrix M. If $n = 0$, there is nothing to prove. Otherwise, if $n \geq 1$, there exists an eigenpair (λ, x):

$$Mx = \lambda x, \quad \|x\|_2 = 1.$$

Since M is normal, $M - \lambda I_n$ is, too. From above, we see that $\|(M^* - \bar{\lambda})x\|_2 = \|(M - \lambda)x\|_2 = 0$, and hence $M^*x = \bar{\lambda}x$. Let V be a unitary matrix such that $V\mathbf{e}^1 = x$. Then the matrix $M_1 := V^*MV$ is normal and satisfies $M_1\mathbf{e}^1 = \lambda\mathbf{e}^1$. Hence it satisfies $M_1^*\mathbf{e}^1 = \bar{\lambda}\mathbf{e}^1$. This amounts to saying that M_1 is block-diagonal, of the form $M_1 = \mathrm{diag}(\lambda, M')$. Obviously, M' inherits the normality of M_1. From the induction hypothesis, M', and therefore M_1 and M, are unitarily diagonalizable.

■

One observes that the same matrix U diagonalizes M^*, because $M = U^{-1}DU$ implies $M^* = U^*D^*U^{-1*} = U^{-1}D^*U$, since U is unitary.

Let us consider the case of a positive semidefinite Hermitian matrix H. If $HX = \lambda X$, then $0 \leq X^*HX = \lambda\|X\|^2$. The eigenvalues are thus nonnegative. Let $\lambda_1, \ldots, \lambda_p$ be the nonzero eigenvalues of H. Then H is unitarily similar to

$$D := \mathrm{diag}(\lambda_1, \ldots, \lambda_p, 0, \ldots, 0).$$

From this, we conclude that $\mathrm{rk}\,H = p$. Let $U \in \mathbf{U}_n$ be such that $H = UDU^*$. Defining the vectors $X_\alpha = \sqrt{\lambda_\alpha}U_\alpha$, where the U_α are the columns of U, we obtain the following statement.

Proposition 3.2.1 *Let $H \in \mathbf{M}_n(\mathbb{C})$ be a positive semidefinite Hermitian matrix. Let p be its rank. Then H has p real, positive eigenvalues, while the eigenvalue $\lambda = 0$ has multiplicity $n - p$. There exist p column vectors X_α, pairwise orthogonal, such that*

$$H = X_1X_1^* + \cdots + X_pX_p^*.$$

Finally, H is positive definite if and only if $p = n$ (in which case, $\lambda = 0$ is not an eigenvalue).

3.3 Normal and Symmetric Real-Valued Matrices

The situation is a bit more involved if M, a normal matrix, has real entries. Of course, one can consider M as a matrix with complex entries and diagonalize it in an orthonormal basis, but we quit in general the field of real numbers when doing so. We prefer to allow bases consisting of only *real* vectors. Since some of the eigenvalues might be nonreal, one cannot in general diagonalize M. The statement is thus the following.

Theorem 3.3.1 *Let $M \in \mathbf{M}_n(\mathbb{R})$ be a normal matrix. There exists an orthogonal matrix O such that OMO^{-1} be block-diagonal, the diagonal blocks being 1×1 (those corresponding to the real eigenvalues of M) or 2×2, the latter being matrices of direct similitude:*[3]

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \quad (b \neq 0).$$

Similarly, OM^TO^{-1} is block-diagonal, the diagonal blocks being eigenvalues or matrices of direct similitude.

Proof

One again proceeds by induction on n. When $n \geq 1$, the proof is the same as in the previous section whenever M has at least one real eigenvalue.

If this is not the case, then n is even. Let us first consider the case $n = 2$. Then

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Since M is normal, we have $b^2 = c^2$ and $(a-d)(b-c) = 0$. However, $b \neq c$, since otherwise M would be symmetric, hence would have two real eigenvalues. Hence $b = -c$ and $a = d$.

Now let us consider the general case, with $n \geq 4$. We know that M has an eigenpair (λ, z), where λ is not real. If the real and imaginary parts of z were colinear, M would have a real eigenvector, hence a real eigenvalue, a contradiction. In other words, the real and imaginary parts of z span a plane P in $\mathbb{R}^n$. As before, $Mz = \lambda z$ implies $M^Tz = \bar{\lambda}z$. Hence we have $MP \subset P$ and $M^TP \subset P$. Now let V be an orthogonal matrix that maps the plane $P_0 := \mathbb{R}\mathbf{e}^1 \oplus \mathbb{R}\mathbf{e}^2$ onto P. Then the matrix $M_1 := V^TMV$ is normal and satisfies

$$M_1P_0 \subset P_0, \quad M_1^TP_0 \subset P_0.$$

This means that M_1 is block-diagonal. Of course, each diagonal block (of sizes 2×2 and $(n-2) \times (n-2)$) inherits the normality of M_1. Applying the induction hypothesis, we know that these blocks are unitarily similar to a

[3] A similitude is an endomorphism of a Euclidean space that preserves angles. It splits as aR, where R is orthogonal and a is a scalar. It is direct if its determinant is positive.

block-diagonal matrix whose diagonal blocks are direct similitudes. Hence M_1 and M are unitarily similar to such a matrix.

■

Corollary 3.3.1 *Real symmetric matrices are diagonalizable over $\mathbb{R}$, through orthogonal conjugation. In other words, given $M \in \mathbf{Sym}_n(\mathbb{R})$, there exists an $O \in \mathbf{O}_n(\mathbb{R})$ such that OMO^{-1} is diagonal.*

In fact, since the eigenvalues of M are real, OMO^{-1} has only 1×1 blocks. We say that real symmetric matrices are *orthogonally diagonalizable*.

The interpretation of this statement in terms of quadratic forms is the following. For every quadratic form Q on $\mathbb{R}^n$, there exists an orthonormal basis $\{e_1, \dots, e_n\}$ in which this form can be written with at most n squares:[4]

$$Q(x) = \sum_{i=1}^{n} a_i x_i^2.$$

Replacing the basis vector e_j by $|a_j|^{1/2} e_j$, one sees that there also exists an orthogonal basis in which the quadratic form can be written

$$Q(x) = \sum_{i=1}^{r} x_i^2 - \sum_{j=1}^{s} x_{j+r}^2,$$

with $r+s \leq n$. This quadratic form is nondegenerate if and only if $r+s = n$. The pair (r, s) is unique and called the *signature* or the *Sylvester index* of the quadratic form. In such a basis, the matrix associated to Q is

$$\begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & 0 & \\ & & 1 & & & & & \\ & & & -1 & & & & \\ & & & & \ddots & & & \\ & & & & & -1 & & \\ & & & & & & 0 & \\ & 0 & & & & & & \ddots \\ & & & & & & & & 0 \end{pmatrix}.$$

3.3.1 Rayleigh Quotients

Let M be a real $n \times n$ symmetric matrix, and let $\lambda_1 \leq \cdots \leq \lambda_n$ be its eigenvalues arranged in increasing order and counted with multiplicity. Let

[4] In solid mechanics, when Q is the matrix of inertia, the vectors of this basis are along the *inertia axes*, and the a_j, which then are positive, are the momenta of inertia.

us denote by $\mathcal{B} = \{v_1, \dots, v_n\}$ an orthonormal eigenbasis ($Mv_j = \lambda_j v_j$). If $x \in \mathbb{R}^n$, let us denote by $y_1, \dots, y_n$ the coordinates of x in the basis $\mathcal{B}$. Finally, let us denote by $\|\cdot\|_2$ the usual Euclidean norm on $\mathbb{R}^n$. Then

$$x^T M x = \sum_j \lambda_j y_j^2 \leq \lambda_n \sum_j y_j^2 = \lambda_n \|x\|_2^2.$$

Since $v_n^T M v_n = \lambda_n \|v_n\|_2^2$, we deduce the value of the largest eigenvalue of M:

$$\lambda_n = \max_{x \neq 0} \frac{x^T M x}{\|x\|_2^2} = \max \left\{ x^T M x \,|\, \|x\|_2^2 = 1 \right\}. \tag{3.1}$$

Similarly, the smallest eigenvalue of a real symmetric matrix is given by

$$\lambda_1 = \min_{x \neq 0} \frac{x^T M x}{\|x\|_2^2} = \min\{ x^T M x \,|\, \|x\|_2^2 = 1 \}. \tag{3.2}$$

For a Hermitian matrix, the formulas (3.1,3.2) remain valid when we replace x^T by x^*.

We evaluate the other eigenvalues of $M \in \mathbf{Sym}_n(\mathbb{R})$ in the following way. For every linear subspace F of $\mathbb{R}^n$ of dimension k, let us define

$$R(F) = \max_{x \in F \setminus \{0\}} \frac{x^T M x}{\|x\|_2^2} = \max \left\{ x^T M x \,|\, x \in F,\ \|x\|_2^2 = 1 \right\}.$$

The intersection of F with the linear subspace spanned by $\{v_k, \dots, v_n\}$ is of dimension greater than or equal to one. There exists, therefore, a nonzero vector $x \in F$ such that $y_1 = \cdots = y_{k-1} = 0$. One has then

$$x^T M x = \sum_{j=k}^{n} \lambda_j y_j^2 \geq \lambda_k \sum_j y_j^2 = \lambda_k \|x\|_2^2.$$

Hence, $R(F) \geq \lambda_k$. Furthermore, if G is the space spanned by $\{v_1, \dots, v_k\}$, one has $R(G) = \lambda_k$. Thus, we have

$$\lambda_k = \min\{ R(F) \,|\, \dim F = k \}.$$

Finally, we may state the following theorem.

Theorem 3.3.2 *Let M be an $n \times n$ real symmetric matrix and $\lambda_1, \dots, \lambda_n$ its eigenvalues arranged in increasing order, counted with multiplicity. Then*

$$\lambda_k = \min_{\dim F = k} \max_{x \in F \setminus \{0\}} \frac{x^T M x}{\|x\|_2^2}.$$

If M is complex Hermitian, one has similarly

$$\lambda_k = \min_{\dim F = k} \max_{x \in F \setminus \{0\}} \frac{x^* M x}{\|x\|_2^2}.$$

This formula generalizes (3.1, 3.2).

3.3.2 Applications

Theorem 3.3.3 *Let $H \in \mathbf{H}_{n-1}$, $x \in \mathbb{C}^{n-1}$, and $a \in \mathbb{R}$ be given. Let $\lambda_1 \leq \cdots \leq \lambda_{n-1}$ be the eigenvalues of H and $\mu_1 \leq \cdots \leq \mu_n$ those of the Hermitian matrix*

$$H' = \begin{pmatrix} H & x \\ x^* & a \end{pmatrix}.$$

One has then $\mu_1 \leq \lambda_1 \leq \cdots \leq \mu_j \leq \lambda_j \leq \mu_{j+1} \leq \cdots$.

Proof

By Theorem 3.3.2, the inequality $\mu_j \leq \lambda_j$ is obvious, because the infimum is taken over a smaller set.

Conversely, let $\pi : x \mapsto (x_1, \ldots, x_{n-1})^T$ be the projection from $\mathbb{C}^n$ on $\mathbb{C}^{n-1}$. If F is a linear subspace of $\mathbb{C}^n$ of dimension $j+1$, its image under π contains a linear subspace G of dimension j (it will often be exactly of dimension j). By Theorem 3.3.2, applied to H, one therefore has

$$R'(F) \geq R(G) \geq \lambda_j.$$

Taking the infimum, we obtain $\mu_{j+1} \geq \lambda_j$.

■

The previous theorem is optimal, in the following sense.

Theorem 3.3.4 *Let $\lambda_1 \leq \cdots \leq \lambda_{n-1}$ and $\mu_1 \leq \cdots \leq \mu_n$ be real numbers satisfying $\mu_1 \leq \lambda_1 \leq \cdots \leq \mu_j \leq \lambda_j \leq \mu_{j+1} \leq \cdots$. Then there exist a vector $x \in \mathbb{R}^n$ and $a \in \mathbb{R}$ such that the real symmetric matrix*

$$H = \begin{pmatrix} \Lambda & x \\ x^T & a \end{pmatrix},$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{n-1})$, has the eigenvalues μ_j.

Proof

Let us compute the characteristic polynomial of H from Schur's complement formula[5] (see Proposition 8.1.2):

$$\begin{aligned} p_n(X) &= \left(X - a - x^T (X I_{n-1} - \Lambda)^{-1} x\right) \det(X I_{n-1} - \Lambda) \\ &= \left(X - a - \sum_j \frac{x_j^2}{X - \lambda_j}\right) \prod_j (X - \lambda_j). \end{aligned}$$

Let us assume for the moment that all the inequalities $\mu_j \leq \lambda_j \leq \mu_{j+1}$ hold strictly. In particular, the λ_j's are distinct. Let us consider the partial fraction decomposition of the rational function

$$\frac{\prod_l (X - \mu_l)}{\prod_j (X - \lambda_j)} = X - a - \sum_j \frac{c_j}{X - \lambda_j}.$$

[5] One may equally (exercise) compute it by induction on n.

One thus obtains

$$a = \sum_l \mu_l - \sum_j \lambda_j ,$$

a formula that could also have been found by comparing the traces of Λ and of H. The inequalities $\lambda_{j-1} < \mu_j < \lambda_j$ ensure that each c_j is positive, because

$$c_j = -\frac{\prod_l (\lambda_j - \mu_l)}{\prod_{k \neq j} (\lambda_j - \lambda_k)}.$$

Let us put, then, $x_j = \sqrt{c_j}$ (or $-x_j = \sqrt{c_j}$). We obtain, as announced,

$$p_n(X) = \prod_l (X - \mu_l).$$

In the general case one may choose sequences $\mu_l^{(m)}$ and $\lambda_j^{(m)}$ that converge to the μ_l's and the λ_j's as $m \to +\infty$ and that satisfy the inequalities in the hypothesis strictly. The first part of the proof (case with strict inequalities) provides matrices $H^{(m)}$. Since the spectral radius is a norm over $\mathbf{Sym}_n(\mathbb{R})$ (the spectral radius is defined in the next Chapter), the sequence $(H^{(m)})_{m \in \mathbb{N}}$ is bounded. In other words, $(a^{(m)}, x^{(m)})$ remains bounded. Let us extract a subsequence that converges to a pair $(a, x) \in \mathbb{R} \times \mathbb{R}^{n-1}$. The matrix H associated to (a, x) solves our problem, since the eigenvalues depend continuously on the entries of the matrix.

■

Corollary 3.3.2 *Let $H \in \mathbf{Sym}_{n-1}(\mathbb{R})$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_{n-1}$. Let $\mu_1, \ldots, \mu_n$ be real numbers satisfying $\mu_1 \leq \lambda_1 \leq \cdots \leq \mu_j \leq \lambda_j \leq \mu_{j+1} \leq \cdots$. Then there exist a vector $x \in \mathbb{R}^n$ and $a \in \mathbb{R}$ such that the real symmetric matrix*

$$H' = \begin{pmatrix} H & x \\ x^T & a \end{pmatrix}$$

has the eigenvalues μ_j.

The proof consists in diagonalizing H through an orthogonal conjugation, then applying the theorem, and finally performing the inverse conjugation.

3.4 The Spectrum and the Diagonal of Hermitian Matrices

Let us begin with an order relation between finite sequences of real numbers. If $a = (a_1, \ldots, a_n)$ is a sequence of n real numbers, and if $1 \leq l \leq n$,

we denote by $s_k(a)$ the number

$$\min\left\{\sum_{j\in J} a_j \mid \operatorname{card} J = k\right\}.$$

Definition 3.4.1 *Let $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ be two sequences of n real numbers. One says that b* majorizes *a, and one writes $a \prec b$, if*

$$s_k(a) \le s_k(b), \quad \forall 1 \le k \le n, \quad s_n(a) = s_n(b).$$

The functions s_k are symmetric:

$$s_k(a) = s_k(a_{\sigma(1)}, \dots, a_{\sigma(n)})$$

for every permutation σ. One thus may always restrict attention to the case of nondecreasing sequences $a_1 \le \cdots \le a_n$. One has then $s_k(a) = a_1 + \cdots + a_k$. The relation $a \prec b$ for nondecreasing sequences, can now be written as

$$\begin{aligned} a_1 + \cdots + a_k &\le b_1 + \cdots + b_k, \quad k = 1, \dots, n-1, \\ a_1 + \cdots + a_n &= b_1 + \cdots + b_n. \end{aligned}$$

The latter equality plays a crucial role in the analysis below. The relation $\prec$ is a partial ordering.

Proposition 3.4.1 *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if for every real number t,*

$$\sum_{j=1}^{n} |x_j - t| \ge \sum_{j=1}^{n} |y_j - t|. \tag{3.3}$$

Proof

We may assume that x and y are nondecreasing. If the inequality (3.3) holds, we write it first for t outside the interval I containing the x_j's and the y_j's. This gives $s_n(x) = s_n(y)$. Then we write it for $t = x_k$. Using $s_n(x) = s_n(y)$, we obtain

$$\begin{aligned} \sum_j |x_j - x_k| &= \sum_1^k (x_k - y_j) + \sum_{k+1}^n (y_j - x_k) + 2(s_k(y) - s_k(x)) \\ &\le \sum_j |y_j - x_k| + 2(s_k(y) - s_k(x)), \end{aligned}$$

which with (3.3) gives $s_k(x) \le s_k(y)$.

Conversely, let us assume that $x \prec y$. Let us define $\phi(t) := \sum_j |x_j - t| - \sum_j |y_j - t|$. This is a piecewise linear function, zero outside I. Its derivative, integer-valued, is piecewise constant. It increases at the points x_j's and decreases at the points y_j's only. If $\min\{\phi(t); t \in \mathbb{R}\} < 0$, this minimum will thus be reached at some x_k, with $\phi'(x_k - 0) \le 0 \le \phi'(x_k + 0)$,

from which one obtains $y_{k-1} \leq x_k \leq y_{k+1}$. Therefore, there are two cases, depending on the position of y_x with respect to x_k. For example, if $y_k \leq x_k$, we compute

$$\sum_j |x_j - x_k| = \sum_{k+1}^{n} (x_j - x_k) + \sum_{1}^{k} (x_k - x_j).$$

From the assumption, it follows that

$$\sum_j |x_j - x_k| \geq \sum_{k+1}^{n} (y_j - x_k) + \sum_{1}^{k} (x_k - y_j) = \sum_{j \neq k} |y_j - x_k|,$$

which means that $\phi(x_k) \geq 0$, which contradicts the hypothesis. Hence, ϕ is a nonnegative function.

■

Our first statement expresses an order between the diagonal and the spectrum of a Hermitian matrix.

Theorem 3.4.1 (Schur) *Let H be a Hermitian matrix with diagonal a and spectrum λ. Then $a \succ \lambda$.*

Proof

Let n be the size of H. We argue by induction on n. We may assume that a_n is the largest component of a. Since $s_n(\lambda) = \operatorname{Tr} A$, one has $s_n(\lambda) = s_n(a)$. In particular, the theorem holds true for order 1. Let us assume that it holds for order $n-1$. Let A be the matrix obtained from H by deleting the nth row and the nth column. Let $\mu = (\mu_1, \ldots, \mu_{n-1})$ be the spectrum of A. Let us arrange λ and μ in increasing order. From Theorem 3.3.3, one has $\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \cdots \leq \mu_{n-1} \leq \lambda_n$. It follows that $s_k(\mu) \geq s_k(\lambda)$ for every $k < n$. The induction hypothesis tells us that $s_k(\mu) \leq s_k(a')$, where $a' = (a_1, \ldots, a_{n-1})$. Finally, we have $s_k(a') = s_k(a)$, and $s_k(\lambda) \leq s_k(a)$ for every $k < n$, which ends the induction.

■.

Here is the converse.

Theorem 3.4.2 *Let a and λ be two sequences of n real numbers such that $a \succ \lambda$. Then there exists a real symmetric matrix of size $n \times n$ whose diagonal is a and spectrum is λ.*

Proof

We proceed by induction on n. The statement is trivial if $n = 1$. If $n \geq 2$, we use the following lemma, which will be proved afterwards.

Lemma 3.4.1 *Let $n \geq 2$ and α, β two nondecreasing sequences of n real numbers, satisfying $\alpha \prec \beta$. Then there exists a sequence γ of $n-1$ real numbers such that*

$$\alpha_1 \leq \gamma_1 \leq \alpha_2 \leq \cdots \leq \gamma_{n-1} \leq \alpha_n$$

and $\gamma \prec \beta' = (\beta_1, \ldots, \beta_{n-1})$.

We apply the lemma to the sequences $\alpha = \lambda$, $\beta = a$. Since $\gamma \prec a'$, the induction hypothesis tells us that there exists a real symmetric matrix S of size $(n-1) \times (n-1)$ with diagonal a' and spectrum γ. From Corollary 3.3.2, there exist a vector $y \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that the matrix

$$\Sigma = \begin{pmatrix} S & y^T \\ y & b \end{pmatrix}$$

has spectrum λ. Since $s_n(a) = s_n(\lambda) = \operatorname{Tr} \Sigma = \operatorname{Tr} S + b = s_{n-1}(a') + b$, we have $b = a_n$. Hence, a is the diagonal of Σ.

■

We prove now Lemma 3.4.1. Let Δ be the set of sequences δ of $n-1$ real numbers satisfying

$$\alpha_1 \leq \delta_1 \leq \alpha_2 \leq \cdots \leq \delta_{n-1} \leq \alpha_n \tag{3.4}$$

together with

$$\sum_{j=1}^{k} \delta_j \leq \sum_{j=1}^{k} \beta_j, \quad \forall k \leq n-2. \tag{3.5}$$

We must show that there exists $\delta \in \Delta$ such that $s_{n-1}(\delta) = s_{n-1}(\beta')$. Since Δ is convex and compact (it is closed and bounded in $\mathbb{R}^n$), it is enough to show that

$$\inf_{\delta \in \Delta} s_{n-1}(\delta) \leq s_{n-1}(\beta') \leq \sup_{\delta \in \Delta} s_{n-1}(\delta). \tag{3.6}$$

On the one hand, $\alpha' = (\alpha_1, \ldots, \alpha_{n-1})$ belongs to Δ and $s_{n-1}(\alpha') \leq s_{n-1}(\beta')$ from the hypothesis, which proves the first inequality in (3.6).

Let us now choose a δ that achieves the supremum of s_{n-1} over Δ. Let r be the largest index less than or equal to $n-2$ such that $s_r(\delta) = s_r(\beta')$, with $r = 0$ if all the inequalities are strict. From $s_j(\delta) < s_j(\beta')$ for $r < j < n-1$, one has $\delta_j = \alpha_{j+1}$, since otherwise, there would exist $\epsilon > 0$ such that $\hat{\delta} := \delta + \epsilon \mathbf{e}^j$ belong to Δ, and one would have $s_{n-1}(\hat{\delta}) = s_{n-1}(\delta) + \epsilon$, contrary to the maximality of δ. Now let us compute

$$\begin{aligned} s_{n-1}(\delta) - s_{n-1}(\beta') &= s_r(\beta) - s_{n-1}(\beta) + \alpha_{r+2} + \cdots + \alpha_n \\ &= s_r(\beta) - s_{n-1}(\beta) + s_n(\alpha) - s_{r+1}(\alpha) \\ &\geq s_r(\beta) - s_{n-1}(\beta) + s_n(\beta) - s_{r+1}(\beta) \\ &= \beta_n - \beta_{r+1} \geq 0. \end{aligned}$$

This proves (3.6) and completes the proof of the lemma.

■

3.4.1 Hadamard's Inequality

Proposition 3.4.2 *Let $H \in \mathbf{H}_n$ be a positive semidefinite Hermitian matrix. Then*

$$\det H \leq \prod_{j=1}^{n} h_{jj}.$$

If $H \in \mathbf{HPD}_n$, the equality holds only if H is diagonal.

Proof

If $\det H = 0$, there is nothing to prove, because the h_{jj} are nonnegative (these are numbers $(\mathbf{e}^j)^* H \mathbf{e}^j$). Otherwise, H is positive definite and one has $h_{jj} > 0$. We restrict attention to the case with a constant diagonal by letting $D := \mathrm{diag}(h_{11}^{-1/2}, \ldots, h_{nn}^{-1/2})$ and writing $(\det H)/(\prod_j h_{jj}) = \det DHD = \det H'$, where the diagonal entries of H' equal one. There remains to prove that $\det H' \leq 1$. However, the eigenvalues $\mu_1, \ldots, \mu_n$ of H' are strictly positive, of sum n. Since the logarithm is concave, one has

$$\frac{1}{n} \log \det H' = \frac{1}{n} \sum_j \log \mu_j \leq \log \frac{1}{n} \sum \mu_j = \log 1 = 0,$$

which proves the inequality. Since the concavity is strict, the equality holds only if $\mu_1 = \cdots = \mu_n = 1$, but then H' is similar, thus equal to I_n. In that case, H is diagonal.

■

Applying proposition 3.4.2 to matrices of the form M^*M or MM^*, one obtains the following result.

Theorem 3.4.3 *For $M \in \mathbf{M}_n(\mathbb{C})$, one has*

$$|\det M| \leq \prod_{i=1}^{n} \left(\sum_{j=1}^{n} |m_{ij}|^2 \right)^{1/2}, \quad |\det M| \leq \prod_{j=1}^{n} \left(\sum_{i=1}^{n} |m_{ij}|^2 \right)^{1/2}.$$

When $M \in \mathbf{GL}_n(\mathbb{C})$, the first (respectively the second) inequality is an equality only if the rows (respectively the columns) of M are pairwise orthogonal.

3.5 Exercises

1. Show that the eigenvalues of skew-Hermitian matrices, or as well those of real skew-symmetric matrices, are pure imaginary.

2. Let $P, Q \in \mathbf{M}_n(\mathbb{R})$ be given. Assume that $P + iQ \in \mathbf{GL}_n(\mathbb{C})$. Show that there exist $a, b \in \mathbb{R}$ such that $aP + bQ \in \mathbf{GL}_n(\mathbb{R})$. Deduce that if $M, N \in \mathbf{M}_n(\mathbb{R})$ are similar in $\mathbf{M}_n(\mathbb{C})$, then these matrices are similar in $\mathbf{M}_n(\mathbb{R})$.

3. Show that a triangular *and* normal matrix is diagonal. Deduce that if U^*TU is a unitary trigonalization of M, and if M is normal, then T is diagonal.

4. For $A \in \mathbf{M}_n(\mathbb{R})$, symmetric positive definite, show that

$$\max_{i,j \leq n} |a_{ij}| = \max_{i \leq n} a_{ii}.$$

5. Given an invertible matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{GL}_2(\mathbb{R}),$$

define a map h_M from $S^2 := \mathbb{C} \cup \{\infty\}$ into itself by

$$h_M(z) := \frac{az+b}{cz+d}.$$

(a) Show that h_M is a bijection.
(b) Show that $h : M \mapsto h_M$ is a group homomorphism. Compute its kernel.
(c) Let $\mathcal{H}$ be the upper half-plane, consisting on those $z \in \mathbb{C}$ with $\Im z > 0$. Compute $\Im h_M(z)$ in terms of $\Im z$ and deduce that the subgroup

$$\mathbf{GL}_2^+(\mathbb{R}) := \{M \in \mathbf{GL}_2(\mathbb{R}) \mid \det M > 0\}$$

acts on $\mathcal{H}$.
(d) Conclude that the group $\mathbf{PSL}_2(\mathbb{R}) := \mathbf{SL}_2(\mathbb{R})/\{\pm I_2\}$, called the *modular group*, acts on $\mathcal{H}$.
(e) Let $M \in \mathbf{SL}_2(\mathbb{R})$ be given. Determine, in terms of $\operatorname{Tr} M$, the number of fixed points of h_M on $\mathcal{H}$.

6. Show that the supremum of a family of convex functions on $\mathbb{R}^N$ is convex. Deduce that the map $M \mapsto \lambda_n$ (largest eigenvalue of M) defined on $\mathbf{H}_n$ is convex.

7. Show that $M \in \mathbf{M}_n(\mathbb{C})$ is normal if and only if there exists a unitary matrix U such that $M^* = MU$.

8. Show that in $\mathbf{M}_n(\mathbb{C})$ the set of diagonalizable matrices is dense. **Hint**: Use Theorem 3.1.3.

9. Let $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ be two sequences of real numbers. Find the supremum and the infimum of $\operatorname{Tr}(AB)$ as A (respectively B) runs over the Hermitian matrices with spectrum equal to $(a_1, \ldots, a_n)$ (respectively $(b_1, \ldots, b_n)$).

10. (Kantorovich inequality)

(a) Let $a_1 \leq \cdots \leq a_n$ be a list of real numbers, with $a_n^{-1} = a_1 > 0$. Define

$$l(u) := \sum_{j=1}^{n} a_j u_j, \quad L(u) := \sum_{j=1}^{n} \frac{u_j}{a_j}.$$

Let K_n be the simplex of $\mathbb{R}^n$ defined by the constraints $u_j \geq 0$ for every $j = 1, \ldots, n$, and $\sum_j u_j = 1$. Show that there exists an element $v \in K_n$ that maximizes $l + L$ and minimizes $|L - l|$ on K_n simultaneously.

(b) Deduce that

$$\max_{u \in K_n} l(u)L(u) = \left(\frac{a_1 + a_n}{2}\right)^2.$$

(c) Let $A \in \mathbf{HPD}_n$ and let a_1, a_n be the smallest and largest eigenvalues of A. Show that for every $x \in C^n$,

$$(x^* Ax)(x^* A^{-1} x) \leq \frac{(a_1 + a_n)^2}{4a_1 a_n} \|x\|^4.$$

11. (Weyl's inequalities)
Let A, B be two Hermitian matrices of size $n \times n$ whose respective eigenvalues are $\alpha_1 \leq \cdots \leq \alpha_n$ and $\beta_1 \leq \cdots \leq \beta_n$. Define $C = A + B$ and let $\gamma_1 \leq \cdots \leq \gamma_n$ be its eigenvalues.

(a) Show that $\alpha_j + \beta_1 \leq \gamma_j \leq \alpha_j + \beta_n$.

(b) Let us recall that if F is a linear subspace of $\mathbb{C}^n$, one writes

$$R_A(F) = \max\{x^* Ax \,|\, x \in F,\ \|x\|_2 = 1\}.$$

Show that if G, H are two linear subspaces of $\mathbb{C}^n$, then $R_C(G \cap H) \leq R_A(G) + R_B(H)$.

(c) Deduce that if $l, m \geq 1$ and $l + m = k + n$ (hence $l + m \geq n + 1$), then

$$\gamma_k \leq \alpha_l + \beta_m.$$

(d) Similarly, show that $l + m = k + 1$ implies

$$\gamma_k \geq \alpha_l + \beta_m.$$

(e) Conclude that the function $A \mapsto \lambda_k(A)$ that associates to a Hermitian matrix its kth eigenvalue (in increasing order) is Lipschitz with ratio 1, meaning that

$$|\lambda_k(B) - \lambda_k(A)| \leq \|B - A\|_2 = \rho(B - A)$$

(see the next chapter for the meaning of the norm $\|M\|_2$ and for the spectral radius $\rho(M)$).

Remark: The description of the set of the $3n$-tuplets $(\vec{\alpha}, \vec{\beta}, \vec{\gamma})$ as A and B run over $\mathbf{H}_n$ is especially delicate. For a complete historical

account of this question, one may read the first section of Fulton's and Bhatia's articles [16, 6]. For another partial result, see Exercise 19 of Chapter 5 (theorem of Lidskii).

12. Let A be a Hermitian matrix of size $n \times n$ whose eigenvalues are $\alpha_1 \leq \cdots \leq \alpha_n$. Let B be a Hermitian positive semidefinite matrix. Let $\gamma_1 \leq \cdots \leq \gamma_n$ be the eigenvalues of $A + B$. Show that $\gamma_k \geq \alpha_k$.

13. Let M, N be two Hermitian matrices such that N and $M - N$ are positive semidefinite. Show that $\det N \leq \det M$.

14. Let $A \in \mathbf{M}_p(\mathbb{C})$, $C \in \mathbf{M}_q(\mathbb{C})$ be given with $p, q \geq 1$. Assume that

$$M := \begin{pmatrix} A & B \\ B^* & C \end{pmatrix}$$

is Hermitian positive definite. Show that $\det M \leq (\det A)(\det C)$. Use the previous exercise and Proposition 8.1.2.

15. For $M \in \mathbf{HPD}_n$, we denote by $P_k(M)$ the product of all the principal minors of order k of M. There are

$$\begin{pmatrix} n \\ k \end{pmatrix}$$

such minors.
Applying Proposition 3.4.2 to the matrix M^{-1}, show that

$$P_n(M)^{n-1} \leq P_{n-1}(M),$$

and then in general that

$$P_{k+1}(M)^k \leq P_k(M)^{n-k}.$$

16. Let $d : \mathbf{M}_n(\mathbb{R}) \to \mathbb{R}^+$ be a multiplicative function; that is,

$$d(MN) = d(M)d(N)$$

for every $M, N \in \mathbf{M}_n(\mathbb{R})$. If $\alpha \in \mathbb{R}$, define $\delta(\alpha) := d(\alpha I_n)^{1/n}$. Assume that d is not constant.

(a) Show that $d(0_n) = 0$ and $d(I_n) = 1$. Deduce that $P \in \mathbf{GL}_n(\mathbb{R})$ implies $d(P) \neq 0$ and $d(P^{-1}) = 1/d(P)$. Show, finally, that if M and N are similar, then $d(M) = d(N)$.
(b) Let $D \in \mathbf{M}_n(\mathbb{R})$ be diagonal. Find matrices $D_1, \ldots, D_{n-1}$, similar to D, such that $DD_1 \cdots D_{n-1} = (\det D)I_n$. Deduce that $d(D) = \delta(\det D)$.
(c) Let $M \in \mathbf{M}_n(\mathbb{R})$ be a diagonalizable matrix. Show that $d(M) = \delta(\det M)$.
(d) Using the fact that M^T is similar to M, show that $d(M) = \delta(\det M)$ for every $M \in \mathbf{M}_n(\mathbb{R})$.

17. Let $B \in \mathbf{GL}_n(\mathbb{C})$. Verify that the inverse and the Hermitian adjoint of $B^{-1}B^*$ are similar. Conversely, let $A \in \mathbf{GL}_n(\mathbb{C})$ be a matrix whose inverse and the Hermitian adjoint are similar: $A^* = PA^{-1}P^{-1}$.

    (a) Show that there exists an invertible Hermitian matrix H such that $H = A^*HA$. Look for an H as a linear combination of P and of P^*.
    (b) Show that there exists a matrix $B \in \mathbf{GL}_n(\mathbb{C})$ such that $A = B^{-1}B^*$. Look for a B of the form $(aI_n + bA^*)H$.

18. Let $A \in \mathbf{M}_n(\mathbb{C})$ be given, and let $\lambda_1, \ldots, \lambda_n$ be its eigenvalues. Show, by induction on n, that A is normal if and only if

$$\sum_{i,j} |a_{ij}|^2 = \sum_{1}^{n} |\lambda_l|^2.$$

    Hint: The left-hand side (whose square root is called *Schur's norm*) is invariant under conjugation by a unitary matrix. It is then enough to restrict attention to the case of a triangular matrix.

19. (a) Show that $|\det(I_n + A)| \geq 1$ for every skew-Hermitian matrix A, and that equality holds only if $A = 0_n$.
    (b) Deduce that for every $M \in \mathbf{M}_n(\mathbb{C})$ such that $H := (M + M^*)/2$ is positive definite,

$$\det H \leq |\det M|$$

    by showing that $H^{-1}(M - M^*)$ is similar to a skew-Hermitian matrix. You may use the *square root* defined at Chapter 7.

20. Describe every positive semidefinite matrix $M \in \mathbf{Sym}_n(\mathbb{R})$ such that $m_{jj} = 1$ for every j and possessing the eigenvalue $\lambda = n$ (first show that M has rank one).

21. If $A, B \in M_{n\times m}(\mathbb{C})$, define the *Hadamard product* of A and B by

$$A \circ B := (a_{ij}b_{ij})_{1\leq i\leq n, 1\leq j\leq m}.$$

    (a) Let A, B be two Hermitian matrices. Verify that $A \circ B$ is Hermitian.
    (b) Assume that A and B are positive semidefinite, of respective ranks p and q. Using Proposition 3.2.1, show that there exist pq vectors $z_{\alpha\beta}$ such that

$$A \circ B = \sum_{\alpha,\beta} z_{\alpha\beta} z_{\alpha\beta}^*.$$

    Deduce that $A \circ B$ is positive semi-definite.
    (c) If A and B are positive definite, show that $A \circ B$ also is positive definite.

(d) Construct an example for which $p, q < n$, but $A \circ B$ is positive definite.

22. (Fiedler and Pták [13]) Given a matrix $A \in \mathbf{M}_n(\mathbb{R})$, we wish to prove the equivalence of the following properties:

P1 For every vector $x \neq 0$ there exists an index k such that $x_k(Ax)_k > 0$.

P2 For every vector $x \neq 0$ there exists a diagonal matrix D with positive diagonal elements such that the scalar product (Ax, Dx) is positive.

P3 For every vector $x \neq 0$ there exists a diagonal matrix D with nonnegative diagonal elements such that the scalar product (Ax, Dx) is positive.

P4 The real eigenvalues of all principal submatrices of A are positive.

P5 All principal minors of A are positive.

We shall use the following notation: if $x \in \mathbb{R}^n$ and if J is the index set of its nonzero components, then x^J denotes the vector in $\mathbb{R}^k$, and k the cardinality of J, where one retains only the nonzero components of x. To the set J one also associates the matrix A^J, retaining only the indices in J.

(a) Prove that **Pj** implies **P(j+1)** for every $j = 1, \ldots, 4$.

(b) Assume **P5**. Show that for every diagonal matrix D with nonnegative entries, one has $\det(A + D) > 0$.

(c) Then prove that **P5** implies **P1**.

4

Norms

4.1 A Brief Review

In this Chapter, the field K will always be $\mathbb{R}$ or $\mathbb{C}$ and E will denote K^n.

If $A \in \mathbf{M}_n(K)$, the *spectral radius* of A, denoted by $\rho(A)$, is defined as the largest modulus of the eigenvalues of A:

$$\rho(A) = \max\{|\lambda|; \lambda \in \mathrm{Sp}(A)\}.$$

When $K = \mathbb{R}$, one takes into account the complex eigenvalues when computing $\rho(A)$.

The scalar (if $K = \mathbb{R}$) or Hermitian (if $K = \mathbb{C}$) product on E is denoted by $(x, y) := \sum_j x_j \bar{y}_j$. The vector space E is endowed with various norms, pairwise equivalent since E has finite dimension (Proposition 4.1.3 below). Among these, the most used norms are the l^p norms:

$$\|x\|_p = \left(\sum_j |x_j|^p \right)^{1/p}, \quad \|x\|_\infty = \max_j |x_j|.$$

Proposition 4.1.1 *For $1 \leq p \leq \infty$, the map $x \mapsto \|x\|_p$ is a norm on E. In particular, one has* Minkowski's inequality

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p. \tag{4.1}$$

Furthermore, one has Hölder's inequality

$$|(x,y)| \leq \|x\|_p \|y\|_{p'}, \quad \frac{1}{p} + \frac{1}{p'} = 1. \tag{4.2}$$

The numbers p, p' are called *conjugate exponents.*

Proof

Everything except the Hölder and Minkowski inequalities is obvious. When $p = 1$ or $p = \infty$, these inequalities are trivial. We thus assume that $1 < p < \infty$.

Let us begin with (4.2). If x or y is null, it is obvious. Indeed, one can even assume, by decreasing the value of n, that none of the x_j, y_j's is null. Likewise, since $|(x,y)| \leq \sum_j |x_j||y_j|$, one can also assume that the x_j, y_j are real and positive. Dividing by $\|x\|_p$ and by $\|y\|_{p'}$, one may restrict attention to the case where $\|x\|_p = \|y\|_{p'} = 1$. Hence, $x_j, y_j \in (0,1]$ for every j. Let us define

$$a_j = p \log x_j, \quad b_j = p' \log y_j.$$

Since the exponential function is convex,

$$e^{a_j/p + b_j/p'} \leq \frac{1}{p} e^{a_j} + \frac{1}{p'} e^{b_j},$$

that is,

$$x_j y_j \leq \frac{1}{p} x_j^p + \frac{1}{p'} y_j^{p'}.$$

Summing over j, we obtain

$$(x,y) \leq \frac{1}{p}\|x\|_p^p + \frac{1}{p'}\|y\|_{p'}^{p'} = \frac{1}{p} + \frac{1}{p'} = 1,$$

which proves (4.2).

We now turn to (4.1). First, we have

$$\|x+y\|_p^p = \sum_k |x_k + y_k|^p \leq \sum_k |x_k||x_k + y_k|^{p-1} + \sum_k |y_k||x_k + y_k|^{p-1}.$$

Let us apply Hölder's inequality to each of the two terms of the right-hand side. For example,

$$\sum_k |x_k||x_k + y_k|^{p-1} \leq \|x\|_p \left(\sum_k |x_k + y_k|^{(p-1)p'} \right)^{1/p'},$$

which amounts to

$$\sum_k |x_k||x_k + y_k|^{p-1} \leq \|x\|_p \|x+y\|_p^{p-1}.$$

Finally,

$$\|x+y\|_p^p \leq (\|x\|_p + \|y\|_p)\|x+y\|_p^{p-1},$$

which gives (4.1).

■

For $p = 2$, the norm $\|\cdot\|_2$ is given by a Hermitian form and thus satisfies the Cauchy–Schwarz inequality:

$$|(x, y)| \leq \|x\|_2 \|y\|_2.$$

This is a particular case of Hölder's inequality.

Proposition 4.1.2 *For conjugate exponents p, p', one has*

$$\|x\|_p = \sup_{y \neq 0} \frac{\Re(x, y)}{\|y\|_{p'}}.$$

Proof

The inequality $\geq$ is a consequence of Hölder's. The reverse inequality is obtained by taking $y_j = \bar{x}_j |x_j|^{p-2}$ if $p < \infty$. If $p = \infty$, choose $y_j = \bar{x}_j$ for an index j such that $|x_j| = \|x\|_\infty$. For $k \neq j$, take $y_k = 0$.

■

Definition 4.1.1 *Two norms N and N' on a (real or complex) vector space are said to be equivalent if there exist two numbers $c, c' \in \mathbb{R}$ such that*

$$N \leq cN', \quad N' \leq c'N.$$

The equivalence between norms is obviously an equivalence relation, as its name implies. As announced above, we have the following result.

Proposition 4.1.3 *All norms on $E = K^n$ are equivalent. For example,*

$$\|x\|_\infty \leq \|x\|_p \leq n^{1/p} \|x\|_\infty.$$

Proof

It is sufficient to show that every norm is equivalent to $\|\cdot\|_1$.

Let N be a norm on E. If $x \in E$, the triangle inequality gives

$$N(x) \leq \sum_i |x_i| N(\mathbf{e}^i),$$

where $(\mathbf{e}^1, \ldots, \mathbf{e}^n)$ is the canonical basis. One thus has $N \leq c\|\cdot\|_1$ for $c := \max_i N(\mathbf{e}^i)$. Observe that this first inequality expresses the fact that N is Lipschitz (hence continuous) on the metric space $X = (E, \|\cdot\|_1)$.

For the reverse inequality, we reduce *ad absurdum*: Let us assume that the supremum of $\|x\|_1 / N(x)$ is infinite for $x \neq 0$. By homogeneity, there would then exist a sequence of vectors $(x^m)_{m \in \mathbb{N}}$ such that $\|x^m\|_1 = 1$ and $N(x^m) \to 0$ when $m \to +\infty$. Since the unit sphere of X is compact, one may assume (up to the extraction of a subsequence) that x^m converges to a vector x such that $\|x\|_1 = 1$. In particular, $x \neq 0$. Since N is continuous on X, one has also $N(x) = \lim_{m \to +\infty} N(x^m) = 0$. Since N is a norm, we deduce $x = 0$, a contradiction.

■

4.1.1 Duality

Definition 4.1.2 *Given a norm $\|\cdot\|$ on $\mathbb{R}^n$, its dual norm on $\mathbb{R}^n$ is defined by*

$$\|x\|' := \sup_{y\neq 0} \frac{y^T x}{\|y\|}.$$

The fact that $\|\cdot\|'$ is a norm is obvious. The dual of a norm on $\mathbb{C}^n$ is defined in a similar way, with $\Re y^* x$ instead of $y^T x$. For every $x, y \in \mathbb{C}^n$, one has

$$\Re y^* x \leq \|x\| \cdot \|y\|'. \tag{4.3}$$

Proposition 4.1.2 shows that the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$ for $1/p+1/q = 1$. This suggests the following property.

Proposition 4.1.4 *The bidual (dual of the dual norm) of a norm is this norm itself:*

$$(\|\cdot\|')' = \|\cdot\|.$$

Proof

From (4.3), one has $(\|\cdot\|')' \leq \|\cdot\|$. The converse is a consequence of the Hahn–Banach theorem: the unit ball B of $\|\cdot\|$ is convex and compact. If x is a point of its boundary (that is, $\|x\| = 1$), there exists an $\mathbb{R}$-affine (that is, of the form constant plus $\mathbb{R}$-linear) function that is zero at x and nonpositive on B. Such a function can be written in the form $z \mapsto \Re z^* y + c$, where c is a constant, necessarily equal to $-\Re z^* x$. Without loss of generality, one may assume that $z^* x$ is real. Hence

$$\|y\|' = \sup_{\|z\|=1} \Re y^* z = y^* x.$$

One deduces

$$(\|x\|')' \geq \frac{y^* x}{\|y\|'} = 1 = \|x\|.$$

By homogeneity, this is true for every $x \in \mathbb{C}^n$.

■

4.1.2 Matrix Norms

Let us recall that $\mathbf{M}_n(K)$ can be identified with the set of endomorphisms of $E = K^n$ by

$$A \mapsto (x \mapsto Ax).$$

Definition 4.1.3 *If $\|\cdot\|$ is a norm on E and if $A \in \mathbf{M}_n(K)$, we define*

$$\|A\| := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Equivalently,

$$\|A\| = \sup_{\|x\|\leq 1} \|Ax\| = \max_{\|x\|\leq 1} \|Ax\|.$$

One verifies easily that $A \mapsto \|A\|$ is a norm on $\mathbf{M}_n(K)$. It is called the *norm induced* by that of E, or the norm *subordinated* to that of E. Though we adopted the same notation $\|\cdot\|$ for the two norms, that on E and that on $\mathbf{M}_n(K)$, these are, of course, distinct objects. In many places, one finds the notation $|||\cdot|||$ for the induced norm. When one does not wish to mention from which norm on E a given norm on $\mathbf{M}_n(K)$ is induced, one says that $A \mapsto \|A\|$ is a *matrix norm*. The main properties of matrix norms are

$$\|AB\| \leq \|A\|\,\|B\|, \quad \|I_n\| = 1.$$

These properties are those of any *algebra norm* (otherwise called *norm of algebra*, see Section 4.4). In particular, one has $\|A^k\| \leq \|A\|^k$ for every $k \in \mathbb{N}$.

Here are a few examples induced by the norms l^p:

$$\begin{aligned} \|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{i=n} |a_{ij}|, \\ \|A\|_\infty &= \max_{1<i<n} \sum_{j=1}^{j=n} |a_{ij}|, \\ \|A\|_2 &= \rho(A^*A)^{1/2}. \end{aligned}$$

To prove these formulas, we begin by proving the inequalities $\geq$, selecting a suitable vector x, and writing $\|A\|_p \geq \|Ax\|_p/\|x\|_p$. For $p = 1$ we choose an index j such that the maximum in the above formula is achieved. Then we let $x_j = 1$, while $x_k = 0$ otherwise. For $p = \infty$, we let $x_j = \bar{a}_{i_0 j}/|a_{i_0 j}|$, where i_0 achieves the maximum in the above formula; For $p = 2$ we choose an eigenvector of A^*A associated to an eigenvalue of maximal modulus. We thus obtain three inequalities. The reverse inequalities are direct consequences of the definitions. The values of $\|A\|_1$ and $\|A\|_\infty$ illustrate a particular case of the general formula

$$\|A^*\|' = \|A\| = \sup_{x\neq 0} \sup_{y\neq 0} \frac{\Re(y^*Ax)}{\|x\| \cdot \|y\|'}.$$

Proposition 4.1.5 *For an induced norm, the condition $\|B\| < 1$ implies that $I_n - B$ is invertible, with inverse given by the sum of the series*

$$\sum_{k=0}^{\infty} B^k.$$

Proof

The series $\sum_k B^k$ is normally convergent, since $\sum_k \|B^k\| \leq \sum_k \|B\|^k$, where the latter series converges because $\|B\| < 1$. Since $\mathbf{M}_n(K)$ is com-

plete, the series $\sum_k B^k$ converges. Furthermore, $(I_n - B)\sum_{k\leq N} B^k = I_n - B^{N+1}$, which tends to I_n. The sum of the series is thus the inverse of $I_n - B$. One has, moreover,

$$\|(I_n - B)^{-1}\| \leq \sum_k \|B\|^k = \frac{1}{1 - \|B\|}.$$

■

One can also deduce Proposition 4.1.5 from the following statement.

Proposition 4.1.6 *For every induced norm, one has*

$$\rho(A) \leq \|A\|.$$

Proof

The case $K = \mathbb{C}$ is easy, because there exists an eigenvector $X \in E$ associated to an eigenvalue of modulus $\rho(A)$:

$$\rho(A)\|X\| = \|\lambda X\| = \|AX\| \leq \|A\|\,\|X\|.$$

If $K = \mathbb{R}$, one needs a more involved trick.

Let us choose a norm on $\mathbb{C}^n$ and let us denote by N the induced norm on $\mathbf{M}_n(\mathbb{C})$. We still denote by N its restriction to $\mathbf{M}_n(\mathbb{R})$; it is a norm. Since this space has finite dimension, any two norms are equivalent: There exists $C > 0$ such that $N(B) \leq C\|B\|$ for every B in $\mathbf{M}_n(\mathbb{R})$. Using the result already proved in the complex case, one has for every $m \in \mathbb{N}$ that

$$\rho(A)^m = \rho(A^m) \leq N(A^m) \leq C\|A^m\| \leq C\|A\|^m.$$

Taking the mth root and letting m tend to infinity, and noticing that $C^{1/m}$ tends to 1, one obtains the announced inequality. ■

In general, the equality does not hold. For example, if A is nilpotent though nonzero, one has $\rho(A) = 0 < \|A\|$ for every matrix norm.

Proposition 4.1.7 *Let $\|\cdot\|$ be a norm on K^n and $P \in \mathbf{GL}_n(K)$. Hence, $N(x) := \|Px\|$ defines a norm on K^n. Denoting still by $\|\cdot\|$ and N the induced norms on K^n, one has $N(A) = \|PAP^{-1}\|$.*

Proof

Using the change of dummy variable $y = Px$, we have

$$N(A) = \sup_{x\neq 0} \frac{\|PAx\|}{\|Px\|} = \sup_{y\neq 0} \frac{\|PAP^{-1}y\|}{\|y\|} = \|PAP^{-1}\|.$$

■

4.2 Householder's Theorem

Householder's theorem is a kind of converse of the inequality $\rho(B) \leq \|B\|$.

Theorem 4.2.1 *For every $B \in \mathbf{M}_n(\mathbb{C})$ and all $\epsilon > 0$, there exists a norm on $\mathbb{C}^n$ such that for the induced norm*

$$\|B\| \leq \rho(B) + \epsilon.$$

In other words, $\rho(B)$ is the infimum of $\|B\|$, as $\|\cdot\|$ ranges over the set of matrix norms.

Proof

From Theorem 2.7.1 there exists $P \in \mathbf{GL}_n(\mathbb{C})$ such that $T := PBP^{-1}$ is upper triangular. From Proposition 4.1.7, one has

$$\inf \|B\| = \inf \|PBP^{-1}\| = \inf \|T\|,$$

where the infimum is taken over the set of induced norms. Since B and T have the same spectra, hence the same spectral radius, it is enough to prove the theorem for upper triangular matrices.

For such a matrix T, Proposition 4.1.7 still gives

$$\inf \|T\| \leq \inf\{\|QTQ^{-1}\|_2; Q \in \mathbf{GL}_n(\mathbb{C})\}.$$

Let us now take $Q(\mu) = \mathrm{diag}(1, \mu, \mu^2, \ldots, \mu^{n-1})$. The matrix $Q(\mu)TQ(\mu)^{-1}$ is upper triangular, with the same diagonal as that of T. Indeed, the entry with indices (i, j) becomes $\mu^{i-j}t_{ij}$. Hence,

$$\lim_{\mu \to \infty} Q(\mu)TQ(\mu)^{-1}$$

is simply the matrix $D = \mathrm{diag}(t_{11}, \ldots, t_{nn})$. Since $\|\cdot\|_2$ is continuous (as is every norm), one deduces

$$\inf \|T\| \leq \lim_{\mu \to \infty} \|Q(\mu)TQ(\mu)^{-1}\|_2 = \|D\|_2 = \sqrt{\rho(D^*D)} = \max |t_{jj}| = \rho(T).$$

■

Remark: The theorem tells us that $\rho(A) = \Lambda(A)$, where

$$\Lambda(A) := \inf \|A\|,$$

the infimum being taken over the set of matrix norms. The first part of the proof tells us that ρ and Λ coincide on the set of diagonalizable matrices, which is a dense subset of $\mathbf{M}_n(\mathbb{C})$. But this is insufficient to conclude, since Λ is a priori only upper semicontinuous, as the infimum of continuous functions. The continuity of Λ is actually a consequence of the theorem.

4.3 An Interpolation Inequality

Theorem 4.3.1 (case $K = \mathbb{C}$) *Let $\|\cdot\|_p$ be the norm on $\mathbf{M}_n(\mathbb{C})$ induced by the norm l^p on $\mathbb{C}^n$. The function*

$$\begin{array}{rcl} 1/p & \mapsto & \log \|A\|_p, \\ [0,1] & \to & I\!R, \end{array}$$

is convex. In other words, if $1/r = \theta/p + (1-\theta)/q$ *with* $\theta \in (0,1)$*, then*

$$\|A\|_r \leq \|A\|_p^{\theta} \|A\|_q^{1-\theta}.$$

Remark:

1. The proof uses the fact that $K = \mathbb{C}$. However, the norms induced by the $\|\cdot\|_p$'s on $\mathbf{M}_n(\mathbb{R})$ and $\mathbf{M}_n(\mathbb{C})$ take the same values on real matrices, even though their definitions are different (see Exercise 6). The statement is thus still true in $\mathbf{M}_n(\mathbb{R})$.

2. The case $(p,q,r) = (1,\infty,2)$ admits a direct proof. See the exercises.

3. The result still holds true in infinite dimension, at the expense of some functional analysis. One even can take different L^p norms at the source and target spaces. Here is an example:

   Theorem 4.3.2 (Riesz–Thorin) *Let* Ω *be an open set in* $\mathbb{R}^D$ *and* ω *an open set in* $\mathbb{R}^d$*. Let* p_0, p_1, q_0, q_1 *be four numbers in* $[1, +\infty]$*. Let* $\theta \in [0,1]$ *and* p, q *be defined by*

   $$\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}, \quad \frac{1}{q} = \frac{1-\theta}{q_0} + \frac{\theta}{q_1}.$$

   Consider a linear operator T *defined on* $L^{p_0} \cap L^{p_1}(\Omega)$*, taking values in* $L^{q_0} \cap L^{q_1}(\omega)$*. Assume that* T *can be extended as a continuous operator from* $L^{p_j}(\Omega)$ *to* $L^{q_j}(\omega)$*, with norm* M_j*,* $j = 1, 2$ *:*

   $$M_j := \sup_{f \neq 0} \frac{\|Tf\|_{q_j}}{\|f\|_{p_j}}.$$

   Then T *can be extended as a continuous operator from* $L^p(\Omega)$ *to* $L^q(\omega)$*, and its norm is bounded above by*

   $$M_0^{1-\theta} M_1^{\theta}.$$

4. A fundamental application is the continuity of the Fourier transform from $L^p(\mathbb{R}^d)$ into its dual $L^{p'}(\mathbb{R}^d)$ when $1 \leq p \leq 2$. We have only to observe that $(p_0, p_1, q_0, q_1) = (1, 2, +\infty, 2)$ is suitable. It can be proved by inspection that every pair (p,q) such that the Fourier transform is continuous from $L^p(\mathbb{R}^d)$ into $L^q(\mathbb{R}^d)$ has the form (p, p') with $1 \leq p \leq 2$.

5. One has analogous results for Fourier series. There lies the origin of Riesz–Thorin theorem.

Proof (due to F. Riesz)
Let us fix x and y in K^n. We have to bound

$$|(Ax, y)| = \left| \sum_{j,k} a_{jk} x_j \bar{y}_k \right|.$$

Let B be the strip in the complex plane defined by $\Re z \in [0,1]$. Given $z \in B$, define (conjugate) exponents $r(z)$ and $r'(z)$ by

$$\frac{1}{r(z)} = \frac{z}{p} + \frac{1-z}{q}, \quad \frac{1}{r'(z)} = \frac{z}{p'} + \frac{1-z}{q'}.$$

Set

$$\begin{aligned} X_j(z) &:= |x_j|^{-1+r/r(z)} x_j = x_j \exp\left(\left(\frac{r}{r(z)} - 1\right) \log |x_j|\right), \\ Y_j(z) &:= |y_j|^{-1+r'/r'(\bar{z})} y_j. \end{aligned}$$

We then have

$$\|X(z)\|_{r(\Re z)} = \|x\|_r^{r/r(\Re z)}, \quad \|Y(z)\|_{r'(\Re z)} = \|y\|_{r'}^{r'/r'(\Re z)}.$$

Next, define a holomorphic map in the strip B by $f(z) := (AX(z), Y(z))$. It is bounded, because the numbers $X_j(z)$ and $Y_k(z)$ are. For example,

$$|X_j(z)| = |x_j|^{r/r(\Re z)}$$

lies between $|x_j|^{r/p}$ and $|x_j|^{r/q}$.

Let us set $M(\theta) = \sup\{|f(z)|; \Re z = \theta\}$. Hadamard's *three lines lemma* (see [29], Chapter 12, exercise 8) expresses that

$$\theta \mapsto \log M(\theta)$$

is convex on $(0,1)$. However, $r(0) = q$, $r(1) = p$, $r'(0) = q'$, $r'(1) = p'$, $r(\theta) = r$, $r'(\theta) = r'$, $X(\theta) = x$, and $Y(\theta) = y$. Hence

$$|(Ax, y)| = |f(\theta)| \le M(\theta) \le M(1)^\theta M(0)^{1-\theta}.$$

Now we have

$$\begin{aligned} M(1) &= \sup\{|f(z)|; \Re z = 1\} \\ &\le \sup\{\|AX(z)\|_{r(1)} \|Y(z)\|_{r(1)'}; \Re z = 1\} \\ &= \sup\{\|AX(z)\|_p \|Y(z)\|_{p'}; \Re z = 1\} \\ &\le \|A\|_p \sup\{\|X(z)\|_p \|Y(z)\|_{p'}; \Re z = 1\} \\ &= \|A\|_p \|x\|_r^{r/p} \|y\|_{r'}^{r'/p'}. \end{aligned}$$

Likewise, $M(0) \le \|A\|_q \|x\|_r^{r/q} \|y\|_{r'}^{r'/q'}$. Hence

$$\begin{aligned} |(Ax, y)| &\le \|A\|_p^\theta \|A\|_q^{1-\theta} \|x\|_r^{r(\theta/p + (1-\theta)/q)} \|y\|_{r'}^{r'(\theta/p' + (1-\theta)/q')} \\ &= \|A\|_p^\theta \|A\|_q^{1-\theta} \|x\|_r \|y\|_{r'}. \end{aligned}$$

Finally,

$$\|Ax\|_r = \sup_{y \neq 0} \frac{|(Ax, y)|}{\|y\|_{r'}} \le \|A\|_p^\theta \|A\|_q^{1-\theta} \|x\|_r,$$

which proves the theorem. ■

4.4 A Lemma about Banach Algebras

Definition 4.4.1 *A* normed algebra *is a K-algebra endowed with a norm satisfying $\|xy\| \leq \|x\| \, \|y\|$. Such a norm is called an* algebra norm. *When a normed algebra is complete (which is always true in finite dimension), it is called a* Banach algebra.

Lemma 4.4.1 *Let $\mathcal{A}$ be a normed algebra and let $x \in \mathcal{A}$. The sequence $u_m := \|x^m\|^{1/m}$ converges to its infimum, denoted by $r(x)$. Additionally, if $K = \mathbb{C}$, and if $\mathcal{A}$ has a unit element and is complete, then $1/r(x)$ is the radius of the largest open ball $B(0;R)$ such that $e - zx$ is invertible for every $z \in B(0;R)$.*

Of course, one may apply the lemma to $\mathcal{A} = \mathbf{M}_n(\mathbb{C})$ endowed with a matrix norm. One then has $r(x) = \rho(x)$, because $e - zx = I - zA$ is invertible, provided that z is not the inverse of an eigenvalue. In the case $K = \mathbb{R}$, one uses an auxiliary norm N that is the restriction to $\mathbf{M}_n(\mathbb{R})$ of an induced norm on $\mathbf{M}_n(\mathbb{C})$. Since $\|\cdot\|$ and N are equivalent, one simply writes

$$\rho(A) = \rho(A^m)^{1/m} \leq \|A^m\|^{1/m} \leq C^{1/m} N(A^m)^{1/m}.$$

The latter sequence converges to $\rho(A)$ from the lemma, which implies the convergence of the former. We thus have the following result.

Proposition 4.4.1 *If $A \in \mathbf{M}_n(K)$, then*

$$\rho(A) = \lim_{m \to \infty} \|A^m\|^{1/m}$$

for every matrix norm.

Proof

Convergence. The result is trivial if $x^m = 0$ for some exponent. In the opposite case, we use the following inequalities, which come directly from the definition:

$$\|x^{ap+r}\| \leq \|x^p\|^a \|x^r\|, \quad \forall a, p, r \in \mathbb{N}.$$

We then define

$$v_m = \frac{1}{m} \log \|x^m\| = \log u_m.$$

Let us fix an integer p and perform Euclidean division of m by p: $m = ap + r$ with $0 \leq r \leq p - 1$. This yields

$$v_{ap+r} \leq \frac{apv_p + rv_r}{ap + r}.$$

As m, hence a, tends to infinity, the right-hand side converges, because rv_r remains bounded:

$$\limsup v_m \leq v_p.$$

Since this holds true for every p, we conclude that

$$\limsup v_m \leq \inf v_p \leq \liminf v_p,$$

which proves the convergence to the infimum.

Characterization (complex case). If $R < 1/r(x)$, the Taylor series

$$\sum_{m \in \mathbb{N}} z^m x^m, \quad z \in \mathbb{C},$$

converges in norm in the ball $B(0;R)$. Its sum equals $(e-zx)^{-1}$ (see the proof of Proposition 4.1.5).
The domain of the map $z \mapsto (e-zx)^{-1}$ is open, since if it contains a point z_0, the previous paragraph shows that $e-(z-z_0)(e-z_0x)^{-1}x$ is invertible for every z satisfying

$$|z-z_0| r\left((e-z_0x)^{-1}x\right) < 1.$$

Denoting by X_z the inverse, we see that $X_z(e-z_0x)^{-1}$ is an inverse of $e-zx$. In particular, $f : z \mapsto (e-z)^{-1}$ is holomorphic.
If f is defined on a ball $B(0;s)$, Cauchy's formula

$$x^m = \frac{1}{m!} f^{(m)}(0) = \frac{1}{2i\pi} \int_{B(0;s)} \frac{f(z)}{z^{m+1}} dz$$

shows that $\|x^m\| = \mathcal{O}(s^{-m})$. Hence, $1/r(x) \geq s$.

■

Corollary 4.4.1 *Let $B \in \mathbf{M}_n(K)$ be given. Then $B^m \overset{m\to+\infty}{\to} 0$ if and only if $\rho(B) < 1$.*

Indeed, $\rho(B) \geq 1$ implies $\|B^m\| \geq \rho(B^m) \geq 1$ for every m. Conversely, $\rho(B) < 1$ implies $\|B^m\| < r^m$ for m large enough, where r is selected in $(\rho(B), 1)$.

■

We observe that this result is also a consequence of Householder's theorem.

4.5 The Gershgorin Domain

Let $A \in \mathbf{M}_n(\mathbb{C})$, and let λ be an eigenvalue and x an associated eigenvector. Let i be an index such that $|x_i| = \|x\|_\infty$. Then $x_i \neq 0$ and

$$|a_{ii} - \lambda| = \left| \sum_{j \neq i} a_{ij} \frac{x_j}{x_i} \right| \leq \sum_{j \neq i} |a_{ij}|.$$

Proposition 4.5.1 (Gershgorin) *The spectrum of A is included in the Gershgorin domain $\mathcal{G}(A)$, defined as the union of the* Gershgorin disks *$D_i := D(a_{ii}; \sum_{j \neq i} |a_{ij}|)$.*

This result can also be deduced from Proposition 4.1.5: Let us decompose $A = D + C$, where D is the diagonal part of A. If $\lambda \neq a_{ii}$ for every i, then $\lambda I_n - A = (\lambda I_n - D)(I_n - B)$ with $B = (\lambda I_n - D)^{-1}C$. Hence, if λ is an eigenvalue, then either λ is an a_{ii}, or $\|B\|_\infty \geq 1$.

One may improve this result by considering the connected components of $\mathcal{G}$. Let G be one of them. It is the union of the D_k's that meet it. Let p be the number of such disks. One then has $G = \cup_{i \in I} D_i$ where I has cardinality p.

Theorem 4.5.1 *There are exactly p eigenvalues of A in G, counted with their multiplicities.*

Proof

For $r \in [0, 1]$, we define a matrix $A(r)$ by the formula

$$a_{ij}(r) := \begin{cases} a_{ii}, & j = i, \\ r a_{ij}, & j \neq i. \end{cases}$$

It is clear that the Gershgorin domain $\mathcal{G}_r$ of $A(r)$ is included in $\mathcal{G}$. We observe that $A(1) = A$, and that $r \mapsto A(r)$ is continuous. Let us denote by $m(r)$ the number of eigenvalues (counted with multiplicity) of $A(r)$ that belong to G.

Since G and $\mathcal{G} \setminus G$ are compact, one can find a Jordan curve, oriented in the trigonometric sense, that separates G from $\mathcal{G} \setminus G$. Let Γ be such a curve. Since $\mathcal{G}_r$ is included in $\mathcal{G}$, the residue formula expresses $m(r)$ in terms of the characteristic polynomial of $A(r)$:

$$m(r) = \frac{1}{2i\pi} \int_\Gamma \frac{P_r'(z)}{P_r(z)}\, dz.$$

Since P_r does not vanish on Γ and $r \mapsto P_r, P_r'$ are continuous, we deduce that $r \mapsto m(r)$ is continuous. Since $m(r)$ is an integer and $[0, 1]$ is connected, $m(r)$ remains constant. In particular, $m(0) = m(1)$.

Finally, $m(0)$ is the number of entries a_{jj} (eigenvalues of $A(0)$) that belong to G. But a_{jj} is in G if and only if $D_j \subset G$. Hence $m(0) = p$, which implies $m(1) = p$, the desired result. ■

An improvement of Gershgorin's theorem concerns irreducible matrices.

Proposition 4.5.2 *Let A be an irreducible matrix. If an eigenvalue of A does not belong to the interior of any Gershgorin disk, then it belongs to all the circles $S(a_{ii}; \sum_{j \neq i} |a_{ij}|)$.*

Proof

Let λ be such an eigenvalue and x an associated eigenvector. By assumption, one has $|\lambda - a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$ for every i. Let I be the set of indices

for which $|x_i| = \|x\|_\infty$ and let J be its complement. If $i \in I$, then

$$\|x\|_\infty \sum_{j\neq i} |a_{ij}| \leq |\lambda - a_{ii}|\, \|x\|_\infty = \left|\sum_{j\neq i} a_{ij}x_j\right| \leq \sum_{j\neq i} |a_{ij}|\, |x_j|.$$

It follows that $\sum_{j\neq i}(\|x\|_\infty - |x_j|)|a_{ij}| \leq 0$, where all the terms in the sum are nonnegative. Each term is thus zero, so that $a_{ij} = 0$ for $j \in J$. Since A is irreducible, J is empty. One has thus $|x_j| = \|x\|_\infty$ for every j, and the previous inequalities show that λ belongs to every circle.

■

Definition 4.5.1 *A square matrix $A \in \mathbf{M}_n(\mathbb{C})$ is said to be*

1. diagonally dominant *if*

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|, \quad 1 \leq i \leq n;$$

2. strongly diagonally dominant *if in addition at least one of these n inequalities is strict;*

3. strictly diagonally dominant *if the inequality is strict for every index i.*

Corollary 4.5.1 *Let A be a square matrix. If A is strictly diagonally dominant, or if A is irreducible and strongly diagonally dominant, then A is invertible.*

In fact, either zero does not belong to the Gershgorin domain, or it is not interior to the disks. In the latter case, A is assumed to be irreducible, and there exists a disk D_j that does not contain zero.

4.6 Exercises

1. Under what conditions on the vectors a, $b \in \mathbb{C}^n$ does the matrix M defined by $m_{ij} = a_i b_j$ satisfy $\|M\|_p = 1$ for every $p \in [1, \infty]$?

2. Under what conditions on x, y, and p does the equality in (4.2) or (4.1) hold?

3. Show that

$$\lim_{p\to+\infty} \|x\|_p = \|x\|_\infty, \quad \forall x \in E.$$

4. A norm on K^n is a *strictly convex* norm if $\|x\| = \|y\| = 1$, $x \neq y$, and $0 < \theta < 1$ imply $\|\theta x + (1-\theta)y\| < 1$.

   (a) Show that $\|\cdot\|_p$ is strictly convex for $1 < p < \infty$, but is not so for $p = 1, \infty$.

(b) Deduce from Corollary 5.5.1 that the induced norm $\|\cdot\|_p$ is not strictly convex on $\mathbf{M}_n(\mathbb{R})$.

5. Let N be a norm on $\mathbb{R}^n$.

(a) For $x \in \mathbb{C}^n$, define

$$N_1(x) := \inf \left\{ \sum_l |\alpha_l| N(x^l) \right\},$$

where the infimum is taken over the set of decompositions $x = \sum_l \alpha_l x^l$ with $\alpha_l \in \mathbb{C}$ and $x^l \in \mathbb{R}^n$. Show that N_1 is a norm on $\mathbb{C}^n$ (as a $\mathbb{C}$-vector space) whose restriction to $\mathbb{R}^n$ is N. **Note:** N_1 is called the *complexification* of N.

(b) Same question as above for N_2, defined by

$$N_2(x) := \frac{1}{2\pi} \int_0^{2\pi} [e^{i\theta} x] d\theta,$$

where

$$[x] := \sqrt{N(\Re x)^2 + N(\Im x)^2}.$$

(c) Show that $N_2 \leq N_1$.

(d) If $N(x) = \|x\|_1$, show that $N_1(x) = \|x\|_1$. Considering then the vector

$$x = \begin{pmatrix} 1 \\ i \end{pmatrix},$$

show that $N_2 \neq N_1$.

6. (continuation of exercise 5)
The norms N (on $\mathbb{R}^n$) and N_1 (on $\mathbb{C}^n$) lead to induced norms on $\mathbf{M}_n(\mathbb{R})$ and $\mathbf{M}_n(\mathbb{C})$, respectively. Show that if $M \in \mathbf{M}_n(\mathbb{R})$, then $N(M) = N_1(M)$. Deduce that Theorem 4.3.1 holds true in $\mathbf{M}_n(\mathbb{R})$.

7. Let $\|\cdot\|$ be an algebra norm on $\mathbf{M}_n(K)$ ($K = \mathbb{R}$ or $\mathbb{C}$), that is, a norm satisfying $\|AB\| \leq \|A\| \cdot \|B\|$. Show that $\rho(A) \leq \|A\|$ for every $A \in \mathbf{M}_n(K)$.

8. In $\mathbf{M}_n(\mathbb{C})$, let D be a diagonalizable matrix and N a nilpotent matrix that commutes with D. Show that $\rho(D) = \rho(D + N)$.

9. Let $B \in \mathbf{M}_n(\mathbb{C})$ be given. Assume that there exists an induced norm such that $\|B\| = \rho(B)$. Let λ be an eigenvalue of maximal modulus and X a corresponding eigenvector. Show that X does not belong to the range of $B - \lambda I_n$. Deduce that the Jordan block associated to λ is diagonal (Jordan reduction is presented in Chapter 6).

10. (continuation of exercise 9)

Conversely, show that if the Jordan blocks of B associated to the eigenvalues of maximal modulus of B are diagonal, then there exists a norm on $\mathbb{C}^n$ such that, using the induced norm, $\rho(B) = \|B\|$.

11. Here is another proof of Theorem 4.2.1. Let $K = \mathbb{R}$ or $\mathbb{C}$, $A \in \mathbf{M}_n(K)$, and let N be a norm on K^n. If $\epsilon > 0$, we define for all $x \in K^n$

$$\|x\| := \sum_{k \in \mathbb{N}} (\rho(A) + \epsilon)^{-k} N(A^k x).$$

(a) Show that this series is convergent (use Corollary 4.4.1).
(b) Show that $\|\cdot\|$ is a norm on K^n.
(c) Show that for the induced norm, $\|A\| \leq \rho(A) + \epsilon$.

12. A matrix norm $\|\cdot\|$ on $\mathbf{M}_n(\mathbb{C})$ is said to be *unitarily invariant* if $\|UAV\| = \|A\|$ for every $A \in \mathbf{M}_n(\mathbb{C})$ and all unitary matrices U, V.

(a) Find, among the most classical norms, two examples of unitarily invariant norms.
(b) Given a unitarily invariant norm, show that there exists a norm N on $\mathbb{R}^n$ such that

$$\|A\| = N(s_1(A), \ldots, s_n(A)),$$

where the $s_j(A)$'s, the eigenvalues of H in the polar decomposition $A = QH$ (see Chapter 7 for this notion), are called the *singular values* of A.

13. (R. Bhatia [5]) Suppose we are given a norm $\|\cdot\|$ on $\mathbf{M}_n(\mathbb{C})$ that is unitarily invariant (see the previous exercise). If $A \in \mathbf{M}_n(\mathbb{C})$, we denote by $D(A)$ the diagonal matrix obtained by keeping only the a_{jj} and setting all the other entries to zero. If σ is a permutation, we denote by A^σ the matrix whose entry of index (j,k) equals a_{jk} if $k = \sigma(j)$, and zero otherwise. For example, $A^{id} = D(A)$, where id is the identity permutation. If r is an integer between $1-n$ and $n-1$, we denote by $D_r(A)$ the matrix whose entry of index (j,k) equals a_{jk} if $k - j = r$, and zero otherwise. For example, $D_0(A) = D(A)$.

(a) Let $\omega = \exp(2i\pi/n)$ and let U be the diagonal matrix whose diagonal entries are the roots of unity $1, \omega, \ldots, \omega^{n-1}$. Show that

$$D(A) = \frac{1}{n} \sum_{j=0}^{n-1} U^{*j} A U^j.$$

Deduce that $\|D(A)\| \leq \|A\|$.
(b) Show that $\|A^\sigma\| \leq \|A\|$ for every $\sigma \in \mathcal{S}_n$. Observe that $\|P\| = \|I_n\|$ for every permutation matrix P. Show that $\|M\| \leq \|I_n\|$ for every bistochastic matrix M (see Section 5.5 for this notion).

(c) If $\theta \in \mathbb{R}$, let us denote by U_θ the diagonal matrix, whose kth diagonal term equals $\exp(ik\theta)$. Show that

$$D_r(A) = \frac{1}{2\pi} \int_0^{2\pi} e^{ir\theta} U_\theta A U_\theta^* d\theta.$$

(d) Deduce that $\|D_r(A)\| \leq \|A\|$.

(e) Let p be an integer between zero and $n-1$ and $r = 2p+1$. Let us denote by $T_r(A)$ the matrix whose entry of index (j,k) equals a_{jk} if $|k-j| \leq p$, and zero otherwise. For example, $T_3(A)$ is a tridiagonal matrix. Show that

$$T_r(A) = \frac{1}{2\pi} \int_0^{2\pi} d_p(\theta) U_\theta A U_\theta^* d\theta,$$

where

$$d_p(\theta) = \sum_{-p}^{p} e^{ik\theta}$$

is the *Dirichlet kernel*.

(f) Deduce that $\|T_r(A)\| \leq L_p \|A\|$, where

$$L_p = \frac{1}{2\pi} \int_0^{2\pi} |d_p(\theta)| d\theta$$

is the *Lebesgue constant* (note: $L_p = 4\pi^{-2} \log p + \mathcal{O}(1)$).

(g) Let $\Delta(A)$ be the upper triangular matrix whose entries above the diagonal coincide with those of A. Using the matrix

$$B = \begin{pmatrix} 0 & \Delta(A)^* \\ \Delta(A) & 0 \end{pmatrix},$$

show that $\|\Delta(A)\|_2 \leq L_n \|A\|_2$ (observe that $\|B\|_2 = \|\Delta(A)\|_2$).

(h) What inequality do we obtain for $\Delta_0(A)$, the strictly upper triangular matrix whose entries lying strictly above the diagonal coincide with those of A?

14. We endow $\mathbb{C}^n$ with the usual Hermitian structure, so that $\mathbf{M}_n(\mathbb{C})$ is equipped with the norm $\|A\| = \rho(A^*A)^{1/2}$.

Suppose we are given a sequence of matrices $(A_j)_{j \in \mathbb{Z}}$ in $\mathbf{M}_n(\mathbb{C})$ and a summable sequence $\gamma \in l^1(\mathbb{Z})$ of positive real numbers. Assume, finally, that for every pair $(j,k) \in \mathbb{Z} \times \mathbb{Z}$,

$$\|A_j^* A_k\| \leq \gamma(j-k)^2, \quad \|A_j A_k^*\| \leq \gamma(j-k)^2.$$

(a) Let F be a finite subset of $\mathbb{Z}$. Let B_F denote the sum of the A_j's as j runs over F. Show that

$$\|(B_F^* B_F)^{2m}\| \leq \text{ card } F \, \|\gamma\|_1^{2m}, \quad \forall m \in \mathbb{N}.$$

(b) Deduce that $\|B_F\| \leq \|\gamma\|_1$.

(c) Show (*Cotlar's lemma*) that for every $x, y \in \mathbb{C}^n$, the series

$$y^T \sum_{j \in \mathbb{Z}} A_j x$$

is convergent, and that its sum $y^T A x$ defines a matrix $A \in \mathbf{M}_n(\mathbb{C})$ that satisfies

$$\|A\| \leq \sum_{j \in \mathbb{Z}} \gamma(j).$$

Hint: For a sequence $(u_j)_{j \in \mathbb{Z}}$ of real numbers, the series $\sum_j u_j$ is absolutely convergent if and only if there exists $M < +\infty$ such that $\sum_{j \in F} |u_j| \leq M$ for every finite subset F.

(d) Deduce that the series $\sum_j A_j$ converges in $\mathbf{M}_n(\mathbb{C})$. May one conclude that it converges normally?

15. Let $\|\cdot\|$ be an induced norm on $\mathbf{M}_n(\mathbb{R})$. We wish to characterize the matrices $B \in \mathbf{M}_n(\mathbb{R})$ such that there exist $\epsilon_0 > 0$ and $\omega > 0$ with

$$(0 < \epsilon < \epsilon_0) \Longrightarrow (\|I_n - \epsilon B\| \leq 1 - \omega\epsilon).$$

(a) For the norm $\|\cdot\|_\infty$, it is equivalent that B be strictly diagonally dominant.

(b) What is the characterization for the norm $\|\cdot\|_1$?

(c) For the norm $\|\cdot\|_2$, it is equivalent that $B^T + B$ be positive definite.

16. If $A \in \mathbf{M}_n(\mathbb{C})$ and $j = 1, \ldots, n$ are given, we define $r_j(A) := \sum_{k \neq j} |a_{jk}|$. For $i \neq j$, define

$$\mathcal{B}_{ij}(A) = \{z \in \mathbb{C}\,;\, |(z - a_{ii})(z - a_{jj})| \leq r_i(A) r_j(A)\}.$$

These sets are *Cassini ovals*. Finally, let

$$\mathcal{B}(A) := \cup_{1 \leq i < j \leq n} \mathcal{B}_{ij}(A).$$

(a) Show that $\operatorname{Sp} A \subset \mathcal{B}(A)$.

(b) Show that this result is sharper than Proposition 4.5.1.

(c) When $n = 2$, show that in fact $\operatorname{Sp} A$ is included in the boundary of $\mathcal{B}(A)$.

17. Let $B \in \mathbf{M}_n(\mathbb{C})$.

(a) Returning to the proof of Theorem 4.2.1, show that for every $\epsilon > 0$ there exists on $\mathbb{C}^n$ a Hermitian norm $\|\cdot\|$ such that for the induced norm $\|B\| \leq \rho(B) + \epsilon$.

(b) Deduce that $\rho(B) < 1$ holds if and only if there exists a matrix $A \in \mathbf{HPD}_n$ such that $A - B^* A B \in \mathbf{HPD}_n$.

18. For $A \in \mathbf{M}_n(\mathbb{C})$, define

$$\epsilon := \max_{i \neq j} |a_{ij}|, \quad \delta := \min_{i \neq j} |a_{ii} - a_{jj}|.$$

We assume in this exercise that $\delta > 0$ and $\epsilon \leq \delta/4n$.

(a) Show that each Gershgorin disk D_j contains exactly one eigenvalue of A.
(b) Let $\rho > 0$ be a real number. Show that A^ρ, obtained by multiplying the ith row of A by ρ and the ith column by $1/\rho$, has the same eigenvalues as A.
(c) Choose $\rho = 2\epsilon/\delta$. Show that the ith Gershgorin disk of A^ρ contains exactly one eigenvalue. Deduce that the eigenvalues of A are simple and that

$$d(\mathrm{Sp}(A), \mathrm{diag}(A)) \leq \frac{2n\epsilon^2}{\delta},$$

where $\mathrm{diag}(A) = \{a_{11}, \ldots, a_{nn}\}$.

19. Let $A \in \mathbf{M}_n(\mathbb{C})$ be a diagonalizable matrix:

$$A = S\,\mathrm{diag}(d_1, \ldots, d_n)S^{-1}.$$

Let $\|\cdot\|$ be an induced norm for which $\|D\| = \max_j |d_j|$ holds, where $D := \mathrm{diag}(d_1, \ldots, d_n)$. Show that for every $E \in \mathbf{M}_n(\mathbb{C})$ and for every eigenvalue λ of $A + E$, there exists an index j such that

$$|\lambda - d_j| \leq \|S\| \cdot \|S^{-1}\| \cdot \|E\|.$$

20. Let $A \in \mathbf{M}_n(K)$, with $K = \mathbb{R}$ or $\mathbb{C}$. Give another proof, using the Cauchy–Schwarz inequality, of the following particular case of Theorem 4.3.1:

$$\|A\|_2 \leq \|A\|_1^{1/2} \|A\|_\infty^{1/2}.$$

21. Show that if $A \in \mathbf{M}_n(\mathbb{C})$ is normal, then $\rho(A) = \|A\|_2$. Deduce that if A and B are normal, $\rho(AB) \leq \rho(A)\rho(B)$.

22. Let N_1 and N_2 be two norms on $\mathbb{C}^n$. Denote by $\mathcal{N}_1$ and $\mathcal{N}_2$ the induced norms on $\mathbf{M}_n(\mathbb{C})$. Let us define

$$R := \max_{x \neq 0} \frac{N_1(x)}{N_2(x)}, \quad S := \max_{x \neq 0} \frac{N_2(x)}{N_1(x)}.$$

(a) Show that

$$\max_{A \neq 0} \frac{\mathcal{N}_1(A)}{\mathcal{N}_2(A)} = RS = \max_{A \neq 0} \frac{\mathcal{N}_2(A)}{\mathcal{N}_1(A)}.$$

(b) Deduce that if $\mathcal{N}_1 = \mathcal{N}_2$, then N_2/N_1 is constant.
(c) Show that if $\mathcal{N}_1 \leq \mathcal{N}_2$, then N_2/N_1 is constant and therefore $\mathcal{N}_2 = \mathcal{N}_1$.

23. (continuation of exercise 22)
Let $\|\cdot\|$ be an algebra norm on $\mathbf{M}_n(\mathbb{C})$. If $y \in \mathbb{C}^n$ is nonzero, we define $\|x\|_y := \|xy^*\|$.

(a) Show that $\|\cdot\|_y$ is a norm on $\mathbb{C}^n$ for every $y \neq 0$.
(b) Let $\mathcal{N}_y$ be the norm induced by $\|\cdot\|_y$. Show that $\mathcal{N}_y \leq \|\cdot\|$.
(c) We say that $\|\cdot\|$ is *minimal* if there exists no other algebra norm less than or equal to $\|\cdot\|$. Show that the following assertions are equivalent:
   i. $\|\cdot\|$ is an induced norm on $\mathbf{M}_n(\mathbb{C})$.
   ii. $\|\cdot\|$ is a minimal norm on $\mathbf{M}_n(\mathbb{C})$.
   iii. For all $y \neq 0$, one has $\|\cdot\| = \mathcal{N}_y$.

24. (continuation of exercise 23)
Let $\|\cdot\|$ be an induced norm on $\mathbf{M}_n(\mathbb{C})$.

(a) Let $y, z \neq 0$ be two vectors in $\mathbb{C}^n$. Show that (with the notation of the previous exercise) $\|\cdot\|_y / \|\cdot\|_z$ is constant.
(b) Prove the equality

$$\|xy^*\| \cdot \|zt^*\| = \|xt^*\| \cdot \|zy^*\|.$$

25. Let $M \in \mathbf{M}_n(\mathbb{C})$ and $H \in \mathbf{HPD}_n$ be given. Show that

$$\|HMH\|_2 \leq \frac{1}{2}\|H^2M + MH^2\|_2.$$

26. We endow $\mathbb{R}^2$ with the Euclidean norm $\|\cdot\|_2$, and $\mathbf{M}_2(\mathbb{R})$ with the induced norm, denoted also by $\|\cdot\|_2$. We denote by Σ the unit sphere of $\mathbf{M}_2(\mathbb{R})$: $M \in \Sigma$ is equivalent to $\|M\|_2 = 1$, that is, to $\rho(M^T M) = 1$. Similarly, B denotes the unit ball of $\mathbf{M}_2(\mathbb{R})$.
Recall that if C is a convex set and if $P \in C$, then P is called an *extremal* point if $P \in [Q, R]$ and $Q, R \in C$ imply $Q = R = P$.

(a) Show that the set of extremal points of B is equal to $\mathbf{O}_2(\mathbb{R})$.
(b) Show that $M \in \Sigma$ if and only if there exist two matrices $P, Q \in \mathbf{O}_2(\mathbb{R})$ and a number $a \in [0, 1]$ such that

$$M = P \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} Q.$$

(c) We denote by $\mathcal{R} = \mathbf{SO}_2(\mathbb{R})$ the set of rotation matrices, and by $\mathcal{S}$ that of matrices of planar symmetry. Recall that $\mathbf{O}_2(\mathbb{R})$ is the disjoint union of $\mathcal{R}$ and $\mathcal{S}$. Show that Σ is the union of the segments $[r, s]$ as r runs over $\mathcal{R}$ and s runs over $\mathcal{S}$.
(d) Show that two such "open" segments (r, s) and (r', s') are either disjoint or equal.
(e) Let $M, N \in \Sigma$. Show that $\|M - N\|_2 = 2$ (that is, (M, N) is a diameter of B) if and only if there exists a segment $[r, s]$ ($r \in \mathcal{R}$ and $s \in \mathcal{S}$) such that $M \in [r, s]$ and $N \in [-r, -s]$.

5

Nonnegative Matrices

In this chapter matrices have real entries in general. In a few specified cases, entries might be complex.

5.1 Nonnegative Vectors and Matrices

Definition 5.1.1 *A vector $x \in \mathbb{R}^n$ is* nonnegative, *and we write $x \geq 0$, if its coordinates are nonnegative. It is* positive, *and we write $x > 0$, if its coordinates are (strictly) positive. Furthermore, a matrix $A \in M_{n\times m}(\mathbb{R})$ (not necessarily square) is* nonnegative *(respectively* positive*) if its entries are nonnegative (respectively positive); we again write $A \geq 0$ (respectively $A > 0$). More generally, we define an order relationship $x \leq y$ whose meaning is $y - x \geq 0$.*

Definition 5.1.2 *Given $x \in \mathbb{C}^n$, we let $|x|$ denote the nonnegative vector whose coordinates are the numbers $|x_j|$. Similarly, if $A \in \mathbf{M}_n(\mathbb{C})$, the matrix $|A|$ has entries $|a_{ij}|$.*

Observe that given a matrix and a vector (or two matrices), the triangle inequality implies

$$|Ax| \leq |A| \cdot |x|.$$

Proposition 5.1.1 *A matrix is nonnegative if and only if $x \geq 0$ implies $Ax \geq 0$. It is positive if and only if $x \geq 0$ and $x \neq 0$ imply $Ax > 0$.*

Proof

Let us assume that $Ax \geq 0$ (respectively > 0) for every $x \geq 0$ (respectively ≥ 0 and $\neq 0$). Then the ith column $A^{(i)}$ is nonnegative (respectively positive), since it is the image of the ith vector of the canonical basis. Hence $A \geq 0$ (respectively > 0).

Conversely, $A \geq 0$ and $x \geq 0$ imply trivially $Ax \geq 0$. If $A > 0$, $x \geq 0$, and $x \neq 0$, there exists an index l such that $x_l > 0$. Then

$$(Ax)_i = \sum_j a_{ij}x_j \geq a_{il}x_l > 0,$$

and hence $Ax > 0$.

■

An important point is the following:

Proposition 5.1.2 *If $A \in \mathbf{M}_n(\mathbb{R})$ is nonnegative and irreducible, then $(I + A)^{n-1} > 0$.*

Proof

Let x be a nonnegative, nonzero vector and define $x^m = (I + A)^m x$, which is nonnegative. Let us denote by P_m the set of indices of the nonzero components of x^m: P_0 is nonempty. Since $x_i^{m+1} \geq x_i^m$, one has $P_m \subset P_{m+1}$. Let us assume that the cardinality $|P_m|$ of P_m is strictly less than n. There are thus one or more zero components, whose indices form a nonempty subset I, complement of P_m. Since A is irreducible, there exists some nonzero entry a_{ij}, with $i \in I$ and $j \in P_m$. Then $x_i^{m+1} \geq a_{ij}x_j^m > 0$, which shows that P_{m+1} is not equal to P_m, and thus $|P_{m+1}| > |P_m|$. By induction, we deduce that $|P_m| \geq \min\{m + 1, n\}$. Hence $|P_{n-1}| = n$.

■

5.2 The Perron–Frobenius Theorem: Weak Form

Theorem 5.2.1 *Let $A \in \mathbf{M}_n(\mathbb{R})$ be a nonnegative matrix. Then $\rho(A)$ is an eigenvalue of A associated to a nonnegative eigenvector.*

Proof

Let λ be an eigenvalue of maximal modulus and v an eigenvector, normalized by $\|v\|_1 = 1$. Then

$$\rho(A)|v| = |\lambda v| = |Av| \leq A|v|.$$

Let us denote by C the subset of $\mathbb{R}^n$ (actually a subset of the unit simplex K_n) defined by the (in)equalities $\sum_i x_i = 1$, $x \geq 0$, and $Ax \geq \rho(A)x$. This is a closed convex set, nonempty, since it contains $|v|$. Finally, it is bounded, because $x \in C$ implies $0 \leq x_j \leq 1$ for every j; thus it is compact. Let us distinguish two cases:

1. There exists $x \in C$ such that $Ax = 0$. Then $\rho(A)x \leq 0$ furnishes $\rho(A) = 0$. The theorem is thus proved in this case.

2. For every x in C, $Ax \neq 0$. Then let us define on C a continuous map f by

$$f(x) = \frac{1}{\|Ax\|_1} Ax.$$

It is clear that $f(x) \geq 0$ and that $\|f(x)\|_1 = 1$. Finally,

$$Af(x) = \frac{1}{\|Ax\|_1} AAx \geq \frac{1}{\|Ax\|_1} A\rho(A)x = \rho(A)f(x),$$

so that $f(C) \subset C$. Then Brouwer's theorem (see [3], p. 217) asserts that a continuous function from a compact convex subset of $\mathbb{R}^N$ into itself has a fixed point. Thus let y be a fixed point of f. It is a nonnegative eigenvector, associated to the eigenvalue $r = \|Ay\|_1$. Since $y \in C$, we have $ry = Ay \geq \rho(A)y$ and thus $r \geq \rho(A)$, which implies $r = \rho(A)$.

■

That proof can be adapted to the case where a real number r and a nonzero vector y are given satisfying $y \geq 0$ and $Ay \geq ry$. Just take for C the set of vectors x such that $\sum_i x_i = 1$, $x \geq 0$, and $Ax \geq rx$. We then conclude that $\rho(A) \geq r$.

5.3 The Perron–Frobenius Theorem: Strong Form

Theorem 5.3.1 *Let $A \in \mathbf{M}_n(\mathbb{R})$ be a nonnegative irreducible matrix. Then $\rho(A)$ is a simple eigenvalue of A, associated to a positive eigenvector. Moreover, $\rho(A) > 0$.*

5.3.1 Remarks

1. Though the Perron–Frobenius theorem says that $\rho(A)$ is a simple eigenvalue, it does not tell anything about the other eigenvalues of maximal modulus. The following example shows that such other eigenvalues may exist:

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

The existence of several eigenvalues of maximal modulus will be studied in Section 5.4.

2. One obtains another proof of the weak form of the Perron–Frobenius theorem by applying the strong form to $A + \alpha J$, where $J > 0$ and $\alpha > 0$, then letting α tend to zero.

3. Without the irreducibility assumption, $\rho(A)$ may be a multiple eigenvalue, and a nonnegative eigenvector may not be positive. This holds for a matrix of size $n = 2m$ that reads blockwise

$$A = \begin{pmatrix} B & 0_m \\ I_m & B \end{pmatrix}.$$

Here, $\rho(A) = \rho(B)$, and every eigenvalue has an even algebraic multiplicity. Moreover, if $\rho(B)$ is a simple eigenvalue of B, associated to the eigenvector $Z \geq 0$, then the kernel of $A - \rho(A)I_n$ is spanned by

$$X = \begin{pmatrix} 0_m \\ Z \end{pmatrix},$$

which is not positive.

Proof

For $r \geq 0$, we denote by C_r the set of vectors of $\mathbb{R}^n$ defined by the (in)equalities

$$x \geq 0, \quad \|x\|_1 = 1, \quad Ax \geq rx.$$

Each C_r is a convex compact set. We saw in the previous section that if λ is an eigenvalue associated to an eigenvector x of unit norm $\|x\|_1 = 1$, then $|x| \in C_{|\lambda|}$. In particular, $C_{\rho(A)}$ is nonempty. Conversely, if C_r is nonempty, then for $x \in C_r$,

$$r = r\|x\|_1 \leq \|Ax\|_1 \leq \|A\|_1\|x\|_1 = \|A\|_1,$$

and therefore $r \leq \|A\|_1$. Furthermore, the map $r \mapsto C_r$ is nonincreasing with respect to inclusion, and is "left continuous" in the following sense. If $r > 0$, one has

$$C_r = \cap_{s<r} C_s.$$

Let us then define

$$R = \sup\{r \,|\, C_r \neq \emptyset\},$$

so that $R \in [\rho(A), \|A\|_1]$. The monotonicity with respect to inclusion shows that $r < R$ implies $C_r \neq \emptyset$.

If $x > 0$ and $\|x\|_1 = 1$, then $Ax \geq 0$ and $Ax \neq 0$, since A is nonnegative and irreducible. From Lemma 5.3.1 it follows that $R > 0$. The set C_R, being the intersection of a totally ordered family of nonempty compacts sets, is nonempty.

Let $x \in C_R$. Lemma 5.3.1 below shows that x is an eigenvector of A associated to the eigenvalue R. We observe that this eigenvalue is not less than $\rho(A)$ and infer that $\rho(A) = R$. Hence $\rho(A)$ is an eigenvalue associated to the eigenvector x, and $\rho(A) > 0$. Lemma 5.3.2 below ensures that $x > 0$.

The proof of the simplicity of the eigenvalue $\rho(A)$ will be given in Section 5.3.3.

5.3.2 A Few Lemmas

Lemma 5.3.1 *Let $r \geq 0$ and $x \geq 0$ such that $Ax \geq rx$ and $Ax \neq rx$. Then there exists $r' > r$ such that $C_{r'}$ is nonempty.*

Proof

Let $y := (I_n + A)^{n-1}x$. Since A is irreducible and $x \geq 0$ is nonzero, one has $y > 0$. Similarly, $Ay - ry = (I_n + A)^{n-1}(Ax - rx) > 0$. Let us define $r' := \min_j (Ay)_j / y_j$, which is strictly larger than r. We then have $Ay \geq r'y$, so that $C_{r'}$ contains the vector $y/\|y\|_1$.

■

Lemma 5.3.2 *The nonnegative eigenvectors of A are positive.*

Proof

Given such a vector x with $Ax = \lambda x$, we observe that $\lambda \in \mathbb{R}^+$. Then

$$x = \frac{1}{(1+\lambda)^{n-1}}(I_n + A)^{n-1}x,$$

and the right-hand side is strictly positive, from Proposition 5.1.2.

■

Finally, we can state the following result.

Lemma 5.3.3 *Let $A, B \in \mathbf{M}_n(\mathbb{C})$ be matrices, with A irreducible and $|B| \leq A$. Then $\rho(B) \leq \rho(A)$.*

In case of equality ($\rho(B) = \rho(A)$), the following hold:

- *$|B| = A$;*
- *for every eigenvector x of B associated to an eigenvalue of modulus $\rho(A)$, $|x|$ is an eigenvector of A associated to $\rho(A)$.*

Proof

In order to establish the inequality, we proceed as above. If λ is an eigenvalue of B, of modulus $\rho(B)$, and if x is a normalized eigenvector, then $\rho(B)|x| \leq |B| \cdot |x| \leq A|x|$, so that $C_{\rho(B)}$ is nonempty. Hence $\rho(B) \leq R = \rho(A)$.

Let us investigate the case of equality. If $\rho(B) = \rho(A)$, then $|x| \in C_{\rho(A)}$, and therefore $|x|$ is an eigenvector: $A|x| = \rho(A)|x| = \rho(B)|x| \leq |B| \cdot |x|$. Hence, $(A - |B|)|x| \leq 0$. Since $|x| > 0$ (from Lemma 5.3.2) and $A - |B| \geq 0$, this gives $|B| = A$.

■

5.3.3 The Eigenvalue $\rho(A)$ Is Simple

Let $P_A(X)$ be the characteristic polynomial of A. It is given as the composition of an n-linear form (the determinant) with polynomial vector-valued functions (the columns of $XI_n - A$). If ϕ is p-linear and if $V_1(X), \ldots, V_p(X)$

are polynomial vector-valued functions, then the polynomial $P(X) := \phi(V_1(X), \dots, V_p(X))$ has the derivative

$$P'(X) = \phi(V_1', V_2, \dots, V_p) + \phi(V_1, V_2', \dots, V_p) + \cdots + \phi(V_1, \dots, V_{p-1}, V_p').$$

One therefore has

$$\begin{aligned} P_A'(X) &= \det(\mathbf{e}^1, a_2, \dots, a_n) + \det(a_1, \mathbf{e}^2, \dots, a_n) + \cdots \\ &\quad + \cdots + \det(a_1, \dots, a_{n-1}, \mathbf{e}^n), \end{aligned}$$

where a_j is the jth column of $XI_n - A$ and $\{\mathbf{e}^1, \dots, \mathbf{e}^n\}$ is the canonical basis of $\mathbb{R}^n$. Developping the jth determinant with respect to the jth column, one obtains

$$P_A'(X) = \sum_{j=1}^{n} P_{A_j}(X), \tag{5.1}$$

where $A_j \in \mathbf{M}_{n-1}(\mathbb{R})$ is obtained from A by deleting the jth row and the jth column. Let us now denote by $B_j \in \mathbf{M}_n(\mathbb{R})$ the matrix obtained from A by replacing the entries of the jth row and column by zeroes. This matrix is block-diagonal, the two diagonal blocks being $A_j \in \mathbf{M}_{n-1}(\mathbb{R})$ and $0 \in \mathbf{M}_1(\mathbb{R})$. Hence, the eigenvalues of B_j are those of A_j, together with zero, and therefore $\rho(B_j) = \rho(A_j)$. Furthermore, $|B_j| \leq A$, but $|B_j| \neq A$ because A is irreducible and B_j is block-diagonal, hence reducible. It follows (Lemma 5.3.3) that $\rho(B_j) < \rho(A)$. Hence $P_{A_j}(\rho(A))$ is nonzero, with the same sign as P_{A_j} in a neighborhood of $+\infty$, which is positive. Finally, $P_A'(\rho(A))$ is positive and $\rho(A)$ is a simple root.

This completes the proof of Theorem 5.3.1. A different proof of the simplicity and another proof of the Perron–Frobenius theorem are given in Exercises 2 and 4.

■

5.4 Cyclic Matrices

The following statement completes Theorem 5.3.1.

Theorem 5.4.1 *Under the assumptions of Theorem 5.3.1, the set $R(A)$ of eigenvalues of A of maximal modulus $\rho(A)$ is of the form $R(A) = \rho(A)\boldsymbol{\mathcal{U}}_p$, where $\boldsymbol{\mathcal{U}}_p$ is the group of pth roots of unity, where p is the cardinality of $R(A)$. Every such eigenvalue is simple. The spectrum of A is invariant under multiplication by $\boldsymbol{\mathcal{U}}_p$. Finally, A is similar, by means of a permutation of coordinates in $\mathbb{R}^n$, to the following cyclic form. In this cyclic matrix each element is a block, and the diagonal blocks (which all vanish) are square*

with nonzero sizes:

$$\begin{pmatrix} 0 & M_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ & & & \ddots & \\ 0 & & & \ddots & M_{p-1} \\ M_p & 0 & \cdots & \cdots & 0 \end{pmatrix}.$$

Remarks:

- The converse is true. For example, the spectrum of a cyclic matrix is stable under multiplication by $\exp(2i\pi/p)$.
- One may show that p divides $n - n_0$, where n_0 is the multiplicity of the zero eigenvalue.
- The nonzero eigenvalues of A are the pth roots of those of the matrix $M_1 M_2 \cdots M_p$, which is square, though its factors might not be square.

Proof

Let us denote by X the unique nonnegative eigenvector of A normalized by $\|X\|_1 = 1$. If Y is a unitary eigenvector, associated to an eigenvalue μ of maximal modulus $\rho(A)$, the inequality $\rho(A)|Y| = |AY| \leq A|Y|$ implies (Lemma 5.3.3) $|Y| = X$. Hence there is a diagonal matrix $D = \operatorname{diag}(e^{i\alpha_1}, \ldots, e^{i\alpha_n})$ such that $Y = DX$. Let us define a unimodular complex number $e^{i\gamma} = \mu/\rho(A)$ and let B be the matrix $e^{-i\gamma} D^{-1} A D$. One has $|B| = A$ and $BX = X$. For every j, one therefore has

$$\left| \sum_{k=1}^{n} b_{jk} x_k \right| = \sum_{k=1}^{n} |b_{jk}| x_k.$$

Since $X > 0$, one deduces that B is real-valued and nonnegative; that is, $B = A$. Hence $D^{-1} A D = e^{i\gamma} A$. The spectrum of A is thus invariant under multiplication by $e^{i\gamma}$.

Let $\mathcal{U} = \rho(A)^{-1} R(A)$, which is included in S^1, the unit circle. The previous discussion shows that $\mathcal{U}$ is stable under multiplication. Since $\mathcal{U}$ is finite, it follows that its elements are roots of unity. Since the inverse of a dth root of unity is its own $(d-1)$th power, $\mathcal{U}$ is stable under inversion. Hence it is a finite subgroup of S^1; that is, it is $\boldsymbol{\mathcal{U}}_p$, for a suitable p.

Let P_A be the characteristic polynomial and let $\omega = \exp(2i\pi/p)$. One may apply the first part of the proof to $\mu = \omega\rho(A)$. One has thus $D^{-1} A D = \omega A$, and it follows that $P_A(X) = \omega^n P_A(X/\omega)$. Therefore, multiplication by ω sends eigenvalues to eigenvalues of the same multiplicities. In particular, the eigenvalues of maximal modulus are simple.

Iterating the conjugation, one obtains $D^{-p} A D^p = A$. Let us set

$$D^p = \operatorname{diag}(d_1, \ldots, d_n).$$

One has thus $d_j = d_k$, provided that $a_{jk} \neq 0$. Since A is irreducible, one can link any two indices j and k by a chain $j_0 = j, \dots, j_r = k$ such that $a_{j_{s-1},j_s} \neq 0$ for every s. It follows that $d_j = d_k$ for every j, k. But since one may choose $Y_1 = X_1$, that is, $\alpha_1 = 0$, one also has $d_1 = 1$ and hence $D^p = I_n$. The α_j are thus pth roots of unity. With a conjugation by a permutation matrix we may limit ourselves to the case where D has the block-diagonal form $\mathrm{diag}(J_0, \omega J_1, \dots, \omega^{p-1} J_{p-1})$, where the J_l are identity matrices of respective sizes $n_0, \dots, n_{p-1}$. Decomposing A into blocks A_{lm} of sizes $n_l \times n_m$, one obtains $\omega^k A_{jk} = \omega^{j+1} A_{jk}$ directly from the conjugation identity. Hence $A_{jk} = 0$ except for the pairs (j, k) of the form $(0,1), (1,2), \dots, (p-2, p-1), (p-1, 0)$. This is the announced cyclic form.

■

5.5 Stochastic Matrices

Definition 5.5.1 *A matrix $M \in \mathbf{M}_n(\mathbb{R})$ is said to be* stochastic *if $M \geq 0$ and if for every $i = 1, \dots, n$, one has*

$$\sum_{j=1}^{n} m_{ij} = 1.$$

One says that M is bistochastic *(or* doubly stochastic*) if both M and M^T are stochastic.*

Denoting by $e \in \mathbb{R}^n$ the vector all of whose coordinates equal one, one sees that M is stochastic if and only if $M \geq 0$ and $Me = e$. Moreover, M is bistochastic if $M \geq 0$, $Me = e$, and $e^T M = e^T$. If M is stochastic, one has $\|Mx\|_\infty \leq \|x\|_\infty$ for every $x \in \mathbb{C}^n$, and therefore $\rho(M) \leq 1$. But since $Me = e$, one has in fact $\rho(M) = 1$.

The stochastic matrices play an important role in the study of Markov chains. A special case of a bistochastic matrix is a permutation matrix $P(\sigma)$ ($\sigma \in \boldsymbol{S}_n$), whose entries are

$$p_{ij} = \delta^j_{\sigma(i)}.$$

The following theorem explains the role of permutation matrices.

Theorem 5.5.1 (Birkhoff) *A matrix $M \in \mathbf{M}_n(\mathbb{R})$ is bistochastic if and only if it is a center of mass (that is, a barycenter with nonnegative weights) of permutation matrices.*

The fact that a center of mass of permutation matrices is a doubly stochastic matrix is obvious, since the set Δ_n of doubly stochastic matrices is convex. The interest of the theorem lies in the statement that if $M \in \Delta_n$, there exist permutation matrices $P_1, \dots, P_r$ and positive real numbers $\alpha_1, \dots, \alpha_r$ with $\alpha_1 + \cdots + \alpha_r = 1$ such that $M = \alpha_1 P_1 + \cdots + \alpha_r P_r$.

Let us recall that a point a of a convex set C is an *extreme point* if the equality $x = \theta y + (1-\theta)z$, with $y, z \in C$ and $\theta \in (0,1)$ implies $y = z = x$. The Krein–Milman theorem (see [30], Theorem 3.23) says that a convex compact subset of $\mathbb{R}^n$ is the convex hull, that is, the set of centers of mass, of its extreme points. Since Δ_n is closed and bounded, hence compact, it is permissible to apply the Krein–Milman theorem.

Proof

To begin with, it is immediate that the permutation matrices are extreme points of Δ_n. From the Krein–Milman theorem, the proof amounts to showing that there is no other extreme point in Δ_n.

Let $M \in \Delta_n$ be given. If M is not a permutation matrix, there exists an entry $m_{i_1 j_1} \in (0,1)$. Since M is stochastic, there also exists $j_2 \neq j_1$ such that $m_{i_1 j_2} \in (0,1)$. Since M^T is stochastic, there exists $i_2 \neq i_1$ such that $m_{i_2 j_2} \in (0,1)$. By this procedure one constructs a sequence $(j_1, i_1, j_2, i_2, \dots)$ such that $m_{i_l j_l} \in (0,1)$ and $m_{i_{l-1} j_l} \in (0,1)$. Since the set of indices is finite, it eventually happens that one of the indices (a row index or a column index) is repeated.

Therefore, one can assume that the sequence $(j_1, i_1, \dots, j_r, i_r, j_{r+1} = j_1)$ has the above property. Let us define a matrix $B \in \mathbf{M}_n(\mathbb{R})$ by $b_{i_l j_l} = 1$, $b_{i_l j_{l+1}} = -1$, $b_{ij} = 0$ otherwise. By construction, $Be = 0$ and $e^T B = 0$. If $\alpha \in \mathbb{R}$, one therefore has $(M \pm \alpha B)e = e$ and $e^T(M \pm \alpha B) = e^T$. If $\alpha > 0$ is small enough, $M \pm \alpha B$ turns out to be nonnegative. Finally, $M + \alpha B$ and $M - \alpha B$ are bistochastic, and

$$M = \frac{1}{2}(M - \alpha B) + \frac{1}{2}(M + \alpha B).$$

Hence M is not an extreme point of Δ_n.

■

Here is a nontrivial consequence (Stoer and Witzgall [32]):

Corollary 5.5.1 *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, invariant under permutation of the coordinates. Then $\|M\| = 1$ for every bistochastic matrix (where by abuse of notation we have used $\|\cdot\|$ for the induced norm on $\mathbf{M}_n(\mathbb{R})$).*

Proof

To begin with, $\|P\| = 1$ for every permutation matrix, by assumption. Since the induced norm is convex (true for every norm), one deduces from Birkhoff's theorem that $\|M\| \leq 1$ for every bistochastic matrix. Furthermore, $Me = e$ implies $\|M\| \geq \|Me\|/\|e\| = 1$.

■

This result applies, for instance, to the norm $\|\cdot\|_p$, providing a nontrivial convex set on which the map $1/p \mapsto \log \|M\|_p$ is constant (compare with Theorem 4.3.1).

The bistochastic matrices are intimately related to the relation $\prec$ (see Section 3.4). In fact, we have the following theorem.

Theorem 5.5.2 *A matrix A is bistochastic if and only if $Ax \succ x$ for every $x \in \mathbb{R}^n$.*

Proof

If A is bistochastic, then

$$\|Ax\|_1 \leq \|A\|_1 \|x\|_1 = \|x\|_1,$$

since A^T is stochastic. Since A is stochastic, $Ae = e$. Applying this inequality to $x - te$, one therefore has $\|Ax - te\|_1 \leq \|x - te\|_1$. Proposition 3.4.1 then shows that $x \prec Ax$.

Conversely, let us assume that $x \prec Ax$ for every $x \in \mathbb{R}^n$. Choosing x as the jth vector of the canonical basis, $\mathbf{e}^j$, the inequality $s_1(\mathbf{e}^j) \leq s_1(A\mathbf{e}^j)$ expresses that A is a nonnegative matrix, while $s_n(\mathbf{e}^j) = s_n(A\mathbf{e}^j)$ yields

$$\sum_{i=1}^{n} a_{ij} = 1. \tag{5.2}$$

One then chooses $x = e$. The inequality $s_1(e) \leq s_1(Ae)$ expresses[1] that $Ae \geq e$. Finally, $s_n(e) = s_n(Ae)$ and $Ae \geq e$ give $Ae = e$. Hence, A is bistochastic.

■

This statement is completed by the following.

Theorem 5.5.3 *Let $x, y \in \mathbb{R}^n$. Then $x \prec y$ if and only if there exists a bistochastic matrix A such that $y = Ax$.*

Proof

From the previous theorem, it is enough to show that if $x \prec y$, there exists A, a bistochastic matrix, such that $y = Ax$. To do so, one applies Theorem 3.4.2: There exists a Hermitian matrix H whose diagonal and spectrum are y and x, respectively. Let us diagonalize H by a unitary conjugation: $H = U^* D U$, with $D = \mathrm{diag}(x_1, \ldots, x_n)$. Then $y = Ax$, where $a_{ij} = |u_{ij}|^2$. Since U is unitary, A is bistochastic.[2]

■

An important aspect of stochastic matrices is their action on the simplex

$$K_n := \left\{ x \in \mathbb{R}^n \,;\, x \geq 0 \text{ and } \sum_i x_i = 1 \right\}.$$

It is clear that M^T is stochastic if and only if $M(K_n)$ is contained in K_n; M is bistochastic if, moreover, $Me = e$.

Considered as a part of the affine subspace whose equation is $\sum_i x_i = 1$, K_n is a convex set with a nonempty interior. Its interior comprises those points that satisfy $x > 0$. One denotes ∂K_n the boundary of K_n. If $x \in K_n$,

---

[1] For another vector y, $s_1(y) \leq s_1(Ay)$ does not imply $Ay \geq y$.

[2] This kind of bistochastic matrix is called *orthostochastic*.

we denote by $O(x)$ the set of indices i such that $x_i = 0$, and by $o(x)$ its cardinality, in such a way that ∂K_n comprises those points satisfying $o(x) \geq 1$. One always has $m_{ij} = 0$ for $(i,j) \in O(Mx) \times O(x)^c$, where I^c denotes the complement of I in $\{1, \ldots, n\}$.

Proposition 5.5.1 *Let $x \in K_n$ and $M \in \Delta_n$ be given. Then one has*

$$o(Mx) \leq o(x).$$

Moreover, if $o(Mx) = o(x)$, one has $m_{ij} = 0$ for every $(i,j) \in O(Mx)^c \times O(x)$.

Proof

Let us compute

$$o(x) - o(Mx) = \sum_{i=1}^{n} \sum_{O(x)} m_{ij} - \sum_{O(Mx)} \sum_{j=1}^{n} m_{ij} = \sum_{O(Mx)^c \times O(x)} m_{ij} \geq 0.$$

The case of equality is immediate.

■

We could have obtained the first part of the proposition by applying Theorem 5.5.2.

Corollary 5.5.2 *Let I and J be two subsets of $\{1, \ldots, n\}$ and let $M \in \Delta_n$ be a matrix satisfying $m_{ij} = 0$ for every $(i,j) \in I \times J^c$. Then one has $|J| \geq |I|$. If, moreover, $|I| = |J|$, then m_{ij} also vanishes for $(i,j) \in I^c \times J$.*

Proof

It is sufficient to choose $x \in K^n$ with $J^c = O(x)$ if J is nonempty. If J is empty, the statement is obvious.

■

We shall denote by $\mathbf{S}\Delta_n$ (S for *strict*) the set of doubly stochastic matrices M for which the conditions $|I| = |J|$ and $m_{ij} = 0$ for every $(i,j) \in I \times J^c$ imply either $I = \emptyset$ or $I = \{1, \ldots, n\}$. These are also the matrices for which $x \in \partial K_n$ implies $o(Mx) < o(x)$. This set does not contain permutation matrices P, since these satisfy $o(Px) = o(x)$ for every $x \in K_n$.

Let $M \in \Delta_n$ be given. A *decomposition* of M consists of two partitions $I_1 \cup \cdots \cup I_r$ and $J_1 \cup \cdots \cup J_r$ of the set $\{1, \ldots, n\}$ such that

$$(i \in I_l,\ j \in J_m,\ l \neq m) \Longrightarrow m_{ij} = 0.$$

From Corollary 5.5.2, we have $|I_l| = |J_l|$ for every l. Eliminating empty parts if necessary, we can always assume that none of the I_l's or J_l's is empty. A decomposition of M furnishes a block structure, in which each row-block has only one nonzero block, and the same for the column-blocks. The blocks of indices $I_l \times J_l$ are themselves stochastic matrices. A matrix of $\mathbf{S}\Delta_n$ admits only the trivial decomposition $r = 1$, $I_1 = J_1 = \{1, \ldots, n\}$.

If M admits two decompositions, one with the sets I_l, J_l, $1 \leq l \leq r$, the other one with I'_l, J'_l, $1 \leq l \leq s$, let us form the partitions $\cup_{l,m} I''_{lm}$ and

$\cup_{l,m} J''_{lm}$, with $I''_{lm} := I_l \cap I'_m$ and $J''_{lm} := J_l \cap J'_m$. If $i \in I''_{lm}$ and $j \in J''_{pq}$, with $(l,m) \neq (p,q)$, we have $m_{ij} = 0$. From Corollary 5.5.2, applied to M and to its transposition, we have $|I''_{lm}| = |J''_{lm}|$. Eliminating the empty parts, we obtain therefore a decomposition of M that is finer than the first two, in the sense of inclusion order: Each I_l (or I'_l) is a union of some parts of the form I''_p.

Since the set of decompositions of M is finite, the previous argument shows that there exists a finest one. We shall call it the *canonical decomposition* of M. It is the only decomposition for which the blocks of indices $I_l \times J_l$ are themselves of class $\mathbf{S\Delta}$.

5.6 Exercises

1. We consider the following three properties for a matrix $M \in \mathbf{M}_n(I\!R)$.

   P1 M is nonnegative.
   P2 $M^T e = e$, where $e = (1, \ldots, 1)^T$.
   P3 $\|M\|_1 \leq 1$.

   (a) Show that **P2** and **P3** imply **P1**.
   (b) Show that **P2** and **P1** imply **P3**.
   (c) Does **P1** and **P3** imply **P2**?

2. Here is another proof of the simplicity of $\rho(A)$ in the Perron–Frobenius theorem, which does not require Lemma 5.3.3.

   (a) We assume that A is irreducible and nonnegative, and we denote by x a positive eigenvector associated to the eigenvalue $\rho(A)$. Let K be the set of nonnegative eigenvectors y associated to $\rho(A)$ such that $\|y\|_1 = 1$. Show that K is compact and convex.
   (b) Show that the geometric multiplicity of $\rho(A)$ equals 1 (**Hint**: Otherwise, K would contain a vector with at least one zero component.)
   (c) Show that the algebraic multiplicity of $\rho(A)$ equals 1 (**Hint**: Otherwise, there would be a nonnegative vector y such that $Ay - \rho(A)y = x > 0$.)

3. Let $M \in \mathbf{M}_n(I\!R)$ be either a strictly diagonally dominant, or an irreducible strongly diagonally dominant, matrix. Assume that $m_{jj} > 0$ for every $j = 1, \ldots, n$ and $m_{ij} \leq 0$ otherwise. Show that M is invertible and that the solution of $Mx = b$, when $b \geq 0$, satisfies $x \geq 0$. Deduce that $M^{-1} \geq 0$.

4. Here is another proof of Theorem 5.3.1, due to Perron himself. We proceed by induction on the size n of the matrix. The statement is obvious if $n = 1$. We therefore assume that it holds for matrices of size n. We give ourselves an irreducible nonnegative matrix $A \in$

$\mathbf{M}_{n+1}(\mathbb{R})$, which we decompose blockwise as

$$A = \begin{pmatrix} a & \xi^T \\ \eta & B \end{pmatrix}, \quad a \in \mathbb{R}, \quad \xi, \eta \in \mathbb{R}^n, \quad B \in \mathbf{M}_n(\mathbb{R}).$$

(a) Applying the induction hypothesis to the matrix $B + \epsilon J$, where $\epsilon > 0$ and $J > 0$ is a matrix, then letting ϵ go to zero, show that $\rho(B)$ is an eigenvalue of B, associated to a nonnegative eigenvector (this avoids the use of Theorem 5.2.1).

(b) Using the formula

$$(\lambda I_n - B)^{-1} = \sum_{k=1}^{\infty} \lambda^{-k} B^{k-1},$$

valid for $\lambda \in (\rho(B), +\infty)$, deduce that the function $h(\lambda) := \lambda - a - \xi^T(\lambda I_n - B)^{-1}\eta$ is strictly increasing on this interval and that on the same interval the vector $x(\lambda) := (\lambda I_n - B)^{-1}\eta$ is positive.

(c) Prove the relation $P_A(\lambda) = P_B(\lambda)h(\lambda)$ between the characteristic polynomials.

(d) Deduce that the matrix A has one and only one eigenvalue in $(\rho(B), +\infty)$, and that it is a simple one, associated to a positive eigenvector. One denotes this eigenvalue by λ_0.

(e) Applying the previous results to A^T, show that there exists $\ell \in \mathbb{R}^n$ such that $\ell > 0$ and $\ell^T(A - \lambda_0 I_n) = 0$.

(f) Let μ be an eigenvalue of A, associated to an eigenvector X. Show that $(\lambda_0 - |\mu|)\ell^T|X| \geq 0$. Conclusion?

5. Let $A \in \mathbf{M}_n(\mathbb{R})$ be a matrix satisfying $a_{ij} \geq 0$ for every pair (i, j) of distinct indices.

(a) Using the Exercise 3, show that

$$R(h; A) := (I_n - hA)^{-1} \geq 0,$$

for $h > 0$ small enough.

(b) Deduce that $\exp(tA) \geq 0$ for every $t > 0$ (the exponential of matrices is presented in Chapter 7). Consider Trotter's formula

$$\exp tA = \lim_{m \to +\infty} R(t/m; A)^m,$$

where exp is the exponential of square matrices, defined in Chapter 7. Trotter's formula is justified by the convergence (see Exercise 10 in Chapter 7) of the implicit Euler method for the differential equation

$$\frac{dx}{dt} = Ax. \tag{5.3}$$

(c) Deduce that if $x(0) \geq 0$, then the solution of (5.3) is nonnegative for every nonnegative t.

(d) Deduce also that

$$\sigma := \sup\{\Re\lambda; \lambda \in \mathrm{Sp}\ A\}$$

is an eigenvalue of A.

6. Let $A \in \mathbf{M}_n(\mathbb{R})$ be a matrix satisfying $a_{ij} \geq 0$ for every pair (i,j) of distinct indices.

(a) Let us define

$$\sigma := \sup\{\Re\lambda; \lambda \in \mathrm{Sp}\ A\}.$$

Among the eigenvalues of A whose real parts equal σ, let us denote by μ the one with the largest imaginary part. Show that for every positive large enough real number τ, $\rho(A + \tau I_n) = |\mu + \tau|$.

(b) Deduce that $\mu = \sigma = \rho(A)$ (apply Theorem 5.2.1).

7. Let $B \in \mathbf{M}_n(\mathbb{R})$ be a matrix whose off-diagonal entries are positive and such that the eigenvalues have strictly negative real parts. Show that there exists a nonnegative diagonal matrix D such that $B' := D^{-1}BD$ is strictly diagonally dominant, namely,

$$b'_{ii} < -\sum_{j \neq i} b'_{ij}.$$

8. Let $B \in \mathbf{M}_n(\mathbb{R})$ be a nonnegative matrix and

$$A := \begin{pmatrix} B & 0_m \\ I_m & B \end{pmatrix}.$$

(a) If an eigenvalue λ of A is associated to a positive eigenvector, show that there exists $\mu > \lambda$ and $Z > 0$ such that $BZ \geq \mu Z$. Deduce that $\lambda < \rho(B)$.

(b) Deduce that A admits no strictly positive eigenvector (first of all, apply Theorem 5.2.1 to the matrix A^T).

9. (a) Let $B \in \mathbf{M}_n(\mathbb{R})$ be given, with $\rho(B) = 1$. Assume that the eigenvalues of B of modulus one are (algebraically) simple. Show that the sequence $(B^m)_{m \geq 1}$ is bounded.

(b) Let $M \in \mathbf{M}_n(\mathbb{R})$ be a nonnegative irreducible matrix, with $\rho(M) = 1$. We denote by x and y^T the left and right eigenvectors for the eigenvalue 1 ($Mx = x$ and $y^T M = y^T$), normalized by $y^T x = 1$. We define $L := xy^T$ and $B = M - L$.

i. Verify that $B - I_n$ is invertible. Determine the spectrum and the invariant subspaces of B by means of those of M.

ii. Show that the sequence $(B^m)_{m \geq 1}$ is bounded. Express M^m in terms of B^m.

iii. Deduce that

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{m=0}^{N-1} M^m = L.$$

iv. Under what additional assumption do we have the stronger convergence

$$\lim_{N\to+\infty} M^N = L?$$

10. Let $B \in \mathbf{M}_n(\mathbb{R})$ be a nonnegative irreducible matrix and let $C \in \mathbf{M}_n(\mathbb{R})$ be a nonzero nonnegative matrix. For $t > 0$, we define $r_t := \rho(B + tC)$ and we let X_t denote the nonnegative unitary eigenvector associated to the eigenvalue r_t.

(a) Show that $t \mapsto r_t$ is strictly increasing.
Define $r := \lim_{t\to+\infty} r_t$. We wish to show that $r = +\infty$. Let X be a cluster point of the sequence X_t. We may assume, up to a permutation of the indices, that

$$X = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \quad Y > 0.$$

(b) Suppose that in fact, $r < +\infty$. Show that $BX \leq rX$. Deduce that $B'Y = 0$, where B' is a matrix extracted from B.
(c) Deduce that $X = Y$; that is, $X > 0$.
(d) Show, finally, that $CX = 0$. Conclude that $r = +\infty$.
(e) Assume, moreover, that $\rho(B) < 1$. Show that there exists one and only one $t \in \mathbb{R}$ such that $\rho(B + tC) = 1$.

11. Show that Δ is stable under multiplication. In particular, if M is bistochastic, the sequence $(M^m)_{m\geq 1}$ is bounded.

12. Let $M \in \mathbf{M}_n(\mathbb{R})$ be a bistochastic irreducible matrix. Show that

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{m=0}^{N-1} M^m = \frac{1}{n} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix} =: J_n$$

(use Exercise 9). Show by an example that the sequence $(M^m)_{m\geq 1}$ may or may not converge.

13. Show directly that for every $p \in [1, \infty]$, $\|J_n\|_p = 1$, where J_n was defined in the previous exercise.

14. Let $P \in \mathbf{GL}_n(\mathbb{R})$ be given such that P, $P^{-1} \in \Delta_n$. Show that P is a permutation matrix.

15. If $M \in \Delta_n$ is given, we define an equivalence relation between indices in the following way: $i'\mathcal{R}i''$ if there exists a sequence $i_1 = i', j_1, i_2, j_2, \dots, i_p = i''$ such that $m_{ij} > 0$ each time that (i, j) is

of the form (i_l, j_l) or (i_{l+1}, j_l) (compare with the proof of Theorem 5.5.1). Show that in the canonical decomposition of M, the I_l are the equivalence classes of $\mathcal{R}$.
Deduce that the following matrix belongs to $\mathbf{S}\Delta_n$:

$$\begin{pmatrix} 1/2 & 1/2 & 0 & \cdots & 0 \\ 1/2 & 0 & 1/2 & \ddots & \vdots \\ 0 & 1/2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1/2 \\ 0 & \cdots & 0 & 1/2 & 1/2 \end{pmatrix}.$$

16. Let $M \in \mathbf{S}\Delta_n$ and $M' \in \Delta_n$ be given. Show that MM', $M'M \in \mathbf{S}\Delta_n$.

17. If $M \in \mathbf{S}\Delta_n$, show that $\lim_{N\to+\infty} M^N$ exists.

18. Consider the induced norm $\|\cdot\|_p$ on $\mathbf{M}_n(\mathbb{C})$. Let M be a bistochastic matrix.
    (a) Compute $\|M\|_1$ and $\|M\|_\infty$.
    (b) Show that $\|M\| \geq 1$ for every induced norm.
    (c) Deduce from Theorem 4.3.1 that $\|M\|_p = 1$. To what extent is this result different from Corollary 5.5.1?

19. Suppose that we are given three real symmetric matrices (or Hermitian matrices) $A, B, C = A + B$.
    (a) If $t \in [0,1]$ consider the matrix $S(t) := A + tB$, so that $S(0) = A$ and $S(1) = C$. Arrange the eigenvalues of $S(t)$ in increasing order $\lambda_1(t) \leq \cdots \leq \lambda_n(t)$. For each value of t there exists an orthonormal eigenbasis $\{X_1(t), \ldots, X_n(t)\}$. We admit the fact that it can be chosen continuously with respect to t, so that $t \mapsto X_j(t)$ is continuous with a piecewise continuous derivative. Show that $\lambda_j'(t) = (BX_j(t), X_j(t))$.
    (b) Let $\alpha_j, \beta_j, \gamma_j$ $(j = 1, \ldots, n)$ be the eigenvalues of A, B, C, respectively. Deduce from part (a) that

$$\gamma_j - \alpha_j = \int_0^1 (BX_j(t), X_j(t))\, dt.$$

    (c) Let $\{Y_1, \ldots, Y_n\}$ be an orthonormal eigenbasis, relative to B. Define

$$\sigma_{jk} := \int_0^1 |(X_j(t), Y_k)|^2 dt.$$

    Show that the matrix $\Sigma := (\sigma_{jk})_{1\leq j,k\leq n}$ is bistochastic.
    (d) Show that $\gamma_j - \alpha_j = \sum_k \sigma_{jk}\beta_k$. Deduce (Lidskii's theorem) that the vector $(\gamma_1 - \alpha_1, \ldots, \gamma_n - \alpha_n)$ belongs to the convex hull of the vectors obtained from the vector $(\beta_1, \ldots, \beta_n)$ by all possible permutations of the coordinates.

20. Let $a \in \mathbb{R}^n$ be given, $a = (a_1, \dots, a_n)$.

(a) Show that

$$C(a) := \{b \in \mathbb{R}^n \mid b \succ a\}$$

is a convex compact set. Characterize its extremal points.

(b) Show that

$$Y(a) := \{M \in \mathbf{Sym}_n(\mathbb{R}) \mid \text{Sp } M \succ a\}$$

is a convex compact set. Characterize its extremal points.

(c) Deduce that $Y(a)$ is the closed convex hull (actually the convex hull) of the set

$$X(a) := \{M \in \mathbf{Sym}_n(\mathbb{R}) \mid \text{Sp } M = a\}.$$

(d) Set $\alpha = s_n(a)/n$ and $a' := (\alpha, \dots, \alpha)$. Show that $a' \in C(a)$, and that $b \in C(a) \Longrightarrow b \prec a'$.

(e) Characterize the set

$$\{M \in \mathbf{Sym}_n(\mathbb{R}) \mid \text{Sp } M \prec a'\}.$$

6

Matrices with Entries in a Principal Ideal Domain; Jordan Reduction

6.1 Rings, Principal Ideal Domains

In this Chapter we consider commutative integral domains A (see Chapter 2). In particular, such a ring A can be embeded in its field of fractions, which is the quotient of $A \times (A \setminus \{0\})$ by the equivalence relation $(a,b)\mathcal{R}(c,d) \Leftrightarrow ad = bc$. The embedding is the map $a \mapsto (a,1)$. In a ring A the set of invertible elements is denoted by A^*. If $a, b \in A$ are such that $b = ua$ with $u \in A^*$, we say that a and b are *associated*, and we write $a \sim b$, which amounts to saying that $aA = bA$. If there exists $c \in A$ such that $ac = b$, we say that a divides b and write $a|b$. Then the quotient c is unique and is denoted by b/a. We say that b is a prime, or irreducible, element if the equality $b = ac$ implies that one of the factors is invertible.

An *ideal* I in a ring A is an additive subgroup of A such that $A \cdot I \subset I$: $a \in A$, $x \in I$ imply $ax \in I$. For example, if $b \in A$, the subset bA is an ideal, denoted by (b). Ideals of the form (b) are called *principal ideals*.

6.1.1 Facts About Principal Ideal Domains

Definition 6.1.1 *A commutative ring A is a* principal ideal domain *if every ideal in A is principal: For every ideal $\mathcal{I}$ there exists $a \in A$ such that $\mathcal{I} = (a)$.*

A field is a principal ideal domain that has only two ideals, (0) and (1). The set $\mathbb{Z}$ of rational integers and the polynomial algebra over a field k,

denoted by $k[X]$, are also principal ideal domains. More generally, every Euclidean domain is a principal ideal domain (see Proposition 6.1.3 below).

In a commutative integral domain one says that d is a *greatest common divisor* (*gcd*) of a and b if d divides a and b, and if every common divisor of a and b divides d. In other words, the set of common divisors of a and b admits d as a greatest element. The gcd of a and b, whenever it exists, is unique up to multiplication by an invertible element. We say that a and b are coprime if all their common divisors are invertible; in that case, $\gcd(a, b) = 1$.

Proposition 6.1.1 *In a principal ideal domain, every pair of elements has a greatest common divisor. The gcd satisfies the Bézout identity: For every $a, b \in A$, there exist $u, v \in A$ such that*

$$\gcd(a, b) = ua + vb.$$

Such u and v are coprime.

Proof

Let A be a principal ideal domain. If $a, b \in A$, the ideal $\mathcal{I} =: (a, b)$ spanned by a and b, which is the set of elements of the form $xa + yb$, $x, y \in A$, is principal: $\mathcal{I} = (d)$, where $d = \gcd(a, b)$. Since $a, b \in \mathcal{I}$, d divides a and b. Furthermore, $d = ua + vb$ because $d \in \mathcal{I}$. If c divides a and b, then c divides $ua + vb$; hence divides d, which happens to be a gcd of a and b.

If m divides u and v, then $md|ua + vb$; hence $d = smd$. If $d \neq 0$, one has $sm = 1$, which means that $m \in A^*$. Thus u and v are coprime. If $d = 0$, then $a = b = 0$, and one may take $u = v = 1$, which are coprime.

■

Let us remark that a gcd of a and b is a generator of the ideal $aA + bA$. It is thus nonunique. Every element associated to a gcd of a and b is another gcd. In certain rings one can choose the gcd in a canonical way, such as being positive in $\mathbb{Z}$, or monic in $k[X]$.

The gcd is associative: $\gcd(a, \gcd(b, c)) = \gcd(\gcd(a, b), c)$. It is therefore possible to speak of the gcd of an arbitrary finite subset of A. In the above example we denote it by $\gcd(a, b, c)$. At our disposal is a generalized Bézout formula: There exist elements $u_1, \ldots, u_r \in A$ such that

$$\gcd(a_1, \ldots, a_r) = a_1u_1 + \cdots + a_ru_r.$$

Definition 6.1.2 *A ring A is* Noetherian *if every nondecreasing (for inclusion) sequence of ideals is constant beyond some index: $I_0 \subset I_1 \subset \cdots \subset I_m \subset \cdots$ implies that there is an l such that $I_l = I_{l+1} = \cdots$.*

Proposition 6.1.2 *The principal ideal domains are Noetherian.*

Observe that in the case of principal ideal domains the Noetherian property means exactly that if a sequence $a_1, \ldots$ of elements of A is such that every element is divisible by the next one, then there exists an index J such that the a_j's are pairwise associated for every $j \geq J$.

This property seems natural because it is shared by all the rings encountered in number theory. But the ring of entire holomorphic functions is not Noetherian: Just take for a_n the function

$$z \mapsto \left(\prod_{k=1}^{n} (z-k)^{-1} \right) \sin 2\pi z.$$

Proof

Let A be a principal ideal domain and let $(I_j)_{j\geq 0}$ be a nondecreasing sequence of ideals in A. Let $\mathcal{I}$ be their union. This sequence is nondecreasing under inclusion, so that $\mathcal{I}$ is an ideal. Let a be a generator: $\mathcal{I} = (a)$. Then a belongs to one of the ideals, say $a \in \mathcal{I}_k$. Hence $\mathcal{I} \subset \mathcal{I}_k$, which implies $\mathcal{I}_j = \mathcal{I}$ for $j \geq k$.

■

We remark that the proof works with slight changes if we know that every ideal in A is spanned by a finite set. For example, the ring of polynomials over a Noetherian ring is itself Noetherian: $\mathbb{Z}[X]$ and $k[X,Y]$ are Noetherian rings.

The principal ideal domains are also *factorial* (a short term for *unique factorization domain*): Every element of A admits a factorization consisting of prime factors. This factorization is unique up to ambiguities, which may be of three types: the order of factors, the presence of invertible elements, and the replacement of factors by associated ones. This property is fundamental to the arithmetic in A.

6.1.2 Euclidean Domains

Definition 6.1.3 *A* Euclidean domain *is a ring A endowed with a map $N : A \mapsto \mathbb{N}$ such that for every $a, b \in A$ with $b \neq 0$, there exists a unique pair $(q, r) \in A \times A$ such that $a = qb + r$ with $N(r) < N(b)$ (Euclidean division).*

A special case of Euclidean division occurs when b divides a. Then $r = 0$ and we conclude that $N(b) > N(0)$ for every $b \neq 0$.

Classical examples of Euclidean domains are the ring of the rational integers $\mathbb{Z}$, with $N(a) = |a|$, the ring $k[X]$ of polynomials over a field k, with $N(P) = 2^{\deg P}$,[1] and the ring of Gaussian integers $\mathbb{Z}[\sqrt{-1}]$, with $N(z) = |z|^2$. Observe that if b is nonzero, the Euclidean division of b by itself shows that $N(b)$ is positive. The function N is often called a *norm*, though it does not resemble the norm on a real or complex vector space. In practice, one may define $N(0)$ in a consistent way by 0 if $b \neq 0 \Longrightarrow N(b) > 0$ (case of $\mathbb{Z}$ and $\mathbb{Z}[\sqrt{-1}]$), and by $-\infty$ otherwise (case of $k[X]$). With that

[1] One may take either $N(P) = 1 + \deg P$ if P is nonzero, and $N(0) = 0$.

extension, the pair (q, r) in the definition is uniquely defined by $a = bq + r$ and $N(r) < N(b)$.

Proposition 6.1.3 *Euclidean domains are principal ideal domains.*

Proof

Let $\mathcal{I}$ be an ideal of a Euclidean domain A. If $\mathcal{I} = (0)$, there is nothing to show. Otherwise, let us select in $\mathcal{I} \setminus \{0\}$ an element a of minimal norm. If $b \in \mathcal{I}$, the remainder r of the Euclidean division of b by a is an element of $\mathcal{I}$ and satisfies $N(r) < N(a)$. The minimality of $N(a)$ implies $r = 0$, that is, $a|b$. Finally, $\mathcal{I} = (a)$.

■

The converse of Proposition 6.1.3 is not true. For example, the quadratic ring $\mathbb{Z}[\sqrt{14}]$ is Euclidean, though not a principal ideal domain. More information about rings of quadratic integers can be found in Cohn's monograph [10].

6.1.3 Elementary Matrices

An *elementary matrix* of order n is a matrix of one of the following forms:

- The transposition matrices: If $\sigma \in \boldsymbol{S}_n$, the matrix P_σ has entries $p_{ij} = \delta^j_{\sigma(i)}$, where δ is the Kronecker symbol.
- The matrices $I_n + aJ_{ik}$, for $a \in A$ and $1 \leq i \neq k \leq n$, with
$$(J_{ik})_{lm} = \delta^l_i \delta^m_k.$$
- The diagonal invertible matrices, that is, those whose diagonal entries are invertible in A.

We observe that the inverse of an elementary matrix is again elementary. For example, $(I_n + aJ_{ik})(I_n - aJ_{ik}) = I_n$.

Theorem 6.1.1 *A square invertible matrix of size n with entries in a Euclidean domain A is a product of elementary matrices with entries in A.*

Proof

We shall prove the theorem for $n = 2$. The general case will be deduced from that particular one and from the proof of Theorem 6.2.1 below, since the matrices used in that proof are block-diagonal with 1×1 and 2×2 diagonal blocks.

Let

$$M = \begin{pmatrix} a & a_1 \\ c & d \end{pmatrix}$$

be given in $\mathbf{SL}_2(A)$: we have $ad - a_1c \in A^*$. If $N(a) < N(a_1)$, we multiply M on the right by

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

We are now in the case $N(a_1) \leq N(a)$. Let $a = a_1q + a_2$ be the Euclidean division of a by a_1. Then

$$M \begin{pmatrix} 1 & 0 \\ -q & 1 \end{pmatrix} =: M' = \begin{pmatrix} a_2 & a_1 \\ \cdot & d \end{pmatrix}.$$

Next, we have

$$M' \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} =: M_1 = \begin{pmatrix} a_1 & a_2 \\ \cdot & \cdot \end{pmatrix},$$

with $N(a_2) < N(a_1)$. We thus construct a sequence of matrices M_k of the form

$$\begin{pmatrix} a_{k-1} & a_k \\ \cdot & \cdot \end{pmatrix},$$

with $a_{k-1} \neq 0$, each one the product of the previous one by elementary matrices. Furthermore, $N(a_k) < N(a_{k-1})$. From Proposition 6.1.2, this sequence is finite, and there is a step for which $a_k = 0$. The matrix M_k, being triangular and invertible, has an invertible diagonal D. Then M_kD^{-1} has the form

$$\begin{pmatrix} 1 & 0 \\ \cdot & 1 \end{pmatrix},$$

which is an elementary matrix.

■

Again, the statement is false in a general principal ideal domain. Whether $\mathbf{GL}_n(A)$ equals the group spanned by elementary matrices is a difficult question of Ktheory.

6.2 Invariant Factors of a Matrix

Theorem 6.2.1 *Let $M \in \mathbf{M}_{n\times m}(A)$ be a matrix with entries in a principal ideal domain. Then there exist two invertible matrices $P \in \mathbf{GL}_n(A)$, $Q \in \mathbf{GL}_m(A)$ and a quasi-diagonal matrix $D \in \mathbf{M}_{n\times m}(A)$ (that is, $d_{ij} = 0$ for $i \neq j$) such that:*

- *on the one hand, $M = PDQ$,*
- *on the other hand, $d_1|d_2, \ldots, d_i|d_{i+1}, \ldots$, where the d_j are the diagonal entries of D.*

Furthermore, if $M = P'D'Q'$ is another decomposition with these two properties, the scalars d_j and d'_j are associated. Up to invertible elements, they are thus unique.

Definition 6.2.1 *For this reason, the scalars $d_1, \ldots, d_r$ ($r = \min(n, m)$) are called the* invariant factors *of M.*

Proof

Uniqueness: for $k \leq r$, let us denote by $D_k(N)$ the gcd of minors of order k of the matrix N. From Corollary 2.1.1, we have $D_k(M) = D_k(D) = D_k(D')$. It is immediate that $D_k(D) = d_1 \cdots d_k$ (because the minors of order k are either null, or products of k terms d_j with distinct subscripts), so that

$$d_1 \cdots d_k = u_k d'_1 \cdots d'_k, \quad 1 \leq k \leq r,$$

for some $u_k \in A^*$. Hence, d_1 and d'_1 are associated. Since A is an integral domain, we also have $d'_k = u_k^{-1} u_{k-1} d_k$. In other words, d_k and d'_k are associated.

Existence: We see from the above that the d_j's are determined by the equalities $d_1 \cdots d_j = D_j(M)$. In particular, d_1 is the gcd of the entries of M. Hence the first step consists in finding a matrix M', equivalent to M, such that m'_{11} is equal to this gcd.

To do so, we construct a sequence of equivalent matrices $M^{(p)}$, with $M^{(0)} = M$, such that $m_{11}^{(p)}$ divides $m_{11}^{(p-1)}$. Given the matrix $N := M^{(p-1)}$, we distinguish four cases:

1. n_{11} divides $n_{11}, \ldots, n_{1,j-1}$, but does not divide n_{1j}. Then $d := \gcd(n_{11}, n_{1j})$ reads $d = un_{11} + vn_{1j}$. Let us define $w := -n_{1j}/d$ and $z := n_{11}/d$ and let us define a matrix $Q \in \mathbf{GL}_m(A)$ by:

   - $q_{11} = u$, $q_{j1} = v$, $q_{1j} = w$, $q_{jj} = z$,
   - $q_{kl} = \delta_k^l$, otherwise.

   Then $M^{(p)} := M^{(p-1)}Q$ is suitable, because $m_{11}^{(p)} = d | n_{11} = m_{11}^{(p-1)}$.

2. n_{11} divides each n_{1j}, as well as $n_{11}, \ldots, n_{i-1,1}$, but does not divide n_{i1}. This case is symmetric to the previous one. Multiplication on the right by a suitable $P \in \mathbf{GL}_n(A)$ furnishes $M^{(p)}$, with $m_{11}^{(p)} = \gcd(n_{11}, n_{i1}) | m_{11}^{(p-1)}$.

3. n_{11} divides each n_{1j} and each n_{i1}, but does not divide some n_{ij} with $i, j \geq 2$. Then $n_{i1} = an_{11}$. Let us define a matrix $P \in \mathbf{GL}_n(A)$ by

   - $p_{11} = a + 1$, $p_{i1} = 1$, $p_{1i} = -1$, $p_{ii} = 0$;
   - $p_{kl} = \delta_k^l$, otherwise;

   If we then set $N' = PN$, we have $n'_{11} = n_{11}$ and $n'_{1j} = (a+1)n_{1j} - n_{ij}$. We have thus returned to the first case, and there exists an equiv-

alent matrix $M^{(p)}$, with $m_{11}^{(p)} = \gcd(n'_{11}, n'_{1j}) = \gcd(n_{11}, n_{ij}) | n_{11} = m_{11}^{(p-1)}$.

4. n_{11} divides all the entries of the matrix N. In that case, $M^{(p)} := M^{(p-1)}$.

It is essential to observe that in the first three cases, $m_{11}^{(p)}$ is not associated to $m_{11}^{(p-1)}$, though it divides it.

From Proposition 6.1.2, the elements of the sequence $\left(m_{11}^{(p)}\right)_{p\geq 0}$ are pairwise associated, once p is large enough. We are then in the last of the four cases above: $m_{11}^{(q)}$ divides all the $m_{ij}^{(q)}$'s. We have $m_{i1}^{(q)} = a_i m_{11}^{(q)}$ and $m_{1j}^{(q)} = b_j m_{11}^{(q)}$. Then let $P \in \mathbf{GL}_n(A)$ and $Q \in \mathbf{GL}_m(A)$ be the matrices defined by:

- $p_{ii} = 1$, $p_{i1} = -a_i$ if $i \geq 2$, $p_{ij} = 0$ otherwise,
- $q_{jj} = 1$, $q_{1j} = -b_j$ if $j \geq 2$, $q_{ij} = 0$ otherwise.

The matrix $M' := PM^{(q)}Q$ is equivalent to $M^{(q)}$, hence to M. It has the form

$$M' = \begin{pmatrix} m & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & M'' & \\ 0 & & & \end{pmatrix},$$

where m divides all the entries of M''. Obviously, $m = D_1(M') = D_1(M)$.

Having shown that every matrix M is equivalent to a matrix of the form described above, one may argue by induction on the size of M (that is, on the integer $r = \min(n, m)$). If $r = 1$, we have just proved the claim. If $r \geq 2$ and if the claim is true up to the order $r - 1$, we apply the induction hypothesis to the factor $M'' \in \mathbf{M}_{(n-1)\times(m-1)}(A)$ in the above reduction: there exist $P'' \in \mathbf{GL}_{n-1}(A)$ and $Q'' \in \mathbf{GL}_{m-1}(A)$ such that $P''M''Q''$ is quasi-diagonal, with diagonal entries $d_2, \ldots, d_r$ ordered by $d_l | d_{l+1}$ for $l \geq 2$. From the uniqueness step, $d_2 = D_1(M'')$. Since m divides the entries of M'', we have $m | d_2$. Let us then define $P' = \mathrm{diag}(1, P'')$ and $Q' = \mathrm{diag}(1, Q'')$, which are invertible: $P'M'Q'$ is quasi-diagonal, with diagonal entries $d_1 = m, d_2, \ldots,$ a nondecreasing sequence (according to the division in A). Since M is equivalent to M', this proves the existence part of the theorem.

■

6.2.1 Comments

In the list of invariant factors of a matrix some d_j's may equal zero. In that case, $d_j = 0$ implies $d_{j+1} = \cdots = d_r = 0$. Moreover, some invariant

factor may occur several times in the list $d_1, \dots, d_r$, up to association. The number of times that a factor d or its associates occur is its *multiplicity*.

If $m = n$ and if the invariant factors of a matrix M are $(1, \cdots, 1)$, then $D = I_n$, and $M = PQ$ is invertible. Conversely, if M is invertible, then the decomposition $M = MI_nI_n$ shows that $d_1 = \cdots = d_n = 1$.

If A is a field, then there are only two ideals: $A = (1)$ itself and (0). The list of invariant factors of a matrix is thus of the form $(1, \dots, 1, 0, \dots, 0)$. Of course, there may be no 1's (for the matrix $0_{m\times n}$), or no 0's. There are thus exactly $\min(n, m) + 1$ classes of equivalent matrices in $\mathbf{M}_n(A)$, two matrices being equivalent if and only if they have the same rank q. The rank is then the number of 1's among the invariant factors. The decomposition $M = PDQ$ is then called the *rank decomposition*.

Theorem 6.2.2 *Let k be a field and $M \in \mathbf{M}_{n\times m}(k)$ a matrix. Let q be the rank of M, that is, the dimension of the linear subspace of k^n spanned by the columns of M. Then there exist two square invertible matrices P, Q such that $M = PDQ$ with $d_{ii} = 1$ if $i \leq q$ and $d_{ij} = 0$ in all other cases.*

6.3 Similarity Invariants and Jordan Reduction

From now on, k will denote a field and $A = k[X]$ the ring of polynomials over k. This ring is Euclidean, hence a principal ideal domain. In the sequel, the results are **effective**, in the sense that the normal forms that we define will be obtained by means of an algorithm that uses right or left multiplications by elementary matrices of $\mathbf{M}_n(A)$, the computations being based upon the Euclidean division of polynomials.

Given a matrix $B \in \mathbf{M}_n(k)$ (a square matrix with constant entries, in the sense that they are not polynomials), we consider the matrix $XI_n - B \in \mathbf{M}_n(A)$, where X is the indeterminate in A.

Definition 6.3.1 *If $B \in \mathbf{M}_n(k)$, the invariant factors of $M := XI_n - B$ are called* invariant polynomials *of B, or* similarity invariants *of B.*

This definition is justified by the following statement.

Theorem 6.3.1 *Two matrices in $\mathbf{M}_n(k)$ are similar if and only if they have the same list of invariant polynomials (counted with their multiplicities).*

This theorem is a particular case of a more general one:

Theorem 6.3.2 *Let A_0, A_1, B_0, B_1 be matrices in $\mathbf{M}_n(k)$, with A_0, A_1. Then the matrices $XA_0 + B_0$ and $XA_1 + B_1$ are equivalent (in $\mathbf{M}_n(A)$) if and only if there exist $G, H \in \mathbf{GL}_n(k)$ such that*

$$GA_0 = A_1H, \quad GB_0 = B_1H.$$

When $A_0 = A_1 = I_n$, Theorem 6.3.2 tells that $XI_n - B_0$ and $XI_n - B_1$ are equivalent, namely that they have the same invariant polynomials, if there exists $P \in \mathbf{GL}_n(k)$ such that $PB_0 = B_1P$, which is the criterion given by Theorem 6.3.1.

Proof

We prove Theorem 6.3.2. The condition is clearly sufficient.

Conversely, if XA_0+B_0 and XA_1+B_1 are equivalent, there exist matrices $P, Q \in \mathbf{GL}_n(A)$, such that $P(XA_0 + B_0) = (XA_1 + B_1)Q$. Since A_1 is invertible, one may perform Euclidean division[2] of P by $XA_1 + B_1$ on the right:

$$P = (XA_1 + B_1)P_1 + G,$$

where G is a matrix whose entries are constant polynomials. We warn the reader that since $\mathbf{M}_n(k)$ is not commutative, Euclidean division may be done either on the right or on the left, with distinct quotients and distinct remainders. Likewise, we have $Q = Q_1(XA_0 + B_0) + H$ with $H \in \mathbf{M}_n(k)$. Let us write, then,

$$(XA_1 + B_1)(P_1 - Q_1)(XA_0 + B_0) = (XA_1 + B_1)H - G(XA_0 + B_0).$$

The left-hand side of this equality has degree (the degree is defined as the supremum of the degrees of the entries of the matrix) $2 + \deg(P_1 - Q_1)$, while the right-hand side has degree less than or equal to one. The two sides, being equal, must vanish, and we conclude that

$$GA_0 = A_1H, \quad GB_0 = B_1H.$$

There remains to show that G and H are invertible. To do so, let us define $R \in \mathbf{M}_n(A)$ as the inverse matrix of P (which exists by assumption). We still have

$$R = (XA_0 + B_0)R_1 + K, \quad K \in \mathbf{M}_n(k).$$

Combining the equalities stated above, we obtain

$$I_n - GK = (XA_1 + B_1)(QR_1 + P_1K).$$

Since the left-hand side is constant and the right-hand side has degree $1 + \deg(QR_1 + P_1K)$, we must have $I_n = GK$, so that G is invertible. Likewise, H is invertible.

■

We conclude this paragraph with a remarkable statement:

Theorem 6.3.3 *If $B \in \mathbf{M}_n(k)$, then B and B^T are similar.*

Indeed, $XI_n - B$ and $XI_n - B^T$ are transposes of each other, and hence have the same list of minors, hence the same invariant factors.

---

[2]The fact that A_1 is invertible is essential, since the ring $\mathbf{M}_n(A)$ is not an integral domain.

6.3.1 Example: The Companion Matrix of a Polynomial

Given a polynomial

$$P(X) = X^n + a_1 X^{n-1} + \cdots + a_n,$$

there exists a matrix $B \in \mathbf{M}_n(k)$ such that the list of invariant factors of the matrix $XI_n - B$ is $(1, \ldots, 1, P)$. We may take the *companion matrix* associated to P to be

$$B_P := \begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_n \\ 1 & \ddots & & \vdots & \vdots \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -a_1 \end{pmatrix}.$$

Naturally, any matrix similar to B_P would do as well, because if $B = Q^{-1}B_P Q$, then $XI_n - B$ is similar, hence equivalent, to $XI_n - B_P$. In order to show that the invariant factors of B_P are the polynomials $(1, \ldots, 1, P)$, we observe that $XI_n - B_P$ possesses a minor of order $n-1$ that is invertible, namely, the determinant of the submatrix

$$\begin{pmatrix} -1 & X & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & X \\ 0 & \cdots & \cdots & 0 & -1 \end{pmatrix}.$$

We thus have $D_{n-1}(XI_n - B_P) = 1$, so that the invariant factors $d_1, \ldots, d_{n-1}$ are all equal to 1. Hence $d_n = D_n(XI_n - B_P) = \det(XI_n - B_P)$, the characteristic polynomial of B_P, namely P.

In this example P is also the minimal polynomial of B_P. In fact, if Q is a polynomial of degree less than or equal to $n-1$,

$$Q(X) = b_0 X^{n-1} + \cdots + b_{n-1},$$

the vector $Q(A)\mathbf{e}^1$ reads

$$b_0 \mathbf{e}^n + \cdots + b_{n-1}\mathbf{e}^1.$$

Hence $Q(A) = 0$ and $\deg Q \leq n-1$ imply $Q = 0$. The minimal polynomial is thus of degree at least n. It is thus equal to the characteristic polynomial.

6.3.2 First Canonical Form of a Square Matrix

Let $M \in \mathbf{M}_n(k)$ be a square matrix and $P_1, \ldots, P_n \in k[X]$ its similarity invariants. The sum of their degrees n_j $(1 \leq j \leq n)$ is n. Let us denote

by $M^{(j)} \in \mathbf{M}_{n_j}(k)$ the companion matrix of the polynomial P_j. Let us form the matrix M', block-diagonal, whose diagonal blocks are the $M^{(j)}$'s. The few first polynomials P_j are generally constant (we shall see below that the only case where P_1 is not constant corresponds to $M = \alpha I_n$), and the corresponding blocks are empty, as are the corresponding rows and columns. To be precise, the actual number m of diagonal blocks is equal to the nuber of nonconstant similarity invariants.

Since the matrix $XI_{n_j} - M^{(j)}$ is equivalent to the matrix $N^{(j)} = \mathrm{diag}(1, \dots, 1, P_j)$, we have

$$XI_{n_j} - M^{(j)} = P^{(j)} N^{(j)} Q^{(j)},$$

where $P^{(j)}, Q^{(j)} \in \mathbf{GL}_{n_j}(k[X])$. Let us form matrices $P, Q \in \mathbf{GL}_n(k[X])$ by

$$P = \mathrm{diag}(P^{(1)}, \dots, P^{(n)}), \quad Q = \mathrm{diag}(Q^{(1)}, \dots, Q^{(n)}).$$

We obtain

$$XI_n - M' = PNQ, \quad N = \mathrm{diag}(N^{(1)}, \dots, N^{(n)}).$$

Here N is a diagonal matrix, whose diagonal entries are the similarity invariants of M, up to the order. In fact, each nonconstant P_j appears in the associated block $N^{(j)}$. The other diagonal terms are the constant 1, which occurs $n - m$ times; these are the polynomials $P_1, \dots, P_{n-m}$, as expected. Conjugating by a permutation matrix, we obtain that $XI_n - M'$ is equivalent to the matrix $\mathrm{diag}(P_1, \dots, P_n)$. Hence $XI_n - M'$ is equivalent to $XI_n - M$. From Theorem 6.3.1, M and M' are similar.

Theorem 6.3.4 *Let k be a field, $M \in \mathbf{M}_n(k)$ a square matrix, and $P_1, \dots, P_n$ its similarity invariants. Then M is similar to the block-diagonal matrix M' whose jth diagonal block is the companion matrix of P_j.*

The matrix M' is called the first canonical form *of M, or the* Frobenius canonical form *of M.*

Remark: If L is an extension of k (namely, a field containing k) and $M \in \mathbf{M}_n(k)$, then $M \in \mathbf{M}_n(L)$. Let $P_1, \dots, P_n$ be the similarity invariants of M as a matrix with entries in k. Then $XI_n - M = P\,\mathrm{diag}(P_1, \dots, P_n)Q$, where $P, Q \in \mathbf{GL}_n(k[X])$. Since P, Q, their inverses, and the diagonal matrix also belong to $\mathbf{M}_n(L[X])$, $P_1, \dots, P_n$ are the similarity invariants of M as a matrix with entries in L. In other words, the similarity invariants depend on M but not on the field k. To compute them, it is enough to place ourselves in the smallest possible field, namely that spanned by the entries of M. The same remark holds true for the first canonical form. As we shall see in the next section, it is no longer true for the second canonical form, which is therefore *less* canonical.

We end this paragraph with a characterization of the minimal polynomial.

Theorem 6.3.5 *Let k be a field, $M \in \mathbf{M}_n(k)$ a square matrix, and $P_1, \ldots, P_n$ its similarity invariants. Then P_n is the minimal polynomial of M. In particular, the minimal polynomial does not depend on the field under consideration, as long as it contains the entries of M.*

Proof

We use the first canonical form M' of M. Since M' and M are similar, they have the same minimal polynomial. One thus can assume that M is in the canonical form $M = \mathrm{diag}(M_1, \ldots, M_n)$, where M_j is the companion matrix of P_j. Since $P_j(M_j) = 0$ (Cayley–Hamilton, theorem 2.5.1) and $P_j | P_n$, we have $P_n(M_j) = 0$ and thus $P_n(M) = 0_n$. Hence, the minimal polynomial Q_M divides P_n. Conversely, $Q(M) = 0_n$ implies $Q(M_n) = 0$. Since P_n is the minimal polynomial of M_n, P_n divides Q. Finally, $P_n = Q_M$.

Finally, since the similarity invariants do not depend on the choice of the field, P_n also does not depend on this choice.

■

Warning: One may draw an incorrect conclusion if one applies Theorem 6.3.5 carelessly. Given a matrix $M \in \mathbf{M}_n(\mathbb{Z})$, one can define a matrix $M_{(p)}$ in $\mathbf{M}_n(\mathbb{Z}/p\mathbb{Z})$ by reduction modulo p (p a prime number). But the minimal polynomial of $M_{(p)}$ is not necessarily the reduction modulo p of Q_M. Here is an example: Let us take $n = 2$ and

$$M = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}.$$

Then Q_M divides $P_M = (X-2)^2$, but $Q_M \neq X - 2$, since $M \neq 2I_2$. Hence $Q_M = (X-2)^2$. On the other hand, $M_{(2)} = 0_2$, whose minimal polynomial is X, which is different from X^2, the reduction modulo 2 of Q_M.

The explanation of this phenomenon is the following. The matrices M and $M_{(2)}$ are composed of scalars of different natures. There is no field L containing simultaneously $\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z}$. There is thus no context in which Theorem 6.3.5 could be applied.

6.3.3 Second Canonical Form of a Square Matrix

We now decompose the similarity invariants of M into products of irreducible polynomials. This decomposition depends, of course, on the choice of the field of scalars. Denoting by $p_1, \ldots, p_t$ the list of distinct irreducible (in $k[X]$) factors of P_n, we have

$$P_j = \prod_{k=1}^{t} p_k^{\alpha(j,k)}, \quad 1 \leq j \leq n$$

(because P_j divides P_n), where the $\alpha(j,k)$ are nondecreasing with respect to j, since P_j divides P_{j+1}.

Definition 6.3.2 *The* elementary divisors *of the matrix $M \in \mathbf{M}_n(k)$ are the polynomials $p_k^{\alpha(j,k)}$ for which the exponent $\alpha(j,k)$ is nonzero. The* multiplicity *of an elementary divisor p_k^m is the number of solutions j of the equation $\alpha(j,k) = m$. The* list *of elementary divisors is the sequence of these polynomials, repeated with their multiplicities.*

Let us begin with the case of the companion matrix N of some polynomial P. Its similarity invariants are $(1, \dots, 1, P)$ (see above). Let $Q_1, \dots, Q_t$ be its elementary divisors (we observe that each has multiplicity one). We then have $P = Q_1 \cdots Q_t$, while the Q_l's are pairwise coprime. To each Q_l we associate its companion matrix N_l, and we form a block-diagonal matrix $N' := \mathrm{diag}(N_1, \dots, N_t)$. Since each $N_l - XI_l$ is equivalent to a diagonal matrix

$$\begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & Q_l \end{pmatrix}$$

in $\mathbf{M}_{n(l)}(k[X])$, the whole matrix $N' - XI_n$ is equivalent to

$$Q := \begin{pmatrix} 1 & & & & & \\ & \ddots & & & O & \\ & & 1 & & & \\ & & & Q_1 & & \\ & O & & & \ddots & \\ & & & & & Q_t \end{pmatrix}.$$

Let us now compute the similarity invariants of N', that is, the invariant factors of Q. It will be enough to compute the greatest common divisor D_{n-1} of the minors of size $n-1$. Taking into account the principal minors of Q, we see that D_{n-1} must divide every product of the form

$$\prod_{l \neq k} Q_l, \quad 1 \leq k \leq t.$$

Since the Q_l's are pairwise coprime, this implies that $D_{n-1} = 1$. This means that the list of similarity invariants of N' has the form $(1, \dots, 1, \cdot)$, where the last polynomial must be the characteristic polynomial of N'. This polynomial is the product of the characteristic polynomials of the N_l's. These being equal to the Q_l's, the characteristic polynomial of N' is P. Finally, N and N' have the same similarity invariants and are therefore similar.

Now let M be a general matrix in $\mathbf{M}_n(k)$. We apply the former reduction to every diagonal block M_j of its Frobenius canonical form. Each M_j is similar to a block-diagonal matrix whose diagonal blocks are companion matrices corresponding to the elementary divisors of M entering into the

factorization of the jth invariant polynomial of M. We have thus proved the following statement.

Theorem 6.3.6 *Let $Q_1, \dots, Q_s$ be the elementary divisors of $M \in \mathbf{M}_n(k)$. Then M is similar to a block-diagonal matrix M' whose diagonal blocks are companion matrices of the Q_l's.*

The matrix M' is called the second canonical form *of M.*

Remark: The exact computation of the second canonical form of a given matrix is impossible in general, in contrast to the case of the first form. Indeed, if there were an algorithmic construction, it would provide an algorithm for factorizing polynomials into irreducible factors via the formation of the companion matrix, a task known to be impossible if $k = \mathbb{R}$ or $\mathbb{C}$. Recall that one of the most important results in Galois theory, known as Abel's theorem, states the impossibility of solving a general polynomial equation of degree at least five with complex coefficients, using only the basic operations and the extraction of roots of any order.

6.3.4 Jordan Form of a Matrix

When the characteristic polynomial splits over k, which holds, for instance, if the field k is algebraically closed, the elementary divisors have the form $(X-a)^r$ for $a \in k$ and $r \geq 1$. In that case, the second canonical form can be greatly simplified by replacing the companion matrix of the monomial $(X-a)^r$ by its *Jordan block*

$$J(a;r) := \begin{pmatrix} a & 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & \cdots & 0 & a \end{pmatrix}.$$

In fact, the characteristic polynomial of $J(a;r)$ (of size $r \times r$) is $(X-a)^r$, while the matrix $XI_r - J(a;r)$ possesses an invertible minor of order $r-1$, namely

$$\begin{pmatrix} -1 & 0 & \cdots & 0 \\ X-a & \ddots & \ddots & \vdots \\ & \ddots & \ddots & 0 \\ & & X-a & -1 \end{pmatrix},$$

which is obtained by deleting the first column and the last row. Again, this shows that $D_{n-1}(XI_r - J) = 1$, so that the invariant factors $d_1, \dots, d_{r-1}$ are equal to 1. Hence $d_r = D_r(XI_r - J) = \det(XI_r - J) = (X-a)^r$. Its

invariant factors are thus $1, \ldots, 1, (X-a)^r$. Hence we have the following theorem.

Theorem 6.3.7 *When an elementary divisor of M is $(X-a)^r$, one may, in the second canonical form of M, replace its companion matrix by the Jordan block $J(a;r)$.*

Corollary 6.3.1 *If the characteristic polynomial of M splits over k, then M is similar to a block-diagonal matrix whose jth diagonal block is a Jordan block $J(a_j;r_j)$. This form is unique, up to the order of blocks.*

Corollary 6.3.2 *If k is algebraically closed (for example if $k=\mathbb{C}$), then every square matrix M is similar to a block-diagonal matrix whose jth diagonal block is a Jordan block $J(a_j;r_j)$. This form is unique, up to the order of blocks.*

6.4 Exercises

See also the exercise 12 in Chapter 7.

1. Show that every principal ideal domain is a unique factorization domain.

2. Verify that the characteristic polynomial of the companion matrix of a polynomial P is equal to P.

3. Let k be a field and $M \in \mathbf{M}_n(k)$. Show that M, M^T have the same rank and that in general, the rank of $M^T M$ is less than or equal to that of M. Show that the equality of these ranks always holds if $k=\mathbb{R}$, but that strict inequality is possible, for example with $k=\mathbb{C}$.

4. Compute the elementary divisors of the matrices

$$\begin{pmatrix} 22 & 23 & 10 & -98 \\ 12 & 18 & 16 & -38 \\ -15 & -19 & -13 & 58 \\ 6 & 7 & 4 & -25 \end{pmatrix}, \quad \begin{pmatrix} 0 & -21 & -56 & -96 \\ 18 & 36 & 52 & -8 \\ -12 & -17 & -16 & 38 \\ 3 & 2 & -2 & -20 \end{pmatrix}$$

and

$$\begin{pmatrix} 44 & 89 & 120 & -32 \\ 0 & -12 & -32 & -56 \\ -14 & -20 & -16 & 49 \\ 8 & 14 & 16 & -16 \end{pmatrix}$$

in $\mathbf{M}_n(\mathbb{C})$. What are their Jordan reductions?

5. (Lagrange's theorem)

Let K be a field and $A \in \mathbf{M}_n(K)$. Let $X, Y \in K^n$ be vectors such that $X^T AY \neq 0$. We normalize by $X^T AY = 1$ and define

$$B := A - (AY)(X^T A).$$

Show that in the factorization

$$PAQ = \begin{pmatrix} I_r & 0 \\ 0 & 0_{n-r} \end{pmatrix}, \quad P, Q \in \mathbf{GL}_n(K),$$

one can choose Y as the first column of Q and X^T as the first row of P. Deduce that $\operatorname{rk} B = \operatorname{rk} A - 1$.
More generally, show that if $X, Y \in \mathbf{M}_{n\times m}(K)$, $X^T AY \in \mathbf{GL}_m(K)$, and if

$$B := A - (AY)(X^T AY)^{-1}(X^T A),$$

then $\operatorname{rk} B = \operatorname{rk} A - m$.
If $A \in \mathbf{Sym}_n(\mathbb{R})$ and if A is positive semidefinite, and if $X = Y$, show that B is also positive semidefinite.

6. For $A \in \mathbf{M}_n(\mathbb{C})$, consider the linear differential equation in $\mathbb{C}^n$

$$\frac{dx}{dt} = Ax. \tag{6.1}$$

   (a) Let $P \in \mathbf{GL}_n(\mathbb{C})$ and let $t \mapsto x(t)$ be a solution of (6.1). What is the differential equation satisfied by $t \mapsto Px(t)$?
   (b) Let $(X-a)^m$ be an elementary divisor of A. Show that for every $k = 0, \dots, m-1$, (6.1) possesses solutions of the form $e^{at}Q_k(t)$, where Q_k is a complex-valued polynomial map of degree k.

7. Consider the following differential equation of order n in $\mathbb{C}$:

$$x^{(n)}(t) + a_1 x^{(n-1)}(t) + \cdots + a_n x(t) = 0. \tag{6.2}$$

   (a) Define $P(X) = X^n + a_1 X^{n-1} + \cdots + a_n$ and let M be the companion matrix of P. Let

$$P(X) = \prod_{a\in A} (X-a)^{n_a}$$

   be the factorization of P into irreducible factors. Compute the Jordan form of M.
   (b) Using either the previous exercise or arguing directly, show that the set of solutions of (6.2) is spanned by the solutions of the form

$$t \mapsto e^{at}R(t), \quad R \in \mathbb{C}[X], \quad \deg R < n_a.$$

8. Consider a linear recursion of order n in a field K

$$u_{m+n} + a_1 u_{m+n-1} + \cdots + a_n u_m = 0, \quad m \in \mathbb{N}. \tag{6.3}$$

With the notation of the previous exercise, show that the set of solutions of (6.3) is spanned by the solutions of the form

$$(a^m R(m))_{m\in\mathbb{N}}, \quad R \in \mathbb{C}[X], \quad \deg R < n_a.$$

9. Let $n \geq 2$ and let $M \in \mathbf{M}_n(\mathbb{Z})$ be the matrix defined by $m_{ij} = i + j - 1$:

$$M = \begin{pmatrix} 1 & 2 & \cdots & n \\ 2 & \iddots & \iddots & \vdots \\ \vdots & \iddots & \iddots & \vdots \\ n & \cdots & \cdots & 2n-1 \end{pmatrix}.$$

   (a) Show that M has rank 2 (you may look for two vectors $x, y \in \mathbb{Z}^n$ such that $m_{ij} = x_i x_j - y_i y_j$).
   (b) Compute the invariant factors of M in $\mathbf{M}_n(\mathbb{Z})$ (the equivalent diagonal form is obtained after five elementary operations).

10. The ground field is $\mathbb{C}$.

   (a) Define

$$N = J(0;n), \quad B = \begin{pmatrix} & \cdots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & \end{pmatrix}.$$

   Compute NB, BN, and BNB. Show that $S := \frac{1}{\sqrt{2}}(I + iB)$ is unitary.
   (b) Deduce that N is similar to

$$\frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} + \frac{i}{2}\begin{pmatrix} 0 & \cdots & 0 & -1 & 0 \\ \vdots & \iddots & \iddots & \iddots & 1 \\ 0 & \iddots & \iddots & \iddots & 0 \\ -1 & \iddots & \iddots & \iddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

   (c) Deduce that every matrix $M \in \mathbf{M}_n(\mathbb{C})$ is similar to a complex symmetric matrix. Compare with the real case.

7

Exponential of a Matrix, Polar Decomposition, and Classical Groups

7.1 The Polar Decomposition

The polar decomposition of matrices is defined by analogy with that in the complex plane: If $z \in \mathbb{C}^*$, there exists a unique pair $(r, q) \in (0, +\infty) \times S^1$ (S^1 denotes the unit circle, the set of complex numbers of modulus 1) such that $z = rq$. If z acts on $\mathbb{C}$ (or on $\mathbb{C}^*$) by multiplication, this action can be decomposed as the product of a rotation of angle θ (where $q = \exp(i\theta)$) with a homothety of ratio $r > 0$. The fact that these two actions commute is a consequence of the commutativity of the multiplicative group $\mathbb{C}^*$; this property does not hold for the polar decomposition in $\mathbf{GL}_n(k)$, $k = \mathbb{R}$ or $\mathbb{C}$, because the general linear group is not commutative.

Let us recall that $\mathbf{HPD}_n$ denotes the (open) cone of matrices of $\mathbf{M}_n(\mathbb{C})$ that are Hermitian positive definite, while $\mathbf{U}_n$ denotes the group of unitary matrices. In $\mathbf{M}_n(\mathbb{R})$, $\mathbf{SPD}_n$ is the set of symmetric positive definite matrices, and $\mathbf{O}_n$ is the orthogonal group. The group $\mathbf{U}_n$ is compact, since it is closed and bounded in $\mathbf{M}_n(\mathbb{C})$. Indeed, the columns of unitary matrices are unit vectors, so that $\mathbf{U}_n$ is bounded. On the other hand, $\mathbf{U}_n$ is defined by an equation $U^*U = I_n$, where the map $U \mapsto U^*U$ is continuous; hence $\mathbf{U}_n$ is closed. By the same arguments, $\mathbf{O}_n$ is compact.

Polar decomposition is a fundamental tool in the theory of finite-dimensional Lie groups and Lie algebras. For this reason, it is intimately related to the *exponential* map. We shall not consider these two notions here in their full generality, but we shall restrict attention to their matricial aspects.

Theorem 7.1.1 *For every $M \in \mathbf{GL}_n(\mathbb{C})$, there exists a unique pair*

$$(H, Q) \in \mathbf{HPD}_n \times \mathbf{U}_n$$

such that $M = HQ$. If $M \in \mathbf{GL}_n(\mathbb{R})$, then $(H, Q) \in \mathbf{SPD}_n \times \mathbf{O}_n$.

The map $M \mapsto (H, Q)$, called the polar decomposition *of M, is a homeomorphism between $\mathbf{GL}_n(\mathbb{C})$ and $\mathbf{HPD}_n \times \mathbf{U}_n$ (respectively between $\mathbf{GL}_n(\mathbb{R})$ and $\mathbf{SPD}_n \times \mathbf{O}_n$).*

Theorem 7.1.2 *Let H be a positive definite Hermitian matrix. There exists a unique positive definite Hermitian matrix h such that $h^2 = H$. If H is real-valued, then so is h. The matrix h is called the* square root *of H, and is denoted by $h = \sqrt{H}$.*

Proof

We prove Theorem 7.1.1 and obtain Theorem 7.1.2 as a by-product.

Existence. Since $MM^* \in \mathbf{HPD}_n$, we can diagonalize MM^* by a unitary matrix

$$MM^* = U^* D U, \quad D = \mathrm{diag}(d_1, \dots, d_n),$$

where $d_j \in (0, +\infty)$. The matrix $H := U^* \mathrm{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})U$ is Hermitian positive definite and satisfies $H^2 = HH^* = MM^*$. Then $Q := H^{-1}M$ satisfies $Q^*Q = M^*H^{-2}M = M^*(MM^*)^{-1}M = I_n$, hence $Q \in \mathbf{U}_n$. If $M \in \mathbf{M}_n(\mathbb{R})$, then clearly MM^* is real symmetric. In fact, U is orthogonal and H is real symmetric. Hence Q is real orthogonal. **Note**: H is called the *square root* of MM^*.

Uniqueness. Let $M = H'Q'$ be another suitable decomposition. Then $N := H^{-1}H' = Q(Q')^{-1}$ is unitary, so that $\mathrm{Sp}(N) \subset S^1$. Let $S \in \mathbf{HPD}_n$ be a positive definite Hermitian square root of H' (we shall prove below that it is unique). Then N is similar to $N' := SH^{-1}S$. However, $N' \in \mathbf{HPD}_n$. Hence N is diagonalizable, with real positive eigenvalues. Hence $\mathrm{Sp}(N) = \{1\}$, and N is therefore similar, and thus equal, to I_n.

This proves that the positive definite Hermitian square root of a matrix of $\mathbf{HPD}_n$ is unique in $\mathbf{HPD}_n$, since otherwise, our construction would provide several polar decompositions. We have thus proved Theorem 7.1.2 in passing.

Smoothness. The map $(H, Q) \mapsto HQ$ is polynomial, hence continuous. Conversely,, it is enough to prove that $M \mapsto (H, Q)$ is sequentially continuous, since $\mathbf{GL}_n(\mathbb{C})$ is a metric space. Let $(M_k)_{k \in \mathbb{N}}$ be a convergent sequence in $\mathbf{GL}_n(\mathbb{C})$ and let M be its limit. Let us denote by $M_k = H_k Q_k$ and $M = HQ$ their respective polar decompositions. Let R be a cluster point of the sequence $(Q_k)_{k \in \mathbb{N}}$, that is, the limit of some subsequence $(Q_{k_l})_{l \in \mathbb{N}}$, with $k_l \to +\infty$. Then $H_{k_l} = M_{k_l} Q_{k_l}^*$ converges to $S := MR^*$. The matrix S is Hermitian positive semidefinite (because it is the limit of the H_{k_l}'s) and invertible (because it is the product of M and R^*). It is thus positive definite. Hence, SR is a polar decomposition of M. The uniqueness part ensures that $R = Q$ and $S = H$. The sequence $(Q_k)_{k \in \mathbb{N}}$,

which is relatively compact and has at most one cluster point (namely Q), converges to Q. Finally, $H_k = M_k Q_k^*$ converges to $MQ^* = H$.

■

Remark: There is as well a polar decomposition $M = QH$ with the same properties. We shall use one or the other depending on the context. We warn the reader, however, that for a given matrix, the two decompositions do not coincide. For example, in $M = HQ$, H is the square root of MM^*, though in $M = QH$, it is the square root of M^*M.

7.2 Exponential of a Matrix

The ground field is here $k = \mathbb{C}$. By restriction, we can also treat the case $k = \mathbb{R}$.

For A in $\mathbf{M}_n(\mathbb{C})$, the series

$$\sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

converges normally (which means that the series of norms is convergent), since for any matrix norm, we have

$$\sum_{k=0}^{\infty} \left\| \frac{1}{k!} A^k \right\| \leq \sum_{k=0}^{\infty} \frac{1}{k!} \|A\|^k = \exp \|A\|.$$

Since $\mathbf{M}_n(\mathbb{C})$ is complete, the series is convergent, and the estimation above shows that it converges uniformly on every compact set. Its sum, denoted by $\exp A$, thus defines a continuous map $\exp : \mathbf{M}_n(\mathbb{C}) \to \mathbf{M}_n(\mathbb{C})$, called the *exponential*. When $A \in \mathbf{M}_n(\mathbb{R})$, we have $\exp A \in \mathbf{M}_n(\mathbb{R})$.

Given two matrices A and B in general position, the binomial formula is not valid: $(A + B)^k$ does not necessarily coincide with

$$\sum_{j=0}^{j=k} \begin{pmatrix} k \\ j \end{pmatrix} A^j B^{k-j}.$$

It thus follows that $\exp(A + B)$ differs in general from $\exp A \cdot \exp B$. A correct statement is the following.

Proposition 7.2.1 *Let $A, B \in \mathbf{M}_n(\mathbb{C})$ be commuting matrices; that is, $AB = BA$. Then* $\exp(A + B) = (\exp A)(\exp B)$.

Proof

The proof proceeds in exactly the same way as for the exponential of complex numbers. We observe that since the series defining the exponential of a matrix is normally convergent, we may compute the product

$(\exp A)(\exp B)$ by multiplying term by term the series

$$(\exp A)(\exp B) = \sum_{j,k=0}^{\infty} \frac{1}{j!k!} A^j B^k.$$

In other words,

$$(\exp A)(\exp B) = \sum_{l=0}^{\infty} \frac{1}{l!} C_l,$$

where

$$C_l := \sum_{j+k=l} \frac{l!}{j!k!} A^j B^k.$$

From the assumption $AB = BA$, we know that the binomial formula holds. Therefore, $C_l = (A+B)^l$, which proves the proposition.

■

Noting that $\exp 0_n = I_n$ and that A and $-A$ commute, we derive the following corollary.

Corollary 7.2.1 *For every $A \in \mathbf{M}_n(\mathbb{C})$, $\exp A$ is invertible, and its inverse is $\exp(-A)$.*

Given two conjugate matrices $B = P^{-1}AP$, we have $B^k = P^{-1}A^kP$ for each integer k and thus

$$\exp(P^{-1}AP) = P^{-1}(\exp A)P. \tag{7.1}$$

If $D = \operatorname{diag}(d_1, \dots, d_n)$ is diagonal, we have

$$\exp D = \operatorname{diag}(\exp d_1, \dots, \exp d_n).$$

Of course, this formula, or more generally (7.1), can be combined with Jordan reduction in order to compute the exponential of a given matrix. Let us keep in mind, however, that Jordan reduction cannot be carried out explicitly.

Let us introduce a real parameter t and let us define a function g by $g(t) = \exp tA$. From Proposition 7.2.1, we see that g satisfies the functional equation

$$g(s+t) = g(s)g(t). \tag{7.2}$$

On the other hand, $g(0) = I_n$, and we have

$$\frac{g(t) - g(0)}{t} - A = \sum_{k=2}^{\infty} \frac{t^{k-1}}{k!} A^k.$$

Using any matrix norm, we deduce that

$$\left\| \frac{g(t) - g(0)}{t} - A \right\| \leq \frac{e^{\|tA\|} - 1 - \|tA\|}{|t|},$$

from which we obtain

$$\lim_{t\to 0} \frac{g(t)-g(0)}{t} = A.$$

We conclude that g has a derivative at $t=0$, with $g'(0)=A$. Using the functional equation (7.2), we then obtain that g is differentiable everywhere, with

$$g'(t) = \lim_{s\to 0} \frac{g(t)g(s)-g(t)}{s} = g(t)A.$$

We observe that we also have

$$g'(t) = \lim_{s\to 0} \frac{g(s)g(t)-g(t)}{s} = Ag(t).$$

From either of these differential equations we see that g is actually infinitely differentiable. We shall retain the formula

$$\frac{d}{dt}\exp tA = A\exp tA = (\exp tA)A. \tag{7.3}$$

This differential equation is sometimes the most practical way to compute the exponential of a matrix. This is of particular relevance when A has real entries but has at least one nonreal eigenvalue if one wishes to avoid the use of complex numbers.

Proposition 7.2.2 *For every $A \in \mathbf{M}_n(\mathbb{C})$,*

$$\det\exp A = \exp \operatorname{Tr} A. \tag{7.4}$$

Proof

We could deduce (7.4) directly from (7.3). Here is a more elementary proof. We begin with a reduction of A of the form $A = P^{-1}TP$, where T is upper triangular. Since T^k is still triangular, with diagonal entries equal to t_{jj}^k, $\exp T$ is triangular too, with diagonal entries equal to $\exp t_{jj}$. Hence

$$\det\exp T = \prod_j \exp t_{jj} = \exp\sum_j t_{jj} = \exp\operatorname{Tr} T.$$

This is the expected formula, since $\exp A = P^{-1}(\exp T)P$.

■

Since $(M^*)^k = (M^k)^*$, we see easily that $(\exp M)^* = \exp(M^*)$. In particular, the exponential of a skew-Hermitian matrix is unitary, for then

$$(\exp M)^*\exp M = \exp(M^*)\exp M = \exp(-M)\exp M = I_n.$$

Similarly, the exponential of a Hermitian matrix is Hermitian positive definite, because

$$\exp M = \left(\exp\frac{1}{2}M\right)^* \exp\frac{1}{2}M.$$

This calculation also shows that if M is Hermitian, then

$$\sqrt{\exp M} = \exp \frac{1}{2} M.$$

We shall use the following more precise statement:

Proposition 7.2.3 *The map* $\exp : \mathbf{H}_n \to \mathbf{HPD}_n$ *is a homeomorphism (that is, a bicontinuous bijection).*

Proof

Injectivity: Let $A, B \in \mathbf{H}_n$ with $\exp A = \exp B =: H$. Then

$$\exp \frac{1}{2} A = \sqrt{H} = \exp \frac{1}{2} B.$$

By induction, we have

$$\exp 2^{-m} A = \exp 2^{-m} B, \quad m \in \mathbb{Z}.$$

Substracting I_n, multiplying by 2^m, and passing to the limit as $m \to +\infty$, we obtain

$$\left. \frac{d}{dt} \right|_{t=0} \exp tA = \left. \frac{d}{dt} \right|_{t=0} \exp tB;$$

that is, $A = B$.

Surjectivity: Let $H \in \mathbf{HPD}_n$ be given. Then $H = U^* \operatorname{diag}(d_1, \ldots, d_n) U$, where U is unitary and $d_j \in (0, +\infty)$. From above, we know that $H = \exp M$ for

$$M := U^* \operatorname{diag}(\log d_1, \ldots, \log d_n) U,$$

which is Hermitian.

Continuity: The continuity of exp has already been proved. Let us investigate the continuity of the reciprocal map. Let $(H^l)_{l \in \mathbb{N}}$ be a sequence in $\mathbf{HPD}_n$ that converges to $H \in \mathbf{HPD}_n$. We denote by $M^l, M \in \mathbf{H}_n$, the Hermitian matrices whose exponentials are H^l and H. The continuity of the spectral radius gives

$$\lim_{l \to +\infty} \rho(H^l) = \rho(H), \quad \lim_{l \to +\infty} \rho\left((H^l)^{-1}\right) = \rho\left((H)^{-1}\right). \tag{7.5}$$

Since $\operatorname{Sp}(M^l) = \log \operatorname{Sp}(M^l)$, we have

$$\rho(M^l) = \log \max\left\{\rho(H^l), \rho\left((H^l)^{-1}\right)\right\}. \tag{7.6}$$

Keeping in mind that the restriction to $\mathbf{H}_n$ of the induced norm $\|\cdot\|_2$ coincides with that of the spectral radius ρ, we deduce from (7.5, 7.6) that the sequence $(M^l)_{l \in \mathbb{N}}$ is bounded. If N is a cluster point of the sequence, the continuity of the exponential implies $\exp N = H$. But the injectivity shown above implies $N = M$. The sequence $(M^l)_{l \in \mathbb{N}}$, bounded with a unique cluster point, is convergent.

■

7.3 Structure of Classical Groups

Proposition 7.3.1 *Let G be a subgroup of $\mathbf{GL}_n(\mathbb{C})$. We assume that G is stable under the map $M \mapsto M^*$ and that for every $M \in G \cap \mathbf{HPD}_n$, the square root $\sqrt{M}$ is an element of G. Then G is stable under polar decomposition. Furthermore, polar decomposition is a homeomorphism between G and*

$$(G \cap \mathbf{U}_n) \times (G \cap \mathbf{HPD}_n).$$

This proposition applies in particular to subgroups of $\mathbf{GL}_n(\mathbb{R})$ that are stable under transposition and under extraction of square roots in $\mathbf{SPD}_n$. One has then

$$G \overset{\text{homeo}}{\sim} (G \cap \mathbf{O}_n) \times (G \cap \mathbf{SPD}_n).$$

Proof

Let $M \in G$ be given and let HQ be its polar decomposition. Since $MM^* \in G$, we have $H^2 \in G$, that is, $H \in G$, by assumption. Finally, we have $Q = H^{-1}M \in G$. An application of Theorem 7.1.1 finishes the proof. ■

We apply this general result to the classical groups $\mathbf{U}(p,q)$, $\mathbf{O}(p,q)$ (where $n = p+q$) and $\mathbf{Sp}_m$ (where $n = 2m$). These are respectively the *unitary* group of the Hermitian form $|z_1|^2 + \cdots + |z_p|^2 - |z_{p+1}|^2 - \cdots - |z_n|^2$, the orthogonal group of the quadratic form $x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_n^2$, and the symplectic group. They are defined by $G = \{M \in \mathbf{M}_n(k) | M^* J M = J\}$, with $k = \mathbb{C}$ for $\mathbf{U}(p,q)$, $k = \mathbb{R}$ otherwise. The matrix J equals

$$\begin{pmatrix} I_p & 0_{p\times q} \\ 0_{q\times p} & -I_q \end{pmatrix},$$

for $\mathbf{U}(p,q)$ and $\mathbf{O}(p,q)$, and

$$\begin{pmatrix} 0_m & I_m \\ -I_m & 0_m \end{pmatrix},$$

for $\mathbf{Sp}_m$. In each case, $J^2 = \pm I_n$.

Proposition 7.3.2 *Let J be a complex $n \times n$ matrix satisfying $J^2 = \pm I_n$. The subgroup G of $\mathbf{M}_n(\mathbb{C})$ defined by the equation $M^* J M = J$ is invariant under polar decomposition. If $M \in G$, then $|\det M| = 1$.*

Proof

The fact that G is a group is immediate. Let $M \in G$. Then $\det J = \det M^* \det M \det J$; that is, $|\det M|^2 = 1$. Furthermore, $M^* J M (J M^*) = J^2 M^* = \pm M^* = M^* J^2$. Simplifying by $M^* J$ on the left, there remains $M J M^* = J$, that is, $M^* \in G$.

Observe that, since G is a group, $M \in G$ implies $(M^*)^k J = JM^{-k}$ for every $k \in \mathbb{N}$. By linearity, it follows that $p(M^*)J = Jp(M^{-1})$ holds for every polynomial $p \in \mathbb{R}[X]$.

Let us now assume that $M \in G \cap \mathbf{HPD}_n$. We then have $M = U^* \operatorname{diag}(d_1, \ldots, d_n)U$, where U is unitary and the d_j's are positive real numbers. Let A be the set formed by the numbers d_j and $1/d_j$. There exists a polynomial p with real entries such that $p(a) = \sqrt{a}$ for every $a \in A$. Then we have $p(M) = \sqrt{M}$ and $p(M^{-1}) = \sqrt{M}^{-1}$. Since $M^* = M$, we have also $p(M)J = Jp(M^{-1})$; that is, $\sqrt{M}J = J\sqrt{M}^{-1}$. Hence $\sqrt{M} \in G$. From Proposition 7.3.1, G is stable under polar decomposition.

■

The main result of this section is the following:

Theorem 7.3.1 *Under the hypotheses of Proposition 7.3.2, the group G is homeomorphic to $(G \cap \mathbf{U}_n) \times \mathbb{R}^d$, for a suitable integer d.*

Of course, if $G = \mathbf{O}(p, q)$ or $\mathbf{Sp}_m$, the subgroup $G \cap \mathbf{U}_n$ can also be written as $G \cap \mathbf{O}_n$. We call $G \cap \mathbf{U}_n$ a *maximal compact subgroup* of G, because one can prove that it is not a proper subgroup of a compact subgroup of G. Another deep result, which is beyond the scope of this book, is that every maximal compact subgroup of G is a conjugate of $G \cap \mathbf{U}_n$. In the sequel, when speaking about *the* maximal compact subgroup of G, we shall always have in mind $G \cap \mathbf{U}_n$.

Proof

The proof amounts to showing that $G \cap \mathbf{HPD}_n$ is homeomorphic to some $\mathbb{R}^d$. To do this, we define

$$\mathcal{G} := \{N \in \mathbf{M}_n(k) | \exp tN \in G, \forall t \in \mathbb{R}\}.$$

Lemma 7.3.1 *The set $\mathcal{G}$ defined above satifies*

$$\mathcal{G} = \{N \in \mathbf{M}_n(k) | N^*J + JN = 0_n\}.$$

Proof

If $N^*J + JN = 0_n$, let us set $M(t) = \exp tN$. Then $M(0) = I_n$ and

$$\frac{d}{dt} M(t)^* J M(t) = M^*(t)(N^*J + JN)M(t) = 0_n,$$

so that $M(t)^* J M(t) \equiv J$. We thus have $N \in \mathcal{G}$. Conversely,, if $M(t) := \exp tN \in G$ for every t, then the derivative at $t = 0$ of $M^*(t)JM(t) = J$ gives $N^*J + JN = 0_n$.

■

Lemma 7.3.2 *The map $\exp : \mathcal{G} \cap \mathbf{H}_n \to G \cap \mathbf{HPD}_n$ is a homeomorphism.*

Proof

We must show that $\exp : \mathcal{G} \cap \mathbf{H}_n \to G \cap \mathbf{HPD}_n$ is onto. Let $M \in G \cap \mathbf{HPD}_n$ and let N be the Hermitian matrix such that $\exp N = M$. Let $p \in \mathbb{R}[X]$ be a polynomial with real entries such that for every $\lambda \in$

$\operatorname{Sp} M \cup \operatorname{Sp} M^{-1}$, we have $p(\lambda) = \log \lambda$. Such a polynomial exists, since the numbers λ are real and positive.

Let $N = U^*DU$ be a unitary diagonalization of N. Then $M = \exp N = U^*(\exp D)U$ and $M^{-1} = \exp(-N) = U^* \exp(-D)U$. Hence, $p(M) = N$ and $p(M^{-1}) = -N$. However, $M \in G$ implies $MJ = JM^{-1}$, and therefore $q(M)J = Jq(M^{-1})$ for every $q \in \mathbb{R}[X]$. With $q = p$, we obtain $NJ = -JN$. ■

These two lemmas complete the proof of the theorem, since $\mathcal{G} \cap \mathbf{H}_n$ is an $\mathbb{R}$-vector space. The integer d mentionned in the theorem is its dimension. ■

We wish to warn the reader that neither $\mathcal{G}$, nor $\mathbf{H}_n$ is a $\mathbb{C}$-vector space. We shall see examples in the next section that show that $\mathcal{G} \cap \mathbf{H}_n$ can be naturally $\mathbb{R}$-isomorphic to a $\mathbb{C}$-vector space, which is a source of confusion. One therefore must be cautious when computing d.

The reader eager to learn more about the theory of classical groups is advised to have a look at the book of R. Mneimné and F. Testard [28] or the one by A. W. Knapp [24].

7.4 The Groups $\mathbf{U}(p, q)$

Let us begin with the study of the maximal compact subgroup of $\mathbf{U}(p, q)$. If $M \in U(p, q) \cap \mathbf{U}_n$, let us write M blockwise:

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where $A \in \mathbf{M}_p(\mathbb{C})$, etc. The following equations express that M belongs to $\mathbf{U}_n$:

$$A^*A + C^*C = I_p, \quad B^*B + D^*D = I_q, \quad A^*B + C^*D = 0_{pq}.$$

Similarly, writing that $M \in U(p, q)$,

$$A^*A - C^*C = I_p, \quad D^*D - B^*B = I_q, \quad A^*B - C^*D = 0_{pq}.$$

Combining these equations, we obtain first $C^*C = 0_p$ and $B^*B = 0_q$. For every vector $X \in \mathbb{C}^n$, we have $\|CX\|_2^2 = X^*C^*CX = 0$; hence $CX = 0$. Finally, $C = 0$ and similarly $B = 0$. There remains $A \in \mathbf{U}_p$ and $D \in \mathbf{U}_q$. The maximal compact subgroup of $\mathbf{U}(p, q)$ is thus isomorphic (not only homeomorphic) to $\mathbf{U}_p \times \mathbf{U}_q$.

Furthermore, $\mathcal{G} \cap \mathbf{H}_n$ is the set of matrices

$$N = \begin{pmatrix} A & B \\ B^* & D \end{pmatrix},$$

where $A \in \mathbf{H}_p$, $D \in \mathbf{H}_q$, which satisfy $NJ + JN = 0_n$; that is, $A = 0_p$, $D = 0_q$. Hence $\mathcal{G} \cap \mathbf{H}_n$ is isomorphic to $\mathbf{M}_{p \times q}(\mathbb{C})$. One therefore has $d = 2pq$.

Proposition 7.4.1 *The unitary group* $\mathbf{U}(p,q)$ *is homeomorphic to* $\mathbf{U}_p \times \mathbf{U}_q \times \mathbb{R}^{2pq}$. *In particular,* $\mathbf{U}(p,q)$ *is connected.*

There remains to show connectivity. It is a straightforward consequence of the following lemma.

Lemma 7.4.1 *The unitary group* $\mathbf{U}_n$ *is connected.*

Since $\mathbf{GL}_n(\mathbb{C})$ is homeomorphic to $\mathbf{U}_n \times \mathbf{HPD}_n$ (via polar decomposition), hence to $\mathbf{U}_n \times \mathbf{H}_n$ (via the exponential), it is equivalent to the following statement.

Lemma 7.4.2 *The linear group* $\mathbf{GL}_n(\mathbb{C})$ *is connected.*

Proof

Let $M \in \mathbf{GL}_n(\mathbb{C})$ be given. Define $A := \mathbb{C} \setminus \{(1-\lambda)^{-1} | \lambda \in \mathrm{Sp}(M)\}$. The arcwise-connected set A does not contain the origin, nor the point $z = 1$, since $0 \notin \mathrm{Sp}(M)$. There thus exists a path γ joining 0 to 1 in A: $\gamma \in \mathcal{C}([0,1]; A)$, $\gamma(0) = 0$ and $\gamma(1) = 1$. Let us define $M(t) := \gamma(t)M + (1 - \gamma(t))I_n$. By construction, $M(t)$ is invertible for every t, and $M(0) = I_n$, $M(1) = M$. The connected component of I_n is thus all of $\mathbf{GL}_n(\mathbb{C})$.

■

7.5 The Orthogonal Groups $\mathbf{O}(p,q)$

The analysis of the maximal compact subgroup and of $\mathcal{G} \cap \mathbf{H}_n$ for the group $\mathbf{O}(p,q)$ is identical to that in the previous paragraph. On the one hand, $\mathbf{O}(p,q) \cap \mathbf{O}_n$ is isomorphic to $\mathbf{O}_p \times \mathbf{O}_q$. On the other hand, $\mathcal{G} \cap \mathbf{H}_n$ is isomorphic to $\mathbf{M}_{p \times q}(\mathbb{R})$, which is of dimension $d = pq$.

Proposition 7.5.1 *Let* $n \geq 1$. *The group* $\mathbf{O}(p,q)$ *is homeomorphic to* $\mathbf{O}_p \times \mathbf{O}_q \times \mathbb{R}^{pq}$. *The number of its connected components is two if* p *or* q *is zero, four otherwise.*

Proof

We must show that $\mathbf{O}_n$ has two connected components. However, $\mathbf{O}_n$ is the disjoint union of $\mathbf{SO}_n$ (matrices of determinant $+1$) and of $\mathbf{O}_n^-$ (matrices of determinant -1). Since $\mathbf{O}_n^- = M \cdot \mathbf{SO}_n$ for any matrix $M \in \mathbf{O}_n^-$ (for example a hyperplane symmetry), there remains to show that the special orthogonal group $\mathbf{SO}_n$ is connected, in fact arcwise connected. We use the following property:

Lemma 7.5.1 *Given $M \in \mathbf{O}_n$, there exists $Q \in \mathbf{O}_n$ such that the matrix $Q^{-1}MQ$ has the form*

$$\begin{pmatrix} (\cdot) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & (\cdot) \end{pmatrix}, \tag{7.7}$$

where the diagonal blocks are of size 1×1 or 2×2 and are orthogonal, those of size 2×2 being rotations matrices:

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{7.8}$$

Let us apply Lemma 7.5.1 to $M \in \mathbf{SO}_n$. The determinant of M, which is the product of the determinants of the diagonal blocks, equals $(-1)^m$, m being the multiplicity of the eigenvalue -1. Since $\det M = 1$, m is even, and we can gather the diagonal -1's pairwise in order to form matrices of the form (7.8), with $\theta = \pi$. Finally, there exists $Q \in \mathbf{O}_n$ such that

$$M = Q^T \begin{pmatrix} R_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & R_r & \ddots & & 0 \\ \vdots & & \ddots & 1 & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} Q,$$

where each diagonal block R_j is a matrix of planar rotation:

$$R_j = \begin{pmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{pmatrix}.$$

Let us now define a matrix $M(t)$ as above, in which we replace the angles θ_j by $t\theta_j$. We thus obtain a path in $\mathbf{SO}_n$, from $M(0) = I_n$ to $M(1) = M$. The connected component of I_n is thus the whole of $\mathbf{SO}_n$.

■

We now prove Lemma 7.5.1: As an orthogonal matrix, M is normal. From Theorem 3.3.1, it decomposes into a matrix of the form (7.7), the 1×1 diagonal blocks being the real eigenvalues. These eigenvalues are ± 1, since $Q^{-1}MQ$ is orthogonal. The diagonal blocks 2×2 are direct similitude matrices. However, they are isometries, since $Q^{-1}MQ$ is orthogonal. Hence they are rotation matrices.

■

7.5.1 Notable Subgroups of $\mathbf{O}(p,q)$

We assume here that $p,q \geq 1$, so that $\mathbf{O}(p,q)$ has four connected components. We first describe them.

Let us recall that if $M \in \mathbf{O}(p,q)$ reads blockwise

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where $A \in \mathbf{M}_p(\mathbb{R})$, etc. Then $A^TA = C^TC + I_p$ is larger than I_p as a symmetric matrix, so that $\det A$ cannot vanish. Similarly, $D^TD = B^TB + I_q$ shows that $\det D$ does not vanish. The continuous map $M \mapsto (\det A, \det D)$ thus sends $\mathbf{O}(p,q)$ to $\mathbb{R}^* \times \mathbb{R}^*$ (in fact, to $(\mathbb{R} \setminus (-1,1))^2$). Since the sign map from $\mathbb{R}^*$ to $\{-,+\}$ is continuous, we may thus define a continuous function

$$\begin{array}{rcl} \mathbf{O}(p,q) & \stackrel{\sigma}{\to} & \{-,+\}^2 \sim (\mathbb{Z}/2\mathbb{Z})^2\,, \\ M & \mapsto & (\operatorname{sgn}\det A, \operatorname{sgn}\det D). \end{array}$$

The diagonal matrices whose diagonal entries are ± 1 belong to $\mathbf{O}(p,q)$. It follows that σ is onto. Since σ is continuous, the preimage G_α of an element α of $\{-,+\}^2$ is the union of some connected components of $\mathbf{O}(p,q)$; let $n(\alpha)$ be the number of these components. Then $n(\alpha) \geq 1$ (σ being onto), and $\sum_\alpha n(\alpha)$ equals 4, the number of connected components of $\mathbf{O}(p,q)$. Since there are four terms in this sum, we obtain $n(\alpha) = 1$ for every α. Finally, the connected components of $\mathbf{O}(p,q)$ are the G_α's, where $\alpha \in \{-,+\}^2$.

The left multiplication by an element M of $\mathbf{O}(p,q)$ is continuous, bijective, whose inverse (another multiplication) is continuous. It thus induces a permutation of the set π_0 of connected components of $\mathbf{O}(p,q)$. Since σ induces a bijection between π_0 and $\{-,+\}^2$, there exists thus a permutation q_M of $\{-,+\}^2$ such that $\sigma(MM') = q_M(\sigma(M'))$. Similarly, the multiplication at right by M' is an homeomorphism, allowing to define a permutation $p_{M'}$ of$\{-,+\}^2$ such that $\sigma(MM') = p_{M'}(\sigma(M))$. The equality

$$p_{M'}(\sigma(M)) = q_M(\sigma(M'))$$

shows that p_M and q_M actually depend only on $\sigma(M)$. In other words, $\sigma(MM')$ depends only on $\sigma(M)$ and $\sigma(M')$. A direct evaluation in the special case of matrices in $\mathbf{O}(p,q) \cap \mathbf{O}_n(\mathbb{R})$ leads to the following conclusion.

Proposition 7.5.2 $(p,q \geq 1)$ *The connected components of* $G = \mathbf{O}(p,q)$ *are the sets* $G_\alpha := \sigma^{-1}(\alpha)$, *defined by* $\alpha_1 \det A > 0$ *and* $\alpha_2 \det D > 0$, *when a matrix* M *is written blockwise as above. The map* $\sigma : \mathbf{O}(p,q) \to \{-,+\}^2$ *is a surjective group homomorphism; that is,* $\sigma(MM') = \sigma(M)\sigma(M')$. *In particular:*

1. $G_\alpha^{-1} = G_\alpha$;
2. $G_\alpha \cdot G_{\alpha'} = G_{\alpha\alpha'}$.

Remark: σ admits a right inverse, namely

$$\alpha \mapsto M^\alpha := \mathrm{diag}(\alpha_1 1, 1, \dots, 1, \alpha_2 1).$$

The group $\mathbf{O}(p,q)$ appears, therefore, as the semidirect product of G_{++} with $(\mathbb{Z}/2\mathbb{Z})^2$.

We deduce immediately from the proposition that $\mathbf{O}(p,q)$ possesses five open and closed normal subgroups, the preimages of the five subgroups of $(\mathbb{Z}/2\mathbb{Z})^2$:

- $\mathbf{O}(p,q)$ itself;
- G_{++}, which we also denote by G_0 (see Exercise 21), the connected component of the unit element I_n,
- $G_{++} \cup G_\alpha$, for the three other choices of an element α.

One of these groups, namely $G_{++} \cup G_{--}$ is equal to the kernel $\mathbf{SO}(p,q)$ of the homomorphism $M \mapsto \det M$. In fact, this kernel is open and closed, thus is the union of connected components of $\mathbf{O}(p,q)$. However the sign of $\det M$ for $M \in G_\alpha$ is that of $\alpha_1\alpha_2$, which can be seen directly from the case of diagonal matrices M^α.

7.5.2 The Lorentz Group $\mathbf{O}(1,3)$

If $p = 1$ and $q = 3$, the group $\mathbf{O}(1,3)$ is isomorphic to the orthogonal group of the Lorentz quadratic form $dt^2 - dx_1^2 - dx_2^2 - dx_3^2$, which defines the space-time distance in special relativity.[1] Each element M of $\mathbf{O}(1,3)$ corresponds to the transformation

$$\begin{pmatrix} t, \\ x \end{pmatrix} \mapsto M \begin{pmatrix} t, \\ x \end{pmatrix},$$

which we still denote by M, by abuse of notation. This transformation preserve the light cone of equation $t^2 - x_1^2 - x_2^2 - x_3^2 = 0$. Since it is a homeomorphism of $\mathbb{R}^4$, it permutes the connected components of the complement $\mathcal{C}$ of that cone. There are three such components (see Figure 7.1):

- the convex set $C_+ := \{(t,x) \,|\, \|x\| < t\}$;
- the convex set $C_- := \{(t,x) \,|\, \|x\| < -t\}$;
- the "ring" $\mathcal{A} := \{(t,x) \,|\, |t| < \|x\|\}$.

Clearly, C_+ and C_- are homeomorphic. For example, they are so via the time reversal $t \mapsto -t$. However, they are not homeomorphic to $\mathcal{A}$, because the latter is homeomorphic to $S^2 \times \mathbb{R}^2$ (here, S^2 denotes the unit sphere),

[1] We have selected a system of units in which the speed of light equals one.

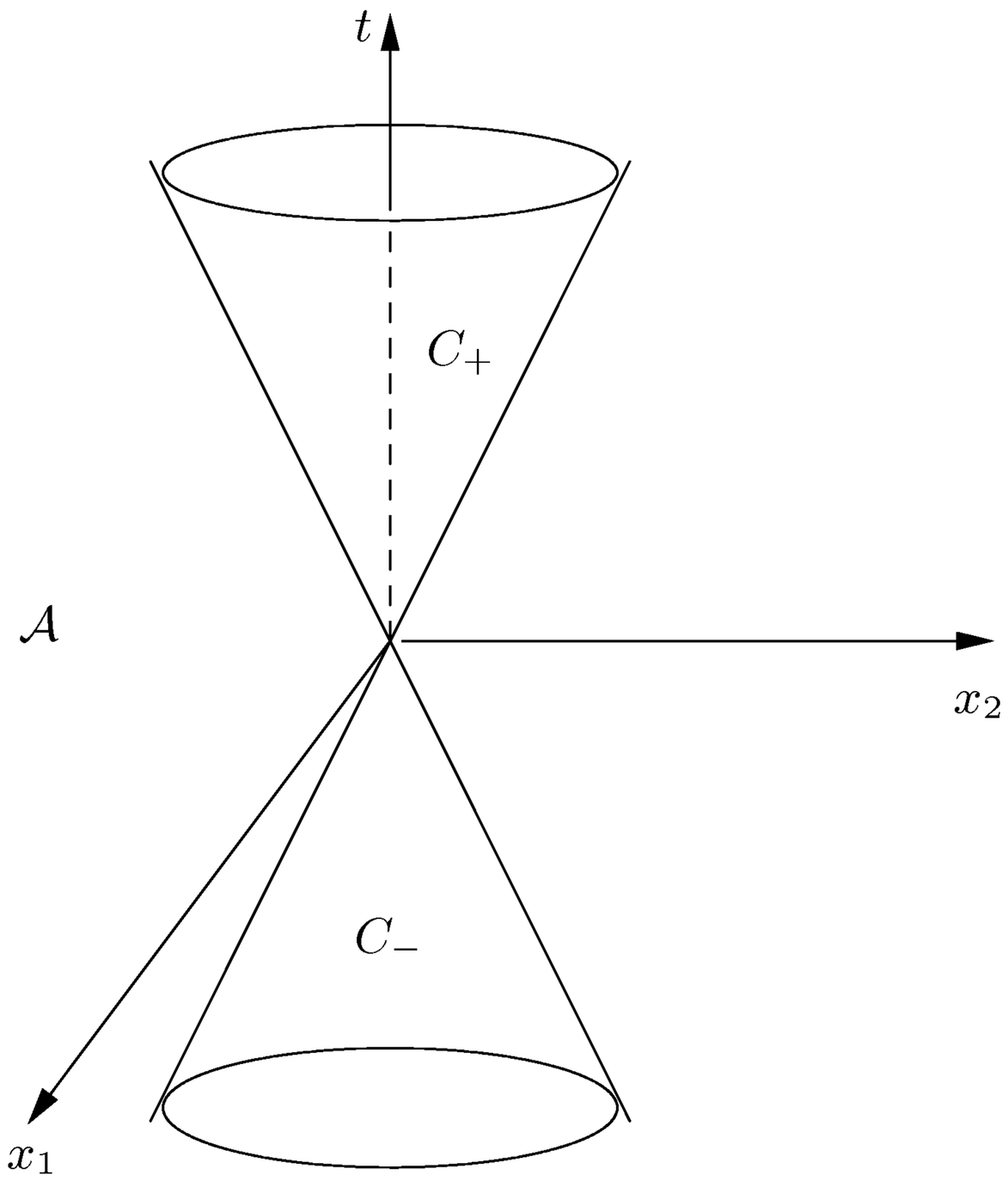

Figure 7.1. The Lorentz cone.

which is not contractible, while a convex set is always contractible. Since M is a homeomorphism, one deduces that necessarily, $M\mathcal{A} = \mathcal{A}$, while $MC_+ = C_\pm$, $MC_- = C_\mp$.

The transformations that preserve C_+, and therefore every connected component of $\mathcal{C}$, form the *orthochronous Lorentz group*. Its elements are those that send the vector $\mathbf{e}_0 := (1,0,0,0)^T$ to C_+; that is, those for which the first component of $M\mathbf{e}_0$ is positive. Since this component is A (here it is nothing but a scalar), this group must be $G_{++} \cup G_{+-}$.

7.6 The Symplectic Group $\mathbf{Sp}_n$

Let us study first of all the maximal compact subgroup $\mathbf{Sp}_n \cap \mathbf{O}_{2n}$. If

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

with blocks of size $n \times n$, then $M \in \mathbf{Sp}_n$ means that

$$A^T C = C^T A, \quad A^T D - C^T B = I_n, \quad B^T D = D^T B,$$

while $M \in \mathbf{O}_{2n}$ yields

$$A^T A + C^T C = I_n, \quad B^T B + D^T D = I_n, \quad B^T A + D^T C = 0_n.$$

But since $M^T \in \mathbf{Sp}_n$, we also have

$$AB^T = BA^T, \quad AD^T - BC^T = I_n, \quad CD^T = DC^T.$$

Let us combine these equations:

$$B = B(A^TA+C^TC) = AB^TA+(AD^T-I_n)C = A(B^TA+D^TC)-C = -C.$$

Similarly,

$$D = D(A^TA+C^TC) = (I_n+CB^T)A+CD^TC = A+C(B^TA+D^TC) = A.$$

Hence

$$M = \begin{pmatrix} A & B \\ -B & A \end{pmatrix}.$$

The remaining conditions are

$$A^TA + B^TB = I_n, \quad A^TB = B^TA.$$

This amounts to saying that $A + iB$ is unitary. One immediately checks that the map $M \mapsto A + iB$ is an isomorphism from $\mathbf{Sp}_n$ onto $\mathbf{U}_n$.

Finally, if

$$N = \begin{pmatrix} A & B \\ B^T & D \end{pmatrix}$$

is symmetric and $NJ + JN = 0_{2n}$, we have, in fact,

$$N = \begin{pmatrix} A & B \\ B & -A \end{pmatrix},$$

where A and B are symmetric. Hence $\mathcal{G} \cap \mathbf{Sym}_{2n}$ is homeomorphic to $\mathbf{Sym}_n \times \mathbf{Sym}_n$, that is, to $\mathbb{R}^{n(n+1)}$.

Proposition 7.6.1 *The symplectic group* $\mathbf{Sp}_n$ *is homeomorphic to* $\mathbf{U}_n \times \mathbb{R}^{n(n+1)}$.

Corollary 7.6.1 *In particular, every symplectic matrix has determinant* $+1$.

Indeed, Proposition 7.6.1 shows that $\mathbf{Sp}_n$ is connected. Since the determinant is continuous, with values in $\{-1, 1\}$, it is constant, equal to $+1$.

7.7 Singular Value Decomposition

As we shall see in Exercise 8 (see also Exercise 12 in Chapter 4), the eigenvalues of the matrix H in the polar decomposition of a given matrix M are of some importance. They are called the *singular values* of M. Since these are the square roots of the eigenvalues of M^*M, one may even speak

of the singular values of an arbitrary matrix, not necessarily invertible. Recalling that (see Exercise 17 in Chapter 2) when M is $n \times m$, M^*M and MM^* have the same nonzero eigenvalues, counting with multiplicities, one may even speak of the singular values of a rectangular matrix, up to an ambiguity concerning the multiplicity of the eigenvalue 0.

The main result of the section is the following.

Theorem 7.7.1 *Let $M \in \mathbf{M}_{n\times m}(\mathbb{C})$ be given. Then there exist two unitary matrices $U \in \mathbf{U}_n$, $V \in \mathbf{U}_m$ and a quasi-diagonal matrix*

$$D = \begin{pmatrix} s_1 & & & & \\ & \ddots & & & \\ & & s_r & & \\ & & & 0 & \\ & & & & \ddots \end{pmatrix},$$

with $s_1, \dots, s_r \in (0, +\infty)$, such that $M = UDV$. The numbers $s_1, \dots, s_r$ are uniquely defined up to permutation; they are the nonzero singular values of M. In particular, r is the rank of M.

If $M \in \mathbf{M}_{n\times m}(\mathbb{R})$, then one may choose U, V to be real orthogonal.

Remark: The factorization given in the theorem is far from being unique, even for invertible square matrices. In fact, the number of real degrees of freedom in that factorization is $n^2+m^2+\min(n,m)$, which is always greater than the dimension $2nm$ of $\mathbf{M}_{n\times m}(\mathbb{C})$ as an $\mathbb{R}$-vector space.

Proof

Since MM^* is positive semidefinite, we may write its eigenvalues as $s_1^2, \dots, s_r^2, 0, \dots$, where the s_j's, the singular values of M, are positive real numbers. The spectrum of M^*M has the same form, except for the multiplicity of 0. Indeed, the multiplicities of 0 as an eigenvalue of MM^* and MM^*, respectively, differ by $n - m$, while the multiplicities of other eigenvalues are the same for both matrices. We set $S = \operatorname{diag}(s_1, \dots, s_r)$.

Since M and MM^* have the same rank, and since $R(MM^*) \subset R(M)$, we have $R(MM^*) = R(M)$. Since MM^* is Hermitian, its kernel is $R(M)^\perp$, where orthogonality is relative to the canonical scalar product; with the duality formula, we conclude that $\ker MM^* = \ker M^*$. Now we are in position to state that

$$\mathbb{C}^n = R(MM^*) \oplus^\perp \ker M^*.$$

Therefore, there exists an orthonormal basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ of $\mathbb{C}^n$ consisting of eigenvectors of MM^*, associated to the s_j^2's, followed by vectors of $\ker M^*$. Let us form the unitary matrix

$$U = (\mathbf{u}_1, \dots, \mathbf{u}_n).$$

Written blockwise, we have $U = (U_R, U_K)$, where

$$MM^*U_R = U_R S^2, \quad M^*U_K = 0.$$

Let now define $V_R := M^* U_R S^{-1}$. From above, we have

$$V_R^* V_R = S^{-1} U_R^* M M^* U_R S^{-1} = I_r.$$

This means that the columns $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of V_R constitute an orthonormal family.

Noting that these column vectors belong to $R(M^*)$, that is, to $(\ker M)^\perp$, a subspace of codimension r, we see that $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ can be extended to an orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ of $\mathbb{C}^m$, where $v_{r+1}, \ldots$ belong to $\ker M$. Let $V =: (V_R, V_K)$ be the unitary matrix whose columns are $\mathbf{v}_1, \ldots$

We now compute blockwise the product $U^* M V$. From $M V_K = 0$ and $M^* U_K^* = 0$, we get

$$U^* M V = \begin{pmatrix} U_R^* M V_R & 0 \\ 0 & 0 \end{pmatrix}.$$

Finally, we obtain

$$U_R^* M V_R = U_R^* M M^* U_R S^{-1} = U_R^* U_R S = S.$$

■

7.8 Exercises

1. Show that the *square root* map from $\mathbf{HPD}_n$ into itself is continuous.

2. Let $M \in \mathbf{M}_n(k)$ be given, with $k = \mathbb{R}$ or $\mathbb{C}$. Show that there exists a polynomial $P \in k(X)$, of degree at most $n-1$, such that $P(M) = \exp M$. However, show that this polynomial cannot be chosen independently of the matrix.
   Compute this polynomial when M is nilpotent.

3. For $t \in \mathbb{R}$, define *Pascal's matrix* $P(t)$ by $p_{ij}(t) = 0$ if $i < j$ (the matrix is lower triangular) and

$$p_{ij}(t) = t^{i-j} \begin{pmatrix} i-1 \\ j-1 \end{pmatrix}$$

   otherwise. Let us emphasize that for just this once in this book, P is an *infinite* matrix, meaning that its indices range over the infinite set $\mathbb{N}^*$. Compute $P'(t)$ and deduce that there exists a matrix L such that $P(t) = \exp(tL)$. Compute L explicitly.

4. Let I be an interval of $\mathbb{R}$ and $t \mapsto P(t)$ be a map of class $\mathcal{C}^1$ with values in $\mathbf{M}_n(\mathbb{R})$ such that for each t, $P(t)$ is a projector: $P(t)^2 = P(t)$.

   (a) Show that the rank of $P(t)$ is constant.
   (b) Show that $P(t)P'(t)P(t) = 0_n$.

(c) Let us define $Q(t) := [P'(t), P(t)]$. Show that $P'(t) = [Q(t), P(t)]$.
(d) Let $t_0 \in I$ be given. Show that the differential equation $U' = QU$ possesses a unique solution in I such that $U(t_0) = I_n$. Show that $P(t) = U(t)P(t_0)U(t)^{-1}$.

5. Show that the set of projectors of given rank p is a connected subset in $\mathbf{M}_n(\mathbb{C})$.

6. (a) Let $A \in \mathbf{HPD}_n$ and $B \in \mathbf{H}_n$ be given. Show that AB is diagonalizable with real eigenvalues (though it is not necessarily Hermitian). Show also that the sum of the multiplicities of the positive eigenvalues (respectively zero, respectively negative) is the same for AB as for B.
(b) Let A, B, C be three Hermitian matrices such that $ABC \in \mathbf{H}_n$. Show that if three of the matrices A, B, C, ABC are positive definite, then the fourth is positive definite too.

7. Let $M \in \mathbf{GL}_n(\mathbb{C})$ be given and $M = HQ$ be its polar decomposition. Show that M is normal if and only if $HQ = QH$.

8. The deformation of an elastic body is represented at each point by a square matrix $F \in \mathbf{GL}_3^+(\mathbb{R})$ (the sign + expresses that $\det F > 0$). More generally, $F \in \mathbf{GL}_n^+(\mathbb{R})$ in other space dimensions. The density of elastic energy is given by a function $F \mapsto W(F) \in \mathbb{R}^+$.
(a) The principle of frame indifference says that $W(QF) = W(F)$ for every $F \in \mathbf{GL}_n^+(\mathbb{R})$ and every rotation Q. Show that there exists a map $w : \mathbf{SPD}_n \to \mathbb{R}^+$ such that $W(F) = w(H)$, where $F = QH$ is the polar decomposition.
(b) When the body is isotropic, we also have $W(FQ) = W(F)$, for every $F \in \mathbf{GL}_n^+(\mathbb{R})$ and every rotation Q. Show that there exists a map $\phi : \mathbb{R}^n \to \mathbb{R}^+$ such that $W(F) = \phi(h_1, \dots, h_n)$, where the h_j are the entries of the characteristic polynomial of H. In other words, $W(F)$ depends only on the singular values of F.

9. We use *Schur's norm* $\|A\| = (\operatorname{Tr} A^*A)^{1/2}$.
(a) If $A \in \mathbf{M}_n(\mathbb{C})$, show that there exists $Q \in \mathbf{U}_n$ such that $\|A - Q\| \leq \|A - U\|$ for every $U \in \mathbf{U}_n$. We shall define $S := Q^{-1}A$. We therefore have $\|S - I_n\| \leq \|S - U\|$ for every $U \in \mathbf{U}_n$.
(b) Let $H \in \mathbf{H}_n$ be a Hermitian matrix. Show that $\exp(itH) \in \mathbf{U}_n$ for every $t \in \mathbb{R}$. Compute the derivative at $t = 0$ of

$$t \mapsto \|S - \exp(itH)\|^2$$

and deduce that $S \in \mathbf{H}_n$.
(c) Let D be a diagonal matrix, unitarily similar to S. Show that $\|D - I_n\| \leq \|DU - I_n\|$ for every $U \in \mathbf{U}_n$. By selecting a suitable U, deduce that $S \geq 0_n$.

(d) If $A \in \mathbf{GL}_n(\mathbb{C})$, show that QS is the polar decomposition of A.
(e) Deduce that if $H \in \mathbf{HPD}_n$ and if $U \in \mathbf{U}_n$, $U \neq I_n$, then $\|H - I_n\| < \|H - U\|$.
(f) Finally, show that if $H \in \mathbf{H}_n$, $H \geq 0_n$ and $U \in \mathbf{U}_n$, then $\|H - I_n\| \leq \|H - U\|$.

10. Let $A \in \mathbf{M}_n(\mathbb{C})$ and $h \in \mathbb{C}$. Show that $I_n - hA$ is invertible as soon as $|h| < 1/\rho(A)$. One then denotes its inverse by $R(h; A)$.

(a) Let $r \in (0, 1/\rho(A))$. Show that there exists a $c_0 > 0$ such that for every $h \in \mathbb{C}$ with $|h| \leq r$, we have

$$\|R(h; A) - e^{hA}\| \leq c_0 |h|^2.$$

(b) Verify the formula

$$\begin{aligned} C^m - B^m &= (C - B)C^{m-1} + \cdots + B^{l-1}(C - B)C^{m-l} + \cdots \\ &\quad + \cdots + B^{m-1}(C - B), \end{aligned}$$

and deduce the bound

$$\|R(h; A)^m - e^{mhA}\| \leq c_0 m |h|^2 e^{c_2 m |h|},$$

when $|h| \leq r$ and $m \in \mathbb{N}$.
(c) Show that for every $t \in \mathbb{C}$,

$$\lim_{m \to +\infty} R(t/m; A)^m = e^{tA}.$$

11. (a) Let $J(a; r)$ be a Jordan block of size r, with $a \in \mathbb{C}^*$. Let $b \in \mathbb{C}$ be such that $a = e^b$. Show that there exists a nilpotent $N \in \mathbf{M}_r(\mathbb{C})$ such that $J(a; r) = \exp(bI_r + N)$.
(b) Show that $\exp : \mathbf{M}_n(\mathbb{C}) \to \mathbf{GL}_n(\mathbb{C})$ is onto, but that it is not one-to-one. Deduce that $X \mapsto X^2$ is onto $\mathbf{GL}_n(\mathbb{C})$. Verify that it is not onto $\mathbf{M}_n(\mathbb{C})$.

12. (a) Show that the matrix

$$J_2 = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$$

is not the square of any matrix of $\mathbf{M}_2(\mathbb{R})$.
(b) Show, however, that the matrix $J_4 := \mathrm{diag}(J_2, J_2)$ is the square of a matrix of $\mathbf{M}_4(\mathbb{R})$.
Show also that the matrix

$$J_3 = \begin{pmatrix} J_2 & I_2 \\ 0_2 & J_2 \end{pmatrix}$$

is not the square of a matrix of $\mathbf{M}_4(\mathbb{R})$.
(c) Show that J_2 is not the exponential of any matrix of $\mathbf{M}_2(\mathbb{R})$. Compare with the previous exercise.

(d) Show that J_4 is the exponential of a matrix of $\mathbf{M}_4(\mathbb{R})$, but that J_3 is not.

13. Let $\mathbf{A}_n(\mathbb{C})$ be the set of skew-Hermitian matrices of size n. Show that $\exp : \mathbf{A}_n(\mathbb{C}) \to \mathbf{U}_n$ is onto. **Hint**: If U is unitary, diagonalize it.

14. (a) Let $\theta \in \mathbb{R}$ be given. Compute $\exp B$, where

$$B = \begin{pmatrix} 0 & \theta \\ -\theta & 0 \end{pmatrix}.$$

(b) Let $\mathbf{A}_n(\mathbb{R})$ be the set of real skew-symmetric matrices of size n. Show that $\exp : \mathbf{A}_n(\mathbb{R}) \to \mathbf{SO}_n$ is onto. **Hint**: Use the reduction of direct orthogonal matrices.

15. Let $\phi : \mathbf{M}_n(\mathbb{R}) \to \mathbb{R}$ be a nonnull map satisfying $\phi(AB) = \phi(A)\phi(B)$ for every $A, B \in \mathbf{M}_n(\mathbb{R})$. If $\alpha \in \mathbb{R}$, we set $\delta(\alpha) = |\phi(\alpha I_n)|^{1/n}$. We have seen, in Exercise 16 of Chapter 3, that $|\phi(M)| = \delta(\det M)$ for every $M \in \mathbf{M}_n(\mathbb{R})$.

(a) Show that on the range of $M \mapsto M^2$ and on that of $M \mapsto \exp M$, $\phi \equiv \delta \circ \det$.
(b) Deduce that $\phi \equiv \delta \circ \det$ on $\mathbf{SO}_n$ (use Exercise 14) and on $\mathbf{SPD}_n$.
(c) Show that either $\phi \equiv \delta \circ \det$ or $\phi \equiv (\operatorname{sgn}(\det))\delta \circ \det$.

16. Let A be a K-Banach algebra ($K = \mathbb{R}$ or $\mathbb{C}$) with a unit denoted by e. If $x \in A$, define $x^0 := e$.

(a) Given $x \in A$, show that the series

$$\sum_{m \in \mathbb{N}} \frac{1}{m!} x^m$$

converges normally, hence converges in A. Its sum is denoted by $\exp x$.
(b) If $x, y \in A$, $[x, y] = xy - yx$ is called the "commutator" of x and y. Show that $[x, y] = 0$ implies

$$\exp(x + y) = (\exp x)(\exp y), \quad [x, \exp y] = 0.$$

(c) Show that the map $t \mapsto \exp tx$ is differentiable on $\mathbb{R}$, with

$$\frac{d}{dt} \exp tx = x \exp tx = (\exp tx)x.$$

(d) Let $x, y \in A$ be given. Assume that $[x, y]$ commutes with x and y.
   i. Show that $(\exp -tx)xy(\exp tx) = xy + t[y, x]x$.
   ii. Deduce that $[\exp -tx, y] = t[y, x] \exp -tx$.

iii. Compute the derivative of $t \mapsto (\exp -ty)(\exp -tx)\exp t(x+y)$. Finally, prove the Campbell–Hausdorff formula

$$\exp(x+y) = (\exp x)(\exp y)\left(\exp \frac{1}{2}[y,x]\right).$$

(e) In $A = \mathbf{M}_3(\mathbb{R})$, construct an example that satisfies the above hypothesis ($[x,y]$ commutes with x and y), where $[x,y]$ is nonzero.

17. Show that the map

$$H \mapsto f(H) := (iI_n + H)(iI_n - H)^{-1}$$

induces a homeomorphism from $\mathbf{H}_n$ onto the set of matrices of $\mathbf{U}_n$ whose spectrum does not contain -1. Find an equivalent of $f(tH) - \exp(-2itH)$ as $t \to 0$.

18. Let G be a group satisfying the hypotheses of Proposition 7.3.2.

(a) Show that $\mathcal{G}$ is a *Lie algebra*, meaning that it is stable under the bilinear map $(A,B) \mapsto [A,B] := AB - BA$.

(b) Show that for $t \to 0+$,

$$\exp(tA)\exp(tB)\exp(-tA)\exp(-tB) = I_n + t^2[A,B] + \mathcal{O}(t^3).$$

Deduce another proof of the stability of $\mathcal{G}$ by $[\cdot,\cdot]$.

(c) Show that the map $M \mapsto [A,M]$ is a derivation, meaning that the Jacobi identity

$$[A,[B,C]] = [[A,B],C] + [B,[A,C]]$$

holds.

19. In the case $p = 1$, $q \geq 1$, show that $G_{++} \cup G_{+-}$ is the set of matrices $M \in \mathbf{O}(p,q)$ such that the image under M of the "time" vector $(1,0,\ldots,0)^T$ belongs to the convex cone whose equation is

$$x_1 > \sqrt{x_2^2 + \cdots + x_n^2}.$$

20. Assume that $p,q \geq 1$ and consider the group $\mathbf{O}(p,q)$. Define $G_0 := G_{++}$. Since $-I_n \in \mathbf{O}(p,q)$, we denote by (μ,β) the indices for which $-I_n \in G_{\mu,\beta}$.
If $H \in \mathbf{GL}_n(\mathbb{R})$, denote by σ_H the conjugation $M \mapsto H^{-1}MH$.

(a) Let $H \in G$ be given. Show that σ_H (or rather its restriction to G_0) is an automorphism of G_0.

(b) Let $H \in \mathbf{M}_n(\mathbb{R})$ be such that $HM = MH$ for every $M \in G_0$. Show that $HN = NH$ for every $N \in \mathcal{G}$. Deduce that H is a homothety.

(c) Let $H \in G$. Show that there exists $K \in G_0$ such that $\sigma_H = \sigma_K$ if and only if $H \in G_0 \cup G_{\mu,\beta}$.

21. A *topological group* is a group G endowed with a topology for which the maps $(g,h) \mapsto gh$ and $g \mapsto g^{-1}$ are continuous. Show that in a topological group, the connected component of the unit element is a normal subgroup. Show also that the open subgroups are closed. Give an example of a closed subgroup that is not open.

22. One identifies $\mathbb{R}^{2n}$ with $\mathbb{C}^n$ by the map

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto x + iy.$$

Therefore, every matrix $M \in \mathbf{M}_{2n}(\mathbb{R})$ defines an $\mathbb{R}$-linear map $\tilde{M}$ form $\mathbb{C}^n$ into itself.

(a) Let

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \mathbf{M}_{2n}(\mathbb{R})$$

be given. Under what condition on the blocks A, B, C, D is the map $\tilde{M}$ $\mathbb{C}$-linear?

(b) Show that $M \mapsto \tilde{M}$ is an isomorphism from $\mathbf{Sp}_n \cap \mathbf{O}_{2n}$ onto $\mathbf{U}_n$.

8

Matrix Factorizations

The direct solution (by Cramer's method) of a linear system $Mx = b$, where $M \in \mathbf{GL}_n(k)$ ($b \in k^n$) is computationally expensive, especially if one wishes to solve the system many times with various values of b. In the next chapter we shall study iterative methods for the case $k = \mathbb{R}$ or $\mathbb{C}$. Here we concentrate on a simple idea: To decompose M as a product PQ in such a way that the resolution of the intermediate systems $Py = b$ and $Qx = y$ is "cheap." In general, at least one of the matrices is triangular. For example, if P is lower triangular ($p_{ij} = 0$ if $i < j$), then its diagonal entries p_{ii} are nonzero, and one may solve the system $Py = b$ step by step:

$$\begin{aligned}
y_1 &= \frac{b_1}{p_{11}}, \\
&\vdots \\
y_i &= \frac{b_i - p_{i1}y_1 - \cdots - p_{i,i-1}y_{i-1}}{p_{ii}}, \\
&\vdots \\
y_n &= \frac{b_n - p_{n1}y_1 - \cdots - p_{n,n-1}y_{n-1}}{p_{nn}}.
\end{aligned}$$

The computation of y_i needs $2i-1$ operations and the final result is obtained in n^2 operations. This is not expensive if one notes that computing the product $x = M^{-1}b$ (assuming that M^{-1} is computed once and for all, an expensive task) needs $2n^2 - n$ operations.

Another example of easily invertible matrices is the orthogonal matrices: If $Q \in \mathbf{O}_n$ (or $Q \in \mathbf{U}_n$), then $Qx = y$ amounts to $x = Q^T y$ (or $x = Q^* y$), which is computed in $\mathcal{O}(n^2)$ operations.

The techniques described below are often called *direct solving methods*.

8.1 The LU Factorization

Definition 8.1.1 *Let $M \in \mathbf{GL}_n(k)$, where k is a field. We say that M admits an LU factorization if there exist in $\mathbf{GL}_n(k)$ two matrices L (lower triangular with 1's on the diagonal) and U (upper triangular) such that $M = LU$.*

Remarks:

- The diagonal entries of U are not equal to 1 in general. The LU factorization is thus asymmetric with respect to L and U.
- The letters L and U recall the shape of the matrices: L for *lower* and U for *upper*.
- If there exists an LU factorization (which is unique, as we shall see below), then it can be computed by induction on the size of the matrix. The algorithm is provided in the proof of the next theorem. Indeed, if $N^{(p)}$ denotes the matrix extracted from N by keeping only the first p rows and columns, we have easily

$$M^{(p)} = L^{(p)} U^{(p)},$$

  where the matrices $L^{(p)}$ and $U^{(p)}$ have the required properties.

Definition 8.1.2 *The* leading principal minors *of M are the determinants of the matrices $M^{(p)}$, for $1 \leq p \leq n$.*

Theorem 8.1.1 *The matrix $M \in \mathbf{GL}_n(k)$ admits an LU factorization if and only if its leading principal minors are nonzero. When this condition is fulfilled, the LU factorization is unique.*

Proof

Let us begin with uniqueness: If $LU = L'U'$, then $(L')^{-1}L = U'U^{-1}$, which reads $L'' = U''$, where L'' and U'' are triangular of opposite types, the diagonal entries of L'' being 1's. We deduce $L'' = U'' = I_n$; that is, $L' = L$, $U' = U$.

We next assume that M admits an LU factorization. Then $\det M^{(p)} = \det L^{(p)} \det U^{(p)} = \prod_{1 \leq j \leq p} u_{jj}$, which is nonzero because U is invertible.

We prove the converse (the existence of an LU factorization) by induction on the size of the matrices. It is clear if $n = 1$. Otherwise, let us assume that the statement is true up to the order $n-1$ and let $M \in \mathbf{GL}_n(k)$ be

given, with nonzero leading principal minors. We look for L and U in the blockwise form

$$L = \begin{pmatrix} L' & 0 \\ X^T & 1 \end{pmatrix}, \quad U = \begin{pmatrix} U' & Y \\ 0 & u \end{pmatrix},$$

with $L', U' \in \mathbf{M}_{n-1}(k)$, etc. We likewise obtain the description

$$M = \begin{pmatrix} M' & R \\ S^T & m \end{pmatrix}.$$

Multiplying blockwise, we obtain the equations

$$L'U' = M', \quad L'Y = R, \quad (U')^T X = S, \quad u = m - X^T Y.$$

By assumption, the leading principal minors of M' are nonzero. The induction hypothesis guarantees the existence of the factorization $M' = L'U'$. Then Y and X are the unique solutions of (triangular) Cramer systems. Finally, u is explicitly given.

■

Let us now compute the number of operations needed in the computation of L and U. We pass from a factorization in $\mathbf{GL}_{n-1}(k)$ to a factorization in $\mathbf{GL}_n(k)$ by means of the computations of X ($(n-1)(n-2)$ operations), Y ($(n-1)^2$ operations) and u ($2(n-1)$ operations), for a total of $(n-1)(2n-1)$ operations. Finally, the computation ex nihilo of an LU factorization costs $P(n)$ operations, where P is a polynomial of degree three, with $P(X) = 2X^3/3 + \cdots$.

Proposition 8.1.1 *The LU factorization is computable in $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ operations.*

One says that the *complexity* of the LU factorization is $\frac{2}{3}n^3$.
Remark: When all leading principal minors but the last ($\det M$) are nonzero, the proof above furnishes a factorization $M = LU$, in which U is not invertible; that is, $u_{nn} = 0$.

8.1.1 Block Factorization

One can likewise perform a *blockwise LU* factorization. If $n = p_1 + \cdots + p_r$ with $p_j \geq 1$, the matrices L and U will be block-triangular. The diagonal blocks are square, of respective sizes $p_1, \ldots, p_r$. Those of L are of the form I_{p_j}, while those of U are invertible. A necessary and sufficient condition for such a factorization to exist is that the leading principal minors of M, of orders $p_1 + \cdots + p_j$ ($j \leq r$), be nonzero. As above, it is not necessary that the last minor $\det M$ be nonzero. Such a factorization is useful for the resolution of the linear system $MX = b$ if the diagonal blocks of U are easily inverted, for instance if their sizes are small enough (say $p_j \approx \sqrt{n}$).

An interesting application of block factorization is the computation of the determinant by the Schur complement formula:

Proposition 8.1.2 *Let $M \in \mathbf{M}_n(k)$ read blockwise*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where the diagonal blocks are square and A is invertible. Then

$$\det M = \det A \det(D - CA^{-1}B).$$

Of course, this formula generalizes $\det M = ad - bc$, which is valid only for 2×2 matrices. The matrix $D - CA^{-1}B$ is called the *Schur complement* of A in M.

Proof

Since A is invertible, M admits a blockwise LU factorization, with the same subdivision. We easily compute

$$L = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix}, \quad U = \begin{pmatrix} A & B \\ 0 & D - CA^{-1}B \end{pmatrix}.$$

Then $\det M = \det L \det U$ furnishes the expected formula.

■

Corollary 8.1.1 *Let $M \in \mathbf{GL}_n(k)$, with $n = 2m$, read blockwise*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad A, B, C, D \in \mathbf{GL}_m(k).$$

Then

$$M^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & (C - DB^{-1}A)^{-1} \\ (B - AC^{-1}D)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

Proof

We can verify the formula by multiplying by M. The only point to show is that the inverses are meaningful, that is, that $A - BD^{-1}C, \ldots$ are invertible. Because of the symmetry of the formulas, it is enough to check it for a single term, namely $D - CA^{-1}B$. However, $\det(D - CA^{-1}B) = \det M / \det A$, which is nonzero by assumption.

■

We might add that as soon as $M \in \mathbf{GL}_n(k)$ and $A \in \mathbf{GL}_p(k)$ (even if $p \neq n/2$), then

$$M^{-1} = \begin{pmatrix} \cdot & \cdot \\ \cdot & (D - CA^{-1}B)^{-1} \end{pmatrix},$$

because M admits the blockwise LU factorization and

$$M^{-1} = U^{-1}L^{-1} = \begin{pmatrix} A^{-1} & \cdot \\ 0 & (D - CA^{-1}B)^{-1} \end{pmatrix} \cdot \begin{pmatrix} I & 0 \\ \cdot & I \end{pmatrix}.$$

8.1.2 Complexity of Matrix Inversion

We can now show that the complexity of inverting a matrix is not higher than that of matrix multiplication, at equivalent sizes. We assume here that $k = \mathbb{R}$ or $\mathbb{C}$.

Notation 8.1.1 *We denote by J_n the number of operations in k used in the inversion of an $n \times n$ matrix, and by P_n the number of operations (in k) used in the product of two $n \times n$ matrices.*

Of course, the number J_n must be understood for generic matrices, that is, for matrices within a dense open subset of $\mathbf{M}_n(k)$. More important, J_n, P_n also depend on the algorithm chosen for inversion or for multiplication. In the sequel we wish to adapt the inversion algorithm to the one used for multiplication.

Let us examine first of all the matrices whose size n has the form 2^k.

We decompose the matrices $M \in \mathbf{GL}_n(k)$ blockwise, with blocks of size $n/2 \times n/2$. The condition $A \in \mathbf{GL}_{n/2}(k)$ defines a dense open set, since $M \mapsto \det A$ is a nonzero polynomial. Suppose that we are given an inversion algorithm for generic matrices in $\mathbf{GL}_{n/2}(k)$ in j_{k-1} operations. Then blockwise LU factorization and the formula $M^{-1} = U^{-1}L^{-1}$ furnish an inversion algorithm for generic matrices in $\mathbf{GL}_n(k)$. We can then bound j_k by means of j_{k-1} and the number $\pi_{k-1} = P_{2^{k-1}}$ of operations used in the computation of the product of two matrices of size $2^{k-1} \times 2^{k-1}$. We shall denote also by $\sigma_k = 2^{2k}$ the number of operations involved in the computation of the sum of matrices in $\mathbf{M}_{2^k}(k)$.

To compute M^{-1}, we first compute A^{-1}, then CA^{-1}, which gives us L^{-1} in $j_{k-1} + \pi_{k-1}$ operations. Then we compute $(D - CA^{-1}B)^{-1}$ (this amounts to $\sigma_{k-1} + \pi_{k-1} + j_{k-1}$ operations) and $A^{-1}B(D - CA^{-1}B)^{-1}$ (cost: $2\pi_{k-1}$), which furnishes U^{-1}. The computation of $U^{-1}L^{-1}$ is done at the cost $\sigma_{k-1} + 2\pi_{k-1}$. Finally,

$$j_k \leq 2j_{k-1} + 2\sigma_{k-1} + 6\pi_{k-1}.$$

In other words,

$$2^{-k}j_k - 2^{1-k}j_{k-1} \leq 2^{k-1} + 3 \cdot 2^{1-k}\pi_{k-1}. \tag{8.1}$$

The complexity of the product in $\mathbf{M}_n(k)$ obeys the inequalities

$$n^2 \leq P_n \leq n^2(2n-1).$$

The first inequality is due to the number of data ($2n^2$) and the fact that each operation involves only two of them. The second is given by the naive algorithm that consists in computing n^2 scalar products.

Lemma 8.1.1 *If $P_n \leq c_\alpha n^\alpha$ (with $2 \leq \alpha \leq 3$), then $j_l \leq C_\alpha \pi_l$, where $C_\alpha = 1 + 3c_\alpha/(2^{\alpha-1} - 1)$.*

It is enough to sum (8.1) from $k = 1$ to l and use the inequality $1 + q + \cdots + q^{l-1} \leq q^l/(q-1)$ for $q > 1$.

When n is not a power of 2, we obtain M^{-1} by computing the inverse of a block-diagonal matrix $\mathrm{diag}(M, I)$, whose size N satisfies $n \leq N = 2^l < 2n$. We obtain $J_n \leq j_l \leq C_\alpha \pi_l$. Finally, we have the following result.

Proposition 8.1.3 *If the complexity P_n of the product in $\mathbf{M}_n(\mathbb{C})$ is bounded by $c_\alpha n^\alpha$, then the complexity J_n of inversion in $\mathbf{GL}_n(\mathbb{C})$ is bounded by $d_\alpha n^\alpha$, where*

$$d_\alpha = \left(1 + \frac{3c_\alpha}{2^{\alpha-1} - 1}\right) 2^\alpha.$$

That can be summarized as follows:

Those who know how to multiply know also how to invert.

8.1.3 Complexity of the Matrix Product

The ideas that follow apply to the product of rectangular matrices, but for the sake of simplicity, we present only the case of square matrices.

As we have seen above, the complexity P_n of matrix multiplication in $M_n(k)$ is $\mathcal{O}(n^3)$. However, better algorithms will allow us to improve the exponent 3. The simplest and oldest one is Strassen's algorithm, which uses a recursion. We note first that there exists a way of computing the product of two 2×2 matrices by means of 7 multiplications and 18 additions. Two features of Strassen's formula are essential. First, the number of multiplications that it involves is stricly less than that (eight) of the naive algorithm. The second is that the method is valid when the matrices have entries in a *noncommutative* ring, and so it can be employed for two matrices $M, N \in \mathbf{M}_n(k)$, considered as elements of $\mathbf{M}_2(A)$, with $A := \mathbf{M}_{n/2}(k)$. This trick yields

$$P_n \leq 7P_{n/2} + 9n^2/2.$$

For $n = 2^l$, we then have

$$7^{-l}\pi_l - 7^{1-l}\pi_{l-1} \leq \frac{9}{2}\left(\frac{4}{7}\right)^l,$$

which, after summation from $k = 1$ to l, gives

$$7^{-l}\pi_l \leq \frac{9}{2}\frac{1}{1 - 4/7},$$

because of $\frac{4}{7} < 1$. Finally,

$$\pi_l \leq \frac{21}{2}7^l.$$

When n is not a power of two, one chooses l such that $n \leq 2^l < 2n$ and we obtain the following result.

Proposition 8.1.4 *The complexity of the multiplication of $n \times n$ matrices is $\mathcal{O}(n^\alpha)$, with $\alpha = \log 7/\log 2 = 2.807\ldots$ More precisely,*

$$P_n \leq \frac{147}{2} n^{\frac{\log 7}{\log 2}}.$$

The exponent α can be improved, at the cost of greater complication and a larger constant c_α. The best exponent known in 1997, due to Coppersmith and Winograd [11], is $\alpha = 2.376\ldots$ A rather complete analysis can be found in the book by P. Bürgisser, M. Clausen, and M. A. Shokrollahi [7].

Here is Strassen's formula [33]. Let $M, N \in \mathbf{M}_2(\mathcal{A})$, with

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad N = \begin{pmatrix} x & y \\ z & t \end{pmatrix}.$$

One first forms the expressions $x_1 = (a+d)(x+t)$, $x_2 = (c+d)x$, $x_3 = a(y-t)$, $x_4 = d(z-x)$, $x_5 = (a+b)t$, $x_6 = (c-a)(x+y)$, $x_7 = (b-d)(z+t)$. Then one computes the product

$$MN = \begin{pmatrix} x_1 + x_4 - x_5 + x_7 & x_3 + x_5 \\ x_2 + x_4 & x_1 - x_2 + x_3 + x_6 \end{pmatrix}.$$

8.2 Choleski Factorization

In this section $k = \mathbb{R}$, and we consider symmetric positive definite matrices.

Theorem 8.2.1 *Let $M \in \mathbf{SPD}_n$. Then there exists a unique lower triangular matrix $L \in \mathbf{M}_n(\mathbb{R})$, with strictly positive diagonal entries, satisfying $M = LL^T$.*

Proof

Let us begin with uniqueness. If L_1 and L_2 have the properties stated above, then $I_n = LL^T$, for $L = L_2^{-1}L_1$, which still has the same form. In other words, $L = L^{-T}$, where both sides are triangular matrices, but of opposite types (lower and upper). The equality shows that L is actually diagonal, with $L^2 = I_n$. Since its diagonal is positive, we obtain $L = I_n$; that is, $L_2 = L_1$.

We shall give two constructions of L.

First method. The matrix $M^{(p)}$ is positive definite (test the quadratic form induced by M on the linear subspace $\mathbb{R}^p \times \{0\}$). The leading principal minors of M are thus nonzero and there exists an LU factorization $M = L_0U_0$. Let D be the diagonal of U_0, which is invertible. Then $U_0 = DU_1$, where the diagonal entries of U_1 equal 1. By transposition, we have $M = U_1^T D_0 L_0^T$. From uniqueness of the LU factorization, we deduce $U_1 = L_0^T$ and $M = L_0 D L_0^T$. Then $L = \sqrt{D}L_0$ satisfies the conditions of the theorem. Observe that $D > 0$ because $D = PMP^T$, with $P = L_0^{-1}$.

Second method. We proceed by induction on n. The statement is clear if $n = 1$. Otherwise, we seek an L of the form

$$L = \begin{pmatrix} L' & 0 \\ X^T & l \end{pmatrix},$$

knowing that

$$M = \begin{pmatrix} M' & R \\ R^T & m \end{pmatrix}.$$

The matrix L' is obtained by Choleski factorization of M', which belongs to $\mathbf{SPD}_{n-1}$. Then X is obtained by solving $L'X = R$. Finally, l is a square root of $m - \|X\|^2$. Since $0 < \det M = (l \det L')^2$, we see that $m - \|X\|^2 > 0$; we thus choose $l = \sqrt{m - \|X\|^2}$. This method again shows uniqueness.

■

Remark: Choleski factorization extends to Hermitian positive definite matrices. In that case, L has complex entries, but its diagonal entries are still real and positive.

8.3 The QR Factorization

In this section $k = \mathbb{R}$ or $\mathbb{C}$, the real case being a particular case of the complex one.

Proposition 8.3.1 *Let $M \in \mathbf{GL}_n(\mathbb{C})$ be given. Then there exist a unitary matrix Q and an upper triangular matrix R, whose diagonal entries are real positive, such that $M = QR$. This factorization is unique.*

We observe that the condition on the numbers r_{jj} is essential for uniqueness. In fact, if D is diagonal with $|d_{jj}| = 1$ for every j, then $Q' := Q\bar{D}$ is unitary, $R' := DR$ is upper triangular, and $M = Q'R'$, which gives an infinity of factorizations "QU." Even in the real case, where Q is orthogonal, there are 2^n "QU" factorizations.

Proof

We first prove uniqueness. If (Q_1, R_1) and (Q_2, R_2) give two factorizations, then $Q = R$, with $Q := Q_2^{-1}Q_1$ and $R := R_2R_1^{-1}$. Since Q is unitary, we deduce $Q^* = R^{-1}$, or $Q = R^{-*}$. This shows (recall that the inverse of a triangular matrix is a triangular matrix of same type) that Q is simultaneously upper and lower triangular, and is therefore diagonal. Additionally, its diagonal part is strictly positive. Then $Q^2 = Q^*Q = I_n$ gives $Q = I_n$. Finally, $Q_2 = Q_1$ and consequently, $R_2 = R_1$.

The existence follows from that of Choleski factorization. If $M \in \mathbf{GL}_n(\mathbb{C})$, the matrix M^*M is Hermitian positive definite, hence admits a Choleski factorization R^*R, where R is upper triangular with real positive

diagonal entries. Defining $Q := MR^{-1}$, we have

$$Q^*Q = R^{-*}M^*MR^{-1} = R^{-*}R^*RR^{-1} = I_n;$$

hence Q is unitary. Finally, $M = QR$ by construction.

■

The method used above is unsatisfactory from a practical point of view, because one can compute Q and R directly, at a lower cost, without computing M^*M or its Choleski factorization. Moreover, the direct method, which we shall present now, is based on a theoretical observation: The QR factorization is nothing but the Gram–Schmidt orthonormalization procedure in $\mathbb{C}^n$, endowed with the canonical scalar product $\langle \cdot, \cdot \rangle$. In fact, if $V^1, \ldots, V^n$ denote the column vectors of M, then giving M in $\mathbf{GL}_n(\mathbb{C})$ amounts to giving a basis of $\mathbb{C}^n$. If $Y^1, \ldots, Y^n$ denote the column vectors of Q, then $\{Y^1, \ldots, Y^n\}$ is an orthonormal basis. Moreover, if $M = QR$, then

$$V^k = \sum_{j=1}^{k} r_{jk} Y^j.$$

Denoting by E_k the linear subspace spanned by $Y^1, \ldots, Y^k$, of dimension k, one sees that $V^1, \ldots, V^k$ are in E_k; that is, $\{V^1, \ldots, V^k\}$ is a basis of E_k. Hence, the columns of Q are obtained by the Gram–Schmidt procedure, applied to the columns of M.

The practical computation of Q and R is done by induction on k. If $k = 1$, then

$$r_{11} = \|V^1\|, \quad Y^1 = \frac{1}{r_{11}} V^1.$$

If $k > 1$, and if $Y^1, \ldots, Y^{k-1}$ are already known, one looks for Y^k and the entries r_{jk} $(j \leq k)$. For $j < k$, we have

$$r_{jk} = \langle V^k, Y^j \rangle.$$

Then

$$r_{kk} = \|Z_k\|, \quad Y^k = \frac{1}{r_{kk}} Z^k,$$

where

$$Z^k := V^k - \sum_{j=1}^{k-1} r_{jk} Y^j.$$

Let us examine the complexity of the procedure described above. To pass from the step $k-1$ to the step k, one computes $k-1$ scalar products, then Z^k, its norm, and finally Y^k. This requires $(4n-1)k + 3n$ operations. Summing from $k = 1$ to n yields $2n^3 + O(n^2)$ operations. This method is not optimal, as we shall see in Section 10.2.3.

The interest of this construction lies also in giving a more complete statement than Proposition 8.3.1:

Theorem 8.3.1 *Let $M \in \mathbf{M}_n(\mathbb{C})$ be a matrix of rank p. There exists $Q \in \mathbf{U}_n$ and an upper triangular matrix R, with $r_{ll} \in \mathbb{R}^+$ for every l and $r_{jk} = 0$ for $j > p$, such that $M = QR$.*

Remarks: The QR factorization of a singular matrix (i.e., a noninvertible one) is not unique. There exists, in fact, a QR factorization for rectangular matrices, in which R is a "quasi-triangular" matrix.

8.4 The Moore–Penrose Generalized Inverse

The resolution of a general linear system $Ax = b$, where A may be singular and may even not be square, is a delicate question, whose treatment is made much simpler by the use of the Moore–Penrose generalized inverse. We begin with the fundamental theorem.

Theorem 8.4.1 *Let $A \in M_{n\times m}(\mathbb{C})$ be given. There exists a unique matrix $A^\dagger \in M_{m\times n}(\mathbb{C})$, called the Moore–Penrose* generalized inverse, *satisfying the following four properties:*

1. $AA^\dagger A = A$;
2. $A^\dagger AA^\dagger = A^\dagger$;
3. $AA^\dagger \in H_n$;
4. $A^\dagger A \in H_m$.

Finally, if A has real entries, then so has $A^\dagger$.

When $A \in \mathbf{GL}_n(\mathbb{C})$, $A^\dagger$ coincides with the standard inverse A^{-1}, since the latter obviously satisfies the four properties. More generaly, if A is onto, then property 1 shows that $AA^\dagger = I_n$; i.e., $A^\dagger$ is a right inverse of A. Likewise, if A is one-to-one, then $A^\dagger A = I_m$; i.e., $A^\dagger$ is a left inverse of A.

Proof

We first remark that if X is a generalized inverse of A, that is, it satisfies these four properties, and if $U \in \mathbf{U}_n$, $V \in \mathbf{U}_m$, then V^*XU^* is a generalized inverse of UAV. Therefore, existence and uniqueness need to be proved for only a single representative D of the equivalence class of A modulo unitary multiplications on the right and the left. From Theorem 7.7.1, we may choose $D = \mathrm{diag}(s_1, \dots, s_r, 0, \dots)$, where $s_1, \dots, s_r$ are the nonzero singular values of A.

We are thus concerned only with quasi-diagonal matrices D. Let $D^\dagger$ be any generalized inverse of D, which we write blockwise as

$$D^\dagger = \begin{pmatrix} G & H \\ J & K \end{pmatrix}.$$

We use the notation of Theorem 7.7.1. From property 1, we obtain $S = SGS$, where $S := \mathrm{diag}(s_1, \ldots, s_r)$. Since S is nonsingular, we obtain $G = S^{-1}$. Next, property 3 implies $SH = 0$, that is, $H = 0$. Likewise, property 4 gives $JS = 0$, that is, $J = 0$. Finally, property 2 yields $K = JSH = 0$. We see, then, that $D^\dagger$ must equal (uniqueness)

$$\begin{pmatrix} S^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

One easily checks that this matrix solves our problem (existence).

■

Some obvious properties are stated in the following proposition. We warn the reader that, contrary to what happens for the standard inverse, the generalized inverse of AB does not need to be equal to $B^\dagger A^\dagger$.

Proposition 8.4.1 *The following equalities hold for the generalized inverse:*

$$(\lambda A)^\dagger = \frac{1}{\lambda} A^\dagger \quad (\lambda \neq 0), \quad \left(A^\dagger\right)^\dagger = A, \quad \left(A^\dagger\right)^* = (A^*)^\dagger.$$

If $A \in GL_n(\mathbb{C})$, then $A^\dagger = A^{-1}$.

Since $(AA^\dagger)^2 = AA^\dagger$, the matrix $AA^\dagger$ is a projector, which can therefore be described in terms of its range and kernel. Since $AA^\dagger$ is Hermitian, these subspaces are orthogonal to each other. Obviously, $R(AA^\dagger) \subset R(A)$. But since $AA^\dagger A = A$, the reverse inclusion holds too. Finally, we have

$$R(AA^\dagger) = R(A),$$

and $AA^\dagger$ is the orthogonal projector onto $R(A)$. Likewise, $A^\dagger A$ is an orthogonal projector. Obviously, $\ker A \subset \ker A^\dagger A$, while the identity $AA^\dagger A = A$ implies the reverse inclusion, so that

$$\ker A^\dagger A = \ker A.$$

Finally, $A^\dagger A$ is the orthogonal projector onto $(\ker A)^\perp$.

8.4.1 Solutions of the General Linear System

Given a matrix $M \in M_{n\times m}(\mathbb{C})$ and a vector $b \in \mathbb{C}^n$, let us consider the linear system

$$Mx = b. \tag{8.2}$$

In (8.2), the matrix M need not be square, even not of full rank. From property 1, a necessary condition for the solvability of (8.2) is $MM^\dagger b = b$. Obviously, this is also sufficient, since it ensures that $x_0 := M^\dagger b$ is a solution. Hence, the generalized inverse plays one of the roles of the standard inverse, namely to provide one solution of (8.2) when it is solvable. To catch every solution of that system, it remains to solve the homogeneous

problem $My = 0$. From the analysis done in the previous section, $\ker M$ is nothing but the range of $I_m - M^\dagger M$. Therefore, we may state the following proposition:

Proposition 8.4.2 *The system (8.2) is solvable if and only if* $b = MM^\dagger b$. *When it is solvable, its general solution is* $x = M^\dagger b + (I_m - M^\dagger M)z$, *where* z *ranges over* $\mathbb{C}^m$. *Finally, the special solution* $x_0 := M^\dagger b$ *is the one of least Hermitian norm.*

There remains to prove that x_0 has the smallest norm among the solutions. That comes from the Pythagorean theorem and from the fact that $R(M^\dagger) = R(M^\dagger M) = (\ker M)^\perp$.

8.5 Exercises

1. Assume that there exists an algorithm for multiplying two $N \times N$ matrices with entries in a noncommutative ring by means of K multiplications and L additions. Show that the complexity of the multiplication in $\mathbf{M}_n(k)$ is $\mathcal{O}(n^\alpha)$, with $\alpha = \log K / \log N$.

2. What is the complexity of Choleski factorization?

3. Let $M \in \mathbf{SPD}_n$ be also tridiagonal. What is the structure of L in the Choleski factorization? More generally, what is the structure of L when $m_{ij} = 0$ for $|i - j| > r$?

4. (continuation of exercise 3)
   For $i \leq n$, denote by $\phi(i)$ the smallest index j such that $m_{ij} \neq 0$. In Choleski factorization, show that $l_{ij} = 0$ for every pair (i, j) such that $j < \phi(i)$.

5. In the QR factorization, show that the map $M \mapsto (Q, R)$ is continuous on $\mathbf{GL}_n(\mathbb{C})$.

6. Let H be an $n \times n$ Hermitian matrix, that blockwise reads

$$H = \begin{pmatrix} A & B^* \\ B & C \end{pmatrix}.$$

Assume that $A \in \mathbf{HPD}_{n-k}$ $(1 \leq k \leq n-1)$.
Find a matrix T of the form

$$T = \begin{pmatrix} I_{n-k} & 0 \\ . & I_k \end{pmatrix}$$

such that THT^* is block-diagonal. Deduce that if $W \in \mathbf{H}_k$, then

$$H - \begin{pmatrix} 0 & 0 \\ 0 & W \end{pmatrix}$$

is positive (semi)definite if and only if $S - W$ is, where S is the Schur complement of A in H.

7. (continuation of exercise 6)
Fix the size k and denote by $S(H)$ the Schur complement in the Hermitian matrix H when $A \in \mathbf{HPD}_{n-k}$. Using the previous exercise, show that:

   (a) $S(H + H') - S(H) - S(H')$ is positive semidefinite.

   (b) If $H - H'$ is positive semidefinite, then so is $S(H) - S(H')$.

   In other words, $H \mapsto S$ is "concave nondecreasing" on the convex set formed of those matrices of $\mathbf{H}_n$ such that $A \in \mathbf{HPD}_{n-k}$ into the ordered set $\mathbf{H}_k$. The article [26] gives a review of the properties of the map $H \mapsto S(H)$.

8. In Proposition 8.3.1, find an alternative proof of the uniqueness part, by inspection of the spectrum of the matrix $Q := Q_2^{-1}Q_1 = R_2R_1^{-1}$.

9. Identify the generalized inverse of row matrices and column matrices.

10. What is the generalized inverse of an orthogonal projector, that is, a Hermitian matrix P satisfying $P^2 = P$? Deduce that the description of $AA^\dagger$ and $A^\dagger A$ as orthogonal projectors does not characterize $A^\dagger$ uniquely.

11. Given a matrix $B \in \mathbf{M}_{p\times q}(\mathbb{C})$ and a vector $a \in \mathbb{C}^p$, let us form the matrix $A := (B, a) \in \mathbf{M}_{p\times(q+1)}(\mathbb{C})$.

   (a) Let us define $d := B^\dagger a$, $c := a - Bd$, and

$$b := \begin{cases} c^\dagger, & \text{if } c \neq 0, \\ (1 + |d|^2)^{-1} d^* B^\dagger, & \text{if } c = 0. \end{cases}$$

   Prove that

$$A^\dagger = \begin{pmatrix} B^\dagger - db \\ b \end{pmatrix}.$$

   (b) Deduce an algorithm (*Greville*'s algorithm in $O(pq^2)$ operations for the computation of the generalized inverse of a $p \times q$ matrix.

Hint: To get started with the algorithm, use Exercise 9.

9

Iterative Methods for Linear Problems

In this chapter the field of scalars is $K = \mathbb{R}$ or $\mathbb{C}$.

We have seen in the previous Chapter a few direct methods for solving a linear system $Ax = b$, when $A \in \mathbf{M}_n(K)$ is invertible. For example, if A admits an LU factorization, the successive resolution of $Ly = b$, $Ux = y$ is called the *Gauss method*. When a leading principal minor of A vanishes, a permutation of the columns allows us to return to the generic case. More generally, the Gauss method with pivoting consists in permuting the columns at each step of the factorization in such a way as to limit the magnitude of round-off errors and that of the conditioning number of the matrices L, U.

The direct computation of the solution of a Cramer's linear system $Ax = b$, by the Gauss method or by any other direct method, is rather costly, on the order of n^3 operations. It also presents several inconveniences. On the one hand, it does not exploit completely the sparse shape of many matrices A; in numerical analysis it happens frequently that an $n \times n$ matrix has only $\mathcal{O}(n)$ nonzero entries, instead of $\mathcal{O}(n^2)$. On the other hand, the computation of an LU factorization is rather unstable, because the round-off errors produced by the computer are amplified at each step of the computation.

For these reasons, one often uses an iterative method to compute an approximate solution x^m, instead of an exact solution. The iterative methods fully exploit the sparse structure of A. The number of operations is $\mathcal{O}(am)$, where a is the number nonzero entries in A. The choice of m depends on the accuracy that one requires a priori. It is, however, modest, because the error $\|x^m - \bar{x}\|$ from the exact solution $\bar{x}$ is of order constant $\times\, k^m$,

where $k < 1$ whenever the method converges. Typically, a dozen iterations give a rather good result, and then $\mathcal{O}(10a) \ll \mathcal{O}(n^3)$. another advantage of the iterative methods is that the round-off errors are damped during the computation, instead of being amplified.

General principle: Choose a decomposition of A of the form $M - N$ and rewrite the system, assuming that M is invertible:

$$x = M^{-1}(Nx + b).$$

Then choosing a starting vector $x^0 \in K^n$, which may be a rather coarse approximation of the solution, one constructs a sequence $(x^m)_{m \in I\!N}$ by induction:

$$x^{m+1} = M^{-1}(Nx^m + b). \tag{9.1}$$

In practice, one does not compute M^{-1} explicitly but one solves the linear systems $Mx^{m+1} = \cdots$. It is thus important that this resolution be cheap. This will be the case when M is triangular. In that case, the invertibility of M can be read from its diagonal, since it occurs precisely when the diagonal entries are nonzero.

9.1 A Convergence Criterion

Definition 9.1.1 *Let us assume that A and M are invertible, $A = M - N$. We say that an iterative method is* convergent *if for every pair $(x^0, b) \in K^n \times K^n$, we have*

$$\lim_{m \to +\infty} x^m = A^{-1}b.$$

Proposition 9.1.1 *An iterative method is convergent if and only if $\rho(M^{-1}N) < 1$.*

Proof

If the method is convergent, then for $b = 0$,

$$\lim_{m \to +\infty} (M^{-1}N)^m x^0 = 0,$$

for every $x^0 \in K^n$. In other words,

$$\lim_{m \to +\infty} (M^{-1}N)^m = 0.$$

From Corollary 4.4.1, this implies $\rho(M^{-1}N) < 1$.

Conversely, if $\rho(M^{-1}N) < 1$, then by Proposition 4.4.1,

$$\lim_{m \to +\infty} (M^{-1}N)^m = 0,$$

and hence

$$x^m - A^{-1}b = (M^{-1}N)^m(x^0 - A^{-1}b) \to 0.$$

■

To be more precise, if $\|\cdot\|$ is a norm on K^n, then

$$\|x^m - A^{-1}b\| \leq \|(M^{-1}N)^m\| \, \|x^0 - A^{-1}b\|.$$

From Householder's theorem (Theorem 4.2.1), there exists for every $\epsilon > 0$ a constant $C(\epsilon) < \infty$ such that

$$\|x^m - A^{-1}b\| \leq C(\epsilon)\|x^0 - \bar{x}\|(\rho(M^{-1}N) + \epsilon)^m.$$

In most cases (in fact, when there exists an induced norm satisfying $\|M^{-1}N\| = \rho(M^{-1}N)$), one can choose $\epsilon = 0$ in this inequality such that

$$\|x^m - A^{-1}b\| = \mathcal{O}(\rho(M^{-1}N)^m).$$

The choice of a vector x^0 such that $x^0 - A^{-1}b$ is an eigenvector associated to an eigenvalue of maximal modulus shows that this inequality cannot be improved in general. For this reason, we call the positive number

$$\tau := -\log \rho(M^{-1}N)$$

the *convergence ratio* of the method. Given two convergent methods, we say that the first one converges faster than the second one if $\tau_1 > \tau_2$. For example, we say that it converges twice as fast if $\tau_1 = 2\tau_2$. In fact, with an error of order $\rho(M^{-1}N)^m = \exp(-m\tau)$, we see that the faster method needs only half as many iterations to obtain the same accuracy.

9.2 Basic Methods

There are three basic iterative methods, of which the first has only a historical or theoretical interest. Each uses the decomposition of A into three parts, a diagonal one D, a lower triangular $-E$, and an upper triangular one $-F$:

$$A = D - E - F = \begin{pmatrix} d_1 & & \\ & \ddots & -F \\ & -E & \ddots \\ & & d_n \end{pmatrix}.$$

In all cases, one assumes that D is invertible: The diagonal entries of A are nonzero.

Jacobi method: One chooses $M = D$; thus $N = E + F$. The iteration matrix is $J := D^{-1}(E + F)$. Knowing the vector x^m, one computes the components of the vector x^{m+1} by the formula

$$x_i^{m+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^m \right).$$

Gauss–Seidel method: One chooses $M = D - E$, and thus $N = F$. The iteration matrix is $G := (D-E)^{-1}F$. As we shall see below, one never computes G explicitly. One computes the approximate solutions by a *double* induction, on m on the one hand, and on $i \in \{1, \dots, n\}$ on the other hand:

$$x_i^{m+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{m+1} - \sum_{j=i+1}^{j=n} a_{ij} x_j^m \right).$$

The difference between the two methods is that in Gauss–Seidel one always uses the most recently computed values of each coordinate.

Relaxation method: It often happens that the Gauss–Seidel method converges exceedingly slowly. We thus wish to improve the Gauss–Seidel method by looking for a "best" approximated value of the x_j (with $j < i$) when computing x_i^{m+1}. Instead of being simply x_j^m, as in the Jacobi method, or x_j^{m+1}, as in that of Gauss–Seidel, this best value will be an interpolation of both (we shall see that it is merely an extrapolation). This justifies the choice of

$$M = \frac{1}{\omega} D - E, \quad N = \left(\frac{1}{\omega} - 1 \right) D + F,$$

where $\omega \in \mathbb{C}$ is a parameter. This parameter remains, in general, constant throughout the calculations. The method is called successive relaxation. When $\omega > 1$, it bears the name successive overrelaxation (SOR). The iteration matrix is

$$\mathcal{L}_\omega := (D - \omega E)^{-1}((1-\omega)D + \omega F).$$

The Gauss–Seidel method is a particular case of the relaxation method, with $\omega = 1$: $\mathcal{L}_1 = G$. Special attention is given to the choice of ω, in order to reach the minimum of $\rho(\mathcal{L}_\omega)$. The computation of the approximate solutions is done through a double induction:

$$x_i^{m+1} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{m+1} - \sum_{j=i+1}^{j=n} a_{ij} x_j^m + \left(\frac{1}{\omega} - 1 \right) a_{ii} x_i^m \right).$$

Without additional assumptions relative to the matrix A, the only result concerning the convergence is the following:

Proposition 9.2.1 *We have $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$. In particular, if the relaxation method converges for a matrix $A \in \mathbf{M}_n(\mathbb{C})$ and a parameter $\omega \in \mathbb{C}$, then*

$$|\omega - 1| < 1.$$

In other words, it is necessary that ω belong to the disk for which $(0, 2)$ is a diameter.

Proof
If the method is convergent, we have $\rho(\mathcal{L}_\omega) < 1$. However,

$$\det \mathcal{L}_\omega = \frac{\det((1-\omega)D + \omega F)}{\det(D - \omega E)} = \frac{\det((1-\omega)D)}{\det D} = (1-\omega)^n.$$

Hence

$$\rho(\mathcal{L}_\omega) \geq |\det \mathcal{L}_\omega|^{1/n} = |1-\omega|.$$

■

9.3 Two Cases of Convergence

In this section and the following one we show that simple and natural hypotheses on A imply the convergence of the classical methods. We also compare their efficiencies.

9.3.1 The Diagonally Dominant Case

We assume here that one of the following two properties is satisfied:

1. A is strictly diagonally dominant,
2. A is irreducible and strongly diagonally dominant.

Proposition 9.3.1 *Under one or the other of the hypotheses (1) and (2), the Jacobi method converges, as well as the relaxation method, with* $\omega \in (0, 1]$.

Proof

Jacobi method: The matrix $J = D^{-1}(E+F)$ is clearly irreducible if A is. Furthermore,

$$\sum_{j=1}^{n} |J_{ij}| \leq 1, \quad i = 1, \dots, n,$$

in which all inequalities are strict if (1) holds, and at least one inequality is strict under the hypothesis (2). Then either Gershgorin's theorem (Theorem 4.5.1) or its improvement, Proposition 4.5.2 for irreducible matrices, yields $\rho(J) < 1$.

Relaxation method: We assume that $\omega \in (0, 1]$. Let $\lambda \in \mathbb{C}$ be a nonzero eigenvalue of $\mathcal{L}_\omega$. It is a root of

$$\det((1-\omega-\lambda)D + \lambda\omega E + \omega F) = 0.$$

Hence, $\lambda+\omega-1$ is an eigenvalue of $A' := \omega D^{-1}(\lambda E+F)$. This matrix is irreducible when A is. Then Gershgorin's theorem (Theorem 4.5.1)

shows that

$$|\lambda+\omega-1| \leq \max\left\{ \frac{\omega}{|a_{ii}|}\left(|\lambda|\sum_{j<i}|a_{ij}|+\sum_{j>i}|a_{ij}|\right) ; 1\leq i\leq n\right\}. \tag{9.2}$$

If $|\lambda| \geq 1$, we deduce that

$$|\lambda+\omega-1| \leq \max\left\{ \frac{\omega|\lambda|}{|a_{ii}|}\sum_{j\neq i}|a_{ij}| ; 1\leq i\leq n\right\}.$$

In case (1), this yields

$$|\lambda+\omega-1| < \omega|\lambda|,$$

so that $|\lambda| \leq |\lambda+\omega-1|+|1-\omega| < |\lambda|\omega+1-\omega$; that is, $(|\lambda|-1)(1-\omega) < 0$, which is a contradiction. In case (2), Proposition 4.5.2 says that inequality (9.2) is strict. One concludes the proof the same way as in case (1).

■

Of course, this result is not fully satisfactory, since $\omega \leq 1$ is not the hypothesis that we should consider. Recall that in practice, one uses over-relaxation (that is, $\omega > 1$), which turns out to be much more efficient than the Gauss–Seidel method for an appropriate choice of the parameter.

9.3.2 The Case of a Hermitian Positive Definite Matrix

Let us begin with an intermediate result.

Lemma 9.3.1 *If A and $M^* + N$ are Hermitian positive definite (in a decomposition $A = M - N$), then $\rho(M^{-1}N) < 1$.*

Proof

Let us remark first that $M^* + N = M^* + M - A$ is necessarily Hermitian when A is.

It is therefore enough to show that $\|M^{-1}Nx\|_A < \|x\|_A$ for every nonzero $x \in \mathbb{C}^n$, where $\|\cdot\|_A$ denotes the norm associated to A:

$$\|x\|_A = \sqrt{x^*Ax}.$$

We have $M^{-1}Nx = x - y$ with $y = M^{-1}Ax$. Hence,

$$\begin{aligned} \|M^{-1}Nx\|_A^2 &= \|x\|_A^2 - y^*Ax - x^*Ay + y^*Ay \\ &= \|x\|_A^2 - y^*(M^*+N)y. \end{aligned}$$

We conclude by observing that y is not zero; hence $y^*(M^*+N)y > 0$.

■

This proof gives a slightly more precise result than what was claimed: By taking the supremum of $\|M^{-1}Nx\|_A$ on the unit ball, which is compact, we obtain $\|M^{-1}N\| < 1$ for the matrix norm induced by $\|\cdot\|_A$.

The main application of this lemma is the following theorem.

Theorem 9.3.1 *If A is Hermitian positive definite, then the relaxation method converges if and only if $|\omega - 1| < 1$.*

Proof

We have seen in Proposition 9.2.1 that the convergence implies $|\omega - 1| < 1$. Let us see the converse. We have $E^* = F$ and $D^* = D$. Thus

$$M^* + N = \left(\frac{1}{\omega} + \frac{1}{\bar{\omega}} - 1\right) D = \frac{1 - |\omega - 1|^2}{|\omega|^2} D.$$

Since D is positive definite, $M^* + N$ is positive definite if and only if $|\omega - 1| < 1$.

■

However, Lemma 9.3.1 does not apply to the Jacobi method, since the hypothesis (A positive definite) does not imply that $M^* + N = D + E + F$ must be positive definite. We shall see in an exercise that this method diverges for certain matrices $A \in \mathbf{HPD}_n$, though it converges when $A \in \mathbf{HPD}_n$ is tridiagonal.

9.4 The Tridiagonal Case

We consider here the case of tridiagonal matrices A, frequently encountered in the approximation of partial differential equations by finite differences or finite elements. The general structure of A is the following:

$$A = \begin{pmatrix} x & x' & 0 & \cdots & 0 \\ x'' & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & y' \\ 0 & \cdots & 0 & y'' & y \end{pmatrix}.$$

In other words, the entries a_{ij} are zero as soon as $|j - i| \geq 2$.

In many cases, these matrices are blockwise tridiagonal, meaning that the a_{ij} are matrices, the diagonal blocks a_{ii} being square matrices. In that case, the iterative methods also read blockwise, the decomposition $A = D - E - F$ being done blockwise. The corresponding iterative methods need the inversion of matrices of smaller sizes, namely the a_{ii}, usually done by a direct method. We shall not detail here this extension of the classical methods.

The structure of the matrix allows us to write a useful algebraic relation:

Lemma 9.4.1 *Let μ be a nonzero complex number and C a tridiagonal matrix, of diagonal C_0, of upper triangular part C_+ and lower triangular part C_-. Then*

$$\det C = \det\left(C_0 + \frac{1}{\mu}C_- + \mu C_+\right).$$

Proof
It is enough to observe that the matrix C is conjugate to

$$C_0 + \frac{1}{\mu}C_- + \mu C_+,$$

through the linear transformation matrix

$$Q_\mu = \begin{pmatrix} \mu & & & \\ & \mu^2 & & 0 \\ & & \ddots & \\ & 0 & & \ddots \\ & & & & \mu^n \end{pmatrix}.$$

■

Let us apply the lemma to the computation of the characteristic polynomial P_ω of $\mathcal{L}_\omega$. We have

$$\begin{aligned}(\det D)P_\omega(\lambda) &= \det((D-\omega E)(\lambda I_n - \mathcal{L}_\omega)) \\ &= \det((\omega+\lambda-1)D - \omega F - \lambda\omega E) \\ &= \det\left((\omega+\lambda-1)D - \mu\omega F - \frac{\lambda\omega}{\mu}E\right),\end{aligned}$$

for every nonzero μ. Let us choose for μ any square root of λ. We then have

$$\begin{aligned}(\det D)P_\omega(\mu^2) &= \det((\omega+\mu^2-1)D - \mu\omega(E+F)) \\ &= (\det D)\det((\omega+\mu^2-1)I_n - \mu\omega J).\end{aligned}$$

Finally, we have the following lemma.

Lemma 9.4.2 *If A is tridiagonal and D invertible, then*

$$P_\omega(\mu^2) = (\mu\omega)^n P_J\left(\frac{\mu^2+\omega-1}{\mu\omega}\right),$$

where P_J is the characteristic polynomial of the Jacobi matrix J.

Let us begin with the analysis of a simple case, that of the Gauss–Seidel method, for which $G = \mathcal{L}_1$.

Proposition 9.4.1 *If A is tridiagonal and D invertible, then:*

1. *$P_G(X^2) = X^n P_J(X)$, where P_G is the characteristic polynomial of the Gauss–Seidel matrix G,*

2. $\rho(G) = \rho(J)^2$,

3. *the Gauss–Seidel method converges if and only if the Jacobi method converges; moreover, in case of convergence, the Gauss–Seidel method converges twice as fast as the Jacobi method;*

4. *the spectrum of* J *is even:* $\operatorname{Sp} J = -\operatorname{Sp} J$.

Proof

Formula (1) comes from Lemma 9.4.2. The spectrum of G is thus formed of $\lambda = 0$ (which is of multiplicity $[(n+1)/2]$ at least) and of squares of the eigenvalues of J, which proves 2). Point 3 follows immediately. Finally, if $\mu \in \operatorname{Sp} J$, then $P_J(\mu) = 0$, and also $P_G(\mu^2) = 0$, so that $(-\mu)^n P_J(-\mu) = 0$. Finally, either $P_J(-\mu) = 0$, or $\mu = 0 = -\mu$, in which case $P_J(-\mu)$ also vanishes.

■

In fact, the comparison given in point 3 of the proposition holds under various assumptions. For example (see Exercises 3 and 8), it holds true when D is positive and E, F are nonnegative.

We go back to the SOR, with an additional hypothesis: The spectrum of J is real, and the Jacobi method converges. This property is satisfied, for instance, when A is Hermitian positive definite, since Theorem 9.3.1 and Proposition 9.4.1 ensure the convergence of the Jacobi method, and since J is similar to the Hermitian matrix $D^{-1/2}(E+F)D^{-1/2}$.

We also select a real ω, that is, $\omega \in (0, 2)$, taking into account Proposition 9.2.1. The spectrum of J is thus formed of the eigenvalues

$$-\lambda_r < \cdots < -\lambda_1 \leq \lambda_1 < \cdots < \lambda_r = \rho(J) < 1,$$

from Proposition 9.4.1. This notation does not mean that n be even: If n is odd, $\lambda_1 = 0$. Aside from the zero eigenvalue, which does not enter into the computation of the spectral radius, the eigenvalues of $\mathcal{L}_\omega$ are the squares of the roots of

$$\mu^2 + \omega - 1 = \mu\omega\lambda_a, \tag{9.3}$$

for $1 \leq a \leq r$. Indeed, taking $-\lambda_a$ instead of λ_a furnishes the sames squares.

Let us define $\Delta(\lambda) := \omega^2\lambda^2 + 4(1-\omega)$, the discriminant of (9.3). If $\Delta(\lambda_a)$ is negative, both roots of (9.3) are complex conjugate, hence have modulus $|\omega - 1|^{1/2}$. The case $\lambda = 0$ furnishes the same modulus. If that discriminant is strictly positive, the roots are real and of distinct modulus. One of them, denoted by μ_a, satisfies $\mu_a^2 > |\omega - 1|$, the other one satisfying the opposite inequality.

From Proposition 9.2.1, $\rho(\mathcal{L}_\omega)$ is thus equal to one of the following:

- $|\omega - 1|$, if $\Delta(\lambda_a) \leq 0$ for every a, that is, if $\Delta(\rho(J)) \leq 0$;
- the maximum of the μ_a^2's defined above, otherwise.

The first case corresponds to the choice $\omega \in [\omega_J, 2)$, where

$$\omega_J = 2\frac{1-\sqrt{1-\rho(J)^2}}{\rho(J)^2} = \frac{2}{1+\sqrt{1-\rho(J)^2}} \in [1,2).$$

Then $\rho(\mathcal{L}_\omega) = \omega - 1$.

The second case is $\omega \in (0, \omega_J)$. If $\Delta(\lambda_a) > 0$, let us denote by $Q_a(X)$ the polynomial $X^2 + \omega - 1 - X\omega\lambda_a$. The sum of its roots being positive, μ_a is the largest one; it is thus positive. Moreover, $Q_a(1) = \omega(1-\lambda_a) > 0$ shows that both roots belong to the same half-line of $\mathbb{R} \setminus \{1\}$. Since their product has modulus less than or equal to one, they are less than or equal to one. In particular,

$$|\omega - 1|^{1/2} < \mu_a < 1.$$

This shows that $\rho(\mathcal{L}_\omega) < 1$ holds for every $\omega \in (0,2)$. Under our hypotheses, the relaxation method is convergent.

If $\lambda_a \neq \rho(J)$, we have $Q_r(\mu_a) = \mu_a\omega(\lambda_a - \rho(J)) < 0$. Hence, μ_a lies between both roots of Q_r, so that $\mu_a < \mu_r$. Finally, the case $\Delta(\rho(J)) \leq 0$ furnishes $\rho(\mathcal{L}_\omega) = \mu_r^2$. We then have

$$(2\mu_r - \omega\rho(J))\frac{d\mu_r}{d\omega} + 1 - \mu_r\rho(J) = 0.$$

Since $2\mu_r$ is larger than the sum $\omega\rho(J)$ of the roots and since $\mu_r, \rho(J) \in [0,1)$, one deduces that $\omega \mapsto \rho(\mathcal{L}_\omega)$ is nonincreasing over $(0, \omega_J)$.

We conclude that $\rho(\mathcal{L}_\omega)$ reaches its minimum at ω_J, that minimum being

$$\omega_J - 1 = \frac{1-\sqrt{1-\rho(J)^2}}{1+\sqrt{1-\rho(J)^2}}.$$

Theorem 9.4.1 *[See Figure 9.1] Suppose that A is tridiagonal, D is invertible, and that the eigenvalues of J are real and belong to $(-1,1)$. Assume also that $\omega \in \mathbb{R}$.*

Then the relaxation method converges if and only if $\omega \in (0,2)$. Furthermore, the convergence ratio is optimal for the parameter

$$\omega_J := \frac{2}{1+\sqrt{1-\rho(J)^2}} \in [1,2),$$

where the spectral radius of $\mathcal{L}_{\omega_J}$ is

$$(\omega_J - 1 =) \quad \frac{1-\sqrt{1-\rho(J)^2}}{1+\sqrt{1-\rho(J)^2}} = \left(\frac{1-\sqrt{1-\rho(J)^2}}{\rho(J)}\right)^2.$$

Remarks:

- We shall see in Exercise 7 that Theorem 9.4.1 extends to complex values of ω: Under the same assumptions, $\rho(\mathcal{L}_\omega)$ is minimal at ω_J, and the relaxation method converges if and only if $|\omega - 1| < 1$.

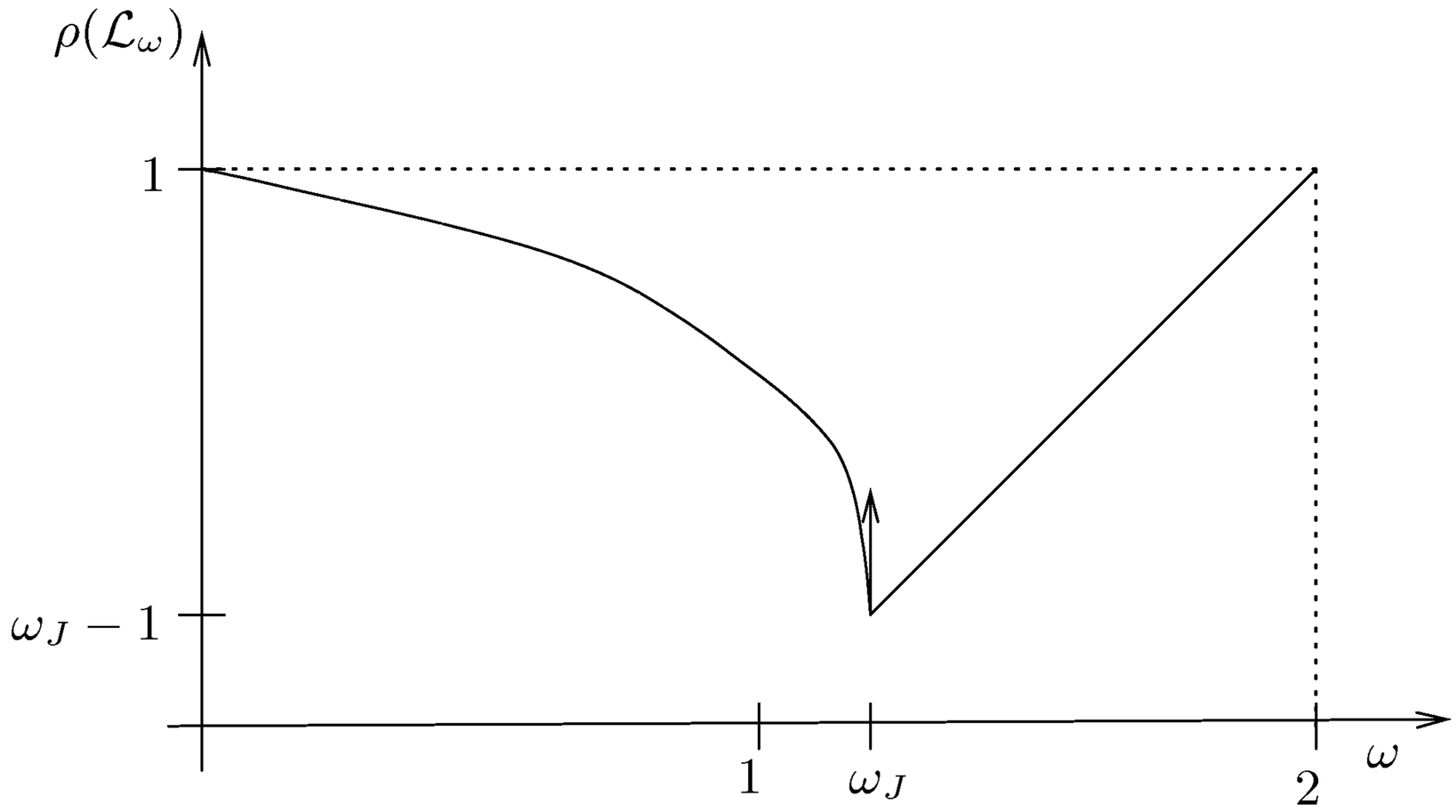

Figure 9.1. $\rho(\mathcal{L}_\omega)$ in the tridiagonal case.

- The Gauss–Seidel method is not optimal in general; $\omega_J = 1$ holds only when $\rho(J) = 0$, though in practice $\rho(J)$ is close to 1. A typical example is the resolution of an elliptic PDE by the finite element method.
  For values of $\rho(J)$ that are not too close to 1, the relaxation method with optimal parameter ω_J, though improving the convergence ratio, is not overwhelmingly more efficient than Gauss–Seidel. In fact,

$$\rho(G)/\rho\left(\mathcal{L}_{\omega_J}\right) = \left(1 + \sqrt{1 - \rho(J)^2}\right)^2$$

lies between 1 (for $\rho(J)$ close to 1) and 4 (for $\rho(J) = 0$), so that the ratio

$$\log \rho(\mathcal{L}_{\omega_J}) / \log \rho(G)$$

remains moderate, as long as $\rho(J)$ keeps away from 1. However, in the realistic case where $\rho(J)$ is close to 1, we have

$$\log \rho(G) / \log \rho(\mathcal{L}_{\omega_J}) \sim \sqrt{\frac{1 - \rho(J)}{2}},$$

which is very small. The number of iterations needed for a prescribed accuracy is multiplied by that ratio when one replaces the Gauss–Seidel method by the relaxation method with the optimal parameter.

9.5 The Method of the Conjugate Gradient

We present here the *conjugate gradient* method in the most appropriate framework, namely that of systems $Ax = b$ where A is real symmetric positive definite ($A \in \mathbf{SPD}_n$). As we shall see below, it is a *direct* method, in the sense that it furnishes the solution $\bar{x}$ after a finite number of iterations

(at most n). However, the round-off errors pollute the final result, and we would prefer to consider the conjugate gradient as an *iterative* method in which the number N of iterations, much less than n, gives a rather good approximation of $\bar{x}$. We shall see that the choice of N is linked to the *condition number* of the matrix A.

We denote by $\langle \cdot, \cdot \rangle$ the canonical scalar product on $\mathbb{R}^n$. When $A \in \mathbf{SPD}_n$ and $b \in \mathbb{R}^n$, the function

$$x \mapsto J(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

is strictly convex and tends to infinity as $\|x\| \to +\infty$. It thus reaches its infimum at a unique point $\bar{x}$, which is the unique vector where the gradient of J vanishes. We shall denote by r (for *residue*) the gradient of J: $r(x) = Ax - b$. Hence $\bar{x}$ is the solution of the linear system $Ax = b$.

If $A\bar{x} = b$ and $x \in \mathbb{R}^n$, $x \neq \bar{x}$, then

$$J(x) = J(\bar{x}) + \frac{1}{2}\langle A(x - \bar{x}), x - \bar{x}\rangle > J(\bar{x}). \tag{9.4}$$

The conjugate gradient is thus a descent method.

We shall denote by E the quadratic form associated to A: $E(x) := \langle Ax, x\rangle$. It is the square of a norm of $\mathbb{R}^n$. The character $\perp_A$ indicates the orthogonality with respect to the scalar product defined by A.

9.5.1 A Theoretical Analysis

Let $x_0 \in \mathbb{R}^n$ be given. We define $e_0 = x_0 - \bar{x}$, $r_0 = r(x_0) = Ae_0$. We may assume that $e_0 \neq 0$; otherwise, we would already have the solution. For $k \geq 1$, let us define the vector space

$$\mathcal{H}_k := \{P(A)r_0 \,|\, P \in \mathbb{R}[X], \deg P \leq k - 1\}, \quad \mathcal{H}_0 = \{0\}.$$

In $\mathcal{H}_{k+1}$, the linear subspace $\mathcal{H}_k$ is of codimension 0 or 1. In the first case, $\mathcal{H}_{k+1} = \mathcal{H}_k$, and it follows that $\mathcal{H}_{k+2} = A\mathcal{H}_{k+1} + \mathcal{H}_{k+1} = A\mathcal{H}_k + \mathcal{H}_k = \mathcal{H}_{k+1} = \mathcal{H}_k$ and thus by induction, $\mathcal{H}_k = \mathcal{H}_m$ for every $m > k$. Let us denote by l the smallest index such that $\mathcal{H}_l = \mathcal{H}_{l+1}$. For $k < l$, $\mathcal{H}_k$ is thus of codimension one in $\mathcal{H}_{k+1}$, while if $k \geq l$, then $\mathcal{H}_k = \mathcal{H}_{k+1}$. It follows that $\dim \mathcal{H}_k = k$ if $k \leq l$. In particular, $l \leq n$.

One can always find, by Gram–Schmidt orthonormalization, an A-orthogonal[1] basis (that is, such that $\langle Ap_j, p_i\rangle = 0$ if $i \neq j$) $\{p_0, \ldots, p_{l-1}\}$ of $\mathcal{H}_l$ such that $\{p_0, \ldots, p_{k-1}\}$ is a basis of $\mathcal{H}_k$ when $k \leq l$. The vectors p_j, which are not necessarily unit vectors, are defined, up to a scalar multiple, by

$$p_k \in \mathcal{H}_{k+1}, \quad p_k \perp_A \mathcal{H}_k.$$

[1] One must distinguish in this section between the two scalar products, namely $\langle \cdot, \cdot \rangle$ and $\langle A\cdot, \cdot \rangle$.

One says that the vectors p_j are pairwise *conjugate*. Of course, conjugation means A-orthogonality. This explains the name of the method.

The quadratic function J, strictly convex, reaches its infimum on the affine subspace $x_0 + \mathcal{H}_k$ at a unique vector, which we denote by x_k. This notation makes sense for $k = 0$. If $x = y+\gamma p_k \in x_0+\mathcal{H}_{k+1}$ with $y \in x_0+\mathcal{H}_k$, then

$$\begin{aligned} J(x) &= J(\bar{x}) + \frac{1}{2}E(x-\bar{x}) \\ &= J(\bar{x}) + \frac{1}{2}E(y-\bar{x}) + \frac{1}{2}\gamma^2 E(p_k) + \gamma\langle Ap_k, y-\bar{x}\rangle \\ &= J(y) + \frac{1}{2}\gamma^2 E(p_k) - \gamma\langle Ap_k, e_0\rangle, \end{aligned}$$

since $\langle Ap_k, y - x_0\rangle = 0$. Hence, minimizing J over $x_0 + \mathcal{H}_{k+1}$ amounts to minimizing J over $x_0 + \mathcal{H}_k$, together with minimizing $\gamma \mapsto \frac{1}{2}\gamma^2 E(p_k) - \gamma\langle p_k, r_0\rangle$ over $\mathbb{R}$. We therefore have

$$x_{k+1} - x_k \in \mathbb{R}p_k. \tag{9.5}$$

By definition of l there exists a nonzero polynomial P of degree l such that $P(A)r_0 = 0$, that is, $AP(A)e_0 = 0$. Since A is invertible, $P(A)e_0 = 0$. Let us assume that $P(0)$ vanishes. Then $P(X) = XQ(X)$ with $\deg Q = l-1$. Therefore, $Q(A)r_0 = 0$: The map $S \mapsto S(A)r_0$ is not one-to-one over the polynomials of degree less than or equal to $l - 1$. Hence $\dim \mathcal{H}_l < l$, a contradiction. Hence $P(0) \neq 1$, and we may assume that $P(0) = 1$. Then $P(X) = 1 - XR(X)$, where $\deg R = l - 1$. Thus $e_0 = R(A)r_0 \in \mathcal{H}_l$ or, equivalently, $\bar{x} \in x_0 + \mathcal{H}_l$. Conversely, if $k \leq l$ and $\bar{x} \in x_0 + \mathcal{H}_k$, then $e_0 \in \mathcal{H}_k$; that is, $e_0 = Q(A)r_0$, where $\deg Q \leq k - 1$. Then $Q_1(A)e_0 = 0$, because $Q_1(X) = 1 - XQ(X)$. Therefore, $Q_1(A)r_0 = 0$, $Q_1(0) \neq 0$, and $\deg Q_1 \leq k$. Hence $k \geq l$; that is, $k = l$. Summing up, we have $\bar{x} \in x_0 + \mathcal{H}_l$ but $\bar{x} \notin x_0 + \mathcal{H}_{l-1}$. Therefore, $x_l = \bar{x}$ and $x_k \neq \bar{x}$ if $k < l$.

Lemma 9.5.1 *Let us denote by $\lambda_n \geq \cdots \geq \lambda_1 (> 0)$ the eigenvalues of A. If $k \leq l$, then*

$$E(x_k - \bar{x}) \leq E(e_0) \cdot \min_{\deg Q \leq k-1} \max_j |1 + \lambda_j Q(\lambda_j)|^2.$$

Proof
Let us compute

$$\begin{aligned} E(x_k - \bar{x}) &= \min\{E(x-\bar{x}) \,|\, x \in x_0 + \mathcal{H}_k\} \\ &= \min\{E(e_0 + y) \,|\, y \in \mathcal{H}_k\} \\ &= \min\{E((I_n + AQ(A))e_0) \,|\, \deg Q \leq k-1\} \\ &= \min\{\|(I_n + AQ(A))A^{1/2}e_0\|_2^2 \,|\, \deg Q \leq k-1\}, \end{aligned}$$

where we have used the equality $\langle Aw, w\rangle = \|A^{1/2}w\|_2^2$. Hence

$$\begin{aligned} E(x_k - \bar{x}) &\leq \min\{\|I_n + AQ(A)\|_2^2 \|A^{1/2}e_0\|_2^2 \mid \deg Q \leq k-1\} \\ &= E(e_0)\min\{\rho(I_n + AQ(A))^2 \mid \deg Q \leq k-1\}, \end{aligned}$$

since $\rho(S) = \|S\|_2$ holds for every real symmetric matrix.

■

From Lemma 9.5.1, we deduce an estimate of the error $E(x_k - \bar{x})$ by bounding the right-hand side by

$$\min_{\deg Q \leq k-1} \max_{t\in[\lambda_1,\lambda_n]} |1 + tQ(t)|^2.$$

Classically, the minimum is reached for

$$1 + XQ(X) = \omega_k T_k\left(\frac{2X - \lambda_1 - \lambda_n}{\lambda_n - \lambda_1}\right),$$

where T_k is a Chebyshev polynomial:

$$T_k(t) = \begin{cases} \cos k \arccos t & \text{if} \quad |t| \leq 1, \\ \cosh\, k \operatorname{arcosh} t & \text{if} \quad t \geq 1, \\ (-1)^k \cosh\, k \operatorname{arcosh} |t| & \text{if} \quad t \leq -1. \end{cases}$$

The number ω_k is the number that furnishes the value 1 at $X = 0$, namely

$$\omega_k = \frac{(-1)^k}{T_k\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)}.$$

Then

$$\max_{[\lambda_1,\lambda_n]} |1 + tQ(t)| = |\omega_k| = \frac{1}{\cosh\, k \operatorname{arcosh}\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}}.$$

Hence $E(x_k - \bar{x}) \leq |\omega_k|^2 E(e_0)$. However, if

$$\theta := \operatorname{arrcosh}\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1},$$

then $|\omega_k| = (\cosh k\theta)^{-1} \leq 2\exp(-k\theta)$, while $\exp(-\theta)$ is the root, less than one, of the quadratic polynomial

$$T^2 - 2\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}T + 1.$$

Setting $K(A) := \|A\|_2\|A^{-1}\|_2 = \lambda_n/\lambda_1$ the *condition number* of A, we obtain

$$e^{-\theta} = \frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1} - \sqrt{\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)^2 - 1} = \frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}} = \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}.$$

The final result is the following.

Theorem 9.5.1 *If $k \leq l$, then*

$$E(x_k - \bar{x}) \leq 4E(x_0 - \bar{x}) \left(\frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1} \right)^{2k}. \tag{9.6}$$

We now set $r_k = r(x_k) = A(x_k - \bar{x})$. We have seen that $r_l = 0$ and that $r_k \neq 0$ if $k < l$. In fact, r_k is the gradient of J at x_k. The minimality of J at x_k over $x_0 + \mathcal{H}_k$ thus implies that $r_k \perp \mathcal{H}_k$ (for the usual scalar product). In other words, we have $\langle r_k, p_j \rangle = 0$ if $j < k$. However, $x_k - \bar{x} \in e_0 + \mathcal{H}_k$ can also be written as $x_k - \bar{x} = Q(A)e_0$ with $\deg Q \leq k$, which implies $r_k = Q(A)r_0$, so that $r_k \in \mathcal{H}_{k+1}$. If $k < l$, one therefore has $\mathcal{H}_{k+1} = \mathcal{H}_k \oplus \mathbb{R}r_k$.

We now normalize p_k (which was not done up to now) by

$$p_k - r_k \in \mathcal{H}_k.$$

In other words, p_k is the A-orthogonal projection of $r_k = r(x_k)$, parallel to $\mathcal{H}_k$. It is actually an element of $\mathcal{H}_{k+1}$, since $r_k \in \mathcal{H}_{k+1}$. It is also nonzero since $r_k \notin \mathcal{H}_k$. We note that r_k is orthogonal to $\mathcal{H}_k$ with respect to the usual scalar product, though p_k is orthogonal to $\mathcal{H}_k$ with respect to the A-scalar product; this explains why p_k and r_k are generally different.

If $j \leq k-2$, we compute $\langle A(p_k - r_k), p_j \rangle = -\langle Ar_k, p_j \rangle = -\langle r_k, Ap_j \rangle = 0$. We have used successively the conjugation of the p_k, the symmetry of A, the fact that $Ap_j \in \mathcal{H}_{j+2}$, and the orthogonality of r_k and $\mathcal{H}_k$. We have therefore $p_k - r_k \perp_A \mathcal{H}_{k-1}$, so that

$$p_k = r_k + \delta_k p_{k-1} \tag{9.7}$$

for a suitable number δ_k.

9.5.2 Implementing the Conjugate Gradient

The main feature of the conjugate gradient is the simplicity of the computation of the vectors x_k, which is done by induction. To begin with, we have $p_0 = r_0 = Ax_0 - b$, where x_0 is at our disposal. Let us assume now that x_k and p_{k-1} are known. Then $r_k = Ax_k - b$. If $r_k = 0$, we already have the solution. Otherwise, the formulas (9.5, 9.7) show that in fact, x_{k+1} minimizes J over the plane $x_k + \mathbb{R}r_k \oplus \mathbb{R}p_{k-1}$. We therefore have $x_{k+1} = x_k + \alpha_k r_k + \beta_k p_{k-1}$, where the entries α_k, β_k are obtained by solving the linear system of two equations

$$\begin{cases} \alpha_k \langle Ar_k, r_k \rangle + \beta_k \langle Ar_k, p_{k-1} \rangle + \|r_k\|^2 = 0, \\ \alpha_k \langle Ar_k, p_{k-1} \rangle + \beta_k \langle Ap_{k-1}, p_{k-1} \rangle = 0 \end{cases}$$

(we have used $\langle r_k, p_{k-1} \rangle = 0$). Then we have $\delta_k = \beta_k / \alpha_k$. Observe that α_k is nonzero, because otherwise β_k would vanish and r_k would too.

Summing up, the algorithm reads as follows

- Choose x_0; then define $p_0 = r_0 = r(x_0) := Ax_0 - b$.

- For $k \geq 0$ with unit increment, do
  - Compute $r_k = r(x_k) = Ax_k - b$. If $r_k = 0$, then $\bar{x} = x_k$.
  - Otherwise, minimize $J(x_k + \alpha r_k + \beta p_{k-1})$, by computing α_k, β_k as above.
  - Define

$$p_{k+1} = r_k + (\beta_k/\alpha_k)p_{k-1}, \quad x_{k+1} = x_k + \alpha_k p_k.$$

A priori, this computation furnishes the exact solution $\bar{x}$ in l iterations. However, l equals n in general, and the cost of each iteration is $O(n^2)$. The conjugate gradient, viewed as a direct method, is thus rather slow. One often uses this method for sparse matrices, whose maximal number of nonzero elements m per rows is small compared to n. The complexity of an iteration is then $O(mn)$. However, that is still rather costly as a direct method ($O(mn^2)$ operations in all), since the complexity of iterative methods is also reduced for sparse matrices.

This explains why one prefers to consider the conjugate gradient as an *iterative* method, in which one makes only a few iterations $N \ll n$. Strictly speaking, Theorem 9.5.1 does not define a convergence rate τ, since one does not have, in general, an inequality of the form

$$\|x_{k+1} - \bar{x}\| \leq e^{-\tau}\|x_k - \bar{x}\|.$$

In particular, one is not certain that $\|x_1 - \bar{x}\|$ is smaller than $\|x_0 - \bar{x}\|$. However, the inequality (9.6) is analogous to what we have for a classical iterative method, up to the factor 4. We shall therefore say that the conjugate gradient admits a *convergence rate* τ_{CG} that satisfies

$$\tau_{\mathrm{CG}} \leq \theta = -\log \frac{\sqrt{K(A)} - 1}{\sqrt{K(A)} + 1}. \tag{9.8}$$

This rate is equivalent to $2K(A)^{-1/2}$ when $K(A)$ is large. This method can be considered as an iterative method when $n\tau_{\mathrm{CG}} \ll 1$ since then it is possible to choose $N \ll n$. Obviously, a sufficient condition is $K(A) \ll n^2$.
Application: Let us consider the resolution of the Laplace equation in an open bounded set Ω of $\mathbb{R}^d$, with a Dirichlet boundary condition, by the finite elements method:

$$\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega.$$

The matrix A is symmetric, reflecting the symmetry of the variational formulation

$$\int_\Omega (\nabla u \cdot \nabla v + fv)\, dx = 0, \quad \forall v \in H_0^1(\Omega).$$

If the diameter of the grid is h with $0 < h \ll 1$, and if that grid is regular enough, the number of degrees of freedom (the size of the matrix) n is of order C/h^d, where C is a constant. The matrix is sparse with $m = O(1)$.

Each iteration thus needs $O(n)$ operations. Finally, the condition number of A is of order c/h^2. Hence, a number of iterations $N \gg 1/h$ is appropriate. This is worthwhile as soon as $d \geq 2$. The method becomes more useful as d grows larger and the threshold $1/h$ is independent of the dimension.

Preconditioning: In practice, the performance of the method is improved by *preconditioning* the matrix A. The idea is to replace the system $Ax = b$ by $B^T ABy = B^T b$, where the inversion of B is easy, for example B is block-triangular or block-diagonal with small blocks. If BB^T is close enough to A^{-1}, the condition number of the new matrix is smaller, and the number of iterations is reduced. Actually, when the condition number reaches its infimum $K = 1$, we have $A = I_n$, and the solution $\bar{x} = b$ is obvious. The simplest preconditioning consists in choosing $B = D^{-1/2}$. Its efficiency is clear in the (trivial) case where A is diagonal, because the matrix of the new system is I_n, and the condition number is lowered to 1. Observe that preconditioning is also used with SOR, because it allows us to diminish the value of $\rho(J)$, hence also the convergence rate. We shall see in Exercise 5 that, if $A \in \mathbf{SPD}_n$ is tridiagonal and if $D = dI_n$ (which corresponds to the preconditioning described above), the conjugate gradient method is twice as slow as the relaxation method with optimal parameter; that is,

$$\theta = \frac{1}{2}\tau_{\mathrm{RL}}.$$

This equality is obtained by computing θ and the optimal convergence rate τ_{RL} of the relaxation method in terms of $\rho(J)$. In the real world, in which A might not be tridiagonal, or be only blockwise tridiagonal, the map $\rho(J) \mapsto \theta$ remains the same, while τ_{RL} deteriorates. The conjugate gradient method becomes more efficient than the relaxation method. It has also the advantage that it does not need the preliminary computation of $\rho(J)$, in contrast to the relaxation method with optimal parameter.

The reader will find a deeper analysis of the method of the conjugate gradient in the article of J.-F. Maître in [1].

9.6 Exercises

1. Let A be a tridiagonal matrix with an invertible diagonal and let J be its Jacobi matrix. Show that J is conjugate to $-J$. Compare with Proposition 9.4.1.

2. We fix $n \geq 2$. Use Theorem 3.4.2 to construct a matrix $A \in \mathbf{SPD}_n$ for which the Jacobi method does not converge. Show in particular that

$$\sup\{\rho(J) \,|\, A \in \mathbf{SPD}_n,\, D = I_n\} = n - 1.$$

3. Let $A \in \mathbf{M}_n(\mathbb{R})$ satisfy $a_{ii} > 0$ for every index i, and $a_{ij} \leq 0$ whenever $j \neq i$. Using (several times) the weak form of the Perron–

Frobenius theorem, prove that either $1 \leq \rho(J) \leq \rho(G)$ or $\rho(G) \leq \rho(J) \leq 1$. In particular, as in point 3 of Proposition 9.4.1, the Jacobi and Gauss–Seidel methods converge or diverge simultaneously, and Gauss–Seidel is faster in the former case. Hint: Prove that

$$(\rho(G) \geq 1) \Longrightarrow (\rho(J) \geq 1) \Longrightarrow (\rho(G) \geq \rho(J))$$

and

$$(\rho(G) \leq 1) \Longrightarrow (\rho(J) \geq \rho(G)).$$

4. Let $n \geq 2$ and $A \in \mathbf{HPD}_n$ be given. Assume that A is tridiagonal.
   (a) Verify that the spectrum of J is real and even.
   (b) Show that the eigenvalues of J satisfy $\lambda < 1$.
   (c) Deduce that the Jacobi method is convergent.

5. Let $A \in \mathbf{HPD}_n$, $A = D - E - E^*$. Use the Hermitian norm $\|\cdot\|_2$.
   (a) Show that $|((E + E^*)v, v)| \leq \rho(J)\|D^{1/2}v\|^2$ for every $v \in \mathbb{C}^n$. Deduce that
   $$K(A) \leq \frac{1 + \rho(J)}{1 - \rho(J)} K(D).$$
   (b) Let us define a function by
   $$g(x) := \frac{\sqrt{x} - 1}{\sqrt{x} + 1}.$$
   Verify that
   $$g\left(\frac{1 + \rho(J)}{1 - \rho(J)}\right) = \frac{1 - \sqrt{1 - \rho(J)^2}}{\rho(J)}.$$
   (c) Deduce that if A is tridiagonal and if $D = dI_n$, then the convergence ratio θ of the conjugate gradient is the half of that of SOR with optimal parameter.

6. Here is another proof of Theorem 9.3.1, when ω is real. Let $A \in \mathbf{HPD}_n$.
   (a) Suppose we are given $\omega \in (0, 2)$.
      i. Assume that $\lambda = e^{2i\theta}$ (θ real) is an eigenvalue of $\mathcal{L}_\omega$. Show that $(1 - \omega - \lambda)e^{-i\theta} \in \mathbb{R}$.
      ii. Deduce that $\lambda = 1$, then show that this case is impossible too.
      iii. Let $m(\omega)$ be the number of eigenvalues of $\mathcal{L}_\omega$ of modulus less than or equal to one (counted with multiplicities). Show that m is constant on $(0, 2)$.
   (b) i. Compute
   $$\lim_{\omega \to 0} \frac{1}{\omega}(\mathcal{L}_\omega - I_n).$$

ii. Deduce that $m = n$, hence that the SOR converges for every $\omega \in (0, 2)$.

7. (Extension of Theorem 9.4.1 to complex values of ω). We still assume that A is tridiagonal, that the Jacobi method converges, and that the spectrum of J is real. We retain the notation of Section 9.4.

   (a) Given an index a such that $\lambda_a > 0$, verify that $\Delta(\lambda_a)$ vanishes for two real values of ω, of which only one, denoted by ω_a, belongs to the open disk $D = D(1; 1)$. Show that $1 < \omega_a < 2$.

   (b) Show that if $\omega \in D \setminus [\omega_a, 2)$, then the roots of $X^2 + \omega - 1 - \omega\lambda_a X$ have distinct moduli, with one and only one of them, denoted by $\mu_a(\omega)$, of modulus larger than $|\omega - 1|^{1/2}$.

   (c) Show that $\omega \mapsto \mu_a$ is holomorphic on its domain, and that

$$\begin{aligned} \lim_{|\omega-1|\to 1} |\mu_a(\omega)|^2 &= 1, \\ \lim_{\omega\to\gamma} |\mu_a(\omega)|^2 &= \gamma - 1 \quad \text{if } \gamma \in [\omega_a, 2). \end{aligned}$$

   (d) Deduce that $|\mu_a(\omega)| < 1$ (use the maximum principle), then that the relaxation method converges for every $\omega \in D$.

   (e) Show, finally, that the spectral radius of $\mathcal{L}_\omega$ is minimal for $\omega = \omega_r$, which previously was denoted by ω_J.

8. Let B be a cyclic matrix of order three. With square diagonal blocks, it reads blockwise as

$$B = \begin{pmatrix} 0 & 0 & M_1 \\ M_2 & 0 & 0 \\ 0 & M_3 & 0 \end{pmatrix}.$$

   We wish to compare the Jacobi and Gauss–Seidel methods for the matrix $A := I - B$. Compute the matrix G. Show that $\rho(G) = \rho(J)^3$. Deduce that both methods converge or diverge simultaneously and that, in case of convergence, Gauss–Seidel is three times faster than Jacobi. Show that for A^T, the convergence or the divergence still holds simultaneously, but that Gauss–Seidel is only one and a half times faster. Generalize to cyclic matrices of any order p.

10

Approximation of Eigenvalues

The computation of the eigenvalues of a square matrix is a problem of considerable difficulty. The naive idea, according to which it is enough to compute the characteristic polynomial and then find its roots, turns out to be hopeless because of Abel's theorem, which states that the general equation $P(x) = 0$, where P is a polynomial of degree $d \geq 5$, is not solvable using algebraic operations and roots of any order. For this reason, there exists no direct method, even an expensive one, for the computation of $\mathrm{Sp}(M)$.

Dropping half of that program, one could compute the characteristic polynomial exactly, then compute an approximation of its roots. But the cost and the instability of the computation are prohibitive. Amazingly, the opposite strategy is often used: A standard algorithm for computing the roots of a polynomial of high degree consists in forming its companion matrix[1] and then applying to this matrix the QR algorithm to compute its eigenvalues with good accuracy.

Hence, all the methods are iterative. In particular, we shall limit ourselves to the cases $K = \mathbb{R}$ or $\mathbb{C}$. The general strategy consists in constructing a sequence of matrices

$$M^{(0)}, M^{(1)}, \ldots, M^{(m)}, \ldots,$$

[1] Fortunately, the companion matrix is a Hessenberg matrix; see below for this notion and its practical aspects.

pairwise similar, whose structure has some convergence property. Each method is conceived in such a way that the sequence converges to a simple form, triangular or diagonal, since then the eigenvalues can be read on the diagonal. Such convergence is not always possible. For example, an algorithm in $\mathbf{M}_n(\mathbb{R})$ cannot converge to a triangular form when the matrix under consideration possesses a pair of nonreal eigenvalues.

There are two strategies for the choice of $M^{(0)}$. One can naively take $M^{(0)} = M$. But since an iteration on a generic matrix is rather costly, one often uses a preliminary reduction to a simple form (for example the Hessenberg form, in the QR algorithm), which is preserved throughout the iterations. With a few such tricks, certain methods can be astonishingly efficient. The danger of iterative methods is the possible growth of round-off errors and errors in the data. Typically, a procedure that doubles the errors at each step transforms an initial error of size 10^{-3} into an $O(1)$ after ten iterations, which is by no means acceptable. For this reason, it is important that the passage of $M^{(m)}$ to $M^{(m+1)}$ be *contracting*, that is, that the errors be damped, or at worst not be amplified. Since $M^{(m+1)}$ is conjugate to $M^{(m)}$ by some matrix P (which in fact depends on m), the growth rate is approximately the number $K(P) := \|P\| \cdot \|P^{-1}\|$, called the *condition number*, which is always greater than or equal to one. Using the induced norm $\|\cdot\|_2$, it equals 1 if and only if P is a similitude matrix; that is, $P \in \mathbb{C}^* \cdot \mathbf{U}_n$. For this reason, each iterative method builds sequences of *unitarily* similar matrices: The conjugation matrices $P^{(m)}$ are unitary (orthogonal if the ground field is $\mathbb{R}$).

10.1 Hessenberg Matrices

Definition 10.1.1 *A square matrix $M \in \mathbf{M}_n(K)$ is called* upper Hessenberg *(one speaks simply of a Hessenberg matrix) if $m_{jk} = 0$ for every pair (j,k) such that $j - k \geq 2$.*

A Hessenberg matrix thus has the form

$$\begin{pmatrix} x & \cdots & \cdots & & \\ y & \ddots & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & z & t \end{pmatrix}.$$

In particular, an upper triangular matrix is a Hessenberg matrix.

When computing the spectrum of a given matrix, we may always restrict ourselves to the case of an irreducible matrix, using a conjugation by a permutation matrix: If M is reducible, we may limit ourselves to a block-triangular matrix whose diagonal blocks are irreducible. It is enough then

to compute the spectrum of each diagonal block. This principle applies as well to a Hessenberg matrix. Hence one may always assume that M is Hessenberg and that the $m_{j+1,j}$'s are nonzero. In that case, the eigenspaces have dimension one. In fact, if $\lambda \in \bar{K}$, let L be the matrix extracted from $M - \lambda I_n$ by deletion of the first row and the last column. It is a triangular matrix of $\mathbf{M}_{n-1}(\bar{K})$, invertible because its diagonal entries, the $m_{j+1,j}$'s, are nonzero. Hence, $M - \lambda I_n$ is of rank at least equal to $n-1$, which implies that the dimension of $\ker(M - \lambda I_n)$ equals at most one.

Proposition 10.1.1 *If $M \in \mathbf{M}_n(K)$ is a Hessenberg matrix with $m_{j+1,j} \neq 0$ for every j, in particular if this matrix is irreducible, then the eigenvalues of M are geometrically simple.*

The example

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

shows that the eigenvalues of an irreducible Hessenberg matrix are not necessarily algebraically simple.

From the point of view of matrix reduction by conjugation, one can attribute two advantages to the Hessenberg class, compared with the class of triangular matrices. First of all, if $K = \mathbb{R}$, many matrices are not trigonalizable in $\mathbb{R}$, though all are trigonalizable in $\mathbb{C}$. Of course, computing with complex numbers is more expensive than computing with real numbers. But we shall see that every square matrix with real entries is similar to a Hessenberg matrix over the real numbers. Next, if K is algebraically closed, the trigonalization of M needs the effective computation of the eigenvalues, which is impossible in view of Abel's theorem. However, the computation of a similar Hessenberg matrix is obtained after a finite number of operations.

Let us observe, finally, that as the trigonalization (see Theorem 3.1.3), the Hessenberg form is obtained through unitary transformations, a well-conditionned process. When $K = \mathbb{R}$, these transformation are obviously real orthogonal.

Theorem 10.1.1 *For every matrix $M \in \mathbf{M}_n(\mathbb{C})$ there exists a unitary transformation U such that $U^{-1}MU$ is a Hessenberg matrix. If $M \in \mathbf{M}_n(\mathbb{R})$, one may take $U \in \mathbf{O}_n$.*

Moreover, the matrix U is computable in $5n^3/3 + O(n^2)$ multiplications and $4n^3/3 + O(n^2)$ additions.

Proof

Let $X \in \mathbb{C}^m$ be a unit vector: $X^*X = 1$. The matrix of the unitary (orthogonal) symmetry with respect to the hyperplane $X^\perp$ is $S = I_m - 2XX^*$. In fact, $SX = X - 2X = -X$, while $Y \in X^\perp$; that is $X^*Y = 0$, implies $SY = Y$.

We construct a sequence $M_1 = M, \ldots, M_{n-1}$ of unitarily similar matrices. The matrix M_{n-r} will be of the form

$$\begin{pmatrix} H & & B \\ 0_{r,n-r-1} & Z & N \end{pmatrix},$$

where $H \in \mathbf{M}_{n-r}(\mathbb{C})$ is Hessenberg and Z is a vector in $\mathbb{C}^r$. Hence, M_{n-1} will be suitable.

One passes from M_{n-r} to M_{n-r+1}, that is, from r to $r-1$, in the following way. Let $\mathbf{e}^1$ be the first vector of the canonical basis of $\mathbb{C}^r$. If Z is colinear to $\mathbf{e}^1$, one does nothing besides defining $M_{n-r+1} = M_{n-r}$. Otherwise, one chooses $X \in \mathbb{C}^r$ so that SZ is parallel to $\mathbf{e}^1$ (we discuss below the possible choices for X). Then one sets

$$V = \begin{pmatrix} I_{n-r} & 0_{n-r,r} \\ 0_{r,n-r} & S \end{pmatrix},$$

which is a unitary matrix, with $V^* = V^{-1} = V$ (such a matrix is called a *Householder* matrix). We then have

$$V^{-1} M_{n-r} V = \begin{pmatrix} H & & BS \\ 0_{n,n-r-1} & SZ & SNS \end{pmatrix}.$$

We thus define $M_{n-r+1} = V^{-1} M_{n-r} V$.

There are two possible choices for S, given by

$$X_\pm := \frac{1}{\|Z \pm \|Z\|_2 \vec{q}\|_2}(Z \pm \|Z\|_2 \mathbf{q}), \quad \mathbf{q} = \frac{z_1}{|z_1|}\mathbf{e}^1.$$

It is always advantageous to choose the sign that gives the largest denominator, namely the positive sign. One thus optimizes the round-off errors when Z is almost aligned with $\mathbf{e}^1$.

Let us consider now the complexity of the $(n-r)$th step. Only the terms of order r^2 and $r(n-r)$ are meaningful. The computation of X, in $O(r)$ operations, is thus negligible, like that of X^* and of $2X$. The computation of $BS = B - (BX)(2X^*)$ needs about $4r(n-r)$ operations. Then $2NX$ needs $2r^2$ operations, as does $2X^*N$. We next compute $4X^*NX$, and then form the vector $T := 4(X^*NX)X - 2NX$ at the cost $O(r)$. The product TX^* takes r^2 operations, as $2X(X^*N)$. Then $N + TX^* - X(2X^*N)$ needs $2r^2$ additions. The complete step is thus accomplished in $7r^2 + 4r(n-r) + O(n)$ operations. A sum from $r = 1$ to $n-2$ yields a complexity of $3n^3 + O(n^2)$, in which one recognizes $5n^3/3 + O(n^2)$ multiplications, $4n^3/3 + O(n^2)$ additions, and $O(n)$ square roots.

■

When M is Hermitian, the matrix $U^{-1}MU$ is still Hermitian. Since it is Hessenberg, it is tridiagonal, with $a_{j,j+1} = \bar{a}_{j+1,j}$ and $a_{jj} \in \mathbb{R}$. The symmetry reduces the complexity to $2n^3/3 + O(n^2)$ multiplications. One can then use the Hessenberg form of M in order to localize its eigenvalues.

Proposition 10.1.2 *If M is tridiagonal Hermitian and if the entries $m_{j+1,j}$ are nonzero (that is, if M is irreducible), then the eigenvalues of M are real and simple. Furthermore, if M_j is the (Hermitian, tridiagonal, irreducible) matrix obtained by keeping only the j last rows and columns of M, the eigenvalues of M_j strictly separate those of M_{j+1}.*

The separation, not necessarily strict, of the eigenvalues of M_{j+1} by those of M_j has already been proved, in a more general framework, in Theorem 3.3.3.

Proof

The eigenvalues of a Hermitian matrix are real. Since this matrix is diagonalizable, Proposition 10.1.1 shows that the eigenvalues are simple. Both properties can be deduced from the following analysis.

We proceed by induction on j. If $j \geq 1$, we decompose the matrix M_{j+1} blockwise:

$$\begin{pmatrix} m & \bar{a} & 0 & \cdots & 0 \\ a & & & & \\ 0 & & M_j & & \\ \vdots & & & & \\ 0 & & & & \end{pmatrix},$$

where $a \neq 0$ and $m \in \mathbb{R}$, $m > 0$. Let P_l be the characteristic polynomial of M_l. We compute that of M_{j+1} by expanding according to the elements of the first column:

$$P_{j+1}(X) = mP_j(X) - |a|^2 P_{j-1}(X), \tag{10.1}$$

where $P_0 \equiv 1$ by convention.

The induction hypothesis is as follows: P_j and P_{j-1} have real entries and have respectively j and $j-1$ real roots $\mu_1, \ldots, \mu_j$ and $\sigma_1, \ldots, \sigma_{j-1}$, with

$$\mu_1 < \sigma_1 < \mu_2 < \cdots < \sigma_{j-1} < \mu_j.$$

In particular, they have no other roots, and their roots are simple. The signs of the values of P_{j-1} at points μ_j thus alternate. Since P_{j-1} is positive over $(\sigma_{j-1}, +\infty)$, we have $(-1)^{j-k} P_{j-1}(\mu_k) > 0$.

This hypothesis clearly holds at step $j = 1$. If $j \geq 2$ and if it holds at step j, then (10.1) shows that $P_{j+1} \in \mathbb{R}[X]$. Furthermore,

$$(-1)^{j-k} P_{j+1}(\mu_k) = -|a|^2 (-1)^{j-k} P_{j-1}(\mu_k) < 0.$$

From the intermediate value theorem, P_{j+1} possesses a root λ_k in (μ_{k-1}, μ_k). Furthermore, $P_{j+1}(\mu_j) < 0$, and $P_{j+1}(x)$ is positive for $x \gg 1$; hence there is also a root in $(\mu_j, +\infty)$. Likewise, P_{j+1} has a root in $(-\infty, \mu_1)$. Hence, P_{j+1} possesses $j+1$ distinct real roots λ_k, with

$$\lambda_1 < \mu_1 < \lambda_2 < \cdots < \mu_j < \lambda_{j+1}.$$

Since P_{j+1} has degree $j+1$, there is no root other than the λ_k's, and these are simple.

■

The sequence of polynomials P_j is a *Sturm sequence*, which allows us to compute the number of roots of P_n in a given interval (a, b). A Sturm sequence is a finite sequence of real polynomials $Q_0, \ldots, Q_n$, with Q_0 a nonzero constant such that $Q_j(x) = 0$ and $0 < j < n$ imply $Q_{j+1}(x)Q_{j-1}(x) < 0$. In particular, Q_j and Q_{j+1} do not share a common root. If $a \in \mathbb{R}$ is not a root of Q_n, we denote by $V(a)$ the number of sign changes in the sequence $(Q_0(a), \ldots, Q_n(a))$, with the zeros playing no role.

Proposition 10.1.3 *If $Q_n(a) \neq 0$ and $Q_n(b) \neq 0$, and if $a < b$, then the number of roots of Q_n in (a, b) is equal to $V(a) - V(b)$.*

Let us remark that it is not necessary of compute the polynomials P_j to apply them to this proposition. Given $a \in \mathbb{R}$, it is enough to compute the sequence of values $P_j(a)$.

Once an interval (a, b) is known to contain an eigenvalue λ and only that one (by means of Proposition 10.1.3 or Theorem 4.5.1), one can compute an approximate value of λ, either by dichotomy, or by computing the numbers $V((a+b)/2), \ldots$, or by the secant or Newton method. In the latter case, one must compute P_n itself. The last two methods are convergent, provided that we have a good initial approximation at our disposal, because $P_n'(\lambda) \neq 0$.

We end this section with an obvious but nevertheless useful remark. If M is Hessenberg and T upper triangular, the products TM and MT are still Hessenberg (that would not be true if both matrices were Hessenberg). For example, if M admits an LU factorization, then L is Hessenberg, and thus has only two nonzero diagonals, because $L = MU^{-1}$. Similarly, if $M \in \mathbf{GL}_n(\mathbb{C})$, then the factor Q of the factorization $M = QR$ is again Hessenberg, because $Q = MR^{-1}$. An elementary compactness and continuity argument shows that the same fact holds true for every $M \in \mathbf{M}_n(\mathbb{C})$.

10.2 The QR Method

The QR method is considered the most efficient one for the approximate computation of the whole spectrum of a square matrix $M \in \mathbf{M}_n(\mathbb{C})$. One employs it only after having reduced M to Hessenberg form, because this form is preserved throughout the algorithm, while each iteration is much cheaper than it is for a generic matrix.

10.2.1 Description of the QR *Method*

Let $A \in \mathbf{M}_n(K)$ be given, with $K = \mathbb{R}$ or $\mathbb{C}$. We construct a sequence of matrices $(A_j)_{j \in \mathbb{N}}$, with $A_1 = A$. The induction $A_j \mapsto A_{j+1}$ consists in performing the QR factorization of A_j, $A_j = Q_j R_j$, and then defining

$A_{j+1} := R_j Q_j$. We then have

$$A_{j+1} = Q_j^{-1} A_j Q_j,$$

which shows that A_{j+1} is unitarily similar to A_j. Hence,

$$A_j = (Q_0 \cdots Q_{j-1})^{-1} A (Q_0 \cdots Q_{j-1}) \tag{10.2}$$

is conjugate to A by a unitary transformation.

Let $P_j := Q_0 \cdots Q_{j-1}$, which is unitary. Since $\mathbf{U}_n$ is compact, the sequence $(P_j)_{j \in \mathbb{N}}$ possesses cluster values. Let P be one of them. Then $A' := P^{-1} A P = P^* A P$ is a cluster point of $(A_j)_{j \in \mathbb{N}}$. Hence, if the sequence $(A_j)_j$ converges, its limit is unitarily similar to A, hence has the same spectrum.

This argument shows that in general, the sequence $(A_j)_j$ does not converge to a diagonal matrix, because then the eigenvectors of A would be the columns of P. In other words, A would have an orthonormal eigenbasis. Namely, A would be normal. Except in this special case, one expects merely that the sequence $(A_j)_j$ converges to a triangular matrix, an expectation that is compatible with Theorem 3.1.3. But even this hope is too optimistic in general. For example, if A is unitary, then $A_j = A$ for every j, with $Q_j = A$ and $R_j = I_n$; in that case, the convergence is useless, since the limit A is not simpler than the data. We shall see later on that the reason for this bad behavior is that the eigenvalues of a unitary matrix have the same modulus: The QR method does not do a good job of separating the eigenvalues of close modulus.

An important case in which a matrix has at least two eigenvalues of the same modulus is that of matrices with real entries. If $A \in \mathbf{M}_n(\mathbb{R})$, then each Q_j is real orthogonal, R_j is real, and A_j is real. This is seen by induction on j. A limit A' will not be triangular if some eigenvalues of A are nonreal, that is, if A possesses a pair of complex conjugate eigenvalues.

Let us sum up what can be expected in a brave new world. If all the eigenvalues of $A \in \mathbf{M}_n(\mathbb{C})$ have distinct moduli, the sequence $(A_j)_j$ might converge to a triangular matrix, or at least its lower triangular part might converge to

$$\begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

When $A \in \mathbf{M}_n(\mathbb{R})$, one makes the following assumption. Let p be the number of real eigenvalues and $2q$ that of nonreal eigenvalues; then there are $p+q$ distinct eigenvalue moduli. In that case, $(A_j)_j$ might converge to a block-triangular form, the diagonal blocks being 2×2 or 1×1. The limits of the diagonal blocks provide trivially the eigenvalues of A.

The assertions made above have never been proved in full generality, to our knowledge. We shall give below a rather satisfactory result in the complex case.

10.2.2 The Case of a Singular Matrix

When A is not invertible, the QR factorization is not unique, raising a difficulty in the definition of the algorithm. The computation of the determinant would detect immediately the case of noninvertibility, but would not provide any solution. However, if the matrix has been first reduced to the Hessenberg form, then a single QR iteration detects the case and *does* provide a solution. Indeed, if A is Hessenberg but not invertible, and if $A = QR$, then Q is Hessenberg and R is not invertible. If $a_{21} = 0$, the matrix A is block-triangular and we are reduced to the case of a matrix of size $(n-1) \times (n-1)$ by deleting the first row and the first column. Otherwise, there exists $j \geq 2$ such that $r_{jj} = 0$. The matrix $A_1 = RQ$ is then block-triangular, because it is Hessenberg and $(A_1)_{j,j-1} = r_{jj} q_{j,j-1} = 0$. We are thus reduced to the computation of the spectra of two matrices of sizes $j \times j$ and $(n-j) \times (n-j)$, the diagonal blocks of A_1. After finitely many such steps (not larger than the multiplicity of the null eigenvalue), there remain only Hessenberg invertible matrices to deal with. We shall assume therefore from now one that $A \in \mathbf{GL}_n(K)$.

10.2.3 Complexity of an Iteration

An iteration of the QR method requires the factorization $A_j = Q_j R_j$ and the computation of $A_{j+1} = R_j Q_j$. Each part costs $O(n^3)$ operations if it is done on a generic matrix (using the naive way of multiplying matrices). Since the reduction to the Hessenberg form has a comparable cost, we loose nothing by reducing A to this form. Actually, we make considerable gains in two aspects. First, the cost of the QR iterations is reduced to $O(n^2)$. Second, the cluster values of the sequence $(A_j)_j$ must have the Hessenberg form too.

Let us examine first the Householder method of QR factorization for a generic matrix A. In practice, one computes only the factor R and matrices of unitary symmetries whose product is Q. One then multiplies these unitary matrices by R on the left to obtain $A' = RQ$.

Let $\mathbf{a}_1 \in \mathbb{C}^n$ be the first column vector of A. We begin by determining a unit vector $v_1 \in \mathbb{C}^n$ such that the hyperplane symmetry $H_1 := I_n - 2v_1 v_1^*$

sends $\mathbf{a}_1$ to $\|\mathbf{a}_1\|_2\mathbf{e}^1$. The matrix H_1A has the form

$$\tilde{A} = \begin{pmatrix} \|\mathbf{a}_1\|_2 & x & \cdots \\ 0 & \vdots & \\ \vdots & \vdots & \\ 0 & y & \cdots \end{pmatrix}.$$

We then perform these operations again on the matrix extracted from $\tilde{A}$ by deleting the first rows and columns, and so on. At the kth step, H_k is a matrix of the form

$$\begin{pmatrix} I_k & 0 \\ 0 & I_{n-k} - 2v_kv_k^* \end{pmatrix},$$

where $v_k \in \mathbb{C}^{n-k}$ is a unit vector. The computation of v_k requires $O(n-k)$ operations. The product $H_kA^{(k)}$, where $A^{(k)}$ is block-triangular, amounts to that of two square matrices of size $n-k$, one of them $I-2v_kv_k^*$. We thus compute a matrix $N - 2vv^*N$ from v and N, which costs about $4(n-k)^2$ operations. Summing from $k=1$ to $k=n-1$, we find that the complexity of the computation of R alone is $4n^3/3 + O(n^2)$. As indicated above, we do not compute the factor Q, but compute all the matrices $RH_{n-1}\cdots H_k$. That necessitates $2n^3 + O(n)$ operations (check this!). The complexity of one step of the QR method on a generic matrix is thus $10n^3/3 + O(n^2)$.

Let us now analyze the situation when A is a Hessenberg matrix. By induction on k, we see that v_k belongs to the plane spanned by $\mathbf{e}^k$ and $\mathbf{e}^{k+1}$. Its computation needs $O(1)$ operations. Then the product of H_k and $A^{(k)}$ can be obtained by simply recomputing the rows of indices k and $k+1$, about $6(n-k)$ operations. Summing from $k=1$ to $n-1$, we find that the complexity of the computation of R alone is $3n^2 + O(n)$. The computation of the product $(RH_{n-1}\cdots H_{k+1})H_k$ needs about $6k$ operations. Finally, the complexity of the QR factorization of a Hessenberg matrix is $6n^2 + O(n)$, in which there are $4n^2 + O(n)$ multiplications.

To sum up, the cost of the preliminary reduction of a matrix to Hessenberg form is less than or equal to what is saved during the first iteration of the QR method.

10.2.4 Convergence of the QR *Method*

As explained above, the best convergence statement assumes that the eigenvalues have distinct moduli.

Let us recall that the sequence A_k is not always convergent. For example, if A is already triangular, its QR factorization is $Q = D$, $R = D^{-1}A$, with $d_j = a_{jj}/|a_{jj}|$. Hence, $A_1 = D^{-1}AD$ is triangular, with the same diagonal as that of A. By induction, A_k is triangular, with the same diagonal as that of A. We have thus $Q_k = D$ for every k, so that $A_k = D^{-k}AD^k$. The

entry of index (l, m) is thus multiplied at each step by a unit number z_{lm}, which is not necessarily equal to one if $l < m$. Hence, the part above the diagonal of A_k does not converge.

Summing up, a convergence theorem may concern only the diagonal of A_k and what is below it.

Lemma 10.2.1 *Let $A \in \mathbf{GL}_n(K)$ be given, with $K = \mathbb{R}$ or $\mathbb{C}$. Let $A_k = Q_k R_k$ be the sequence of matrices given by the* QR *algorithm. Let us define $P_k = Q_0 \cdots Q_{k-1}$ and $U_k = R_{k-1} \cdots R_0$. Then $P_k U_k$ is the* QR *factorization of the kth power of A:*

$$A^k = P_k U_k.$$

Proof
From (10.2), we have $A_k = P_k^{-1} A P_k$; that is, $P_k A_k = A P_k$. Then

$$P_{k+1} U_{k+1} = P_k Q_k R_k U_k = P_k A_k U_k = A P_k U_k.$$

By induction, $P_k U_k = A^k$. However, $P_k \in \mathbf{U}_n$ and U_k is triangular, with a positive real diagonal, as a product of such matrices.

■

Theorem 10.2.1 *Let $A \in \mathbf{GL}_n(\mathbb{C})$ be given. Assume that the moduli of the eigenvalues of A are distinct:*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| \quad (> 0).$$

In particular, the eigenvalues are simple, and thus A is diagonalizable:

$$A = Y^{-1} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) Y.$$

Assume also that Y admits an LU factorization. Then the strictly lower triangular part of A_k converges to zero, and the diagonal of A_k converges to $D := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Proof
Let $Y = LU$ be the factorization of Y. We also make use of the QR factorization of Y^{-1}: $Y^{-1} = QR$. Since $A^k = Y^{-1} D^k Y$, we have $P_k U_k = Y^{-1} D^k Y = QRD^k LU$.

The matrix $D^k L D^{-k}$ is lower triangular with unit numbers on its diagonal. By assumption, its strictly lower part tends to zero (because each term is multiplied by $(\lambda_i/\lambda_j)^k$, where $|\lambda_i/\lambda_j| < 1$). Therefore, $D^k L D^{-k} = I_n + E_k$ with $E_k \to 0_n$ as $k \to +\infty$. Hence, $P_k U_k = QR(I_n + E_k)D^k U = Q(I_n + RE_k R^{-1})RD^k U = Q(I_n + F_k)RD^k U$, where $F_k \to 0_n$. Let $O_k T_k = I_n + F_k$ be the QR factorization of $I_n + F_k$. By continuity, O_k and T_k both tend to I_n. Then

$$P_k U_k = (QO_k)(T_k R D^k U).$$

The first product is a unitary matrix, while the second is a triangular one. Let $|D|$ be the "modulus" matrix of D (whose entries are the moduli

of those of D), and let D_1 be $|D|^{-1}D$, which is unitary. We also define $D_2 = \mathrm{diag}(u_{jj}/|u_{jj}|)$ and $U' = D_2^{-1}U$. Then D_2 is unitary and the diagonal of U' is positive real. From the uniqueness of the QR factorization of an invertible matrix we obtain

$$P_k = QO_kD_1^kD_2, \quad U_k = (D_1^kD_2)^{-1}T_kRD_1^kD_2|D|^kU',$$

which yields

$$\begin{aligned} Q_k &= P_k^{-1}P_{k+1} = D_2^{-1}D_1^{-k}O_k^{-1}O_{k+1}D_1^{k+1}D_2, \\ R_k &= U_{k+1}U_k^{-1} = D_2^{-1}D_1^{-k-1}T_{k+1}RDR^{-1}T_k^{-1}D_1^kD_2. \end{aligned}$$

Since D_1^{-k} and D_1^{k+1} are bounded, we deduce that Q_k converges, to D_1. Similarly, $R_k - R'_k \to 0_n$, where

$$R'_k = D_2^{-1}D_1^{-k}RDR^{-1}D_1^{k-1}D_2. \tag{10.3}$$

The fact that the matrix R'_k is upper triangular shows that the strict lower triangular part of $A_k = Q_kR_k$ tends to zero (observe that the sequence $(R_k)_{k\in\mathbb{N}}$ is bounded, because the set of unitary matrices conjugate to A is bounded). Similarly, the diagonal of R'_k is $|D|$, which shows that the diagonal of A_k converges to $D_1|D| = D$.

■

Remark: Formula (10.3) shows that the sequence A_k does not converge, at least when the eigenvalues have distinct complex arguments. However, if the eigenvalues have equal complex arguments, for example if they are real and positive, then $D_1 = \alpha I_n$ and $R_k \to T := D_2^{-1}R|D|R^{-1}D_2$; hence A_k converges to αT. Note that this limit is not diagonal in this case.

The situation is especially favorable for tridiagonal Hermitian matrices. To begin with, we may assume that A is positive definite, up to the change of A into $A + \mu I_n$ with $\mu > -\rho(A)$. Next, we can write A in block-diagonal form, where the diagonal blocks are tridiagonal irreducible Hermitian matrices. The QR method then treats each block separately. We are thus reduced to the case of a Hermitian positive definite, tridiagonal and irreducible matrix. Its eigenvalues are real, strictly positive, and simple, from Proposition 10.1.2: we have $\lambda_1 > \cdots > \lambda_n > 0$. We can then use the following statement.

Theorem 10.2.2 *Let $A \in \mathbf{GL}_n(\mathbb{C})$ be an irreducible Hessenberg matrix whose eigenvalues are of distinct moduli:*

$$|\lambda_1| > \cdots > |\lambda_n| \quad (> 0).$$

Then the QR *method converges; that is, the lower triangular part of* A_k *converges to*

$$\begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Proof

In the light of Theorem 10.2.1, it is enough to show that the matrix Y in the previous proof admits an LU factorization. We have $YA = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)Y$. The rows of Y are thus the left eigenvectors: $l_j A = \lambda_j l_j$.

If $x \in \mathbb{C}^n$ is nonzero, there exists a unique index r such that $x_r \neq 0$, while $j > r$ implies $x_j = 0$. By induction, quoting the Hessenberg form and the irreducibility of A, we obtain $(A^m x)_{r+m} \neq 0$, while $j > r + m$ implies $(A^m x)_j = 0$. Hence, the vectors $x, Ax, \ldots, A^{n-r}x$ are linearly independent. A linear subspace, stable under A and containing x, is thus of dimension greater than or equal to $n - r + 1$.

Let F be a linear subspace, stable under A, of dimension $p \geq 1$. Let r be the smallest integer such that F contains a nonzero vector x with $x_{r+1} = \cdots = x_n = 0$. The minimality of r implies that $x_r \neq 0$. Hence, we have $p \geq n - r + 1$. By construction, the intersection of F and of linear subspace $[\mathbf{e}^1, \ldots, \mathbf{e}^{n-p}]$ spanned by $\mathbf{e}^1, \ldots, \mathbf{e}^{r-1}$ reduces to $\{0\}$. Thus we also have $p + (r - 1) \leq n$. Finally, $r = n - p + 1$, and we see that

$$F \oplus [\mathbf{e}^1, \ldots, \mathbf{e}^{n-p}] = \mathbb{C}^n.$$

Let us choose $F = [l_1, \ldots, l_q]^\perp$, which is stable under A. Then $p = n - q$, and we have

$$[l_1, \ldots, l_q]^\perp \oplus [\mathbf{e}^1, \ldots, \mathbf{e}^q] = \mathbb{C}^n.$$

This amounts to saying that $\det(l_j \mathbf{e}^k)_{1 \leq j,k \leq q} \neq 0$. In other words, the leading principal minor of order q of Y is nonzero. From Theorem 8.1.1, Y admits an LU factorization.

■

Corollary 10.2.1 *If* $A \in \mathbf{HPD}_n$ *and if* A_0 *is a Hessenberg matrix, unitarily similar to* A *(for example, a matrix obtained by Householder's method), then the sequence* A_k *defined by the* QR *method converges to a diagonal matrix whose diagonal entries are the eigenvalues of* A*.*

Indeed, A_0 is block-diagonal with irreducible diagonal blocks. We are thus reduced to the case of a Hermitian positive definite tridiagonal irreducible matrix. Such a matrix satisfies the hypotheses of Theorem 10.2.2. The lower triangular part converges, hence the whole matrix, since it is Hermitian.
Implementing the *QR* method: The QR method converges faster as λ_n, or merely λ_n/λ_{n-1}, becomes smaller. We can obtain this situation

by translating $A_k \mapsto A_k - \alpha_k I_n$. The strategies for the choice of α_k are described in [25]. This procedure is called *Rayleigh translation*. It allows for a observeable improvement of the convergence of the QR method. If the eigenvalues of A are simple, a suitable translation allows us to restrict ourselves to the case of distinct moduli. But this trick has a nonnegligible cost if A is a real matrix with a pair of complex conjugate eigenvalues, since it requires a translation by a nonreal number α. As mentioned above, the computations become much more costly than they are in the domain of real numbers.

As k increases, the triangular form of A_k appears first at the last row. In other words, the sequence $(A_k)_{nn}$ converges more rapidly thanother sequences $(A_k)_{jj}$. When the last row is sufficiently close to $(0, \dots, 0, \lambda_n)$, the Rayleigh translation must be selected in such a way as to bring λ_{n-1}, instead of λ_n, to the origin; and so on.

With a clever choice of Rayleigh translations, the QR method, when it converges, is of order two for a generic matrix, and is of order three for a Hermitian matrix.

10.3 The Jacobi Method

The Jacobi method allows for the approximate computation of the whole spectrum of a real symmetric matrix $A \in \mathbf{Sym}_n$. As in the QR method, one constructs a sequence of matrices, unitarily similar to A. In particular, the round-off errors are not amplified. Each iteration is cheap ($O(n)$ operations), and the convergence is quadratic when the eigenvalues are distinct. It is thus a rather efficient method.

10.3.1 Conjugating by a Rotation Matrix

Let $1 \leq p, q \leq n$ be two distinct indices and $\theta \in [-\pi, \pi)$ an angle. We denote by $R_{p,q}(\theta)$ the rotation matrix through the angle θ in the plane spanned by $\mathbf{e}^p$ and $\mathbf{e}^q$. For example, if $p < q$, then

$$R = R_{p,q}(\theta) := \begin{pmatrix} I_{p-1} & \vdots & 0 & \vdots & 0 \\ \cdots & \cos\theta & \cdots & \sin\theta & \cdots \\ 0 & \vdots & I_{q-p-1} & \vdots & 0 \\ \cdots & -\sin\theta & \cdots & \cos\theta & \cdots \\ 0 & \vdots & 0 & \vdots & I_{n-q} \end{pmatrix}.$$

If H is a symmetric matrix, we compute $K := R^{-1}HR = R^T HR$, which is also symmetric. Setting $c = \cos\theta$, $s = \sin\theta$ the following formulas hold:

$$\begin{aligned}
k_{ij} &= h_{ij} \quad \text{if } i,j \neq p,q,\\
k_{ip} &= ch_{ip} - sh_{iq} \quad \text{if } i \neq p,q,\\
k_{iq} &= ch_{iq} + sh_{ip} \quad \text{if } i \neq p,q,\\
k_{pp} &= c^2 h_{pp} + s^2 h_{qq} - 2csh_{pq},\\
k_{qq} &= c^2 h_{qq} + s^2 h_{pp} + 2csh_{pq},\\
k_{pq} &= cs(h_{pp} - h_{qq}) + (c^2 - s^2)h_{pq}.
\end{aligned}$$

The cost of the computation of entries k_{ij} for $i,j \neq p,q$ is zero; that of k_{pp}, k_{qq}, and k_{pq} is $O(1)$. The cost of this conjugation is thus $6n + O(1)$ operations, keeping in mind the symmetry $K^T = K$.

Let us remark that the conjugation by the rotation through the angle $\theta \pm \pi$ yields the same matrix K. For this reason, we limit ourselves to angles $\theta \in [-\pi/2, \pi/2)$.

10.3.2 Description of the Method

One constructs a sequence $A^{(0)} = A, A^{(1)}, \ldots$ of symmetric matrices, each one conjugate to the previous one by a rotation as above: $A^{(k+1)} = (R^{(k)})^T A^{(k)} R^{(k)}$. At step k, we choose two distinct indices p and q (in fact, p_k, q_k) in such a way that $a_{pq}^{(k)} \neq 0$ (if it is not possible, $A^{(k)}$ is already a diagonal matrix similar to A). We then choose θ (in fact θ_k) in such a way that $a_{pq}^{(k+1)} = 0$. From the formulas above, this is equivalent to

$$cs(a_{pp}^{(k)} - a_{qq}^{(k)}) + (c^2 - s^2)a_{pq}^{(k)} = 0.$$

This amounts to solving the equation

$$\cot 2\theta = \frac{a_{qq}^{(k)} - a_{pp}^{(k)}}{2a_{pq}^{(k)}} =: \sigma_k. \tag{10.4}$$

This equation possesses two solutions in $[-\pi/2, \pi/2)$, namely $\theta_k \in [-\pi/4, \pi/4)$ and $\theta_k \pm \pi/2$. There are thus two possible rotation matrices, which yield to two distinct results. Once the angle has been selected, its computation is useless (it would be actually rather expensive). In fact, $t := \tan\theta_k$ solves

$$\frac{2t}{1-t^2} = \tan 2\theta;$$

that is,

$$t^2 + 2t\sigma_k - 1 = 0.$$

The two angles correspond to the two possible roots of this quadratic equation. We then obtain

$$c = \frac{1}{\sqrt{1+t^2}}, \quad s = tc.$$

We shall see below that the best choice is the angle $\theta_k \in [-\pi/4, \pi/4)$, which corresponds to the unique root t in $[-1, 1)$.

The computation of c, s needs only $O(1)$ operations, so that the cost of an iteration of the Jacobi method is $6n + O(1)$. Observe that an entry that has vanished at an iteration becomes in general nonzero after a few more iterations.

10.3.3 Convergence of the Jacobi Method

We use here the Schur norm $\|M\| = (\operatorname{Tr} M^T M)^{1/2}$, also called the Frobenius norm, denoted elsewhere by $\|M\|_2$. Since it amounts to showing that $A^{(k)}$ converges to a diagonal matrix, we decompose this matrix in the form $A^{(k)} = D_k + E_k$, where $D_k = \operatorname{diag}(a_{11}^{(k)}, \ldots, a_{nn}^{(k)})$. To begin with, since the sequence is formed of unitarily similar matrices, we have $\|A^{(k)}\| = \|A\|$.

Lemma 10.3.1 *We have*

$$\|E_{k+1}\|^2 = \|E_k\|^2 - 2\left(a_{pq}^{(k)}\right)^2.$$

Proof

It is sufficient to redo the calculations of Section 10.3.1, noting that

$$k_{ip}^2 + k_{iq}^2 = h_{ip}^2 + h_{iq}^2$$

whenever $i \neq p, q$, while $k_{pq}^2 = 0$.

■

We deduce from the lemma that $\|D_{k+1}\|^2 = \|D_k\|^2 + 2\left(a_{pq}^{(k)}\right)^2$. The convergence of the Jacobi method depends, then, on the choice of the pair (p, q) at each step. For example, the choice of the same pair at two consecutive iterations is stupid, since it yields $A^{(k+1)} = A^{(k)}$. A first strategy (the so-called *optimal choice*) consists in taking the pair (p, q) that optimizes the instantaneous decay of $\|E_k\|$, that is, maximizes the number $|a_{pq}^{(k)}|$. Since this method involves the sorting of $n(n-1)/2$ entries, it is rather expensive. Other strategies are available. One can, for instance, range over every pair (p, q) with $p < q$, or choose a (p, q) for which $|a_{pq}^{(k)}|$ is larger than some threshold. Here we shall study only the method with optimal choice.

Theorem 10.3.1 *With the "optimal choice" of (p_k, q_k) and with the choice $\theta_k \in [-\pi/4, \pi/4)$, the Jacobi method converges in the following sense. There exists a diagonal matrix D such that*

$$\|A^{(k)} - D\| \leq \frac{\sqrt{2}\|E_0\|}{1-\rho}\rho^k, \quad \rho := \sqrt{1 - \frac{2}{n^2 - n}}.$$

In particular, the spectrum of A consists of the diagonal terms of D, and the Jacobi method is of order one at least.

Proof
With the optimal choice of (p, q), we have

$$(n^2 - n)\left(a_{pq}^{(k)}\right)^2 \geq \|E_k\|^2.$$

Hence,

$$\|E_{k+1}\|^2 \leq \left(1 - \frac{2}{n^2 - n}\right)\|E_k\|^2.$$

It follows that $\|E_k\| \leq \rho^k \|E_0\|$. In particular, E_k tends to zero as $k \to +\infty$.

It remains to show that D_k converges too. A calculation using the notation of Section 10.3.1 and the fact that $k_{pq} = 0$ yield

$$k_{pp} - h_{pp} = t h_{pq}.$$

Since $|\theta_k| \leq \pi/4$, we have $|t| \leq 1$, so that $|a_{pp}^{(k+1)} - a_{pp}^{(k)}| \leq |a_{pq}^{(k)}|$. Likewise, $|a_{qq}^{(k+1)} - a_{qq}^{(k)}| \leq |a_{pq}^{(k)}|$. Since the other diagonal entries are unchanged, we have $\|D_{k+1} - D_k\| \leq \|E_k\|$.

We have seen that $\|E_k\| \leq \rho^k \|E_0\|$. Therefore,

$$\|D_l - D_k\| \leq \|E_0\| \frac{\rho^k}{1 - \rho}, \quad l > k.$$

The sequence $(D_k)_{k \in \mathbb{N}}$ is thus Cauchy, hence convergent. Since E_k tends to zero, A_k converges to the same limit D. This matrix is diagonal, with the same spectrum as A, since this is true for each A_k. Finally, we obtain

$$\|A^{(k)} - D\|^2 = \|D_k - D\|^2 + \|E_k\|^2 \leq \frac{2}{(1 - \rho)^2}\|E_k\|^2.$$

■

10.3.4 Quadratic Convergence

The following statement shows that the Jacobi method compares rather well with other methods.

Theorem 10.3.2 *The Jacobi method with optimal choice of (p, q) is of order two when the eigenvalues of A are simple, in the following sense. Let $N = n(n-1)/2$ be the number of elements under the diagonal. Then there exists a number $c > 0$ such that*

$$\|E_{k+N}\| \leq c\|E_k\|^2,$$

for every $k \in \mathbb{N}$.

Proof

We first remark that if $i \neq j$ with $\{i,j\} \neq \{p_l, q_l\}$, then

$$|a_{ij}^{(l+1)} - a_{ij}^{(l)}| \leq |t_l|\sqrt{2}\|E_l\|, \tag{10.5}$$

where $t_l = \tan\theta_l$. To see this, observe that $1 - c \leq t$ and $|s| \leq t$ whenever $|t| \leq 1$. However, Theorem 10.3.1 ensures that D_k converges to $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where the λ_j's are the eigenvalues of A. Since these are distinct, there exist $K \in \mathbb{N}$ and $\delta > 0$ such that, if $k \geq K$, then

$$\min_{i \neq j} |a_{ii}^{(k)} - a_{jj}^{(k)}| \geq \delta$$

for $k \geq K$. We have therefore

$$|\sigma_k| \geq \frac{\delta}{\sqrt{2}\|E_k\|} \overset{k \to +\infty}{\longrightarrow} +\infty.$$

It follows that t_k tends to zero and, more precisely, that

$$t_k \sim -\frac{1}{2\sigma_k}.$$

Finally, there exists a constant c_1 such that

$$|t_k| \leq c_1\|E_k\|.$$

Let us fix then k larger than K, and let us denote by J the set of pairs (p_l, q_l) when $k \leq l \leq k + N - 1$. For such an index, we have $\|E_l\| \leq \rho^{l-k}\|E_k\| \leq \|E_k\|$. In particular, $|t_l| \leq c_1\|E_k\|$.

If $(p,q) \in J$ and if $l < k+N$ is the largest index such that $(p,q) = (p_l, q_l)$, a repeated application of (10.5) shows that

$$|a_{pq}^{(k+N)}| \leq c_1 N\sqrt{2}\|E_k\|^2.$$

If J is equal to the set of pairs (i,j) such that $i < j$, these inequalities ensure that $\|E_{k+N}\| \leq c_2\|E_k\|^2$. Otherwise, there exists a pair (p,q) that one twice sets to zero: $(p,q) = (p_l, q_l) = (p_m, q_m)$ with $k \leq l < m < k+N$. In that case, the same argument as above shows that

$$\|E_{k+N}\| \leq \|E_m\| \leq \sqrt{2N}|a_{pq}^{(m)}| \leq 2\sqrt{N}c_1(m-l)\|E_k\|^2.$$

■

Remarks: Exercise 18 shows that the distance between the diagonal and the spectrum of A is $O(\|E_k\|^2)$, and not $O(\|E_k\|)$ as naively expected. We shall also analyze, in Exercise 10, the (bad) behavior of D_k when we make the opposite choice $\pi/4 \leq |\theta_k| \leq \pi/2$.

10.4 The Power Methods

The power methods allow only for the approximation of a single eigenvalue. Of course, their cost is significantly lower than that of the previous ones.

The standard method is especially designed for the search for the optimal parameter in the SOR method for a tridiagonal matrix, where we have to compute the spectral radius of the Jacobi iteration matrix (Theorem 9.4.1).

10.4.1 The Standard Method

Let $M \in \mathbf{M}_n(\mathbb{C})$ be a matrix. We search for an approximation of its eigenvalue of maximum modulus, whenever only one such exists. The standard method consists in choosing a norm on $\mathbb{C}^n$, a unit vector $x^0 \in \mathbb{C}^n$, and then computing successively the vectors x^k by the formula

$$x^{k+1} := \frac{1}{\|Mx^k\|} Mx^k.$$

The justification of this method is given in the following theorem.

Theorem 10.4.1 *One assumes that* $\operatorname{Sp} M$ *contains only one element* λ *of maximal modulus (that modulus is thus equal to* $\rho(M)$*).*

If $\rho(M) = 0$*, the method stops because* $Mx^k = 0$ *for some* $k < n$*.*

Otherwise, let $\mathbb{C}^n = E \oplus F$ *be the decomposition of* $\mathbb{C}^n$*, where* E, F *are stable linear subspaces under* M*, with* $\operatorname{Sp}(M|_E) = \{\lambda\}$ *and* $\lambda \notin \operatorname{Sp}(M|_F)$*. Assume that* $x^0 \notin F$*. Then* $Mx^k \neq 0$ *for every* $k \in \mathbb{N}$ *and:*

1.
$$\lim_{k \to +\infty} \|Mx^k\| = \rho(M). \tag{10.6}$$

2.
$$V := \lim_{k \to +\infty} \left(\frac{\bar{\lambda}}{\rho(M)} \right)^k x^k$$

is a unit eigenvector of M*, associated to the eigenvalue* λ*.*

3. *If* $V_j \neq 0$*, then*
$$\lim_{k \to +\infty} \frac{(Mx^k)_j}{x_j^k} = \lambda.$$

Proof

The case $\rho(M) = 0$ is obvious because M is then nilpotent. We may thus assume that $\rho(M) > 0$.

Let $x^0 = y^0 + z^0$ be the decomposition of x^0 with $y^0 \in E$ and $z^0 \in F$. By assumption, $y^0 \neq 0$. Since $M|_E$ is invertible, $M^k y^0 \neq 0$. Since $M^k x^0 = M^k y^0 + M^k z^0$, $M^k y^0 \in E$, and $M^k z^0 \in F$, we conclude that $M^k x^0 \neq 0$.

The algorithm may be rewritten as[2]

$$x^k = \frac{1}{\|M^k x^0\|} M^k x^0.$$

We therefore have $x^k \neq 0$.

If $F \neq \{0\}$, then $\rho(M|_F) < \rho(M)$ by construction. Hence there exist (from Theorem 4.2.1) $\eta < \rho(M)$ and $C > 0$ such that $\|(M|_F)^k\| \leq C\eta^k$ for every k. Then $\|(M|_F)^k z^0\| \leq C_1\eta^k$. On the other hand, $\rho((M|_E)^{-1}) = 1/\rho(M)$, and the same argument as above ensures that $\|(M|_E)^{-k}\| \leq 1/C_2\mu^k$, for some $\mu \in (\eta, \rho(M))$, so that $\|M^k y^0\| \geq C_3\mu^k$. Hence,

$$\|M^k z^0\| \ll \|M^k y^0\|,$$

so that

$$x^k \sim \frac{1}{\|M^k y^0\|} M^k y^0.$$

We are thus reduced to the case where $z^0 = 0$, that is, where M has no eigenvalue but λ. That will be assumed from now on.

Let r be the degree of the minimal polynomial of M. The vector space spanned by the vectors $x^0, Mx^0, \dots, M^{r-1}x^0$ contains all the x^k's. Up to the replacement of $\mathbb{C}^n$ by this linear subspace, one may assume that it equals $\mathbb{C}^n$. Then we have $r = n$. Furthermore, since $\ker(M-\lambda)^{n-1}$, a nontrivial linear subspace, is stable under A, we see that $x^0 \notin \ker(M - \lambda)^{n-1}$.

The vector space $\mathbb{C}^n$ then admits the basis

$$\{v^1 = x^0, v^2 = (M-\lambda)x^0, \dots, v^n = (M-\lambda)^{n-1}x^0\}.$$

With respect to this basis, M becomes the Jordan matrix

$$\tilde{M} = \begin{pmatrix} \lambda & 0 & \dots & \dots & \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & \dots & 0 & 1 & \lambda \end{pmatrix}.$$

The matrix $\lambda^{-k}\tilde{M}^k$ depends polynomially on k. The coefficient of highest degree, as $k \to +\infty$, is at the intersection of the first column and the last row. It equals

$$\begin{pmatrix} k \\ n-1 \end{pmatrix} \lambda^{1-n},$$

---

[2]One could normalize x^k at the end of the computation, but we prefer doing it at each step in order to avoid overflows, and also to ensure (10.6).

which is equivalent to $(k/\lambda)^{n-1}/(n-1)!$. We deduce that

$$M^k x^0 \sim \frac{k^{n-1}\lambda^{k-n+1}}{(n-1)!} v^n.$$

Hence,

$$x^k \sim \left(\frac{\lambda}{|\lambda|}\right)^{k-n+1} \frac{v^n}{\|v^n\|}.$$

Since v^n is an eigenvector of M, the claims of the theorem have been proved. ■

The case where the algebraic and geometric multiplicities of λ are equal (that is, $M|_E = \lambda I_E$), for example if λ is a simple eigenvalue, is especially favorable. Indeed, $M^k y^0 = \lambda^k y^0$, and therefore

$$x^k = \frac{1}{\|y^0\|} y^0 + O\left(\frac{\|M^k z^0\|}{|\lambda|^k}\right).$$

Theorem 4.2.1 thus shows that the error

$$x^k - \frac{1}{\|y^0\|} y^0$$

tends to zero faster than

$$\left(\frac{\rho(M|_F) + \epsilon}{\rho(M)}\right)^k,$$

for every $\epsilon > 0$. The convergence is thus of order one, and becomes faster as the ratio $|\lambda_2|/|\lambda_1|$ becomes smaller (arranging the eigenvalues by nonincreasing moduli). However, the convergence is much slower when the Jordan blocks of M relative to λ are nontrivial. The error decays then like $1/k$ in general.

The situation is more delicate when $\rho(M)$ is the modulus of several distinct eigenvalues. The vector x^k, suitably normalized, does not converge in general but "spins" closer and closer to the sum of the corresponding eigenspaces. The observation of the asymptotic behavior of x^k allows us to identify the eigendirections associated to the eigenvalues of maximal modulus. The sequence $\|Mx^k\|$ does not converge and depends strongly on the choice of the norm. However, $\log \|Mx^k\|$ converges in the Cesaro sense, that is, in the mean, to $\log \rho(M)$ (Exercise 12).

Remark: The hypothesis on x_0 is generic, in the sense that it is satisfied for every choice of x_0 in an open dense subset of $\mathbb{C}^n$. If by chance x^0 belongs to F, the power method furnishes theoretically another eigenvalue, of smaller modulus. In practice, a large enough number of iterations always allows for the convergence to λ. In fact, the number λ is rarely exactly representable in a computer. If it is not, the linear subspace F does not contain any nonzero representable vector. Thus the vector x^0, or its computer representation, does not belong to F, and Theorem 10.4.1 applies.

10.4.2 The Inverse Power Method

Let us assume that M is invertible. The standard power method, applied to M^{-1}, furnishes the eigenvalue of least modulus, whenever it is simple, or at least its modulus in the general case. Since the inversion of a matrix is a costly operation, we involve ourselves with that idea only if M has already been inverted, for example if we had previously had to make an LU or a QR factorization. That is typically the situation when one begins to implement the QR algorithm for M. It might look strange to involve a method giving only one eigenvalue in the course of a method that is expected to compute the whole spectrum.

The inverse power method is thus subtle. Here is the idea. One begins by implementing the QR method, until one gets coarse approximations $\mu_1, \dots, \mu_n$ of the eigenvalues $\lambda_1, \dots, \lambda_n$. If one persists in the QR method, the proof of Theorem 10.2.1 shows that the error is at best of order σ^k with $\sigma = \max_j |\lambda_{j+1}/\lambda_j|$. When n is large, σ is in general close to 1 and this convergence is rather slow. Similarly, the method with Rayleigh translations, for which σ is replaced by $\sigma(\eta) := \max_j |(\lambda_{j+1} - \eta)/(\lambda_j - \eta)|$, is not satisfactory. However, if one wishes to compute a *single* eigenvalue, say λ_p, with full accuracy, the power method, applied to $M - \mu_p I_n$, produces an error on the order of θ^k, where $\theta := |\lambda_p - \mu_p| / \min_{j \neq p} |\lambda_j - \mu_p|$ is a small number, since $\lambda_p - \mu_p$ is small.

In practice, the inverse power method is used mainly to compute an approximate eigenvector, associated to an eigenvalue for which one already has a good approximate value.

10.5 Leverrier's Method

The method of Leverrier allows for the computation of the characteristic polynomial of a square matrix. Though inserted in this Chapter, this method *is not* suitable for computing approximate values of the eigenvalues of a matrix. First of all, it furnishes only the characteristic polynomial which, as mentioned at the opening if this chapter, is not a good technique for computing the eigenvalues. Its interest is purely academic. Observe, however, that it is of great generality, applying to matrices with entries in any field of characteristic 0.

10.5.1 Description of the Method

Let K be a field of characteristic 0 and $M \in \mathbf{M}_n(K)$ be given. Let us denote by $\lambda_1, \dots, \lambda_n$ the eigenvalues of M, counted with multiplicity. Let us define the two following lists of n numbers:

Elementary symmetric polynomials

$$\begin{aligned}
\sigma_1 &:= \lambda_1 + \cdots + \lambda_n = \operatorname{Tr} M, \\
\sigma_2 &:= \sum_{j<k} \lambda_j \lambda_k, \\
&\vdots \\
\sigma_r &:= \sum_{j_1<\cdots<j_r} \lambda_{j_1} \cdots \lambda_{j_r}, \\
&\vdots \\
\sigma_n &:= \prod_j \lambda_j = \det M.
\end{aligned}$$

Newton sums

$$s_m := \sum_j \lambda_j^m, \quad 1 \le m \le n.$$

The numbers $(-1)^j \sigma_j$ are the coefficients of the characteristic polynomial of M:

$$P_M(X) = X^n - \sigma_1 X^{n-1} + \sigma_2 X^{n-2} - \cdots + (-1)^n \sigma_n.$$

Furthermore, the s_m are the traces of the powers M^m. One can obtain them by computing $M^2, \ldots, M^n$. Each of these matrices is obtained in $O(n^\alpha)$ operations, with $2 \le \alpha \le 3$ ($\alpha = 3$, using the naive method for the product of two matrices). In all, the computation of $s_1, \ldots, s_n$ needs $O(n^{\alpha+1})$ operations, which is a lot, compared to iterative methods (QR, Jacobi), for which each iteration is made in $O(n^2)$ operations at worst.

The passage from Newton sums to elementary symmetric polynomials is done through Newton's formulas. If $\Sigma_j = (-1)^j \sigma_j$ and $\Sigma_0 := 1$, we have

$$m\Sigma_m + \sum_{k=1}^{m} s_k \Sigma_{m-k} = 0, \quad 1 \le n.$$

One uses these formulas in increasing order, beginning with $\Sigma_1 = -s_1$. When $\Sigma_1, \ldots, \Sigma_{m-1}$ are known, one computes

$$\Sigma_m = -\frac{1}{m}(s_1 \Sigma_{m-1} + \cdots + s_m \Sigma_0).$$

This computation, which needs only $O(n^2)$ operations, has a negligible cost.

Besides the high cost of this method, its instability is unfortunate when $k = \mathbb{R}$ or $k = \mathbb{C}$: when n is large, s_k increases like $\rho(M)^k$, thus much more rapidly than σ_k. The eigenvalues of smaller modulus are thus much perturbed by the round-off errors, and this is reinforced by the large number of operations.

When the field is of nonzero characteristic p, the Leverrier method may be employed only if $n < p$. Since $s_p = \sigma_1^p$, the computation of the s_m's for $m \geq p$ does not bring any new information about the σ_j's.

10.6 Exercises

1. Given a polynomial $P \in \mathbb{R}[X]$, use the Euclidean division in order to define a sequence of nonzero polynomials P_j in the following way. Set $P_0 = P$, $P_1 = P'$. If P_j is not constant, $-P_{j+1}$ is the remainder of the division of P_{j-1} by P_j: $P_{j-1} = Q_j P_j - P_{j+1}$, $\deg P_{j+1} < \deg P_j$.
   (a) Assume that P has only simple roots. Show that the sequence $(P_j)_j$ is well-defined, that it has only finitely many terms, and that it is a Sturm sequence.
   (b) Use Proposition 10.1.3 to compute the number of real roots of the real polynomials $X^2 + aX + b$ or $X^3 + pX + q$ in terms of their discriminants.

2. (J. Wilkinson [35], Section 5.45) Let $n = 2p - 1$ be an odd number and $W_n \in \mathbf{M}_n(\mathbb{R})$ be the symmetric tridiagonal matrix

$$\begin{pmatrix} p & 1 & & & & \\ 1 & \ddots & \ddots & & & \\ & \ddots & 1 & \ddots & & \\ & & \ddots & \ddots & 1 & \\ & & & 1 & p \end{pmatrix}.$$

The diagonal entries are thus $p, p-1, \ldots, 2, 1, 2, \ldots, p-1, p$, and the subdiagonal entries are equal to 1.

   (a) Show that the linear subspace

$$E' = \{X \in \mathbb{R}^n \,|\, x_{p+j} = x_{p-j}, 1 \leq j < p\}$$

   is stable under W_n. Similarly, show that the linear subspace

$$E'' = \{X \in \mathbb{R}^n \,|\, x_{p+j} = -x_{p-j}, 0 \leq j < p\}$$

   is stable under W_n.
   (b) Deduce that the spectrum of W_n is the union of the spectra of the matrices

$$W_n' = \begin{pmatrix} p & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 2 & 1 \end{pmatrix} \quad (\in \mathbf{M}_p(\mathbb{R}))$$

and

$$W_n'' = \begin{pmatrix} p & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 3 & 1 \\ & & & 1 & 2 \end{pmatrix} \quad (\in \mathbf{M}_{p-1}(\mathbb{R})).$$

(c) Show that the eigenvalues of W_n'' separate strictly those of W_n'.

3. For $a_1, \ldots, a_n \in \mathbb{R}$, with $\sum_j a_j = 1$, form the matrix

$$M(a) := \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & & a_n \\ a_2 & b_2 & a_3 & \vdots & \vdots & \vdots \\ a_3 & a_3 & b_3 & & \vdots & \vdots \\ a_4 & \cdots & & & & \vdots \\ & \cdots & \cdots & & & a_n \\ a_n & \cdots & \cdots & \cdots & a_n & b_n \end{pmatrix},$$

where $b_j := a_1 + \cdots + a_{j-1} - (j-2)a_j$.

(a) Compute the eigenvalues and the eigenvectors of $M(a)$.
(b) We limit ourselves to n-uplets a that belong to the simplex S defined by $0 \leq a_n \leq \cdots \leq a_1$ and $\sum_j a_j = 1$. Show that for $a \in S$ $M(a)$ is bistochastic and $b_2 - a_2 \leq \cdots \leq b_n - a_n \leq 1$.
(c) Let $\mu_1, \ldots, \mu_n$ be an n-uplet of elements in $[0,1]$ with $\mu_n = 1$. Show that there exists a unique a in S such that $\{\mu_1, \ldots, \mu_n\}$ is equal to the spectrum of $M(a)$ (counting with multiplicity).
(d) Consider the unit sphere Σ of $\mathbf{M}_n(\mathbb{R})$, when this space is endowed with the norm $\|M\|_2 = \sqrt{\rho(M^T M)}$. Show that if $P \in \Sigma$, then there exists a convex polytope T, of dimension $(n-1)^2$, included in Σ and containing P. **Hint**: Use Corollary 5.5.1, with unitary invariance of the norm $\|\cdot\|_2$.

4. Show that the cost of an iteration of the QR method for a Hermitian tridiagonal matrix is $20n + O(1)$.

5. Show that the reduction to the Hessenberg form (in this case, tridiagonal form) of a Hermitian matrix costs $7n^3/6 + O(n^2)$ operations.

6. (Invariants of the algorithm QR) For $M \in \mathbf{M}_n(\mathbb{R})$ and $1 \leq k \leq n-1$, let us denote by $(M)_k$ the matrix of size $(n-k) \times (n-k)$ obtained by deleting the first k rows and the last k columns. For example, $(I)_1$ is the Jordan matrix $J(0; n-1)$. We shall denote also by $K \in \mathbf{M}_n(\mathbb{R})$ the matrix defined by $k_{1n} = 1$ and $k_{ij} = 0$ otherwise.

(a) For an upper triangular matrix T, compute explicitly KT and TK.
(b) Let $M \in \mathbf{M}_n(\mathbb{R})$. Prove the equality

$$\det(M - \lambda I - \mu K) = (-1)^n \mu \det(M - \lambda I)_1 + \det(M - \lambda I).$$

(c) Let $A \in \mathbf{GL}_n(\mathbb{R})$ be given, with factorization $A = QR$. Prove that

$$\det(A - \lambda I)_1 = \frac{\det R}{r_{nn}} \det(Q - \lambda R^{-1})_1.$$

(d) Let $A' = RQ$. Show that

$$r_{nn} \det(A' - \lambda I)_1 = r_{11} \det(A - \lambda I)_1.$$

(e) Generalize the previous calculation by replacing the index 1 by k. Deduce that the roots of the polynomial $\det(A - \lambda I)_k$ are conserved throughout the QR algorithm. How many such roots do we have for a general matrix? How many for a Hessenberg matrix?

7. (Invariants; continuing) For $M \in \mathbf{M}_n(\mathbb{R})$, let us define $P_M(h; z) := \det((1-h)M + hM^T - zI_n)$.
(a) Show that $P_M(h; z) = P_M(1-h; z)$. Deduce that there exists a polynomial Q_M such that $P_M(h; z) = Q_M(h(1-h); z)$.
(b) Show that Q_M remains constant throughout the QR algorithm: If $Q \in \mathbf{O}_n(\mathbb{R})$, R is upper triangular, and $M = QR$, $N = RQ$, then $Q_M = Q_N$.
(c) Deduce that there exist polynomial functions J_{rk} on $\mathbf{M}_n(\mathbb{R})$, defined by

$$P_M(h; z) = \sum_{r=0}^{n} \sum_{k=0}^{[r/2]} (h(1-h))^k z^{n-r} J_{rk}(M),$$

that are invariant throughout the QR algorithm. Verify that the J_{r0}'s can be expressed in terms of invariants that we already know.
(d) Compute explicitly J_{21} when $n = 2$. Deduce that in the case where Theorem 10.2.1 applies and $\det A > 0$, the matrix A_k converges.
(e) Show that for $n \geq 2$,

$$J_{21}(M) = -\frac{1}{2} \operatorname{Tr}\left((M - M^T)^2\right).$$

Deduce that if A_k converges to a diagonal matrix, then A is symmetric.

8. In the Jacobi method, show that if the eigenvalues are simple, then the product $R^1 \cdots R^m$ converges, to an orthogonal matrix R such that $R^* A R$ is diagonal.

9. Extend the Jacobi method to Hermitian matrices. **Hint**: Replace the rotation matrices

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

by unitary matrices

$$\begin{pmatrix} z_1 & z_2 \\ z_3 & z_4 \end{pmatrix}.$$

10. Let $A \in \mathbf{Sym}_n(\mathbb{R})$ be a matrix whose eigenvalues, of course real, are simple. Apply the Jacobi method, but selecting the angle θ_k so that $\pi/4 \leq |\theta_k| \leq \pi/2$.

    (a) Show that E_k tends to zero, that the sequence D_k is relatively compact, and that its cluster values are diagonal matrices whose diagonal terms are the eigenvalues of A.

    (b) Show that an iteration has the effect of permuting, asymptotically, $a_{pp}^{(k)}$ and $a_{qq}^{(k)}$, where $(p,q) = (p_k, q_k)$. In other words

$$\lim_{k \to +\infty} |a_{pp}^{(k+1)} - a_{qq}^{(k)}| = 0,$$

    and vice versa, permuting p and q.

11. The Bernoulli method computes an approximation of the root of largest modulus for a polynomial $a_0 X^n + \cdots + a_n$, when that root is unique. To do so, one defines a sequence by a linear induction of order n:

$$z_k = -\frac{1}{a_0}(a_1 z_{k-1} + \cdots + a_n z_{k-n}).$$

    Compare this method with the power method for a suitable matrix.

12. Consider the power method for a matrix $M \in \mathbf{M}_n(\mathbb{C})$ of which several eigenvalues are of modulus $\rho(M) \neq 0$. Again, $\mathbb{C}^n = E \oplus F$ is the decomposition of $\mathbb{C}^n$ into linear subspaces stable under M, such that $\rho(M|_F) < \rho(M)$ and $\lambda \in \mathrm{Sp}(M|_E) \Longrightarrow |\lambda| = \rho(M)$. Finally, $x^0 = y^0 + z^0$ with $y^0 \in E$, $z^0 \in F$, and $y^0 \neq 0$.

    (a) Express

$$\frac{1}{m} \sum_{k=0}^{m-1} \log \|M x^k\|$$

    in terms of $\|M^m x^0\|$.

(b) Show that if $0 < \mu < \rho(M) < \eta$, then there exist constants C, C' such that

$$C\mu^k \leq \|M^k x^0\| \leq C'\eta^k, \quad \forall k \in I\!N.$$

(c) Deduce that $\log \|Mx^k\|$ converges in the mean to $\log \rho(M)$.

13. Let $M \in \mathbf{M}_n(\mathbb{C})$ be given. Assume that the Gershgorin disk D_l is disjoint from the other disks D_m, $m \neq l$. Show that the inverse power method, applied to $M - m_{ll} I_n$, provides an approximate computation of the unique eigenvalue of M that belongs to D_l.

References

[1] Jacques Baranger. *Analyse numérique*. Hermann, Paris, 1991.

[2] G. R. Belitskii and Yurii. I. Lyubich. *Matrix norms and their applications*, volume 36 of *Operator theory : advances and applications*. Birkhauser, Bâle, 1988.

[3] M. Berger and B. Gostiaux. *Differential geometry : manifold, curves and surfaces*, volume 115 of *Graduate text in Mathematics*. Springer-Verlag, New York, 1988.

[4] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate text in Mathematics*. Springer-Verlag, Heidelberg, 1996.

[5] Rajendra Bhatia. Pinching, trimming, truncating, and averaging of matrices. *Amer. Math. Monthly*, 107(7):602–608, 2000.

[6] Rajendra Bhatia. Linear algebra to quantum cohomology : the story of Alfred Horn's inequalities. *Amer. Math. Monthly*, 108(4):289–318, 2001.

[7] Peter Bürgisser, Michael Clausen, and M. Amin Shokrollahi. *Algebraic complexity theory*. Springer-Verlag, Berlin, 1997. With the collaboration of Thomas Lickteig.

[8] Philippe Ciarlet. *Introduction to numerical linear algebra and optimisation*. Cambridge texts in Applied Mathematics. Cambridge University Press, Cambridge, 1989.

[9] Philippe Ciarlet and Jean-Marie Thomas. *Exercices d'analyse numérique matricielle et d'optimisation*. Mathématiques appliquées pour la maîtrise. Masson, Paris, 1982.

[10] Harvey Cohn. *Advanced number theory*. Dover Publications Inc., New York, 1980. Reprint of *A second course in number theory*, 1962, Dover Books on Advanced Mathematics.

[11] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *J. Symbolic Comput.*, 9(3):251–280, 1990.

[12] J. Davis, Philip. *Circulant matrices*. Chelsea Publishing, New York, 1979.

[13] M. Fiedler and V. Pták. On matrices with non-positive off-diagonal elements and positive principal minors. *Czech. Math. Journal*, 12:382–400, 1962.

[14] Edward Formanek. Polynomial identities and the Cayley-Hamilton theorem. *Math. Intelligencer*, 11(1):37–39, 1989.

[15] Edward Formanek. *The polynomial identities and invariants of $n \times n$ matrices*. Number 78 in CBMS Regional Conf. Ser. Math. Amer. Math. Soc., Providence, RI, 1991.

[16] William Fulton. Eigenvalues, invariant factors, highest weights, and Schubert calculus. *Bull. Amer. Math. Soc. (N.S.)*, 37(3):209–249 (electronic), 2000.

[17] F. R. Gantmacher. *The theory of matrices. Vol. 1*. Chelsea Publish. Co., New York, 1959.

[18] F. R. Gantmacher. *The theory of matrices. Vol. 2*. Chelsea Publish. Co., New York, 1959.

[19] Gene H. Golub and Charles F. Van Loan. *Matrix computations*, volume 3 of *Series in the mathematical sciences*. John Hopkins University Press, Baltimore, 1983.

[20] Nicholas Higham. *Accuracy and stability of numerical algorithms*. SIAM, Philadelphia, PA, 1996.

[21] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1985.

[22] Alston S. Householder. *The theory of matrices in numerical analysis*. Dover, New York, 1975.

[23] Nicholas M. Katz and Peter Sarnak. *Random matrices, Frobenius eigenvalues and monodromy*. Number 45 in Colloquium publ. Amer. Math. Soc., Providence, RI, 1999.

[24] Anthony W. Knapp. *Representation of semisimple groups. An overview based on examples*. Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1986.

[25] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Masson, Paris, 1987.

[26] Chi-Wang Li and Roy Mathias. Extremal characterization of the Schur complement and resulting inequalities. *SIAM Review*, 42:233–246, 2000.

[27] Helmut Lütkepohl. *Handbook of matrices*. J. Wiley & Sons, New York, 1996.

[28] Mneimné, Rached and Testard, Frédéric. *Introduction à la théorie des groupes de Lie classiques*. Hermann, Paris, 1986.

[29] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book co, NY, third edition, 1987.

[30] Walter Rudin. *Functional analysis*. McGraw-Hill Book Co, NY, second edition, 1991.

[31] E. Seneta. *Non-negative matrices and Markov chains*. Springer series in statistics. Springer-Verlag, New York-Berlin, 1981.

[32] Joseph Stoer and Christoph Witzgall. Transformations by diagonal matrices in a normed space. *Numer. Math.*, 4:158–171, 1962.

[33] Volker Strassen. Gaussian elimination is not optimal. *Numer. Math.*, 13:354–356, 1969.

[34] J. H. M. Wedderburn. *Lectures on matrices*, volume XVII of *Colloquium publications*. American Math. Society, New York, 1934.

[35] J. H. Wilkinson. *The algebraic eigenvalue problem*. Oxford Science Publications, Oxford, 1965.

This page intentionally left blank

Index

听雨尘心@含藏识

免费电子书下载