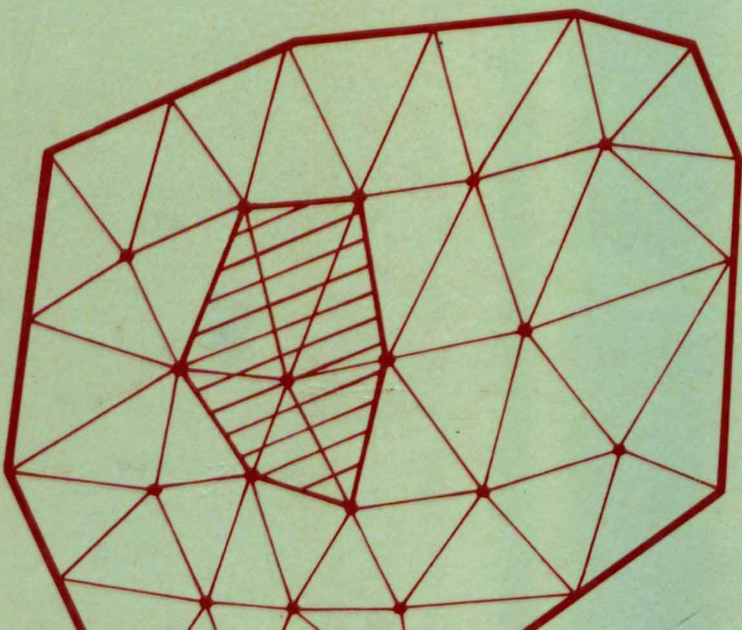
CAMBRIDGE TEXTS IN APPLIED MATHEMATICS

Introduction to numerical linear algebra and optimisation



PHILIPPE G. CIARLET

Introduction to numerical linear algebra and optimisation

PHILIPPE G. CIARLET

Université Pierre et Marie Curie, Paris

with the assistance of

Bernadette Miara and Jean-Marie Thomas

for the exercises

Translated by A. Buttigieg, S.J.

Campion Hall, Oxford

The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS

Cambridge

New York Port Chester

Melbourne Sydney

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

Originally published in French as *Introduction à l'analyse numérique matricielle et à l'optimisation* and *Exercises d'analyse numérique matricielle et d'optimisation* in the series *Mathématiques appliquées pour la maîtrise*, editors P.G. Ciarlet and J.L. Lions, Masson, Pairs, 1982 and

First published in English by Cambridge University Press 1989 as
Introduction to numerical linear algebra and optimisation

Printed in Great Britain at the University Press, Cambridge

British Library cataloguing in publication data

Ciarlet, Philippe, G.
Introduction to numerical linear algebra and optimisation. –
(Cambridge texts in applied mathematics; v. 2).
1. Equations – Numerical solutions
2. Mathematical optimisation
I. Title II. Miara, Bernadette
III. Thomas, Jean-Marie IV. Introduction à l'analyse
numérique matricielle et à l'optimisation. *English*
512.9'42 QA218

Library of Congress cataloguing in publication data

Ciarlet, Philippe, G.
[Introduction à l'analyse numérique matricielle et à l'optimisation. English]
Introduction to numerical linear algebra and optimisation/Philippe G. Ciarlet, with the assistance of Bernadette Miara and Jean-Marie Thomas for the exercises; translated by A. Buttigieg. p. cm.
Translation of: Introduction à l'analyse numérique matricielle et à l'optimisation and Exercises d'analyse numérique matricielle et d'optimisation.
Bibliography: p.
Includes index.
ISBN 0 521 32788 1. ISBN 0 521 33984 7 (pbk.)
1. Algebra, Linear. 2. Numerical calculations. 3. Mathematical optimization.
I. Miara, Bernadette. II. Thomas, Jean-Marie. III. Title. IV. Title:
Numerical linear algebra and optimisation.
QA184.C525 1988
512'.5–dc 19 87–35931 CIP

ISBN 0 521 32788 1 hard covers
ISBN 0 521 33984 7 paperback

TM

I dedicate this English edition to Richard S. Varga

1 Preface to the English edition

My main purpose in writing this textbook was to give, within reasonable limits, a thorough description, and a rigorous mathematical analysis, of some of the most commonly used methods in Numerical Linear Algebra and Optimisation.

Its contents should illustrate not only the remarkable efficiency of these methods, but also the interest *per se* of their mathematical analysis. If the first aspect should especially appeal to the more practically oriented readers and the second to the more mathematically oriented readers, it may be also hoped that both kinds of readers could develop a common interest in these two complementary aspects of Numerical Analysis.

This textbook should be of interest to advanced undergraduate and beginning graduate students in Pure or Applied Mathematics, Mechanics, and Engineering. It should also be useful to practising engineers, physicists, biologists, economists, etc., wishing to acquire a basic knowledge of, or to implement, the basic numerical methods that are constantly used today.

In all cases, it should prove easy for the instructor to adapt the contents to his or her needs and to the level of the audience. For instance, a three hours per week, one-semester, course can be based on Chapters 1 to 6, or on Chapters 7 to 10, or on Chapters 4 to 8.

The mathematical prerequisites are relatively modest, especially in the first part. More specifically, I assumed that the readers are already reasonably familiar with the basic properties of matrices (including matrix computations) and of finite-dimensional vector spaces (continuity and differentiability of functions of several variables, compactness, linear mappings). In the second part, where various results are presented in the more general settings of Banach or Hilbert spaces, and where differential calculus in general normed vector spaces is often used, all relevant definitions and results are precisely stated wherever they are needed. Besides, the text is written in such a way that, in each case, the reader not familiar with these more abstract situations can, without any difficulty, 'stay in finite-dimensional spaces' and thus ignore these generalisations (in this spirit, weak convergence is used for proving only one 'infinite-dimensional' result, whose elementary 'finite-dimensional' proof is also given).

This textbook has some features which, in my opinion, are worth mentioning.

The *combination in a single volume of Numerical Linear Algebra and Optimisation*, with a progressive transition, and many cross-references, between these two themes;
A *mathematical level slowly increasing with the chapter number*;
A considerable space devoted to *reviews of pertinent background material*;
A *description of various practical problems*, originating in *Physics, Mechanics*, or *Economics*, whose numerical solution requires methods from Numerical Linear Algebra or Optimisation;
Complete proofs are given of each theorem;
Many *exercises* or *problems* conclude each section.

The *first part* (Chapters 1 to 6) is essentially devoted to *Numerical Linear Algebra*. It contains:

A review of all those results about *matrices* and *vector or matrix norms* that will be subsequently used (Chapter 1);
Basic notions about the *conditioning* of linear systems and eigenvalue problems (Chapter 2);
A review of various *approximate methods* (finite-difference methods, finite element methods, polynomial and spline interpolations, least square approximations, approximation of 'small' vibrations) that eventually lead to the solution of a linear system or of a matrix eigenvalue problem (Chapter 3);
A description and a mathematical analysis of some of the fundamental *direct methods* (Gauss, Cholesky, Householder; cf. Chapter 4) and *iterative methods* (Jacobi, Gauss–Seidel, relaxation; cf. Chapter 5) for solving *linear systems*;
A description and a mathematical analysis of some of the fundamental methods (Jacobi, Givens–Householder, QR, inverse method) for computing the *eigenvalues and eigenvectors of matrices* (Chapter 6).

The *second part* (Chapters 7 to 10) is essentially devoted to *Optimisation*. It contains:

A thorough review of all relevant prerequisites about *differential calculus in normed vector spaces* (Chapter 7) and about *Hilbert spaces* (Chapter 8);

A progressive introduction to Optimisation, through analyses of *Lagrange multipliers*, of *extrema* and *convexity* of real functions, and of *Newton's method* (Chapter 7);

A description of various linear and nonlinear problems whose approximate solution leads to *minimisation problems in* $\mathbb{R}^n$, *with or without constraints* (Chapters 8 and 10);

A description and mathematical analysis of some of the *fundamental algorithms of Optimisation theory* – relaxation methods, gradient methods (with optimal, fixed, or variable, parameter), conjugate gradient methods, penalty methods (Chapter 8), Uzawa's method (Chapter 9), simplex method (Chapter 10);

An introduction to *duality theory* – Farkas lemma, Kuhn and Tucker relations, Lagrangians and saddle-points, duality in linear programming (Chapters 9 and 10).

More complete descriptions of the topics treated are found in the *introductions* to each chapter.

Important results are stated as *theorems, which thus constitute the core of the text* (there are no lemmas, propositions, or corollaries).

Although the many *remarks* may be in principle skipped during a first reading, they should nevertheless prove to be helpful, by mentioning various special cases of interest, possible generalisations, counter-examples, etc.

The numerous *exercises* and *problems* that conclude each section provide often important, and sometimes challenging to prove, additions to the text.

In addition to *'local' references* (about a specific result, a particular extension, etc.) found at some places, *references of a more general nature are listed by subject and commented upon in a special section, titled 'Bibliography and comments'*, at the end of the book. The reader interested by more in-depth treatments of the various topics considered here, or by the *practical implementation* of the methods, should definitely refer to this section.

While I wrote this text, many colleagues and students were kind enough to make various comments, remarks, suggestions, etc., that substantially contributed to its improvement. In this respect, particular thanks are due to Alain Bamberger, Claude Basdevant, Michel Bernadou, Michel Crouzeix, David Feingold, Srinivasan Kesavan, Colette Lebaud, Jean Meinguet, Annie Raoult, Pierre-Arnaud Raviart, François Robert, Ulrich

Tulowitzki, Lars Wahlbin. Above all, my sincere thanks are due to Bernadette Miara and Jean-Marie Thomas, who not only carefully read the entire manuscript, but also significantly contributed to devising many exercises and problems.

It is also my pleasure to thank David Tranah of Cambridge University Press, and the translator, Alfred Buttigieg, S.J., whose friendly and efficient co-operation made this edition possible.

In 1964, at Case Institute of Technology (now Case Western Reserve University), I had the honour of having an outstanding teacher, who communicated to me his enthusiasm for Numerical Analysis. It is indeed a great privilege to dedicate this English edition to this teacher: Richard S. Varga.

Philippe G. Ciarlet
July 1988

Contents

1

A summary of results on matrices

Introduction

The purpose of this chapter is to recall, and to prove, a number of results relating to matrices and finite-dimensional vector spaces, of which frequent use will be made in the sequel.

It is assumed that the reader is familiar with the elementary properties of finite-dimensional vector spaces (and, in particular, with the theory of matrices). In section 1.1, we give the central definitions and notation relevant to these properties, as also the notion of *block partitioning of a matrix*, which is of outstanding importance in the area of the Numerical Analysis of Matrices.

In order to make this volume as 'self-contained' as possible, all results which are required subsequently are proved: in particular, *the reduction of a general matrix to triangular form, the diagonalisation of normal matrices* (Theorem 1.2-1), and *the equivalence of a matrix to the diagonal matrix of its singular values* (Theorem 1.2-2). (In this respect, it is relevant to point out that we will have no call to make use of Jordan's theorem.) We then examine (Theorem 1.3-1) the *characterisations of the eigenvalues of symmetric or Hermitian matrices* through the use of *Rayleigh's quotient*, and notably the characterisations in terms of 'min-max' and 'max-min'.

We next review the *vector norms* which are the most frequently utilised in the Numerical Analysis of Matrices. These are particular cases of the '*l_p-norms*' (Theorem 1.4-1). We then determine the corresponding *subordinate matrix norms* (Theorem 1.4-2), an example of a matrix norm which is not subordinate to a vector norm being given in Theorem 1.4-4. A reminder is given in Theorem 1.4-5 of the conditions for the invertibility of matrices of the form I + B, and it is shown (Theorem 1.4-3) that *the spectral radius of a matrix is the lower bound of the values of its norms.* This last result is in turn used to prove *two results about the sequence of successive powers of a matrix* (Theorems 1.5-1 and 1.5-2). These play a fundamental role in the study of iterative methods for the solution of linear systems, which are studied in Chapter 5.

1.1 Key definitions and notation

Let V be a vector space of finite dimension n, over the field $\mathbb{R}$ of real numbers, or the field $\mathbb{C}$ of complex numbers; if there is no need to distinguish between the two, we will speak of the field $\mathbb{K}$ of *scalars*.

A *basis* of V is a set $\{e_1, e_2, \dots, e_n\}$ of n linearly independent vectors of V, denoted by $(e_i)_{i=1}^n$, or quite simply by (e_i) if there is no risk of confusion. Every vector $v \in V$ then has the unique representation

$$v = \sum_{i=1}^{n} v_i e_i,$$

the scalars v_i, which we will sometimes denote by $(v)_i$, being the *components* of the vector v relative to the basis (e_i). *As long as a basis is fixed unambiguously, it is thus always possible to identify V with $\mathbb{K}^n$*; that is why it will turn out to be just as likely for us to write $v = (v_i)_{i=1}^n$, or simply (v_i), for a vector v whose components are v_i.

In matrix notation, the vector $v = \sum_{i=1}^n v_i e_i$ will *always* be represented by the *column vector*

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

while v^{T} and v^* will denote the following *row vectors*:

$$v^{\mathrm{T}} = (v_1 v_2 \cdots v_n), \quad v^* = (\bar{v}_1 \bar{v}_2 \cdots \bar{v}_n),$$

where, in general, $\bar{\alpha}$ is the complex conjugate of α. The row vector v^{T} is the *transpose* of the column vector v, and the row vector v^* is the *conjugate transpose* of the column vector v.

The function $(\cdot,\cdot)$: $V \times V \to \mathbb{K}$ defined by

$$(u, v) = v^{\mathrm{T}} u = u^{\mathrm{T}} v = \sum_{i=1}^{n} u_i v_i \quad \text{if} \quad \mathbb{K} = \mathbb{R},$$

$$(u, v) = v^* u = \overline{u^* v} = \sum_{i=1}^{n} u_i \bar{v}_i \quad \text{if} \quad \mathbb{K} = \mathbb{C},$$

will be called the *Euclidean scalar product* if $\mathbb{K} = \mathbb{R}$, the *Hermitian scalar product* if $\mathbb{K} = \mathbb{C}$ and the *canonical scalar product* if the underlying field is left unspecified. When it is desired to keep in mind the dimension of the vector space, we shall write

$$(u, v) = (u, v)_n.$$

Let V be a vector space which is provided with a canonical scalar product. Two vectors u and v of V are *orthogonal* if $(u, v) = 0$. By extension, the vector v is said to be *orthogonal to the subset U of V* (in symbols, $v \perp U$), if the vector v is orthogonal to all the vectors in U. Lastly, a set $\{v_1, \ldots, v_k\}$ of vectors belonging to the space V is said to be *orthonormal* if

$$(v_i, v_j) = \delta_{ij}, \quad 1 \leqslant i, j \leqslant k,$$

where δ_{ij} is the *Kronecker delta*: $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$.

Let V and W be two vector spaces over the same field, equipped with bases $(e_j)_{j=1}^n$ and $(f_i)_{i=1}^m$ respectively. *Relative to these bases*, a linear transformation

$$\mathscr{A}: V \to W$$

is represented by the *matrix* having m *rows* and n columns:

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

the *elements* a_{ij} of the matrix A being defined uniquely by the relations

$$\mathscr{A} e_j = \sum_{i=1}^{m} a_{ij} f_i, \quad 1 \leqslant j \leqslant n.$$

Equivalently, the jth *column vector*

$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

of the matrix A represents the vector $\mathscr{A} e_j$ relative to the basis $(f_i)_{i=1}^m$. We call

$$(a_{i1} a_{i2} \cdots a_{in})$$

the ith *row vector of the matrix* A.

A matrix with m rows and n columns is called a *matrix of type* (m, n), and the vector space over the field $\mathbb{K}$ consisting of matrices of type (m, n) with elements in $\mathbb{K}$ is denoted by $\mathscr{A}_{m,n}(\mathbb{K})$ or simply $\mathscr{A}_{m,n}$. A column vector is then a matrix of type $(m, 1)$ and a row vector a matrix of type $(1, n)$. A matrix is called *real* or *complex* according as its elements are in the field $\mathbb{R}$ or the field $\mathbb{C}$.

A matrix A with elements a_{ij} is written as

$$\mathrm{A} = (a_{ij}),$$

the first index i *always* designating the row and the second, j, the column. Given a matrix A, $(\mathrm{A})_{ij}$ denotes the element in the ith row and jth column.

The *null matrix* and the *null vector* are represented by the same symbol 0.

Given a matrix $A\in\mathscr{A}_{m,n}(\mathbb{C})$, $A^*\in\mathscr{A}_{n,m}(\mathbb{C})$ denotes the *adjoint* of the matrix A and is defined uniquely by the relations

$$(Au, v)_m = (u, A^*v)_n \quad \text{for every } u\in\mathbb{C}^n, \quad v\in\mathbb{C}^m,$$

which imply that $(A^*)_{ij} = \bar{a}_{ji}$. In the same way, given a matrix $A = \mathscr{A}_{m,n}(\mathbb{R})$, $A^T\in\mathscr{A}_{n,m}(\mathbb{R})$ denotes the *transpose* of the matrix A and is defined uniquely by the relations

$$(Au, v)_m = (u, A^Tv)_n \quad \text{for every } u\in\mathbb{R}^n, \quad v\in\mathbb{R}^m,$$

which imply that $(A^T)_{ij} = a_{ji}$.

Remarks

(1) One could also define the transpose of a complex matrix. However, that would provide a concept of limited interest, since the function $u, v \to \sum_{i=1}^n u_i v_i$ is not a scalar product in $\mathbb{C}^n$.

(2) The notation A^T has been given preference over the notation ${}^tA$, this latter being more suitably linked to the notion of a dual basis. The notation A^T keeps in mind the dependence of the notion of transpose upon a particular scalar product, the canonical scalar product.

To the composition of linear transformations there corresponds the multiplication of matrices. If $A = (a_{ik})$ is a matrix of type (m, l) and $B = (b_{kj})$ of type (l, n), their product AB is the matrix of type (m, n) defined by

$$(AB)_{ij} = \sum_{k=1}^{l} a_{ik}b_{kj}.$$

Recall that $(AB)^T = B^TA^T$, $(AB)^* = B^*A^*$.

Let $A = (a_{ij})$ be a matrix of type (m, n). We shall use the term *submatrix* of A for every matrix of the form

$$\begin{pmatrix} a_{i_1j_1} & a_{i_1j_2} & \cdots & a_{i_1j_q} \\ a_{i_2j_1} & a_{i_2j_2} & \cdots & a_{i_2j_q} \\ \vdots & \vdots & & \vdots \\ a_{i_pj_1} & a_{i_pj_2} & \cdots & a_{i_pj_q} \end{pmatrix},$$

provided the integers i_k and j_l satisfy

$$1 \leqslant i_1 < i_2 < \cdots < i_p \leqslant m, \quad 1 \leqslant j_1 < j_2 < \cdots < j_q \leqslant n.$$

Let $A = (a_{ij})$ be the matrix representing a linear transformation from V into W and let

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_N, \quad W = W_1 \oplus W_2 \oplus \cdots \oplus W_M$$

be decompositions of the spaces V and W into the direct sum of subspaces V_J and W_I, of dimensions n_J and m_I respectively, each spanned by a set of basis vectors. With this decomposition of the spaces V and W is associated the *block decomposition of the matrix* A:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & \vdots & & \vdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} \end{pmatrix} = (A_{IJ})$$

each submatrix A_{IJ}, of type (m_I, n_J), representing a linear transformation from the space V_J into the space W_I. What is of interest in these block decompositions is the fact that some of the operations defined on matrices remain *formally* the same, 'the coefficients a_{ij} being replaced by the submatrices A_{IJ}'. However, care is required over the *order* of the factors!

Thus, let $A = (A_{IK})$ and $B = (B_{KJ})$ be two matrices, of type (m, l) and (l, n) respectively, decomposed into blocks, the *decomposition corresponding to the index K being the same* for each matrix. *The matrix* AB *then admits the following block decomposition*

$$AB = (C_{IJ}), \quad \text{with} \quad C_{IJ} = \sum_K A_{IK} B_{KJ},$$

and in this way one is said to have carried out the *block multiplication* of the two matrices.

In the same way, let v be a vector in the space V and let $v = \sum_{J=1}^{N} v_J$, $v_J \in V_J$, be the (unique) representation associated with the decomposition of the space V into a direct sum. The vector $Av \in W$ then has the representation

$$Av = \sum_{I=1}^{M} w_I, \quad \text{with} \quad w_I = \sum_{J=1}^{N} A_{IJ} v_J,$$

as the unique representation associated with the decomposition of the space W into a direct sum. This is equivalent to considering *the vectors v and Av as decomposed into blocks*

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix}, \quad Av = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix}, \quad w_I = \sum_{J=1}^{N} A_{IJ} v_J,$$

the last equation embodying *the block multiplication of the matrix* A *by the vector v.*

A matrix of type (n, n) is said to be *square*, or *a matrix of order n* if it is desired to make explicit the integer n; it is convenient to speak of a matrix as *rectangular* if it is not necessarily square. One denotes by

$$\mathscr{A}_n = \mathscr{A}_{n,n} \quad \text{or} \quad \mathscr{A}_n(\mathbb{K}) = \mathscr{A}_{n,n}(\mathbb{K})$$

the ring of square matrices of order n, with elements in the field $\mathbb{K}$.

Unless anything is said to the contrary, *the matrices to be considered up to the end of this section will be square.*

If $A = (a_{ij})$ is a square matrix, the elements a_{ii} are called *diagonal elements*, and the elements $a_{ij}, i \neq j$, are called *off-diagonal elements*. The *identity matrix* is the matrix

$$I = (\delta_{ij}).$$

A matrix A is *invertible* if there exists a matrix (which is unique, if it does exist), written as A^{-1} and called *the inverse* of the matrix A, which satisfies $AA^{-1} = A^{-1}A = I$. Otherwise, the matrix is said to be *singular*. Recall that if A and B are invertible matrices

$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A^T)^{-1} = (A^{-1})^T, \quad (A^*)^{-1} = (A^{-1})^*.$$

A matrix A is

symmetric if A is real and $A = A^T$,
Hermitian if $A = A^*$,
orthogonal if A is real and $AA^T = A^TA = I$,
unitary if $AA^* = A^*A = I$,
normal if $AA^* = A^*A$.

A matrix $A = (a_{ij})$ is *diagonal* if $a_{ij} = 0$ for $i \neq j$ and is written as

$$A = \operatorname{diag}(a_{ii}) = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

The *trace* of a matrix $A = (a_{ij})$ is defined by

$$\operatorname{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

Let $\mathfrak{S}_n$ be the group of permutations of the set $\{1, 2, \ldots, n\}$. To every element $\sigma \in \mathfrak{S}_n$ there corresponds the *permutation matrix*

$$P_\sigma = (\delta_{i\sigma(j)}).$$

Observe that *every permutation matrix is orthogonal.*

The *determinant* of a matrix A is defined by

$$\det(A) = \sum_{\sigma \in \mathfrak{S}_n} \varepsilon_\sigma a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n},$$

where $\varepsilon_\sigma = 1$, resp. -1, if the permutation σ is even, resp. odd.

The *eigenvalues* $\lambda_i = \lambda_i(A)$, $1 \leqslant i \leqslant n$, of a matrix A of order n are the n roots, real or complex, simple or multiple, of the *characteristic polynomial*

$$p_A : \lambda \in \mathbb{C} \to p_A(\lambda) = \det(A - \lambda I)$$