Course notes

LINMA2380 — Matrix Computations

Professor — Raphaël Jungers

Teaching Assistants — Guillaume Berger
Julien Calbert

Contact — julien.calbert@uclouvain.be

Université catholique de Louvain (UCLouvain)
Louvain school of Engineering (L.S.E. – E.P.L.)

November 17, 2023

LINMA2380 Notes 2023–2024 v1.0

Contents

Introduction

(This section is to be completed...)

Linear algebra is the language of mathematics.

Matrices are canonical representations of linear mappings between finite dimensional vector fields.

Invariants under equivalence relations allow to understand matrices by chopping off redundant information in order to focus on the unique information needed. *This is exactly what this course is about.*

Machine learning, Artificial Intelligence, and modern computations are often bound to very simple Linear Algebraic computations, because, even if the advances of technology allow to store and process huge data, most problems require an even bigger computational effort, and often, only Linear Algebraic problems are simple enough, or well-enough understandable, so as to allow for an efficient resolution.

As every year, there will be a bonus of 0.5 points to the final grade of the students that will have communicated to me the most important typos/mistakes in the current version of these lecture notes. Please contact Julien (put me cc) if you find any.

Thanks to Paul Van Dooren, Vladimir Gusev and Guillaume Berger.

Chapter 1

Matrix algebras

In this introductory chapter, we will give an overview of basic notions required for further development of the theory of matrices. The presented concepts are assumed to be known from the previous courses. Thus, we present only a high-level overview.

A *rectangular matrix of dimensions* $m \times n$ is a collection of mn elements a_{ij} $(i = 1, 2, \dots, m;$ $j = 1, 2, \dots, n)$ organized in a table A (sometimes denoted by $A_{m\times n}$ when we want to emphasize the dimensions):

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \tag{1.1}$$

An equivalent notation to (1.1), that we will use as well, is the following:

$$A = [a_{ij}]_{i,j=1}^{m,n} \qquad \text{(or } A = [a_{ij}] \text{ if the dimensions are clear from the context).}$$

In the course, we will typically assume that the elements a_{ij} belong to a ring $\mathcal{R}$ (e.g., integers, polynomials, etc.) or to a field $\mathcal{F}$ (e.g., reals, complex numbers, rationals, etc.). We refer the reader to Appendix A for the definitions of these mathematical structures. We denote by $\mathcal{R}^{m\times n}$ and $\mathcal{F}^{m\times n}$ the set of $m \times n$ matrices with elements in $\mathcal{R}$ and $\mathcal{F}$ respectively.

We will say that two matrices of equal dimensions are *equal* if their corresponding elements are equal:

$$A_{m\times n} = B_{m\times n} \quad \Longleftrightarrow \quad a_{ij} = b_{ij}, \quad 1 \le i \le m, \quad 1 \le j \le n.$$

1.1 Sums and products

The addition of two matrices A and B of equal dimensions is defined as follows :

$$A_{m\times n} + B_{m\times n} \coloneqq [a_{ij} + b_{ij}]_{i,j=1}^{m,n}.$$

Exercise 1.1. *Show that addition of matrices is commutative, associative, and that the neutral element is the zero matrix:* $0_{m\times n} = [0]_{i,j=1}^{m,n}$.

We define the scalar multiplication of a matrix A by a scalar α as follows:

$$\alpha A_{m\times n} \coloneqq [\alpha a_{ij}]_{i,j=1}^{m,n}. \tag{1.2}$$

Note that the scalar α can belong to a different set than the elements of A, provided that the multiplication is well-defined (for example $\alpha \in \mathbb{C}$ and $a_{ij} \in \mathbb{C}[z]$, the set of polynomials of a single variable z with complex coefficients).

Exercise 1.2. *Verify that, for all matrices A and B belonging to $\mathcal{M}$, the set of matrices of fixed dimensions $m \times n$, and for all scalars α, β belonging to a field $\mathcal{F}$ or a ring $\mathcal{R}$, it holds that*

$$\begin{aligned} 0A &= 0, \\ 1A &= A, \\ (\alpha+\beta)A &= \alpha A + \beta A, \\ \alpha(A+B) &= \alpha A + \alpha B, \\ \alpha(\beta A) &= (\alpha\beta)A. \end{aligned} \tag{1.3}$$

If the properties (1.3) are satisfied, then $(\mathcal{F}, \mathcal{M}, +)$ forms a vector space, and $(\mathcal{R}, \mathcal{M}, +)$ a module (see Appendix A). Typical examples are $(\mathbb{R}, \mathbb{R}^{m\times n}, +)$ and $(\mathbb{C}, \mathbb{C}^{m\times n}, +)$ for the vector spaces, and $(\mathbb{R}[z], \mathbb{R}^{m\times n}[z], +)$ and $(\mathbb{C}[z], \mathbb{C}^{m\times n}[z], +)$ for the modules.

The definition (1.2) allows us to define matrix subtraction, denoted by $A - B$, as the sum of A and the matrix $(-1)B$:

$$A_{m\times n} - B_{m\times n} = A_{m\times n} + (-1)B_{m\times n} = [a_{ij} - b_{ij}]_{i,j=1}^{m,n}.$$

The multiplication of matrices A and B is defined only when their "internal" dimensions are equal:

$$A_{m\times l}B_{l\times n} := \left[\sum_k a_{ik}b_{kj}\right]_{i,j=1}^{m,n}.$$

The main motivation for this definition is that the product of A and B represents the *composition* of the corresponding linear applications.

We define the *identity matrix* I as the following square diagonal matrix:

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

It can be easily shown that the following properties hold true:

Exercise 1.3.

$$\begin{aligned} &AI = IA = A && \textit{(neutral element)} \\ &A(B+C) = AB + AC && \textit{(distributivity)} \\ &(B+C)D = BD + CD && \textit{(distributivity)} \\ &A(BC) = (AB)C && \textit{(associativity)} \end{aligned}$$

If $\mathcal{M} = \mathcal{R}^{n\times n}$ or $\mathcal{M} = \mathcal{F}^{n\times n}$, then the product of two matrices in $\mathcal{M}$ is well defined and is also an element of $\mathcal{M}$. In the case of $\mathcal{M} = \mathcal{F}^{n\times n}$, if we add the matrix product "$\cdot$" to the vector space $(\mathcal{F}, \mathcal{M}, +)$, then we obtain the structure of an *algebra* (see Appendix A).

Note that the addition of matrices is commutative. The multiplication on contrary is not, since the matrices AB and BA can even have different dimensions.

Exercise 1.4. *Verify that the matrices*

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 0 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}$$

do not satisfy $AB = BA$.

There exist many special classes of matrices with interesting and important properties. The first such class is the class of *square matrices* that we will denote in the following way:

$$A = [a_{ij}]_{i,j=1}^{n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

Exercise 1.5. *Show that the set of square matrices of dimension n forms a ring.*

We can define other matrix products as well. For example, the *Hadamard product* of two matrices A and B of equal dimensions is the matrix whose elements are the products of the corresponding elements of A and B:

$$A_{m\times n} \odot B_{m\times n} := [a_{ij} \cdot b_{ij}]_{i,j=1}^{m,n}.$$

The *Kronecker product* of two arbitrary matrices $A_{m\times n}$ and $B_{p\times q}$ is the matrix of size $mp \times nq$ whose elements are all possible products between the elements of A and B arranged in the following way:

$$A \otimes B := \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

Exercise 1.6. *Show that if AC and BD are well defined, then*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

As we will see later, these products play an important role in the study of row and column transformations of a matrix.

The property that the set of square matrices of fixed dimensions is closed under multiplication (i.e., the product of two matrices from the class belongs to the class as well) allows us to introduce the notion of matrix powers:

$$A^p := \underbrace{AA \ldots A}_{p \text{ times}}.$$

Trivially, $A^1 = A$.

Exercise 1.7. *Show that for the matrix powers the classical laws of exponentiation hold true: that is, for all nonnegative integers p, q, and letting $A^0 := I$, we have*

$$A^pA^q = A^{p+q} = A^qA^p \qquad \textit{and} \qquad (A^p)^q = A^{pq}.$$

The notion of matrix powers allows us to define the polynomial of a matrix starting from a scalar polynomial $p(\lambda) = p_0 + p_1\lambda + \cdots + p_d\lambda_d$ in the following way:

$$p(A) := p_0I + p_1A^1 + \cdots + p_dA^d.$$

We will see in Section 4.9 that for any function given by a convergent Taylor series

$$f(\lambda) = \sum_{i=0}^{\infty} f_i\lambda^i,$$

we can define the corresponding matrix function as well:

$$f(A) = \sum_{i=0}^{\infty} f_iA^i$$

provided that certain conditions on the spectrum of A are satisfied. Observe that two functions of the same matrix commute, since the powers of the same matrix commute.

Two basic subclasses of square matrices are the triangular matrices (upper and lower) and the diagonal matrices:

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \quad \text{lower triangular matrix,}$$

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & u_{nn} \end{bmatrix} \quad \text{upper triangular matrix,}$$

$$D = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_{nn} \end{bmatrix} \quad \text{diagonal matrix.}$$

For diagonal matrices, we will often use the shorter notation:

$$D = \operatorname{diag}\{d_{11}, \ldots, d_{nn}\}.$$

We will later see that the triangular and diagonal matrices play an important role in various matrix decompositions.

We can already note that each of these three classes form, for a fixed dimension n, a ring under the matrix addition and multiplication. Indeed, the sum and the matrix product preserve their structure. Furthermore, the product of diagonal matrices is actually commutative ($D_1D_2 = D_2D_1$) and therefore gives rise to a commutative ring, which is not the case for the general square or triangular matrices.

Exercise 1.8. *Consider the set of upper triangular Toeplitz matrices, i.e., the upper triangular matrices with equal elements along the diagonals ($t_{ij} = t_{i+k,j+k}$):*

$$T = \begin{bmatrix} t_1 & t_2 & \cdots & t_n \\ 0 & t_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & t_2 \\ 0 & \cdots & 0 & t_1 \end{bmatrix}.$$

Show that these matrices commute.

Exercise 1.9. *Consider the set of square, circulant matrices:*

$$C = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \\ c_n & c_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & c_2 \\ c_2 & \cdots & c_n & c_1 \end{bmatrix}.$$

Show that these matrices commute.

Exercise 1.10. *Show that a square matrix of dimension n commuting with all the other matrices of the same dimension is necessarily a "scalar" matrix, i.e., it has the form cI.*

Exercise 1.11. *Show that a square matrix of dimension n commuting with a diagonal matrix $\operatorname{diag}\{a_1, \ldots, a_n\}$, where $a_i \neq a_j$ for all $i \neq j$, is also diagonal.*

1.2 Transpose and conjugate transpose

In this section, we restrict our attention to complex (or real) matrices: $A \in \mathbb{C}^{m\times n}$ (or $A \in \mathbb{R}^{m\times n}$). The *transpose* and the *conjugate transpose* of a complex matrix A of dimensions $m \times n$ are the matrices of dimensions $n \times m$ defined respectively by

$$A^\top = [a_{ji}]_{i,j=1}^{n,m} \quad \text{(transpose)} \qquad \text{and} \qquad A^* = [\bar{a}_{ji}]_{i,j=1}^{n,m} \quad \text{(conjugate transpose)}.$$

One can easily verify the following properties:

$$\begin{aligned} (A^\top)^\top &= A & \qquad (A^*)^* &= A \\ (\alpha A)^\top &= \alpha A^\top & (\bar{\alpha} A)^* &= \alpha A^* \\ (A+B)^\top &= A^\top + B^\top & (A+B)^* &= A^* + B^* \\ (AB)^\top &= B^\top A^\top & (AB)^* &= B^* A^*. \end{aligned}$$

Relying on these operations, we can define new classes of matrices. A matrix is called *symmetric*, *antisymmetric (or skew-symmetric)*, *Hermitian*, *anti-Hermitian (or skew-Hermitian)* if it satisfies respectively

$$A = A^\top, \qquad A = -A^\top, \qquad A = A^*, \qquad A = -A^*.$$

Exercise 1.12. *Show that for every matrix $A \in \mathbb{C}^{m\times n}$, the matrices $AA^\top$ and $A^\top A$ are symmetric, and the matrices AA^* and A^*A are Hermitian.*

Exercise 1.13. *Show that for a square matrix $A \in \mathbb{C}^{n\times n}$, the matrix $A + A^\top$ is symmetric, the matrix $A + A^*$ is Hermitian, the matrix $A - A^\top$ is antisymmetric and the matrix $A - A^*$ is anti-Hermitian.*

Exercise 1.14. *Show that every complex matrix can be written as the sum of a symmetric matrix and an antisymmetric matrix, and as the sum of a Hermitian matrix and an anti-Hermitian matrix.*

A square matrix $A \in \mathbb{C}^{n\times n}$ is *unitary* if it satisfies the relations

$$AA^* = I_n = A^*A.$$

If A is a real matrix, this equation can be rewritten as

$$AA^\top = I_n = A^\top A$$

and we call A *orthogonal*.

A *normal* matrix is a square matrix satisfying

$$\begin{aligned} AA^* &= A^*A \qquad \text{if} \quad A \in \mathbb{C}^{n\times n}, \\ AA^\top &= A^\top A \qquad \text{if} \quad A \in \mathbb{R}^{n\times n}. \end{aligned}$$

All these matrices have special properties related to their eigenvalues and singular values. We will discuss them in details later.

Exercise 1.15. *Show that the Kronecker product satisfies*

$$(A \otimes B)^* = A^* \otimes B^*.$$

Exercise 1.16. *Show with the help of the previous exercise that if U_1 and U_2 are unitary matrices, then $U_1 \otimes U_2$ is unitary as well.*

1.3 Determinant of a matrix

1.3.1 Definition and elementary properties

In order to introduce the concept of determinant of a square matrix $A_{n\times n}$ defined on a ring $\mathcal{R}$ or a field $\mathcal{F}$ (i.e., $A \in \mathcal{R}^{n\times n}$ or $A \in \mathcal{F}^{n\times n}$), we have to first define its "*quasi-diagonals*" that are n-tuples of elements of matrix A:

$$a_{1j_1},\ a_{2j_2},\ \ldots\ , a_{nj_n} \tag{1.4}$$

where the indices

$$\mathbf{j} \coloneqq (j_1, j_2, \ldots, j_n)$$

constitute a permutation of the set $\{1, 2, \ldots, n\}$. Hence, we can see that a quasi-diagonal always consists of n elements of the matrix A in such a way that no two of them lie in the same row or column of A. The fact that row indices are ascending in (1.4) is actually an arbitrary choice made to state the definition.

For each quasi-diagonal, we define its *parity*, denoted by $t(\mathbf{j})$, as the number of inversions $j_k > j_p$ for $k < p$ in $\mathbf{j}$. It can be shown that the parity is equal, up to an even integer, to the number of transpositions, i.e., permutations of two elements, needed to bring $\mathbf{j}$ to the standard order. The determinant of A is finally defined as follows:

Definition 1.1

With the notation above, we define the *determinant* of a square matrix $A_{n\times n}$ as

$$\det(A) = \sum_{\mathbf{j}} (-1)^{t(\mathbf{j})}\, a_{1j_1}\cdot a_{2j_2}\cdot \ldots \cdot a_{nj_n}. \tag{1.5}$$

Observe that the parity in the expression (1.5) can be defined up to an even integer, thus, we could have defined it as the number of transpositions instead of the number of inversions. We can further remark that the determinant is multilinear with respect to the rows $\mathbf{a}_{i:}$ and with respect to the columns $\mathbf{a}_{:i}$ of the matrix A, since every term in the right-hand side of the expression (1.5) has a factor corresponding to every row and column of A. We will now present a series of properties of the function $\det(A)$. These properties exist in the "row" version and in the "column" version. Since they are similar, we will restrict ourselves to the "column" version.

Proposition 1.2: Properties of the determinant

1. $\det\left[\mathbf{a}_{:1},\ldots,\mathbf{b}_{:j}+k\mathbf{c}_{:j},\ldots,\mathbf{a}_{:n}\right] = \det\left[\mathbf{a}_{:1},\ldots,\mathbf{b}_{:j},\ldots,\mathbf{a}_{:n}\right] + k\det\left[\mathbf{a}_{:1},\ldots,\mathbf{c}_{:j},\ldots,\mathbf{a}_{:n}\right]$.

   This property simply states that the determinant is multilinear in the columns of A.

2. $\det(kA) = k^n\det(A)$.

   This property easily follows from the previous one.

3. $\det(A^\top) = \det(A)$, if $A \in \mathbb{C}^{n\times n}$.

   It is easy to see that the quasi-diagonals of A are also quasi-diagonals of $A^\top$ and vice-versa. Thus, the same terms appear in formula (1.5) for $\det(A)$ and $\det(A^\top)$. It remains to show that their parity is the same (up to a multiple of 2). This is left to the reader.

4. $\det(A^*) = \overline{\det(A)}$, if $A \in \mathbb{C}^{n\times n}$.

   Indeed, $\det(A^*) = \det(\bar{A}^\top) = \det(\bar{A})$.

5. $\det\left[\mathbf{a}_{:1},\ldots,\mathbf{a}_{:j},\ldots,\mathbf{a}_{:i},\ldots,\mathbf{a}_{:n}\right] = -\det\left[\mathbf{a}_{:1},\ldots,\mathbf{a}_{:i},\ldots,\mathbf{a}_{:j},\ldots,\mathbf{a}_{:n}\right]$.

   The *permutation of two columns* gives rise to an additional permutation for each quasi-diagonal. Therefore, the sign of the determinant changes as well.

6. $\mathbf{a}_{:i} = \mathbf{a}_{:j} \implies \det(A) = 0$.

   The matrix A does not change after the permutation of columns i and j. Thus, by the previous property we have $\det(A) = -\det(A)$, which immediately implies that $\det(A) = 0$.

7. $\mathbf{a}_{:i} = k\mathbf{a}_{:j} \implies \det(A) = 0$.

   Consequence of properties 1 and 6.

8. $\det\left[\mathbf{a}_{:1},\ldots,\mathbf{a}_{:i},\ldots,\mathbf{a}_{:j},\ldots,\mathbf{a}_{:n}\right] = \det\left[\mathbf{a}_{:1},\ldots,\mathbf{a}_{:i},\ldots,\mathbf{a}_{:j}+k\mathbf{a}_{:i},\ldots,\mathbf{a}_{:n}\right]$.

   This operation is usually called an *elementary (column) transformation*. Simply speaking, we add a multiple of column j to another column i. The proof easily follows from properties 1 and 7.

Starting with these properties, we will easily derive a series of important results that are stated as exercises.

Exercise 1.17. *Show that for arbitrary matrices $A_{m\times n}$ and $B_{n\times m}$, we have*

$$\det\begin{bmatrix} A & 0 \\ -I_n & B \end{bmatrix} = \det\begin{bmatrix} A & AB \\ -I_n & 0 \end{bmatrix}.$$

Exercise 1.18. *Using the previous exercise, show that for two arbitrary square matrices $A_{n\times n}$ and $B_{n\times n}$, it holds that*

$$\det(A)\det(B) = \det(AB).$$

Let us now introduce the notions of minor and cofactor. The *minor* $A_{(k\ell)}$ of dimension $n-1$ of a matrix $A_{n\times n}$ is the determinant of the submatrix obtained by removing the kth row and the ℓth column:

$$A_{(k\ell)} \coloneqq \det\left[\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1,\ell-1} & a_{1,\ell+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \cdots & & \vdots \\ a_{k-1,1} & \cdots & a_{k-1,\ell-1} & a_{k-1,\ell+1} & \cdots & a_{k-1,n} \\ \hline a_{k+1,1} & \cdots & a_{k+1,\ell-1} & a_{k+1,\ell+1} & \cdots & a_{k+1,n} \\ \vdots & & \vdots & \cdots & & \vdots \\ a_{n1} & \cdots & a_{n,\ell-1} & a_{n,\ell+1} & \cdots & a_{nn} \end{array}\right]. \tag{1.6}$$

Since $\det(A)$ is linear in each column j and in each row i of A, we can write $\det(A)$ as a linear combination of its elements (**Question**: What is the formal property of linear applications on vector spaces that we are using here?):

$$\det(A) = a_{1j}A^c_{1j} + a_{2j}A^c_{2j} + \cdots + a_{nj}A^c_{nj}, \tag{1.7}$$

$$\det(A) = a_{i1}A^c_{i1} + a_{i2}A^c_{i2} + \cdots + a_{in}A^c_{in}, \tag{1.8}$$

where the coefficients $A^c_{k\ell}$ are called the *cofactors* of the corresponding elements $a_{k\ell}$. In the expressions (1.7) and (1.8) of $\det(A)$, the term $a_{k\ell}A^c_{k\ell}$ consists of all the quasi-diagonals "passing" through $a_{k\ell}$ and thus the cofactor $A^c_{k\ell}$ clearly contains the quasi-diagonals of the submatrix appearing in the definition (1.6). Hence, we can expect that there is a relation between $A^c_{k\ell}$ and $A_{(k\ell)}$. In fact, we will show that:

Theorem 1.3

With the definitions above,

$$A^c_{k\ell} = (-1)^{k+\ell}A_{(k\ell)}. \tag{1.9}$$

Proof. It is enough to use $(k-1)$ permutations of adjacent rows and $(\ell-1)$ permutations of adjacent columns to bring the element $a_{k\ell}$ to position $(1,1)$ without modifying the order of other rows and columns. For this new matrix, we have

$$\hat{A}^c_{11} = (-1)^{k+\ell}A^c_{k\ell}, \qquad \hat{A}_{(11)} = A_{(k\ell)}.$$

Observe now that for the element $\hat{a}_{11}$, we have $\hat{A}^c_{11} = \hat{A}_{(11)}$, and the thesis follows. □

Equation (1.9) is a powerful tool that allows to derive many identities involving the determinant of a matrix. Some of them are given below as exercises.

Exercise 1.19. *Show that we have*

$$\det(\lambda I_n - C) = \det \begin{bmatrix} \lambda & -1 & & & 0 \\ & \lambda & \ddots & & \\ & & \ddots & \ddots & \\ 0 & & & \lambda & -1 \\ a_0 & a_1 & \cdots & a_{n-2} & \lambda + a_{n-1} \end{bmatrix}$$

$$= a_0 + a_1\lambda + \cdots + a_{n-1}\lambda^{n-1} + \lambda^n$$

where the matrix $C_{n\times n}$ above is called the companion matrix *of the polynomial.*

Exercise 1.20. *Show that the determinant of a tridiagonal matrix (sometimes referred to as a* Jacobi matrix*)*

$$J_n = \begin{bmatrix} a_1 & b_2 & & & 0 \\ c_2 & a_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & b_n \\ 0 & & & c_n & a_n \end{bmatrix}$$

satisfies the following recurrence relation:

$$\det(J_n) = a_n \det(J_{n-1}) - b_n c_n \det(J_{n-2}).$$

Exercise 1.21. *Verify that, for a* Vandermonde matrix*, the following identity holds true:*

$$\det \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{bmatrix} = \prod_{j<i} (x_i - x_j) \,.$$

1.3.2 Generalization: the Laplace and Binet–Cauchy relations

It is possible to generalize the cofactor formulas (1.7) and (1.8) by considering the minors of order smaller than $n-1$. In order to do so, we will introduce an appropriate notation. For a pair of p-tuples

$$\mathbf{i}_p := (i_1, i_2, \ldots, i_p) \quad \text{and} \quad \mathbf{j}_p := (j_1, j_2, \ldots, j_p)$$

satisfying

$$1 \le i_1 < i_2 < \cdots < i_p \le n \quad \text{and} \quad 1 \le j_1 < j_2 < \cdots < j_p \le n,$$

we define the *minors of order* p of A as

$$A \binom{\mathbf{i}_p}{\mathbf{j}_p} := \det \left[a_{i_k, j_\ell} \right]_{k,\ell=1}^{p}.$$

We define a *complementary minor of order* $(n-p)$ in the following way:

$$A\left(\begin{smallmatrix}\mathbf{i}_p^c\\ \mathbf{j}_p^c\end{smallmatrix}\right)$$

where the $(n-p)$-tuple $\mathbf{i}_p^c$ is the set complement of the p-tuple $\mathbf{i}_p$ (it means that the complementary minor is obtained by computing the determinant of the matrix after *removing* the rows $\mathbf{i}_p$ and the columns $\mathbf{j}_p$). Finally, we define the *complementary cofactors* of A as

$$A^c\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right) := (-1)^s A\left(\begin{smallmatrix}\mathbf{i}_p^c\\ \mathbf{j}_p^c\end{smallmatrix}\right)$$

where

$$s = \sum_{k=1}^{p}(i_k + j_k).$$

These definitions allow us to generalize the expansions (1.7) and (1.8) of the determinant in terms of cofactors.

Theorem 1.4: Laplace

Let A be a matrix of dimensions $n \times n$ and $\mathbf{i}_p$ be a p-tuple of rows (or a p-tuple $\mathbf{j}_p$ of columns). Then $\det(A)$ is equal to the sum of the $\binom{n}{p}$ products of all possible minors located in these rows (columns) with their complementary cofactors:

$$\det(A) = \sum_{\mathbf{j}_p} A\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right) A^c\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right), \tag{1.10}$$

$$\det(A) = \sum_{\mathbf{i}_p} A\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right) A^c\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right). \tag{1.11}$$

Proof. The proof is similar to the previous one about cofactor expansion, but requires a little bit more care. We will present only a rough sketch of the proof. First of all, it is easily seen that the sum has exactly $\binom{n}{p}$ terms, since there are exactly $\binom{n}{p}$ different minors for a given choice of p rows or columns.

Now, observe that the products of any of the $p!$ quasi-diagonals of the term $A\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right)$ with any of the $(n-p)!$ quasi-diagonals of $A^c\left(\begin{smallmatrix}\mathbf{i}_p\\ \mathbf{j}_p\end{smallmatrix}\right)$ are actually quasi-diagonals of A. Since their total number is equal to

$$\binom{n}{p}\, p!\,(n-p)! = n!,$$

we conclude that all quasi-diagonals of A are present in the sums (1.10) and (1.11). Finally, it remains to show that the parities are equal as well. It can be done in a way similar to the case of $p = 1$. □

Exercise 1.22. *Derive from the Laplace theorem the following identities:*

$$\det\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22}\end{bmatrix} = \det\begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22}\end{bmatrix} = \det(A_{11})\det(A_{22})$$

if A_{11} and A_{22} are square submatrices with dimension p and $n-p$ respectively; and

$$\det\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0\end{bmatrix} = \det\begin{bmatrix} 0 & A_{12} \\ A_{21} & A_{22}\end{bmatrix} = (-1)^{p(n+1)}\det(A_{12})\det(A_{21})$$

if A_{21} and A_{12} are square submatrices with dimension p and $n-p$ respectively.

Exercise 1.23. *Show that if A_i are square submatrices of dimension p, then it holds that*

$$\det \begin{bmatrix} \lambda I_p & -I_p & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ 0 & & & \lambda I_p & -I_p \\ A_0 & A_1 & \cdots & A_{n-2} & \lambda I_p + A_{n-1} \end{bmatrix} =$$

$$\det\Big(A_0 + A_1\lambda + \cdots + A_{n-1}\lambda^{n-1} + I_p\lambda^n\Big).$$

An important application of the Laplace theorem is the theorem of Binet–Cauchy that allows to express the determinant of a product of *non-square* matrices A and B as long as their product is a square matrix.

Theorem 1.5: Binet–Cauchy

Let $\mathbf{m}$ be the m-tuple $(1,\ldots,m)$. Let A and B be matrices of dimensions $m \times n$ and $n \times m$ respectively. If $m \le n$, then

$$\det(AB) = \sum_{\mathbf{j}_m} A\left(\begin{smallmatrix}\mathbf{m}\\ \mathbf{j}_m\end{smallmatrix}\right) B\left(\begin{smallmatrix}\mathbf{j}_m\\ \mathbf{m}\end{smallmatrix}\right).$$

Proof. Since the matrix AB is a square matrix, we can apply the results of Exercises 1.17 and 1.22 to the matrices A and B:

$$\begin{aligned} \det\begin{bmatrix} A & 0 \\ -I_n & B \end{bmatrix} &= \det\begin{bmatrix} A & AB \\ -I_n & 0 \end{bmatrix} \\ &= (-1)^n(-1)^{m(m+n+1)}\det(AB) \\ &= (-1)^{(m+n)(m+1)}\det(AB) \\ &= (-1)^r\det(AB). \end{aligned} \tag{1.12}$$

Now we can apply the theorem of Laplace to the first m rows of the matrix on the left-hand side of (1.12). The only nonzero minors for these rows are the minors of A, namely $A\left(\begin{smallmatrix}\mathbf{m}\\ \mathbf{j}_m\end{smallmatrix}\right)$, and their complement cofactors are

$$(-1)^s\det\Big[-I\left(\begin{smallmatrix}\mathbf{n}\\ \mathbf{j}_m^c\end{smallmatrix}\right)\Big|\,B\Big], \qquad s = \sum_{k=1}^m (j_k + k).$$

Now we apply the theorem of Laplace again to the last m columns of this matrix. The only nonzero terms are

$$(-1)^{s+t}B\left(\begin{smallmatrix}\mathbf{j}_m\\ \mathbf{m}\end{smallmatrix}\right), \qquad s + t = n(m+1) + 2\sum_{k=1}^m j_k\ ,$$

and therefore,

$$(-1)^r\det(AB) = \sum_{\mathbf{j}_m} (-1)^{s+t} A\left(\begin{smallmatrix}\mathbf{m}\\ \mathbf{j}_m\end{smallmatrix}\right) B\left(\begin{smallmatrix}\mathbf{j}_m\\ \mathbf{m}\end{smallmatrix}\right).$$

It remains to note that $s + t - r$ remains even and the result follows. □

1.4 Inverse and rank of a matrix

The expansion in terms of cofactors allows us to obtain another important result. The *adjugate matrix* of a square matrix $A_{n\times n}$ is defined as

$$\operatorname{adj}(A) := \left[A^c_{ji}\right]_{i,j=1}^{n} . \tag{1.13}$$

Note the inversion of indices!

Theorem 1.6

For every square matrix $A_{n\times n}$, we have

$$A \cdot \operatorname{adj}(A) = \det(A)\, I_n = \operatorname{adj}(A) \cdot A.$$

Proof. The result is a straightforward corollary of the definition (1.13), property 6 of the determinant (Proposition 1.2) and the expansion of $\det(A)$ in terms of cofactors by rows or columns. □

Definition 1.7

We say that a matrix is *non-singular* if it is a square matrix and its determinant is not equal to zero.

By Theorem 1.6, we can conclude that a non-singular matrix defined over a field $\mathcal{F}$ has an inverse, since $B := \det(A)^{-1}\operatorname{adj}(A)$ clearly satisfies

$$AB = I = BA.$$

Furthermore, given that

$$AB = I \quad \Longrightarrow \quad \det(AB) = \det(A)\det(B) = 1,$$

we can see that all matrices possessing an inverse B have a nonzero determinant. Therefore, if we are working over a field, then the class of invertible matrices coincides with the class of non-singular matrices.

Exercise 1.24. *Show that the set of square invertible matrices forms a multiplicative group.*

Exercise 1.25. *Show that the following properties hold true:*

$$\begin{aligned}
&\operatorname{adj}(A^\top) = \operatorname{adj}(A)^\top, &&\operatorname{adj}(A^*) = \operatorname{adj}(A)^*,\\
&\operatorname{adj}(I) = I, &&\operatorname{adj}(kA) = k^{n-1}\operatorname{adj}(A),\\
&(A^\top)^{-1} = (A^{-1})^\top, &&(A^*)^{-1} = (A^{-1})^*,\\
&\det(A^{-1}) = \det(A)^{-1}, &&(AB)^{-1} = B^{-1}A^{-1}.
\end{aligned}$$

We have already dealt with the elementary operations of matrices when we talked about the properties of the determinant. Let us formalize their definitions in terms of their matrix representations.

A *permutation matrix* is a matrix that, when applied on the left (resp. right) of A, permutes two rows (resp. columns) of A:

$$E^{(1)} = \left[\begin{array}{ccccc} I & & & & \\ & 0 & \cdots & 1 & \\ & \vdots & I & \vdots & \\ & 1 & \cdots & 0 & \\ & & & & I \end{array}\right] \begin{array}{l} \\ \leftarrow i \\ \\ \leftarrow j \\ \\ \end{array} .$$

A *scaling matrix* is a matrix that, when applied on the left (resp. right) of A, multiplies a row (resp. column) of A by a scalar k:

$$E^{(2)} = \left[\begin{array}{ccc} I & & \\ & k & \\ & & I \end{array}\right] \begin{array}{l} \\ \leftarrow j \\ \\ \end{array} .$$

An *elimination matrix* is a matrix that, when applied on the left (resp. right) of A, adds to a row (resp. column) of A some other row (resp. column) multiplied by k:

$$E^{(3)} = \left[\begin{array}{ccccc} I & & & & \\ & 1 & & & \\ & \vdots & I & & \\ & k & \cdots & 1 & \\ & & & & I \end{array}\right] \begin{array}{l} \\ \leftarrow i \\ \\ \leftarrow j \\ \\ \end{array} \qquad \text{(resp. its transpose).}$$

We assume that the elementary matrix operations are defined over a field $\mathcal{F}$. Thus, we can say that

$$\det(E^{(1)}) = -1, \qquad \det(E^{(2)}) = k \neq 0, \qquad \det(E^{(3)}) = 1,$$

and therefore, $E^{(1)}$, $E^{(2)}$ et $E^{(3)}$ are invertible. Furthermore, we can show that their inverses have the same form. Thus, we have at our disposal a group (in the mathematical sense of it) of operations that allow us to manipulate a matrix, and eventually to simplify its form. This is all we need to quotient the set of matrices by the equivalence relation implied by this group of transformations, and thus to exhibit our *first invariant for matrices:*

Theorem 1.8

Every matrix $A_{m\times n}$ whose elements belong to a field $\mathcal{F}$ can be brought to the following form by means of invertible (or elementary) transformations of rows and columns:

$$RAQ = \left[\begin{array}{c|c} I_r & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right] . \tag{1.14}$$

Proof. We can bring A to such form by means of the following recursive algorithm:

Step 1: Set $B := A$ and $i := 1$.

Step 2: If $B = 0$, then stop; otherwise, we permute the rows and columns of B in order to place a nonzero element in position $(1, 1)$.

Step 3: We perform a scaling transformation of rows or columns to make $b_{1,1}$ equal to 1.

Step 4: By means of column elimination transformations, we make the elements to the right of $b_{1,1}$ equal to zero; and by means of row elimination transformations $b_{1,1}$, we make the elements below $b_{1,1}$ equal to zero.

Step 5: If $i = \min(m, n)$, we stop; otherwise we set $B \coloneqq B(2 : m - i, 2 : n - i)$ and $i \coloneqq i + 1$, and go to step 2.

□

This theorem allows us to define an equivalence relation (i.e., a reflexive, symmetric and transitive relation) on the set of matrices over $\mathcal{F}$ with dimensions $m \times n$:

$$A \sim B \qquad \text{iff} \qquad RAQ = B \tag{1.15}$$

where R and Q are products of elementary transformations.

If two matrices belong to the same equivalence class (i.e., $A \sim B$), then by means of the reduction of Theorem 1.8, they can be brought to the same form (1.14), that can be seen as "the simplest" form of this equivalence class. We will also say that (1.14) is the *canonical form of $m \times n$ matrices under elementary transformations* of rows and columns. This form is completely characterized by the integer r, that we will call the *rank* of the matrix.

Theorem 1.9

The rank of a matrix $A_{m\times n}$ whose elements belong to a field $\mathcal{F}$ is equal to the largest size of its nonzero minors.

Proof. Observe that the elementary transformations do not change the size of the largest nonzero minor. Thus, it is enough to consider the canonical form of A, for which the statement of the theorem holds trivially. □

A square non-singular matrix $A_{n\times n}$ has a nonzero determinant, and therefore its rank is equal to n (the determinant is the minor of size n). Thus, from Theorem 1.8, we obtain the following:

Corollary 1.10

Any non-singular matrix whose elements belong to a field $\mathcal{F}$ can be written as a product of elementary transformations.

That is, the equivalence defined in (1.15) can be seen as equivalence under (left and right) *invertible transformations.*

Exercise 1.26. *The LDU decomposition of a matrix A (whose rows and columns can be permuted) is given by:*

$$P_1AP_2 = LDU,$$

where P_1 and P_2 are permutation matrices, L and U are triangular (lower and upper respectively) matrices with ones on the diagonal and D is a diagonal matrix. Such a decomposition is the result of Gaussian elimination with complete pivoting. Compare it to the canonical form (1.14)*.*

Theorem 1.11: Schur complement

Let $A_{n\times n}$ be an invertible submatrix of the matrix (whose elements belong to a field $\mathcal{F}$)

$$M_{(n+p)\times(n+m)} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Then the rank of M satisfies

$$\operatorname{rank}(M) = n + \operatorname{rank}\left(D - CA^{-1}B\right).$$

The matrix $D - CA^{-1}B$ is called the *Schur complement* of M.

Proof. By means of elementary (or invertible) transformations, we obtain

$$\begin{bmatrix} I_n & 0 \\ -CA^{-1} & I_p \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_n & -A^{-1}B \\ 0 & I_m \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix}.$$

Since A is invertible, the result of the theorem follows immediately. □

Chapter 2

Linear applications, orthogonalization and the QR factorization

2.1 Vector spaces

In this chapter, we will first study some basic properties of linear subspaces, and then apply them to the resolution of systems of linear equations.

Remember that a *vector space* (or *linear space*) $\mathcal{V}$ is a set $(\mathcal{V}, +)$ of *vectors* that forms a commutative group under vector addition and that is equipped with a product satisfying certain special properties and allowing the vectors to be multiplied by the elements of a field of *scalars* $(\mathcal{F}, +, \cdot)$ (see Appendix A). Classical examples of vector spaces are the sets $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$, where the underlying fields are $\mathbb{R}$ or $\mathbb{C}$. Other examples are the set of triangular matrices of fixed dimension or the set of vectors with rational entries.

In the sequel, we will assume that the set of vector components coincides with the field of scalars $\mathcal{F}$, i.e., $\mathcal{V} = \mathcal{F}^n$ (actually, this can be done without loss of generality in the case of finite-dimensional spaces). We remark that, in this setting, the set of vectors on the ring of polynomials, that is, $\mathcal{V} = \mathbb{R}^n[x]$, do not constitute a vector space, but a module (see Appendix A).

A *linear subspace* $\mathcal{S} \subset \mathcal{V}$ is a subset of $\mathcal{V}$ closed under linear combinations:

$$\alpha, \beta \in \mathcal{F}, \quad \mathbf{a}, \mathbf{b} \in \mathcal{S} \qquad \Longrightarrow \qquad \alpha\mathbf{a} + \beta\mathbf{b} \in \mathcal{S}.$$

We can easily show that the *intersection* of two linear subspaces

$$\mathcal{S}_1 \cap \mathcal{S}_2 := \{\mathbf{x} \mid \mathbf{x} \in \mathcal{S}_1,\ \mathbf{x} \in \mathcal{S}_2\}$$

is a linear subspace. We can also observe that the *union* of two linear subspaces

$$\mathcal{S}_1 \cup \mathcal{S}_2 := \{\mathbf{x} \mid \mathbf{x} \in \mathcal{S}_1 \ \text{ or } \ \mathbf{x} \in \mathcal{S}_2\}$$

is not necessarily a subspace. On the other hand, the *sum* of two subspaces, defined as

$$\mathcal{S}_1 + \mathcal{S}_2 := \{\mathbf{x} + \mathbf{y} \mid \mathbf{x} \in \mathcal{S}_1,\ \mathbf{y} \in \mathcal{S}_2\},$$

is a linear subspace.

Given a matrix $A \in \mathcal{F}^{m\times n}$, we define the following two linear subspaces:

$$\operatorname{Ker}(A) := \{\mathbf{x} \mid A\mathbf{x} = 0,\ \mathbf{x} \in \mathcal{V}\},$$

$$\operatorname{Im}(A) := \{\mathbf{y} \mid \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathcal{V}\},$$

called the *kernel* of A and the *image* of A respectively. The following two subspaces are linear subspaces as well:

- the *image of* $\mathcal{S}$ under the application of A:

$$A\mathcal{S} \coloneqq \{\mathbf{y} \mid \mathbf{y} = A\mathbf{x},\ \mathbf{x} \in \mathcal{S}\},$$

- the *space generated* by the vectors $\mathbf{a}_i$ $(i = 1, \dots, k)$.

$$\operatorname{span}\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k\} \coloneqq \left\{ \mathbf{y} \,\middle|\, \mathbf{y} = \sum_{i=1}^{k} \alpha_i \mathbf{a}_i,\ \alpha_i \in \mathcal{F} \right\}.$$

Exercise 2.1. *Verify that* $\mathcal{S}_1 \cap \mathcal{S}_2$, $\mathcal{S}_1 + \mathcal{S}_2$, $\operatorname{Ker}(A)$, $\operatorname{Im}(A)$ *and* $A\mathcal{S}$ *are linear subspaces.*

Exercise 2.2. *Show that for all matrices* R *(with suitable dimensions), we have*

$$\operatorname{Ker}(RA) \supseteq \operatorname{Ker}(A), \quad \operatorname{Im}(AR) \subseteq \operatorname{Im}(A).$$

Intuitively, there seems to be a link between the subspaces $\operatorname{Ker}(A)$, $\operatorname{Im}(A)$ and the solutions of systems of linear equations. This link is made explicit in the following theorem.

Theorem 2.1

If $A \in \mathcal{F}^{n\times n}$ is invertible, then the system

$$A\mathbf{x} = \mathbf{y} \tag{2.1}$$

has a unique solution $\mathbf{x}$ for every vector $\mathbf{y} \in \mathcal{F}^n$, and

$$\operatorname{Ker}(A) = \{0\}, \quad \operatorname{Im}(A) = \mathcal{F}^n.$$

Proof. Since the matrix A is invertible, then a vector $\mathbf{x} = A^{-1}\mathbf{y}$ is defined for all $\mathbf{y} \in \mathcal{F}^n$. Thus, every vector $\mathbf{y}$ of $\mathcal{F}^n$ has a representation in the form of (2.1) and $\operatorname{Im}(A) = \mathcal{F}^n$.

Suppose now that there exist two solutions $\mathbf{x}_1$ et $\mathbf{x}_2$ for the same right-hand side $\mathbf{y}$. This implies that $A(\mathbf{x}_1 - \mathbf{x}_2) = 0$ and $\mathbf{x}_1 - \mathbf{x}_2 = A^{-1}0 = 0$, i.e., $\mathbf{x}_1 = \mathbf{x}_2$. Therefore, the solution is unique.

Finally, the equation $A^{-1}0 = 0$ also implies that the only element in $\operatorname{Ker}(A)$ is 0. □

This theorem deals only with the case of square matrices. In order to discuss the general case of matrices $A_{m\times n}$ of arbitrary ranks and dimensions, we need some additional basic concepts that we will briefly describe here.

We say that k vectors from $\mathcal{F}^n$ are *linearly independent* if the only linear combination equal to the zero vector is trivial:

$$\sum_{i=1}^{k} \alpha_i \mathbf{a}_i = 0, \quad \alpha_i \in \mathcal{F} \qquad \Longrightarrow \qquad \alpha_i = 0.$$

These vectors form a *basis* of the space $\mathcal{S} = \operatorname{span}\{\mathbf{a}_1, \dots, \mathbf{a}_k\}$.

Exercise 2.3. *Show that all vectors* $\mathbf{x} \in \mathcal{S} = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ *have a unique representation*

$$\mathbf{x} = \sum_{i=1}^{k} \alpha_i \mathbf{a}_i$$

if $\{\mathbf{a}_i \mid i = 1, \ldots, k\}$ *is a basis of* $\mathcal{S}$*.*

Exercise 2.4. *Show that two bases of a same space* $\mathcal{S}$ *have the same number of elements.*

Exercise 2.4 allows us to define the notion of *dimension* of a vector space $\mathcal{V}$, denoted by $\dim(\mathcal{V})$, as the number of elements in any of its bases.

Lemma 2.2

If $\mathcal{S}_k \subseteq \mathcal{S}_\ell$ are subspaces of dimensions k and ℓ, where $k < \ell$, then every basis $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ of $\mathcal{S}_k$ can be extended to a basis $\{\mathbf{a}_1, \ldots, \mathbf{a}_\ell\}$ of $\mathcal{S}_\ell$.

Proof. Since $k < \ell$, there exists a vector $\mathbf{a}_{k+1} \in \mathcal{S}_\ell$ that does not belong to $\mathcal{S}_k = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$. Thus,

$$\mathcal{S}_{k+1} \coloneqq \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_{k+1}\}$$

is a vector space of dimension $k+1$ generated by these $k+1$ linearly independent vectors. Furthermore,

$$\mathcal{S}_k \subseteq \mathcal{S}_{k+1} \subseteq \mathcal{S}_\ell.$$

If $k+1 < \ell$, then we repeat the same reasoning inductively until we obtain a basis of size ℓ. □

Theorem 2.3

Let $R \in \mathcal{F}^{n\times n}$ be invertible and $\mathcal{S} \subseteq \mathcal{F}^n$ be a linear subspace. It holds that

$$\begin{aligned}
&\dim(R\mathcal{S}) = \dim(\mathcal{S}), && \text{(a)}\\
&\operatorname{Ker}(RA) = \operatorname{Ker}(A), && \text{(b)}\\
&\operatorname{Im}(AR^{-1}) = \operatorname{Im}(A), && \text{(c)}\\
&\operatorname{Im}(RA) = R\operatorname{Im}(A), && \text{(d)}\\
&\operatorname{Ker}(AR^{-1}) = R\operatorname{Ker}(A). && \text{(e)}
\end{aligned}$$

Proof. (a) Let us consider a basis $\{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$ of $\mathcal{S}$. Since R is invertible, $\{R\mathbf{s}_1, \ldots, R\mathbf{s}_k\}$ is a basis of $R\mathcal{S}$. Indeed, the linear combination

$$\sum_{i=1}^{k} \alpha_i R\mathbf{s}_i = R\sum_{i=1}^{k} \alpha_i \mathbf{s}_i$$

is equal to the zero vector only in the case of the trivial solution $\alpha_i = 0$.

(b)–(e) The fact that R is invertible implies the following statements:

$$\begin{array}{rcl} RA\mathbf{x} = 0 & \iff & A\mathbf{x} = 0, \\ \mathbf{y} = AR^{-1}\mathbf{x} & \iff & \mathbf{y} = A\mathbf{z}\ , \quad \mathbf{x} = R\mathbf{z}, \\ \mathbf{y} = RA\mathbf{x} & \iff & \mathbf{y} = R\mathbf{z}\ , \quad \mathbf{z} = A\mathbf{x}, \\ AR^{-1}\mathbf{x} = 0 & \iff & A\mathbf{z} = 0\ , \quad \mathbf{x} = R\mathbf{z}. \end{array}$$

Each of them leads to the desired identity. □

The above theorem allows us to make a link between the canonical form (1.14) of a matrix A and certain bases of $\mathrm{Ker}(A)$ and $\mathrm{Im}(A)$.

Theorem 2.4

If

$$A_{m\times n} = R \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^{-1}$$

for some invertible matrices $R \in \mathcal{F}^{m\times m}$ and $Q \in \mathcal{F}^{n\times n}$, then

$$\mathrm{Im}(A) = \mathrm{span}\,\{\mathbf{r}_{:1}, \ldots, \mathbf{r}_{:r}\}, \quad \mathrm{Ker}(A) = \mathrm{span}\,\{\mathbf{q}_{:,r+1}, \ldots, \mathbf{q}_{:n}\}.$$

Proof. By the previous theorem, we have

$$\mathrm{Im}(A) = \mathrm{Im}\left(R \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\right) = R\,\mathrm{Im} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix},$$
$$\mathrm{Ker}(A) = Q\,\mathrm{Ker}\left(R \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\right) = Q\,\mathrm{Ker} \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Simplifying further, we derive

$$\mathrm{Im}(A) = R\,\mathrm{Im} \begin{bmatrix} I_r \\ 0 \end{bmatrix} = \mathrm{span}\,\{\mathbf{r}_{:1}, \ldots, \mathbf{r}_{:r}\},$$
$$\mathrm{Ker}(A) = Q\,\mathrm{Im} \begin{bmatrix} 0 \\ I_{n-r} \end{bmatrix} = \mathrm{span}\,\{\mathbf{q}_{:r+1}, \ldots, \mathbf{q}_{:n}\}.$$

□

Corollary 2.5

The rank of a matrix $A \in \mathcal{F}^{m\times n}$ is equal to the dimension of its image $\mathrm{Im}(A)$.

2.2 Euclidean and unitary spaces

In this section, we endow our vector space with a simple but yet extremely powerful tool: an *inner product* (or *dot product*). This allows us to define a metric on the vector space, and later will allow us to use the notion of orthogonality. It turns out that linear applications go along very well with these notions: one can very well understand (theoretically and numerically) the way matrices interact with a Euclidean (unitary) metric, and with the notion of orthogonality. This will lead us to the QR decomposition and, later, to the Singular Value Decomposition.

Before introducing the concept of orthogonality, we define the inner product as a map $\langle\cdot,\cdot\rangle : \mathcal{V}\times\mathcal{V}\to\mathcal{F}$ satisfying the following properties:

- $\langle\mathbf{x},\mathbf{x}\rangle \geq 0 \quad \forall\mathbf{x}\in\mathcal{V}$;
- $\langle\mathbf{x},\mathbf{x}\rangle = 0 \;\Leftrightarrow\; \mathbf{x}=0$;
- $\langle\alpha\mathbf{x}+\beta\mathbf{y},\mathbf{z}\rangle = \alpha\langle\mathbf{x},\mathbf{z}\rangle+\beta\langle\mathbf{y},\mathbf{z}\rangle \quad \forall\mathbf{x},\mathbf{y},\mathbf{z}\in\mathcal{V}$;
- $\langle\mathbf{x},\mathbf{y}\rangle = \overline{\langle\mathbf{y},\mathbf{x}\rangle} \quad \forall\mathbf{x},\mathbf{y}\in\mathcal{V}$.

Note that this requires that the field of scalars is the set of real or complex numbers (for the third and fourth property to be well defined). *Hence, in the rest of these notes (except in Chapter 6, where polynomial matrices are studied), we will restrict our attention to matrices defined over the field of real or complex numbers: i.e., $\mathcal{F}=\mathbb{R}$ or $\mathcal{F}=\mathbb{C}$, and $\mathcal{V}=\mathbb{R}^n$ or $\mathcal{V}=\mathbb{C}^n$.*

A vector space $(\mathcal{V},\mathcal{F})$ equipped with an inner product $\langle\cdot,\cdot\rangle : \mathcal{V}\times\mathcal{V}\to\mathcal{F}$ is called *Euclidean* if $\mathcal{F}=\mathbb{R}$, or *unitary* if $\mathcal{F}=\mathbb{C}$.

For example, observe that the following commonly used products are inner products:

$$\langle\mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{n} x_i\overline{y_i} = \mathbf{y}^*\mathbf{x}, \quad \mathbf{x},\mathbf{y}\in\mathbb{C}^n,$$
$$\langle\mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{y}^\top\mathbf{x}, \quad \mathbf{x},\mathbf{y}\in\mathbb{R}^n.$$

We can associate a vector norm $\|\cdot\| : \mathcal{V}\to\mathbb{R}$ with every inner product in the following way:

$$\|\mathbf{x}\|^2 \coloneqq \langle\mathbf{x},\mathbf{x}\rangle.$$

The following theorem states an important connection between the inner products and their associated norm.

Theorem 2.6: Schwarz inequality

For any vectors $\mathbf{x},\mathbf{y}\in\mathcal{V}$, if $\langle\cdot,\cdot\rangle$ is an inner product, then one has

$$|\langle\mathbf{x},\mathbf{y}\rangle| \leq \|\mathbf{x}\|\,\|\mathbf{y}\|.$$

Proof. Let us define the scalars $\beta\coloneqq\langle\mathbf{x},\mathbf{x}\rangle$, $\alpha\coloneqq-\langle\mathbf{y},\mathbf{x}\rangle$ and the vector $\mathbf{z}\coloneqq\alpha\mathbf{x}+\beta\mathbf{y}$. Since the norm of $\mathbf{z}$ is clearly nonnegative, we have

$$0\leq\|\mathbf{z}\|^2 = |\alpha|^2\langle\mathbf{x},\mathbf{x}\rangle+\alpha\overline{\beta}\langle\mathbf{x},\mathbf{y}\rangle+\overline{\alpha}\beta\langle\mathbf{y},\mathbf{x}\rangle+|\beta|^2\langle\mathbf{y},\mathbf{y}\rangle.$$

Now, observe that, by definition, β is real and nonnegative, and thus

$$\begin{aligned} 0 &\leq \beta|\alpha|^2+\alpha\beta(-\overline{\alpha})+\overline{\alpha}\beta(-\alpha)+\beta^2\|\mathbf{y}\|^2 \\ &= -\beta|\alpha|^2+\beta^2\|\mathbf{y}\|^2 \\ &= \beta\Big(\|\mathbf{x}\|^2\|\mathbf{y}\|^2-|\langle\mathbf{x},\mathbf{y}\rangle|^2\Big). \end{aligned}$$

If $\beta=0$, then the statement trivially holds; otherwise, we derive the required inequality after the division by β. □

Exercise 2.5. *Using Schwarz's inequality, show that*

$$|\mathrm{trace}(Y^* X)| \leq \|X\|_F \|Y\|_F$$

where $X, Y \in \mathbb{C}^{m \times n}$ and $\|\cdot\|_F$ is the Frobenius norm (see Appendix B).

The Schwarz inequality allows us to define the angle θ between two vectors $\mathbf{x}$ and $\mathbf{y}$ as follows:

$$\theta := \mathrm{angle}(\mathbf{x}, \mathbf{y}) \qquad \text{if} \qquad \cos(\theta) = \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\|\mathbf{x}\| \, \|\mathbf{y}\|}, \quad 0 \leq \theta \leq \frac{\pi}{2}.$$

We say that two vectors are *orthogonal* if the angle between them is $\pi/2$, or equivalently, if their inner product is equal to zero:

$$\mathbf{x} \perp \mathbf{y} \qquad \Longleftrightarrow \qquad \langle \mathbf{x}, \mathbf{y} \rangle = 0 \qquad \Longleftrightarrow \qquad \mathrm{angle}(\mathbf{x}, \mathbf{y}) = \frac{\pi}{2}.$$

2.3 Orthogonalization and the QR decomposition

Given two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, we can perform their *orthogonalization*: indeed, define the vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ as follows:

$$\begin{aligned} \mathbf{y}_1 &:= \mathbf{x}_1, \\ \mathbf{y}_2 &:= \mathbf{x}_2 - \frac{\langle \mathbf{y}_1, \mathbf{x}_2 \rangle}{\langle \mathbf{y}_1, \mathbf{y}_1 \rangle} \mathbf{y}_1 = \mathbf{x}_2 - \alpha \mathbf{x}_1, \end{aligned} \tag{2.2}$$

then it is clear that

$$\mathbf{y}_1 \perp \mathbf{y}_2, \quad \mathrm{span}\{\mathbf{x}_1, \mathbf{x}_2\} = \mathrm{span}\,\{\mathbf{y}_1, \mathbf{y}_2\}.$$

If the vectors $\{\mathbf{x}_i\}$ form a basis, i.e., they are linearly independent, then $\{\mathbf{y}_i\}$ is a basis of the same subspace as well. This procedure can be generalized to any basis of an arbitrary r-dimensional subspace of $\mathcal{V}$ and is typically referred to as the *Gram–Schmidt orthogonalization* process.

Theorem 2.7

If $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ is a basis of a subspace $\mathcal{S} \subseteq \mathcal{V}$, then the vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_r\}$ defined by the following recurrence:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{x}_1, \\ \mathbf{y}_p &= \mathbf{x}_p - \sum_{j=1}^{p-1} \frac{\langle \mathbf{y}_j, \mathbf{x}_p \rangle}{\langle \mathbf{y}_j, \mathbf{y}_j \rangle} \mathbf{y}_j, \quad p = 2, \ldots, r. \end{aligned} \tag{2.3}$$

form an orthogonal basis of $\mathcal{S}$.

Proof. The proof is based on recursive application of (2.2). The details are left to the reader. □

The vectors $\{\mathbf{y}_i\}$ in an orthogonal basis can be arranged as the columns of a matrix $Y_{n \times r}$. The orthogonality of the basis implies that

$$Y^* Y = D = \mathrm{diag}\,\{n_1^2, \ldots, n_r^2\} \quad \text{where} \quad n_i^2 = \|\mathbf{y}_i\|^2 \neq 0$$

since a basis cannot contain the zero vector.

If we divide each vector $\mathbf{y}_i$ by its norm n_i, then we obtain an *orthonormal* basis:

$$\mathbf{u}_i = \frac{\mathbf{y}_i}{\|\mathbf{y}_i\|}, \qquad \langle \mathbf{y}_i, \mathbf{y}_j \rangle = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \quad \text{if} \quad i \neq j.$$

The matrix U with columns given by the vectors $\mathbf{u}_i$ satisfies

$$U^* U = I_r, \tag{2.4}$$

and therefore is an *isometry*. Conversely, we can also note that the columns of every isometry (2.4) form an orthonormal basis.

The concepts of orthogonal basis $\{\mathbf{y}_i\}$ and orthonormal basis $\{\mathbf{u}_i\}$ gain much of their importance from the simplicity of the representation of a vector $\mathbf{x} \in \operatorname{colspan}(Y) = \operatorname{Im}(Y) = \operatorname{colspan}(U) = \operatorname{Im}(U)$. In fact,

$$\mathbf{x} = \sum_{i=1}^{r} \frac{\langle \mathbf{x}, \mathbf{y}_i \rangle}{\langle \mathbf{y}_i, \mathbf{y}_i \rangle} \mathbf{y}_i = \sum_{i=1}^{r} c_i \mathbf{y}_i, \tag{2.5}$$

thus it requires only the computation of inner products. This representation easily follows from the expression $Y^* Y = D$:

$$D\mathbf{c} = Y^* Y \mathbf{c} \qquad \Longleftrightarrow \qquad D^{-1} Y^* \mathbf{x} = \mathbf{c} \qquad \Longleftrightarrow \qquad c_i = \frac{\langle \mathbf{x}, \mathbf{y}_i \rangle}{\langle \mathbf{y}_i, \mathbf{y}_i \rangle}.$$

For an orthonormal basis, (2.5) can be simplified even further:

$$\mathbf{x} = \sum_{i=1}^{r} \langle \mathbf{x}, \mathbf{u}_i \rangle \mathbf{u}_i.$$

The matrix representation of the Gram–Schmidt orthogonalization leads us to the QR factorization of a matrix $A_{n\times r}$ of rank r.

Theorem 2.8: QR factorization

Every matrix $A \in \mathbb{C}^{n\times r}$ of full column-rank admits a factorization

$$A = QR \tag{2.6}$$

where $Q \in \mathbb{C}^{n\times r}$ is an isometry (i.e., $Q^* Q = I_r$) and $R \in \mathbb{C}^{r\times r}$ is an upper triangular matrix with positive diagonal.

Proof. We will treat the columns of A as the vectors $\mathbf{x}_j$. These vectors form a basis of $\operatorname{Im}(A)$, since the rank of A is equal to r. The Gram–Schmidt procedure (2.3) allows us to construct the columns $\mathbf{y}_j$ of a matrix Y such that

$$A = YC$$

where

$$\begin{aligned} c_{p,p} &= 1, \\ c_{j,p} &= \frac{\langle \mathbf{x}_p, \mathbf{y}_j \rangle}{\langle \mathbf{y}_j, \mathbf{y}_j \rangle} \quad \forall j < p, \\ c_{j,p} &= 0 \quad \forall j > p. \end{aligned}$$

The matrix C and its inverse are upper triangular matrices with only ones on the main diagonal. By introducing the following positive diagonal matrix:

$$N = \operatorname{diag}\{n_1, \dots, n_r\}, \quad n_i = \|\mathbf{y}_i\|,$$

we have

$$(YN^{-1})(NC) = A = QR$$

where Q is an isometry:

$$Q^*Q = N^{-1}Y^*YN^{-1} = N^{-1}N^2N^{-1} = I_r,$$

and $R = NC$ is an upper triangular matrix with a positive diagonal. □

Exercise 2.6. *Show that R is the* Cholesky factor *of the positive definite matrix A^*A.*

We can extend the concept of orthogonality to subspaces as well.

Definition 2.9

The subspaces $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{V}$ are orthogonal if each vector in $\mathcal{S}_1$ is orthogonal to every vector in $\mathcal{S}_2$:

$$\mathcal{S}_1 \perp \mathcal{S}_2 \qquad \text{if} \qquad \langle \mathbf{x}, \mathbf{y} \rangle = 0 \quad \forall \mathbf{x} \in \mathcal{S}_1,\ \forall \mathbf{y} \in \mathcal{S}_2.$$

Exercise 2.7. *Let X and Y be two matrices whose columns form bases of the subspaces $\mathcal{X}$ and $\mathcal{Y}$ respectively. Show that $\mathcal{X} \perp \mathcal{Y}$ if and only if $Y^*X = 0$.*

We have already seen how to construct an orthogonal basis of an r-dimensional subspace $\mathcal{S}_1 \subset \mathcal{V}$ with the Gram–Schmidt procedure. It is also easy to see that this basis can be completed to an orthogonal basis of the whole space $\mathcal{V}$:

$$\operatorname{Im}(S_1) = \mathcal{S}_1, \quad S_1^*S_1 = D_1 \quad \Longrightarrow \quad \exists S_2 \ \text{ s.t. } \ S_2^*S_1 = 0, \quad \operatorname{Im}[\,S_1 \,|\, S_2\,] = \mathcal{V}. \tag{2.7}$$

This allows us to define the *orthogonal complement* of a subspace $\mathcal{S}$:

$$\mathcal{S}^\perp := \{\mathbf{x} \mid \langle \mathbf{x}, \mathbf{y} \rangle = 0 \ \ \forall \mathbf{y} \in \mathcal{S}\}.$$

Corollary 2.10

If $\mathcal{S}$ is a subspace of dimension r, then $\mathcal{S}^\perp$ is a subspace of dimension $n - r$.

Proof. Straightforward from (2.7): if the columns of $[\,S_1 \,|\, S_2\,]$ form an orthogonal basis of $\mathcal{V}$, then it is clear that $\operatorname{Im}(S_2) = \mathcal{S}^\perp$ and $\dim(\operatorname{Im}(S_2)) = \operatorname{rank}(S_2) = n - r$. □

Exercise 2.8. *Show that the orthogonal complement satisfies the following properties:*

$$\begin{aligned}
&(\mathcal{S}^\perp)^\perp = \mathcal{S}, \\
&(\mathcal{S}_1 + \mathcal{S}_2)^\perp = \mathcal{S}_1^\perp \cap \mathcal{S}_2^\perp, \\
&(\mathcal{S}_1 \cap \mathcal{S}_2)^\perp = \mathcal{S}_1^\perp + \mathcal{S}_2^\perp.
\end{aligned}$$

If we choose the bases (given by the columns of S_1 and S_2 in (2.7)) to be orthonormal, then we immediately obtain the following lemma:

Corollary 2.11

Let $U_1 \in \mathbb{C}^{n\times r}$ be an isometry (i.e., $U_1^* U_1 = I_r$), then there is always a matrix $U_2 \in \mathbb{C}^{n\times(n-r)}$ such that $U = [\, U_1 \,|\, U_2 \,]$ is a unitary matrix.

2.4 Computational aspects

In this section, we will present algorithms that are used to construct the orthogonal matrix appearing in the QR factorization. In fact, the transformations that we will describe are the basis of all modern algorithms performing orthogonal decompositions of a matrix. Although these transformations are well defined for $\mathcal{F} = \mathbb{C}$, we will restrict ourselves to the case $\mathcal{F} = \mathbb{R}$ to simplify the notation.

2.4.1 Givens transformations

Although these transformations can be applied in vector spaces of arbitrary dimension $n \geq 2$, they are initially defined in $\mathbb{R}^2$. For every vector $\mathbf{x} \in \mathbb{R}^2$, there exists a transformation $G \in \mathbb{R}^{2\times 2}$ such that $GG^\top = G^\top G = I_2$ and

$$G\mathbf{x} = \begin{bmatrix} \|\mathbf{x}\|_2 \\ 0 \end{bmatrix}.$$

In fact, this transformation is a *rotation* in $\mathbb{R}^2$:

$$G = \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

with

$$c = \cos(\theta), \qquad s = \sin(\theta) \qquad\qquad \text{for some } \theta \in \mathbb{R}$$

or

$$c = x_1/\|\mathbf{x}\|_2, \qquad s = x_2/\|\mathbf{x}\|_2 \qquad\qquad \text{for some } \mathbf{x} \in \mathbb{R}^2.$$

It is easy to see that a Givens transformation corresponds to a rotation with angle θ (see Figure 2.1).

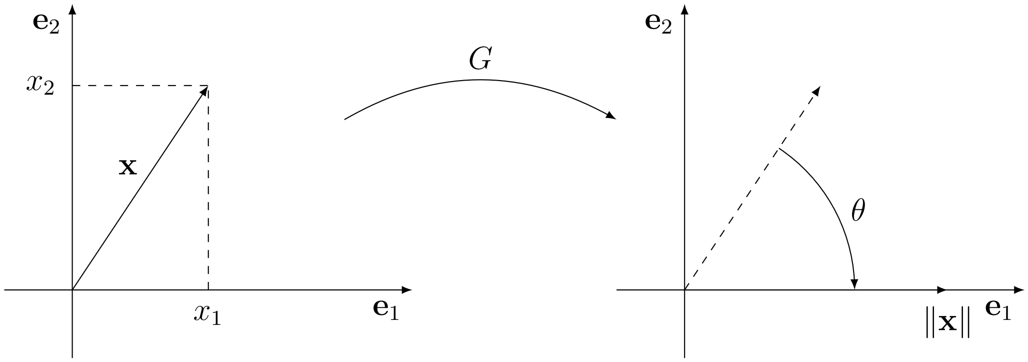

Figure 2.1: Geometrical interpretation of the Givens transformation.

Exercise 2.9. *How to construct the complex Givens transformation $G \in \mathbb{C}^{2\times 2}$ that transforms an arbitrary vector $\mathbf{x} \in \mathbb{C}^2$ into $\begin{bmatrix} \|\mathbf{x}\| \\ 0 \end{bmatrix}$?*

We apply the Givens rotations in $\mathbb{R}^n$ simply by applying it to the coordinates i and j:

$$G_{i,j} = \begin{bmatrix} I_{i-1} & & & & \\ & c & & s & \\ & & I_{j-i-1} & & \\ & -s & & c & \\ & & & & I_{n-j} \end{bmatrix} \quad \in \quad \mathbb{R}^{n\times n}.$$

It follows that Givens transformations are very cheap to apply, since they require only $O(1)$ operations.

2.4.2 Householder transformations

Another classical orthogonal transformation is the Householder transformation $H \in \mathbb{R}^{n\times n}$, which acts on a given vector $\mathbf{x} \in \mathbb{R}^n$ and satisfies

$$H^\top H = HH^\top = I_n, \qquad H\mathbf{x} = \begin{bmatrix} \pm\|\mathbf{x}\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We can construct this transformation as follows:

$$H = I_n - 2\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}} \qquad \text{where} \quad \mathbf{v} = \mathbf{x} \mp \|\mathbf{x}\|_2\mathbf{e}_1.$$

Observe that by construction H is symmetric. It is orthogonal as well, since

$$HH^\top = H^2 = I_n - 4\frac{\mathbf{v}\mathbf{v}^\top}{\mathbf{v}^\top\mathbf{v}} + 4\frac{\mathbf{v}(\mathbf{v}^\top\mathbf{v})\mathbf{v}^\top}{(\mathbf{v}^\top\mathbf{v})^2} = I_n.$$

We note that the above is valid for any $\mathbf{v} \neq 0$ and that H does not change if $\mathbf{v}$ is scaled with a nonzero constant α. In order to choose $\mathbf{v}$, we impose that

$$H\mathbf{x} = \mathbf{x} - 2\mathbf{v}\left(\frac{\mathbf{v}^\top\mathbf{x}}{\mathbf{v}^\top\mathbf{v}}\right) = \pm\mathbf{e}_1\|\mathbf{x}\|_2,$$

If we scale $\mathbf{v}$ such that $2\mathbf{v}^\top\mathbf{x} = \mathbf{v}^\top\mathbf{v}$, we get the desired form for $\mathbf{v}$.

Geometrically, a Householder transformation can be seen as a reflection. Every vector $\mathbf{x}$ can be written as a vector $\mathbf{x}_1$ parallel to $\mathbf{v}$ and a vector $\mathbf{x}_2$ orthogonal to $\mathbf{v}$:

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2 \qquad \text{with} \qquad \mathbf{x}_1 \in \operatorname{span}\{\mathbf{v}\}, \quad \mathbf{x}_2 \in \operatorname{span}\{\mathbf{v}\}^\perp.$$

From the definition of H, we have

$$H\mathbf{x} = -\mathbf{x}_1 + \mathbf{x}_2,$$

in other words, the transformation H reflects the vector $\mathbf{x}$ about the hyperplane defined by $(\operatorname{span}\{\mathbf{v}\})^\perp$. The effect of a Householder transformation in $\mathbb{R}^3$ is illustrated in Figure 2.2.

Note that the sign of $\pm\|\mathbf{x}\|_2$ is not fixed. The sign determines whether the reflected vector $H\mathbf{x}$ is going to be in the direction of $\mathbf{e}_1$ or in the opposite direction. For numerical reasons, the sign of x_1, the first component of $\mathbf{x}$, is often chosen:

$$\mathbf{v} = \mathbf{x} + \operatorname{sign}(x_1)\|\mathbf{x}\|_2\mathbf{e}_1.$$

In fact, it guarantees the smallest possible rounding errors (see, e.g., [Wilkinson, 1965]).

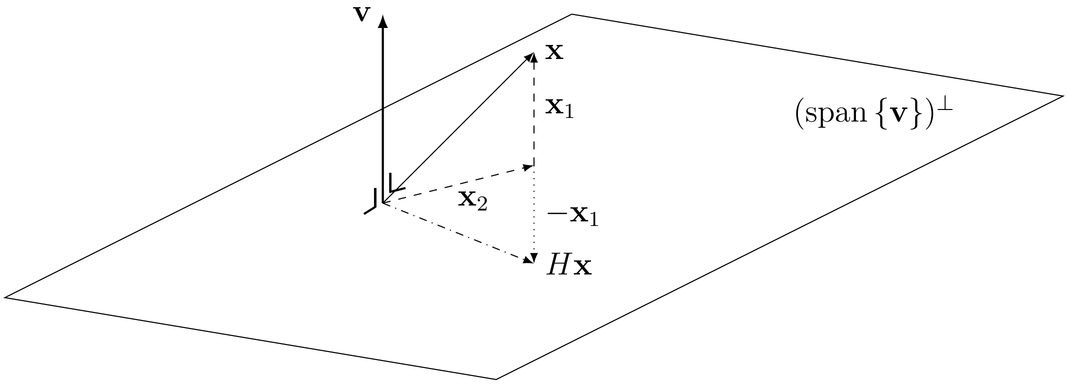

Figure 2.2: Geometrical interpretation of Householder transformation

Exercise 2.10. *How to construct the complex Householder transformation $H \in \mathbb{C}^{n\times n}$ that transforms an arbitrary vector $\mathbf{x} \in \mathbb{C}^n$ into*

$$H\mathbf{x} = \begin{bmatrix} \pm\|\mathbf{x}\| \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and satisfies $H^*H = HH^* = I_n$?

2.4.3 QR factorization

We will now use these transformations to recursively construct a unitary matrix $Q \in \mathbb{C}^{m\times m}$ such that

$$Q^* A_{m\times n} = \begin{bmatrix} R_{n\times n} \\ 0_{(m-n)\times n} \end{bmatrix},$$

where R is upper triangular. Since $Q^*Q = I_m$, we can also write

$$A = Q \begin{bmatrix} R \\ 0_{(m-n)\times n} \end{bmatrix} = Q_1 R$$

where $Q_1 \in \mathbb{C}^{m\times n}$ consists of the first n columns of Q, and thus is an isometry (i.e., $Q_1^* Q_1 = I_n$). We will refer to this notation as the *compact QR* factorization.

The construction of Q is done via recursive application of Householder transformations. First, we find the transformation H_1 such that the first column $\mathbf{a}_{:1}$ of A becomes a vector parallel to $\mathbf{e}_1$:

$$H_1 \mathbf{a}_{:1} = \begin{bmatrix} x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

If we assume that the matrix A has full rank, then clearly $\mathbf{a}_{:1} \neq 0$ and $x_1 \neq 0$. If we apply this transformation to the matrix A, we obtain

$$H_1 A = \underbrace{\left.\left[\begin{array}{c|ccc} x_1 & \times & \cdots & \times \\ \hline 0 & & & \\ \vdots & & \hat{A}_2 & \\ 0 & & & \end{array}\right]\right\} m}_{n}$$

where $\hat{A}_2$ has rank $n-1$ since $H_1 A$ has rank n. Then we apply a transformation of the form

$$\left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \hat{H}_2 & \\ 0 & & & \end{array}\right]\left[\begin{array}{c|ccc} x_1 & \times & \cdots & \times \\ \hline 0 & & & \\ \vdots & & \hat{A}_2 & \\ 0 & & & \end{array}\right] = \left[\begin{array}{c|c|ccc} x_1 & \times & \times & \cdots & \times \\ \hline 0 & x_2 & \times & \cdots & \times \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & \hat{A}_3 & \\ 0 & 0 & & & \end{array}\right] \qquad (2.8)$$

where we choose the Householder transformation $\hat{H}_2$ in such a way that the first column of $\hat{A}_2$ becomes parallel to $\mathbf{e}_1$. If we take $H_2 := \operatorname{diag}\{1, \hat{H}_2\}$, then the product $H_2 H_1 A$ is equal to the right-hand side of (2.8). We continue this process inductively until we obtain

$$H_n \cdots H_2 H_1\, A = \begin{bmatrix} x_1 & \times & \cdots & \times \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \times \\ \vdots & & \ddots & x_n \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} = \begin{bmatrix} R \\ 0_{(m-n)\times n} \end{bmatrix}, \qquad x_i \neq 0,$$

where each transformation H_i brings the corresponding column of A to the desired form. Since the product of unitary transformations is unitary, we have constructed $Q = H_1^* \cdots H_n^*$ such that

$$Q^* A = \begin{bmatrix} R \\ 0_{(m-n)\times n} \end{bmatrix}.$$

Observe now that the matrix R does not necessarily have a positive diagonal, contrary to Theorem 2.8. This can be easily corrected by applying an additional transformation $D = \operatorname{diag}\{\pm 1\}$, which is unitary.

If $A_{m\times n}$ is not a full column-rank matrix, but has rank $r < n$, we can easily modify the previous algorithm by adding permutations of the columns of A (and of its submatrices $\hat{A}_i$) to make sure that the leading column at every step remains nonzero. In this way, we obtain a factorization of the form

$$H_r \cdots H_2 H_1 \, A \, P = \left[\begin{array}{cccc|ccc} x_1 & \times & \cdots & \times & \times & \cdots & \times \\ 0 & x_2 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \times & \vdots & & \vdots \\ 0 & \cdots & 0 & x_r & \times & \cdots & \times \\ \hline 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \end{array}\right], \qquad x_i \neq 0$$

where P is a permutation matrix.

Let us digress for the rest of this subsection in order to emphasize a corollary of independent interest. We can see that the first r rows of the resulting matrix are linearly independent and the others are zero. This leads to the following theorem:

Theorem 2.12

Every matrix $A \in \mathbb{C}^{m\times n}$ can be transformed by a unitary transformation on the left into a matrix

$$Q_\ell^* A = \left[\begin{array}{c} A_1 \\ \hline 0_{(m-r)\times n} \end{array}\right]$$

where the rows of $A_1 \in \mathbb{C}^{r\times n}$ are linearly independent.

Proof. We have just seen that the desired form can be achieved with Q^*AP, but the permutation P does not affect the independence of rows, so it can be omitted. □

By means of transposition, we immediately obtain the dual result:

Theorem 2.13

Every matrix $A \in \mathbb{C}^{m\times n}$ can be transformed by a unitary transformation on the right into a matrix

$$AQ_r = [\, A_1 \,|\, 0_{m\times(n-r)} \,]$$

where the columns of $A_1 \in \mathbb{C}^{m\times r}$ are linearly independent.

By combining these two theorems, we finally arrive to the following result:

Theorem 2.14

Every matrix $A \in \mathbb{C}^{m\times n}$ can be transformed by unitary transformations on the right and on the left into a matrix

$$U^*AV = \left[\begin{array}{c|c} A_{11} & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right]$$

where $A_{11} \in \mathbb{C}^{r\times r}$ has full rank.

Proof. It is enough to first apply Theorem 2.13 and then Theorem 2.12 to the matrix A_1. □

We want to mention that, in many practical cases, this construction leads to an upper triangular matrix A_{11}. Such a decomposition is called a URV decomposition and it allows us to estimate the *singular values* of A introduced in Chapter 3 (see, e.g., [Stewart, 1973]).

2.4.4 Complexity and numerical issues

Let us now compare Householder and Gram–Schmidt methods for the construction of the QR factorization. Therefore, we will first formalize the computations by means of a pseudocode written in the MATLAB language.

The construction of a Householder transformation H satisfying $H\mathbf{x} = -\mathrm{sign}(x_1)\|\mathbf{x}\|_2\mathbf{e}_1$ relies on the computation of the vector $\mathbf{v} = \mathbf{x} + \mathrm{sign}(x_1)\|\mathbf{x}\|_2\mathbf{e}_1$. This can be done by means of the following function that normalizes $\mathbf{v}$ to $\mathbf{v}^\top\mathbf{v} = 2$ and thus $H = I - \mathbf{v}\mathbf{v}^\top$. One may count (can you?) that the complexity of this function is $5m + O(1)$ flops[1].

```
function v = Householder(x)
% this routine computes the vector $\mathbf{v}$ for the Householder transformation
% $H = I - \mathbf{v}\mathbf{v}^\top$ with $\mathbf{v}^\top\mathbf{v} = 2$ such that $H\mathbf{x} = -\mathrm{sign}(x_1)\|\mathbf{x}\|_2\mathbf{e}_1$
n = length(x);
normx = norm(x,2);
v(1) = x(1)+sign(x(1))*normx;
v(2:n) = x(2:n);
v = v*sqrt(2/v'*v);
end
```

In order to apply this transformation on the left of a matrix $A_{m\times n}$, we will call the following function:

```
function A = col.Householder(A,v)
% this routine applies the Householder transformation $H = I - \mathbf{v}\mathbf{v}^\top$
% to the rows of $A$ where $\mathbf{v}$ is normalised such that $\mathbf{v}^\top\mathbf{v} = 2$
s = v'*A;
A = A-v*s;
end
```

Observe that this clever implementation of the product only costs $4mn + O(m + n)$ instead of $\Omega(m^2n)$ flops.

The complexity of these functions is defined as the number of flops needed for the computations. The function `Householder` requires $5m + O(1)$ flops, while the functions `col.Householder` require $4mn + O(m + n)$ flops. Therefore, most of the work is done during the application of the transformation H and not during its construction.

One possible implementation of the Householder algorithm for the QR factorization of a matrix $A_{m\times n}$ (with column pivoting) is given below.

```
for j = 1:n
    c(j) = norm(A(1:m,j),2)^2;
end
r = 0;
```

[1] a flop = one addition/subtraction or one multiplication/division

```
t = max(c);
k = index(t,c);
while t>0
    r = r+1;
    swap(A(1:m,r),A(1:m,k));
    swap(c(r),c(k));
    v(r:m) = Householder(A(r:m,r));
    A(r:m,r:n) = col.Householder(A(r:m,r:n),v(r:m));
    % update c,t,k
    if r<n
        for j = r+1:n
            c(j) = c(j)-A(r,j)^2;
        end
        t = max(c(r+1:n));
        k = index(t,c(r+1:n));
    else
        t = 0;
    end
end
```

Up to some relatively small modifications, this algorithm is the one implemented in the MATLAB function qr. The total number of operations required by the procedure is

$$\sum_{i=1}^{r} 4(m-i+1)(n-i+1) + O(rm+rn)$$

since, at every iteration of the while loop, we apply the Householder transformation to a matrix of size $(m-i+1)\times(n-i+1)$. The complexity is bounded by $4mnr$ if $r \ll n$ or by $2mn^2$ if r tends to n (we assume $m \gg n$). Regarding the propagation of the rounding errors, it can be shown that this algorithm has in general better numerical properties than the Gram–Schmidt algorithm (even though the latter is easier to implement):

```
for k = 1:n
    R(k,k) = norm(A(1:m,k),2);
    Q(1:m,k) = A(1:m,k)/R(k,k);
    for j = k+1:n
        R(k,j) = Q(1:m,k)'*A(1:m,j);
        A(1:m,j) = A(1:m,j)-Q(1:m,k)*R(k,j);
    end
end
```

The complexity is $4m$ flops for every cycle of the inner for loop, and thus

$$4m\sum_{k=1}^{n}(n-k) \cong 2mn^2$$

for the algorithm in total. The complexity is of the same order as for the Householder algorithm, but the numerical properties are worse.

Chapter 3

Unitary transformations and the Singular Value Decomposition

3.1 Diagonalization by unitary transformations

The goal of this section is to obtain a matrix decomposition of the form

$$A = R \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} Q^{-1} \tag{3.1}$$

for an arbitrary matrix $A_{m \times n}$, but with constraints on the transformations R and Q: they have to be unitary in the case $A \in \mathbb{C}^{m \times n}$ and orthogonal in the case $A \in \mathbb{R}^{m \times n}$. Such restrictions make it more difficult to derive than the decomposition (2.6), but it can still be done in a similar fashion.

The unitary and orthogonal transformations form a transformation group, and as a consequence, we will obtain a new canonical form and new invariants. Only this time, our transformation group is more limited: we restrict it to *isometries.* Thus, we will obtain invariants that characterize *the way our matrices act on the norm of vectors.* This is the fundamental geometric meaning of the Singular Value Decomposition.

Since the cases of real and complex numbers are more or less the same, we will provide the details only for the most general case of complex numbers.

We will start with the case of a *Hermitian* matrix $A = A^*$. Indeed for these matrices, we will observe that simple algebraic manipulations allow us to achieve the above-described task, and moreover in this case one can pick $Q = R$! In order to derive the required diagonal decomposition, we will borrow tools that will be central in the next chapter: the *eigenvalues and eigenvectors of a square matrix.*

Definition 3.1

An *eigenvector* and an *eigenvalue* of a square matrix $A \in \mathbb{C}^{n \times n}$ are a pair consisting of a vector $\mathbf{x} \neq 0 \in \mathbb{C}^n$ and a scalar $\lambda \in \mathbb{C}$ satisfying the equation

$$A\mathbf{x} = \lambda \mathbf{x}. \tag{3.2}$$

Note that (3.2) is equivalent to

$$(\lambda I_n - A)\mathbf{x} = 0, \quad \mathbf{x} \neq 0, \tag{3.3}$$

and implies that the matrix $(\lambda I_n - A)$ is singular, i.e.,

$$\det(\lambda I_n - A) = 0.$$

Let us return to the aforementioned problem: the diagonal decomposition of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$. Observe that the polynomial $\chi(\lambda) := \det(\lambda I_n - A)$ has degree n (since the coefficient of λ^n is equal to 1). Thus, it has at least one root $\lambda_1 \in \mathbb{C}$. We know, from our hard work in Chapter 1, that (3.3) will admit a nonzero solution $\mathbf{x}$, which directly gives us an eigenvector-eigenvalue pair! After normalization, we have $\mathbf{u}_1 := \mathbf{x}/\|\mathbf{x}\|$, i.e., $\mathbf{u}_1$ is an eigenvector with norm equal to 1.

By Lemma 2.11, we can extend the vector $\mathbf{u}_1$ with a matrix $U_1^\perp$ whose columns provide an orthonormal basis of the orthogonal complement of span $\{\mathbf{u}_1\}$. This gives the matrix

$$U_1 = \left[\mathbf{u}_1 \,\middle|\, U_1^\perp\right] \in \mathbb{C}^{n\times n}, \qquad U_1^* U_1 = I_n = U_1 U_1^*.$$

Therefore,

$$\hat{A} := U_1^* A U_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1^\top \\ \hline 0 & \\ \vdots & A_2 \\ 0 & \end{array}\right].$$

The first column of $\hat{A}$ is indeed equal to

$$\left[\begin{array}{c} \mathbf{u}_1^* \\ \hline (U_1^\perp)^* \end{array}\right] A\mathbf{u}_1 = \left[\begin{array}{c} \mathbf{u}_1^* \\ \hline (U_1^\perp)^* \end{array}\right] \mathbf{u}_1 \lambda_1 = \left[\begin{array}{c} \lambda_1 \\ \hline 0 \\ \vdots \\ 0 \end{array}\right].$$

Since we assumed that A is Hermitian ($A = A^*$), we can conclude that $\hat{A}^* = U_1^* A^* U_1 = \hat{A}$. Therefore, λ_1 has to be a real number, $\mathbf{a}_1 = 0$ and $A_2^* = A_2$:

$$U_1^* A U_1 = \left[\begin{array}{c|ccc} \lambda_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{array}\right]. \tag{3.4}$$

This decomposition is actually the base case of an inductive proof allowing us to diagonalize a Hermitian matrix, as formalized in the following theorem:

Theorem 3.2

Every Hermitian matrix $A \in \mathbb{C}^{n\times n}$ can be diagonalized by a unitary transformation $U \in \mathbb{C}^{n\times n}$:

$$U^* A U = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}. \tag{3.5}$$

with $\lambda_i \in \mathbb{R}$.

Proof. It suffices to apply the decomposition (3.4) inductively: since A_2 is Hermitian, we can find again a unitary transformation $\hat{U}_2$ such that

$$\hat{U}_2^* A_2 \hat{U}_2 = \left[\begin{array}{c|ccc} \lambda_2 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & A_3 & \\ 0 & & & \end{array}\right]. \tag{3.6}$$

where λ_2 is real and A_3 is Hermitian. If we define the matrix

$$U_2 := \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \hat{U}_2 & \\ 0 & & & \end{array}\right],$$

it is not hard to see that U_2 is unitary as well (can you prove it?) and

$$U_2^* U_1^* A U_1 U_2 = \left[\begin{array}{cc|ccc} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{array}\right].$$

By repeating this procedure with A_3 and subsequent matrices, we obtain

$$U_{n-1}^* \cdots U_2^* U_1^* \, A \, U_1 U_2 \cdots U_{n-1} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$

Since the product of unitary transformations is unitary as well, the statement of the theorem follows. □

Remark 3.1.

1. *Note that*

$$\begin{aligned} \det(\lambda I_n - U^* A U) &= \det(U^*(\lambda I_n - A)U) \\ &= \det(U^*)\det(\lambda I_n - A)\det(U) \\ &= \det(\lambda I_n - A), \end{aligned}$$

due to

$$\det(I_n) = \det(U^* U) = \det(U^*)\det(U) = 1.$$

The eigenvalues are not changed by the similarity transformation. Furthermore, $\det(\lambda I_n - U^* A U) = \prod_{i=1}^n (\lambda - \lambda_i)$*, and the diagonal elements of* $U^* A U$ *are the eigenvalues of* A*.*

2. *The decomposition (3.5) shows that the eigenvalues of a Hermitian matrix are real.*
3. *If the matrix A is real, then its eigenvectors are real as well as the corresponding unitary matrix U (i.e., U is orthogonal: $U \in \mathbb{R}^{n\times n}$, $U^\top U = UU^\top = I_n$).*
4. *If at each step of the construction of (3.5) we choose the largest remaining eigenvalue of A, then we derive a matrix with decreasing λ_i's.*

We will restate all these remarks in the form of a theorem:

Theorem 3.3

The eigenvalues of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ are invariant under unitary similarity transformations:

$$B = U^* A U.$$

Every class of equivalence defined by this transformation group has a unique canonical representative which is the diagonal matrix Λ with the eigenvalues of A decreasing along the diagonal.

Note that the above theorem does not answer the question whether a transformation U diagonalizing a matrix A is unique or not. In fact, this transformation is unique up to a transformation U_{up} that commutes with Λ.

Exercise 3.1. *Assuming that all the diagonal elements of Λ are distinct, show that the matrix U_{up} is necessarily diagonal and consists only of phases:*

$$U_{up} = \operatorname{diag}\{e^{i\psi_1}, \ldots, e^{i\psi_n}\}.$$

Consider now the most general case of arbitrary matrices $A_{m\times n}$. We will show in the following theorem that it is possible to obtain a quasi-diagonalization under unitary transformations of rows and columns.

Theorem 3.4: Singular Value Decomposition

For every matrix $A \in \mathbb{C}^{m\times n}$, there exist unitary transformations $U \in \mathbb{C}^{m\times m}$ $(U^*U = I_m)$ and $V \in \mathbb{C}^{n\times n}$ $(V^*V = I_n)$ such that

$$A = U\Sigma V^* \qquad \text{where} \qquad \Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0_{r\times(n-r)} \\ 0 & & \sigma_r & \\ \hline & 0_{(m-r)\times r} & & 0_{(m-r)\times(n-r)} \end{array}\right],$$

with real positive *singular values*:

$$\sigma_1 \geq \ldots \geq \sigma_r > 0.$$

The value r and the r-tuple $(\sigma_1, ..., \sigma_r)$ are uniquely defined, and as a consequence, the matrix Σ constitutes a canonical form under unitary transformations, that is, under transformations of the form

$$B = \tilde{U}^* A \tilde{V}$$

where $\tilde{U}$ and $\tilde{V}$ are two unitary matrices.

Proof. Observe that the matrix A^*A is Hermitian and we can apply Theorem 3.2. Thus, it is possible to diagonalize A^*A using a unitary matrix:

$$V^*A^*AV = \Lambda, \qquad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n, \tag{3.7}$$

where λ_i are the eigenvalues of A^*A. It implies that $\lambda_i = \|A\mathbf{v}_{:i}\|_2^2 \geq 0$ and we can define

$$\sigma_i^2 := \lambda_i.$$

For $\sigma_i \neq 0$, we can define the vectors $\mathbf{u}_{:i}$ in the following manner:

$$A\mathbf{v}_{:i} = \sigma_i \mathbf{u}_{:i}. \tag{3.8}$$

The equation (3.7) immediately implies that the vectors $\mathbf{u}_{:i}$ are orthonormal.

Assume now that there are r nonzero singular values σ_i. We extend the basis $\mathbf{u}_{:i}$ $(i = 1, \ldots, r)$ to obtain an orthonormal basis of the whole space and arrange it in a unitary matrix:

$$U = [\mathbf{u}_{:1}, \ldots, \mathbf{u}_{:r} | \mathbf{u}_{:r+1}, \ldots, \mathbf{u}_{:m}]. \tag{3.9}$$

Using (3.8) and (3.9), it can be easily shown that

$$AV = U\Sigma \qquad \text{where} \qquad \Sigma = \left[\begin{array}{c|c} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{matrix} & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right],$$

Now it remains to post-multiply the first equation by V^*. From (3.7), we conclude that we have $\sigma_1 \geq \ldots \geq \sigma_r > 0$ as well.

It remains to prove that if we pre- or post-multiply our matrix A by a unitary transformation, it preserves the Σ term (the singular values), but not the left and right unitary matrices U or V.

Let $B = \tilde{U}^*A\tilde{V}$ where $\tilde{U}$ and $\tilde{V}$ are two unitary matrices.

Since the singular values of a matrix A are given by the square root of the eigenvalues of A^*A and the eigenvalues of a Hermitian matrix are invariant under similarity transformation (Theorem 3.3), we have

$$\sigma_i(B) = \sqrt{\lambda_i(B^*B)} = \sqrt{\lambda_i(\tilde{V}^*A^*A\tilde{V})} = \sqrt{\lambda_i(A^*A)} = \sigma_i(A).$$

□

Remark 3.2.

1. *If the matrix A is real, then the vectors $\mathbf{u}_{:i}$ $(i = 1, \ldots, m)$ and $\mathbf{v}_{:i}$ $(i = 1, \ldots, n)$ are real as well, and U and V are orthogonal matrices.*

2. *The transformations U and V diagonalize the matrices AA^* and A^*A respectively, since*

$$U^*AA^*U = \Sigma\Sigma^\top, \qquad V^*A^*AV = \Sigma^\top\Sigma.$$

Furthermore, the columns of U and V are the eigenvectors of AA^* and A^*A respectively.

3. *If H is Hermitian, then it can be decomposed in the following way:*

$$H = U\Lambda U^*,$$

and thus $HH^* = H^*H = U\Lambda^2 U^*$. Moreover, we have:

$$H = U \cdot |\Lambda| \cdot \operatorname{sign}(\Lambda) \cdot U^* = U\Sigma V^*.$$

In particular, the singular values of a Hermitian matrix are the absolute values of the eigenvalues: $\Sigma = |\Lambda|$.

4. *The transformations U and V are not uniquely defined. It is easy to see that all pairs of unitary matrices $U_{up} \in \mathbb{C}^{m\times m}$, $V_{up} \in \mathbb{C}^{n\times n}$ satisfying*

$$U_{up}\Sigma = \Sigma V_{up}$$

lead to

$$UU_{up}\Sigma V_{up}^* V^* = U\Sigma V^* = A$$

which represents another SVD of A. It is possible to show that these are the only possible degrees of freedom of this decomposition. Namely, if $m = n = r$ and Σ has distinct diagonal elements, then $U_{up} = V_{up} = \operatorname{diag}\{e^{i\phi_1}, \dots, e^{i\phi_n}\}$.

3.2 Applications of the SVD

3.2.1 Orthonormal bases for $\mathrm{Ker}(A)$ and $\mathrm{Im}(A)$

There are many applications of the singular value decomposition (SVD). The most important one comes from its combination with Theorem 2.4:

$$\begin{aligned}\mathrm{Im}(A) &= \operatorname{span}\{\mathbf{u}_{:1}, \dots, \mathbf{u}_{:r}\},\\ \mathrm{Ker}(A) &= \operatorname{span}\{\mathbf{v}_{:r+1}, \dots, \mathbf{v}_{:n}\},\\ \operatorname{rank}(A) &= r.\end{aligned}$$

Thus, we can construct, from the SVD, orthonormal bases for $\mathrm{Ker}(A)$ and $\mathrm{Im}(A)$.

3.2.2 Linear transformations and the four fundamental subspaces

For every matrix $A_{m\times n}$ mapping a Euclidean (or unitary) vector space $\mathcal{X} = \mathcal{F}^n$ to another Euclidean (or unitary) vector space $\mathcal{Y} = \mathcal{F}^m$, we can define its *dual matrix* $\tilde{A}$ with respect to the inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ defined in the corresponding spaces: $\tilde{A}$ is the unique $n \times m$ matrix satisfying

$$\langle A\mathbf{x}, \mathbf{y}\rangle_{\mathcal{Y}} = \langle \mathbf{x}, \tilde{A}\mathbf{y}\rangle_{\mathcal{X}}, \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \mathbf{y} \in \mathcal{Y}.$$

This definition is very general and is valid even in the case of operators between infinite-dimensional spaces. In the matrix case with the classical inner products (x^*y for $\mathcal{F} = \mathbb{C}$ and $x^\top y$ for $\mathcal{F} = \mathbb{R}$), we have encountered the dual already many times: for $\mathcal{F} = \mathbb{C}$, we have

$$\mathbf{y}^* A\mathbf{x} = \langle A\mathbf{x}, \mathbf{y}\rangle_{\mathcal{Y}} = \langle \mathbf{x}, \tilde{A}\mathbf{y}\rangle_{\mathcal{X}} = \mathbf{y}^*\tilde{A}^*\mathbf{x}, \qquad \forall \mathbf{x} \in \mathcal{X},\ \forall \mathbf{y} \in \mathcal{Y}$$

which implies that

$$\tilde{A} = A^*.$$

In the case of $\mathcal{F} = \mathbb{R}$, by a similar reasoning, we obtain

$$\tilde{A} = A^\top.$$

The following theorem states the fundamental connection between the spaces Im and Ker of dual operators A and $\tilde{A}$.

Theorem 3.5

For every linear mapping $A : \mathcal{X} \to \mathcal{Y}$ between finite-dimensional spaces $\mathcal{X}$ and $\mathcal{Y}$, we have

$$\begin{aligned}
\operatorname{Ker}(A) &= \operatorname{Im}(\tilde{A})^\perp, \\
\operatorname{Ker}(\tilde{A}) &= \operatorname{Im}(A)^\perp, \\
\operatorname{Im}(A) &= \operatorname{Ker}(\tilde{A})^\perp, \\
\operatorname{Im}(\tilde{A}) &= \operatorname{Ker}(A)^\perp.
\end{aligned}$$

Proof. In order to prove the first identity, we make the following observation:

$$\mathbf{x} \in \operatorname{Ker}(A) \qquad \Longleftrightarrow \qquad \langle A\mathbf{x}, \mathbf{y}\rangle_{\mathcal{Y}} = 0 \quad \forall \mathbf{y} \in \mathcal{Y},$$

but $\langle A\mathbf{x}, \mathbf{y}\rangle_{\mathcal{Y}} = \langle \mathbf{x}, \tilde{A}\mathbf{y}\rangle_{\mathcal{X}}$, and therefore

$$\begin{aligned}
\mathbf{x} \in \operatorname{Ker}(A) \qquad &\Longleftrightarrow \qquad \langle \mathbf{x}, \tilde{A}\mathbf{y}\rangle_{\mathcal{X}} = 0, \quad \forall \mathbf{y} \in \mathcal{Y} \\
&\Longleftrightarrow \qquad \langle \mathbf{x}, \mathbf{z}\rangle_{\mathcal{X}} = 0, \quad \forall \mathbf{z} \in \operatorname{Im}(\tilde{A}) \\
&\Longleftrightarrow \qquad \mathbf{x} \in \operatorname{Im}(\tilde{A})^\perp.
\end{aligned}$$

The other identities are derived by observing that $(\mathcal{X}^\perp)^\perp = \mathcal{X}$ (this is where the “finite dimension” assumption is required) for all spaces $\mathcal{X}$ and that $(\tilde{A})\tilde{} = A$ for all matrices A. □

This theorem eventually will allow us to establish a connection between the characterization of the spaces Ker(A) and Im(A), the singular value decomposition and the solutions $\mathbf{x} \in \mathcal{X} := \mathcal{F}^n$ of a linear system of equations

$$A_{m\times n}\mathbf{x} = \mathbf{y}$$

for a matrix A of arbitrary rank r and a vector $\mathbf{y} \in \mathcal{Y} := \mathcal{F}^m$.

Lemma 3.6

1. $\operatorname{Im}(A) = \mathcal{Y} \quad \Longleftrightarrow \quad r = m \quad \Longleftrightarrow \quad \exists A^r \quad \text{s.t.} \quad AA^r = I_m$
   $\Longleftrightarrow$ there exists a solution $\mathbf{x}$ to $A\mathbf{x} = \mathbf{y}$ for every $\mathbf{y} \in \mathcal{Y}$ (surjectivity).
2. $\operatorname{Ker}(A) = \{0\} \quad \Longleftrightarrow \quad r = n \quad \Longleftrightarrow \quad \exists A^\ell \quad \text{s.t.} \quad A^\ell A = I_n$
   $\Longleftrightarrow$ the system $A\mathbf{x} = \mathbf{y}$ has a unique solution (injectivity).

Proof.

1. $\mathrm{Im}(A) = \mathcal{Y}$ implies $r = m$ (Theorem 2.4), and thus $A = R\,[\,I_m\,|\,0\,]Q^{-1}$ (Theorem 1.8). By taking $A^r := Q\begin{bmatrix} I_m \\ 0 \end{bmatrix} R^{-1}$, we have $AA^r = I_m$. Therefore, $\mathbf{x} := A^r\mathbf{y}$ satisfies $A\mathbf{x} = \mathbf{y}$ for all vectors $\mathbf{y}$. The latter observation can be restated as follows: all vectors $\mathbf{y}$ can be written as a linear combination of the columns of A. Thus, $\mathrm{Im}(A) = \mathcal{Y}$ and the equivalence has been shown.

2. $\mathrm{Ker}(A) = \{0\}$ implies $r = n$ (Theorem 2.4), and thus $A = R\begin{bmatrix} I_n \\ 0 \end{bmatrix} Q^{-1}$ (Theorem 1.8). By taking $A^\ell := Q\,[\,I_n\,|\,0\,]\,R^{-1}$, we have $A^\ell A = I_n$. If we consider two solutions $A\mathbf{x}_i = \mathbf{y}$ $(i = 1, 2)$, then $A(\mathbf{x}_1 - \mathbf{x}_2) = 0$ and $\mathbf{x}_1 - \mathbf{x}_2 = A^\ell(0) = 0$. In other words, the solutions are unique. Applying the last statement to $\mathbf{y} = 0$, we obtain that $\mathrm{Ker}(A) = \{0\}$, and the required equivalence has been shown.

□

Note that if both conditions in Lemma 3.6 hold at the same time, then A is *bijective* and $m = n = r$ (we are thus in the case of Theorem 2.1).

Before stating the connections with the singular value decomposition, we will introduce the concept of *direct sum* of two subspaces.

Definition 3.7

If two subspaces have only the zero vector in common, then their sum is called the direct sum and is denoted by $\oplus$:

$$\mathcal{X}_1 \cap \mathcal{X}_2 = 0 \qquad \Longrightarrow \qquad \mathcal{X}_1 \oplus \mathcal{X}_2 := \mathcal{X}_1 + \mathcal{X}_2.$$

If the subspaces are orthogonal as well, then we say that their sum is orthogonal:

$$\mathcal{X}_1 \perp \mathcal{X}_2 \qquad \Longrightarrow \qquad \mathcal{X}_1 \oplus^\perp \mathcal{X}_2 := \mathcal{X}_1 + \mathcal{X}_2.$$

The following lemma describes an important property of these concepts.

Lemma 3.8

Every vector $\mathbf{x} \in \mathcal{X}_1 \oplus \mathcal{X}_2$ has a unique decomposition $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_i \in \mathcal{X}_i$ $(i = 1, 2)$.

Proof. It is enough to find a basis $\{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_{r_i}^{(i)}\}$ for each of the subspaces $\mathcal{X}_i$ and observe that their union is a basis for $\mathcal{X}_1 \oplus \mathcal{X}_2$. Since the representation of a vector is unique with respect to the basis (see Exercise 2.3), the result follows. □

For the orthogonal sum, the trivial case is given by

$$\mathcal{X} = \mathcal{S} \oplus^\perp \mathcal{S}^\perp$$

for all subspaces $\mathcal{S} \subseteq \mathcal{X}$.

Let us apply this to the subspaces of Theorem 3.5 associated with the following system of equations:

$$A_{m\times n}\mathbf{x} = \mathbf{y}, \qquad \mathbf{x} \in \mathcal{X},\ \mathbf{y} \in \mathcal{Y}.$$

Let

$$\begin{aligned} \mathcal{X}_1 &:= \operatorname{Im}(\tilde{A}), & \mathcal{X}_2 &= \operatorname{Ker}(A), \\ \mathcal{Y}_1 &:= \operatorname{Im}(A), & \mathcal{Y}_2 &= \operatorname{Ker}(\tilde{A}), \end{aligned} \tag{3.10}$$

then

$$\mathcal{X} = \mathcal{X}_1 \oplus^{\perp} \mathcal{X}_2, \qquad \mathcal{Y} = \mathcal{Y}_1 \oplus^{\perp} \mathcal{Y}_2. \tag{3.11}$$

If we choose the coordinate system compatible with the decompositions (3.11), then the matrix A has the form

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \qquad \text{where} \quad A_{ij} : \mathcal{X}_j \mapsto \mathcal{Y}_i.$$

Theorem 3.9

In the coordinate system (3.10) and (3.11), the matrix $A_{m\times n}$ has the form

$$A = \left[\begin{array}{c|c} A_{11} & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right]$$

where $A_{11} \in \mathcal{F}^{r\times r}$ is bijective.

Proof. Since

$$\begin{cases} A_{11}\mathbf{x}_1 + A_{12}\mathbf{x}_2 &= \mathbf{y}_1 \\ A_{21}\mathbf{x}_1 + A_{22}\mathbf{x}_2 &= \mathbf{y}_2 \end{cases},$$

we can conclude that $\mathbf{y}_i = A_{i2}\mathbf{x}_2 = 0$ for all vectors $\mathbf{x}_2 \in \operatorname{Ker}(A)$. Thus, $A_{12} = A_{22} = 0$. Furthermore, $\operatorname{Im}\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \mathcal{Y}_1$ requires that $A_{21} = 0$ and A_{11} has full rank (thus, is invertible). □

The connection with the singular value decomposition should be clear now. In the following coordinate system:

$$U^*AV = \left[\begin{array}{c|c} \Sigma_r & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right], \qquad \Sigma_r = \operatorname{diag}\{\sigma_1, \dots, \sigma_r\},$$

we clearly have

$$\begin{aligned} \mathcal{X}_1 &= \operatorname{Im}\begin{bmatrix} I_r \\ 0 \end{bmatrix}, & \mathcal{X}_2 &= \mathcal{X}_1^{\perp}, \\ \mathcal{Y}_1 &= \operatorname{Im}\begin{bmatrix} I_r \\ 0 \end{bmatrix}, & \mathcal{Y}_2 &= \mathcal{Y}_1^{\perp}. \end{aligned}$$

It implies that, within the initial coordinate system of the matrix A,

$$A = U\Sigma V^* = U_1 \Sigma_r V_1^*,$$

with

$$U = [\,\underbrace{U_1}_{r \text{ columns}}\,|\,U_2\,], \qquad V = [\,\underbrace{V_1}_{r \text{ columns}}\,|\,V_2\,],$$

the respective subspaces are given by

$$\begin{aligned} \mathcal{X}_1 &= \mathrm{Im}(V_1), & \mathcal{X}_2 &= \mathrm{Im}(V_2), \\ \mathcal{Y}_1 &= \mathrm{Im}(U_1), & \mathcal{Y}_2 &= \mathrm{Im}(U_2). \end{aligned}$$

Furthermore, this choice of bases for $\mathcal{Y}_1 = \mathrm{Im}(A)$ and $\mathcal{X}_1 = \mathrm{Ker}(A)^\perp$ gives a “canonical” representation of the bijection

$$A_{11} : \mathrm{Ker}(A)^\perp \to \mathrm{Im}(A)$$

since it is diagonal. The singular values σ_i $(i = 1, \ldots, r)$ can be interpreted as follows. We construct a ball within the space $\mathcal{X}_1$ of radius 1 centered at the origin:

$$B_{\mathcal{X}_1}(0,1) = \{\mathbf{x} \mid \mathbf{x} \in \mathcal{X}_1,\ \|\mathbf{x}\|_2 \leq 1\}.$$

This set is mapped by $A_{11} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_r\}$ to an ellipsoid in $\mathcal{Y}_1$:

$$E_{\mathcal{Y}_1}(0,\Sigma) = \{\mathbf{y} \mid \mathbf{y} \in \mathcal{Y}_1,\ \|\Sigma_r^{-1}\mathbf{y}\|_2 \leq 1\}.$$

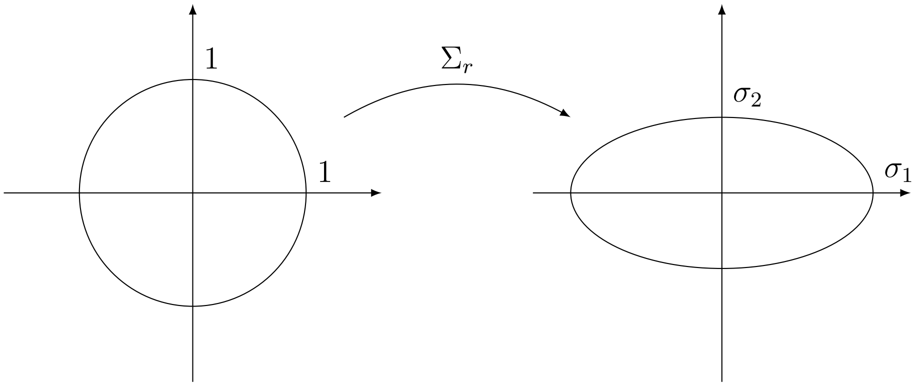

Figure 3.1: Arbitrary orthogonal bases of $\mathcal{S}_1$ and $\mathcal{S}_2$.

An illustration in the two-dimensional case ($\mathcal{F} = \mathbb{R}$ and $r = 2$) is given in Figure 3.1. On the left of the figure, we have the disk representing the set $B_{\mathcal{X}_1}(0,1)$ and on the right, we have the ellipsoid representing $E_{\mathcal{Y}_1}(0,\Sigma)$. Note that the singular values measure the deformation of vectors in $\mathcal{X}_1$ by A_{11} (and A as well).

3.2.3 Projections and generalized inverses

A *projection* P is a square matrix that satisfies $P^2 = P$. In other words, if we project a vector $\mathbf{x}$ twice, then we obtain the same result as if we projected it only once: $P(P\mathbf{x}) = P\mathbf{x}$. The image of P is the space on which we project. The kernel of P is the set of vectors projected to 0. For projections defined on $\mathbb{C}^n$ (i.e., $P \in \mathbb{C}^{n\times n}$), we say that the projection is *orthogonal* if $\mathrm{Ker}(P) \perp \mathrm{Im}(P)$. In this case, by Theorem 3.5, we have $\mathrm{Ker}(P) = \mathrm{Im}(P^*)^\perp$ and $\mathrm{Ker}(P^*) = \mathrm{Im}(P)^\perp$, so that $\mathrm{Im}(P) = \mathrm{Im}(P^*)$ and $\mathrm{Ker}(P) = \mathrm{Ker}(P^*)$, and thus $P = P^*$.

Let $A \in \mathbb{C}^{m\times n}$ be an arbitrary matrix. If A is not square or if $\det(A) = 0$, then A is not invertible. Nevertheless, we would like to define a matrix X which is “as close as possible” to an inverse of A.

Definition 3.10

Given a matrix $A \in \mathbb{C}^{m\times n}$, we will call any matrix $X \in \mathbb{C}^{n\times m}$ satisfying the equations

$$\begin{array}{lrcl} (1) & AXA &=& A, \\ (2) & XAX &=& X, \\ (3) & AX &=& (AX)^*, \\ (4) & XA &=& (XA)^*, \end{array} \tag{3.12}$$

a *pseudoinverse* or *Moore-Penrose inverse* of A.

As we will soon see, it is uniquely defined for every matrix A. If only part of the conditions hold, e.g. only (1) and (3), then we call any matrix satisfying them a generalized (1,3) inverse (we will always assume that at least one of the top two conditions holds). Since such generalized inverses have to satisfy a smaller number of constraints, they might not be uniquely defined anymore. For example,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x \\ 0 & 0 \end{bmatrix}$$

satisfy equations (1), (2) and (4). Therefore, for all values of x, the matrix X is a generalized (1,2,4) inverse of A.

Exercise 3.2. *Show that if X satisfies*

- *(1), then AX is a projection;*
- *(1) and (3), then AX is an orthogonal projection;*
- *(2), then XA is a projection;*
- *(2) and (4), then XA is an orthogonal projection.*

As we have seen, the Singular Value Decomposition allows us to separate A in an invertible and a nilpotent part. We will then leverage it in order to construct the pseudoinverse: given a matrix A, a matrix X satisfies one of the equations (3.12) if and only if the transformed matrices $\hat{A} = U^*AV$ and $\hat{X} = V^*XU$ satisfy it as well. Thus, we can study these equations in the "appropriate" coordinate system and restrict ourselves to the case of a diagonal matrix $\hat{A}$:

$$\hat{A} = U^*AV = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Theorem 3.11

The Moore-Penrose inverse A^I of a matrix $A \in \mathbb{C}^{m\times n}$ is unique and is equal to

$$A^I = V \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*.$$

Proof. Let us consider the matrix $\hat{A} = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$. For this diagonal matrix, we can easily see that

$$\hat{A}^I = \left[\begin{array}{c|c} \hat{X}_{11} & \hat{X}_{12} \\ \hline \hat{X}_{21} & \hat{X}_{22} \end{array}\right]$$

must satisfy $\hat{X}_{11} = \Sigma_r^{-1}$ due to (1), $\hat{X}_{12} = 0$ due to (3), $\hat{X}_{21} = 0$ due to (4), and finally $\hat{X}_{22} = 0$ due to (2). Therefore, $\hat{A}^I$ is well defined and unique. After the transformation, we have

$$A^I = V\hat{A}^I U^*$$

that is unique as well. Although the transformations U and V of the SVD are not unique, the degrees of freedom disappear in the product $A^I = V\hat{A}^I U^*$. □

The singular value decomposition is sometimes given in a more compact form, which gets rid of the nilpotent part, and is handy in the algebraic expression of projectors:

Definition 3.12: Compact SVD

Given the Singular Value Decomposition of an arbitrary matrix $A \in \mathbb{C}^{m\times n}$:

$$A = U \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^*,$$

we define its *compact SVD* decomposition

$$A = U_1 \Sigma_r V_1^*$$

where

$$U = [\underbrace{\quad U_1 \quad}_{r \text{ columns}} | U_2], \qquad V = [\underbrace{\quad V_1 \quad}_{r \text{ columns}} | V_2].$$

In particular, $U_1 \in \mathbb{C}^{m\times r}$ and $V_1 \in \mathbb{C}^{n\times r}$ are isometries ($U_1^* U_1 = I_r = V_1^* V_1$).

Note that this notation allows us to write the Moore-Penrose inverse in a more compact form:

$$A^I = V_1 \Sigma_r^{-1} U_1^*.$$

We immediately see that the projections AA^I and $A^I A$ (see Exercise 3.2) can be written as follows:

$$P_A := AA^I = U_1 U_1^*,$$
$$P_{\tilde{A}} := A^I A = V_1 V_1^*.$$

The projection spaces of these orthogonal projections are $\mathrm{Im}(A)$ and $\mathrm{Im}(\tilde{A})$ respectively.

Note that if P is a projection, then $(I - P)$ is a projection as well:

$$(I - P)^2 = I - 2P + P^2 = I - P.$$

Furthermore, if P is orthogonal (i.e., if $P = P^*$), then $I - P$ is also orthogonal. By applying it to the projections P_A and $P_{\tilde{A}}$, we have

$$I - P_A = U_2 U_2^*,$$
$$I - P_{\tilde{A}} = V_2 V_2^*.$$

Note that these four projections are orthogonal projections on the fundamental spaces of A and its dual $\tilde{A}$.

The following proposition shows that every projection P admits a particular representation.

Proposition 3.13

Every projection $P \in \mathbb{C}^{n\times n}$ has a representation $P = XY^*$ where $X, Y \in \mathbb{C}^{n\times r}$, $Y^*X = I_r$, and r is the rank of P. If P is an orthogonal projection matrix, then we can choose $X = Y$.

Proof. We start with the compact SVD of P:

$$P = U_1 \Sigma_r V_1^*.$$

Since $P^2 = P$, we have that $\Sigma_r^{-1} U_1^* P^2 V_1 = \Sigma_r^{-1} U_1^* P V_1$ and thus

$$V_1^* U_1 \Sigma_r = I_r.$$

It suffices to set $X = U_1 \Sigma_r$ and $Y = V_1$.
If P is an orthogonal projection matrix ($P^2 = P$ and $P^* = P$) then $U_1 = V_1$ and $\Sigma_r = I_r$. □

3.2.4 Least squares problems and Tikhonov regularization

We will analyze now the solutions of a system of equations

$$A\mathbf{x} = \mathbf{y}$$

in the general case of a matrix $A \in \mathbb{C}^{m\times n}$ of rank r. Let $A = U \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} V^*$ be the SVD of A. If we take

$$\hat{\mathbf{x}} = V^*\mathbf{x}, \qquad \hat{\mathbf{y}} = U^*\mathbf{y},$$

then the system is reduced to

$$\left\{ \begin{array}{rcl} \Sigma_r \hat{\mathbf{x}}_1 + 0\hat{\mathbf{x}}_2 & = & \hat{\mathbf{y}}_1 \\ 0\hat{\mathbf{x}}_1 + 0\hat{\mathbf{x}}_2 & = & \hat{\mathbf{y}}_2 \end{array} \right. \tag{3.13}$$

with the proper partitioning of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. It is clear that the system (3.13) admits a solution if and only if $\hat{\mathbf{y}}_2 = 0$. In this case, it suffices to take $\hat{\mathbf{x}}_1 = \Sigma_r^{-1}\hat{\mathbf{y}}_1$ as a solution. This necessary and sufficient condition can be stated as

$$\begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \end{bmatrix} \in \operatorname{Im} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

or

$$\mathbf{y} \in \operatorname{Im}(A)$$

in the original coordinate system. Observe that the choice $\hat{\mathbf{x}}_2$ is arbitrary in (3.13), and putting $\hat{\mathbf{x}}_2 = 0$ guarantees that the solution has the smallest possible norm, since $\|\mathbf{x}\|_2^2 = \|\hat{\mathbf{x}}\|_2^2 = \|\hat{\mathbf{x}}_1\|_2^2 + \|\hat{\mathbf{x}}_2\|_2^2$. Such a solution with the smallest norm can also be written as

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{y}}_1 \\ 0 \end{bmatrix} = \hat{A}^I \hat{\mathbf{y}}$$

and

$$\mathbf{x} = V\hat{\mathbf{x}} = V\hat{A}^I\hat{\mathbf{y}} = V \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*\mathbf{y} = A^I\mathbf{y}$$

where A^I is the Moore-Penrose inverse of A.

If $\hat{\mathbf{y}}_2 \neq 0$, then the system is not compatible. What should we do in this case? In many applications, one would like to obtain the *least squares* solution, i.e., find a vector $\mathbf{x}$ minimizing the error $\|A\mathbf{x} - \mathbf{y}\|_2$:

$$\begin{aligned} \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{y}\|_2^2 &= \min_{\hat{\mathbf{x}}} \left\| \begin{bmatrix} \Sigma_r\hat{\mathbf{x}}_1 - \hat{\mathbf{y}}_1 \\ -\hat{\mathbf{y}}_2 \end{bmatrix} \right\|_2^2 \\ &= \min_{\hat{\mathbf{x}}_1} \|\Sigma_r\hat{\mathbf{x}}_1 - \hat{\mathbf{y}}_1\|_2^2 + \|\hat{\mathbf{y}}_2\|_2^2 \\ &= \|\hat{\mathbf{y}}_2\|_2^2 \quad \text{for} \quad \hat{\mathbf{x}}_1 = \Sigma_r^{-1}\hat{\mathbf{y}}_1. \end{aligned}$$

Again, the choice of $\hat{\mathbf{x}}_2$ is arbitrary and does not affect the minimal value. But, clearly, $\hat{\mathbf{x}}_2 = 0$ gives the solution with the smallest norm. It can be written as

$$\hat{\mathbf{x}} = \begin{bmatrix} \hat{\mathbf{x}}_1 \\ \hat{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \end{bmatrix} = \hat{A}^I \hat{\mathbf{y}}$$

and

$$\mathbf{x} = V\hat{\mathbf{x}} = V\hat{A}^I\hat{\mathbf{y}} = V \begin{bmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} U^*\mathbf{y} = A^I\mathbf{y}$$

where A^I is the Moore-Penrose inverse of A. To summarize the discussion, we state the following general theorem.

Theorem 3.14

Let $A \in \mathbb{C}^{m\times n}$ be a matrix of rank r. The solution of the system of linear equations

$$A\mathbf{x} = \mathbf{y}$$

has the following properties:

- if $m = n = r$: $\mathbf{x} = A^{-1}\mathbf{y}$ is *unique*;
- if $m = r < n$: $\mathbf{x} = A^r\mathbf{y}$ is a solution for all matrices A^r such that $AA^r = I_m$; moreover, $\mathbf{x} = A^I\mathbf{y}$ is the solution with the smallest norm;
- if $m > r = n$: for all matrices A^ℓ such that $A^\ell A = I_n$, the vector $\mathbf{x} = A^\ell\mathbf{y}$ is a solution if and only if $\mathbf{y} \in \mathrm{Im}(A)$; moreover, $\mathbf{x} = A^I\mathbf{y}$ is the unique least squares solution;
- if $r < m, n$: $\mathbf{x} = A^I\mathbf{y}$ is the least squares solution and has the minimal norm among all the least squares solutions.

Proof. It suffices to transform the system

$$U^*AV = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}$$

and apply the results of Lemma 3.6. □

Tikhonov regularization

In many applications, the encountered least squares problems

$$A_{m\times n}\mathbf{x} = \mathbf{y},$$

where $m > r = n$, are such that the singular values are very different: $\sigma_1 \gg \sigma_n$. One such example is *the polynomial interpolation problem.* Let

$$p(x) = \sum_{i=0}^{n-1} a_i x^i$$

be a polynomial for which its values at the points x_i are known:

$$y_i = p(x_i), \qquad i = 1, \ldots, m, \qquad x_i \neq x_j \quad \text{if } \ i \neq j.$$

If $m \geq n$, then we can uniquely reconstruct the polynomial from the measurements (x_i, y_i) since

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{n-1} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

is a compatible system of rank n (why?). The vector $\mathbf{a} \in \mathbb{R}^n$ of the coefficients is the solution of the least squares problem

$$A_{m\times n}\mathbf{a} = \mathbf{y}.$$

It is possible to show that if we take many measurements ($m \gg n$), the difference between the singular values σ_i of the matrix A becomes very large since σ_1 grows with m while σ_n remains bounded.

Thus, the least squares solution

$$\mathbf{x} = A^I\mathbf{y} = V \left[\begin{array}{ccc|c} \sigma_1^{-1} & & 0 & \\ & \ddots & & 0_{n\times(m-n)} \\ 0 & & \sigma_n^{-1} & \end{array}\right] U^*\mathbf{y}$$

$$= V \begin{bmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-1} \end{bmatrix} \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = V\Sigma_n^{-1}\hat{\mathbf{y}}_1$$

is very sensitive to the noise $\Delta\mathbf{y}$. In fact, the perturbed solution $\mathbf{x} + \Delta\mathbf{x}$ satisfies

$$(\mathbf{x} + \Delta\mathbf{x}) = V\Sigma_n^{-1}(\hat{\mathbf{y}}_1 + \Delta\hat{\mathbf{y}}_1)$$

and therefore,

$$\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \frac{\|\Sigma_n^{-1}\Delta\hat{\mathbf{y}}_1\|_2}{\|\Sigma_n^{-1}\hat{\mathbf{y}}_1\|_2} \le \frac{\sigma_1}{\sigma_n}\frac{\|\Delta\hat{\mathbf{y}}_1\|_2}{\|\hat{\mathbf{y}}_1\|_2} = \frac{\sigma_1}{\sigma_n}\frac{\|\Delta\mathbf{y}_1\|_2}{\|\mathbf{y}_1\|_2}.$$

This inequality follows from

$$\|M\mathbf{y}\|_2 \le \|M\|_2\|\mathbf{y}\|_2, \qquad \|M\mathbf{y}\|_2 \ge \|M^{-1}\|_2^{-1}\|\mathbf{y}\|_2.$$

Concluding, if $\sigma_1 \gg \sigma_n$, the signal to noise ratio in the solution can be much worse than in the measurements. We can "control" this phenomenon by introducing a term proportional to $\|\mathbf{x}\|_2$ in order to penalize the growth of $\mathbf{x}$:

$$\min\left\{\|A\mathbf{x}-\mathbf{y}\|_2^2 + \delta^2\|\mathbf{x}\|_2^2\right\} = \min\left\|\begin{bmatrix} A \\ \delta I\end{bmatrix}\mathbf{x} - \begin{bmatrix}\mathbf{y}\\ 0\end{bmatrix}\right\|_2^2. \tag{3.14}$$

It turns out that in this modified least squares problem, all the singular values are lower bounded by δ, which attenuates the bad behaviour exhibited above. This is a consequence of the following lemma:

Lemma 3.15

For any $A \in \mathbb{C}^{m\times n}$,

$$\sigma_i\left(\begin{bmatrix} A \\ \delta I\end{bmatrix}\right) = \sqrt{\sigma_i^2(A)+\delta^2} \ge \max\left\{\sigma_i(A), \delta\right\}.$$

Proof. Simply notice that $\begin{bmatrix} A^* & \delta I\end{bmatrix}\begin{bmatrix} A \\ \delta I\end{bmatrix} = A^*A + \delta^2 I.$ □

Of course this approach comes at a price: we do not really compute the solution minimizing the error, but only an approximated version of this. However in practice, it is much wiser to compute a robust solution to an approximate problem, than a numerically unstable solution to an exact problem! The technique (3.14) above is commonly known in the numerical linear algebra community as the *Tikhonov regularization.*

3.2.5 Unitarily invariant norms

In this section, we are going to establish a link between some matrix norms and the singular value decomposition. The basic notions and definitions about the norms of vectors and matrices are given in Appendix B. We will just recall that a matrix norm (on $\mathbb{C}^{m\times n}$) is *unitarily invariant* if for every matrix $A \in \mathbb{C}^{m\times n}$, we have

$$\|A\| = \|U^*AV\| \qquad \text{if} \quad U, V \ \text{ are unitary.}$$

Proposition 3.16

The 2-norm and the Frobenius norm of $A \in \mathbb{C}^{m\times n}$:

$$\|A\|_2 := \sup_{\mathbf{x}\neq 0}\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}, \qquad \|A\|_F := \left[\sum_{i,j}|a_{i,j}|^2\right]^{1/2}$$

are unitarily invariant.

Proof. Observe first that the 2-norm of a vector is unitarily invariant:

$$\|U\mathbf{x}\|_2 = (\mathbf{x}^*U^*U\mathbf{x})^{1/2} = (\mathbf{x}^*\mathbf{x})^{1/2} = \|\mathbf{x}\|_2.$$

It immediately implies that

$$\|U^*AV\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|U^*AV\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{y}\neq 0} \frac{\|U^*A\mathbf{y}\|_2}{\|V^*\mathbf{y}\|_2} = \sup_{\mathbf{y}\neq 0} \frac{\|A\mathbf{y}\|_2}{\|\mathbf{y}\|_2},$$

where we put $y \coloneqq Vx$. For the Frobenius norm, note first that

$$\|A\|_F = \left[\sum_j \|\mathbf{a}_{:j}\|_2^2\right]^{1/2} = \left[\sum_i \|\mathbf{a}_{i:}\|_2^2\right]^{1/2}.$$

It immediately implies that $\|A\|_F = \|U^*A\|_F = \|AV\|_F$. □

This result allows us to express these norms solely in terms of the singular values of A:

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1 = \sigma_{\max},$$

$$\|A\|_F = \|\Sigma\|_F = \left[\sum_i \sigma_i^2\right]^{1/2}.$$

Proposition 3.17

If $A \in \mathbb{C}^{n\times n}$ is invertible, then

$$\|A^{-1}\|_2 = \sigma_n^{-1} = \sigma_{\min}^{-1}.$$

Proof. Straightforward from $\|A^{-1}\|_2 = \|V\Sigma^{-1}U^*\|_2 = \|\Sigma^{-1}\|_2 = \sigma_n^{-1}$. □

The following theorem (given without proof) due to John von Neumann states that the norms that can be expressed as a special function of singular values are exactly the unitarily invariant norms.

Theorem 3.18

A matrix norm $\|\cdot\|$ (on $\mathbb{C}^{m\times n}$) is unitarily invariant if and only if it is a *symmetric gauge function* ϕ of the singular values $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_s)$ ($s = \min\{m, n\}$), i.e., a function which is a norm on $\mathbb{R}^s$, and such that it is

- permutationally invariant: $\phi(P\boldsymbol{\sigma}) = \phi(\boldsymbol{\sigma})$ for all permutation P
- absolute: $\phi(D\boldsymbol{\sigma}) = \phi(\boldsymbol{\sigma})$ for all diagonal unitary matrices D (diagonal matrices whose diagonal elements are ±1).

Note, the last condition is mentioned for the completeness of the definition of symmetric gauge function but it is not necessary for this theorem because the singular values are always non-negative.

Typical examples of such norms are constructed from the vector $\boldsymbol{\sigma}$ of singular values:

$$\|A\| = \|\boldsymbol{\sigma}\|_p = \left[\sum_i |\sigma_i|^p\right]^{1/p}, \quad 1 \le p \le \infty.$$

3.2.6 The canonical angles

The singular value decomposition allows us to introduce the concept of *canonical (or principal) angles* between two subspaces leading to various applications in statistics and signal processing.

Starting from orthonormal bases of subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, given by the columns of S_1 and S_2 respectively:

$$S_1^* S_1 = I_{r_1}, \qquad S_2^* S_2 = I_{r_2},$$

we can construct other orthonormal bases of the same spaces. By Theorem 2.3, we have

$$\operatorname{Im}(S_i U_i) = \operatorname{Im}(S_i) \qquad \text{if} \qquad U_i \text{ is invertible, } i = 1, 2.$$

If, on top of that, the matrices U_i are unitary ($U_i^* U_i = U_i U_i^* = I_{r_i}$), then $S_i U_i$ are orthonormal bases as well:

$$U_i U_i^* = I_{r_i}, \quad \hat{S}_i = S_i U_i \qquad \Longrightarrow \qquad \hat{S}_i^* \hat{S}_i = I_{r_i}, \quad i = 1, 2.$$

This brings us to the following theorem.

Theorem 3.19

Given two subspaces $\mathcal{S}_i \subseteq \mathbb{C}^n$ ($i = 1, 2$). There exist orthonormal bases, given by the columns of $\hat{S}_i$ respectively, and satisfying

$$\hat{S}_1^* \hat{S}_2 = \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0_{r\times(r_2-r)} \\ 0 & & \sigma_r & \\ \hline & 0_{(r_1-r)\times r} & & 0_{(r_1-r)\times(r_2-r)} \end{array}\right], \qquad 1 \geq \sigma_1 \geq \ldots \geq \sigma_r > 0.$$

Proof. Fix some orthonormal bases: $\mathcal{S}_i = \operatorname{Im}(S_i)$ ($i = 1, 2$). Let

$$S_1^* S_2 = U_1 \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0_{r\times(r_2-r)} \\ 0 & & \sigma_r & \\ \hline & 0_{(r_1-r)\times r} & & 0_{(r_1-r)\times(r_2-r)} \end{array}\right] U_2^*$$

be the singular value decomposition of the matrix $S_1^* S_2$. Note that the columns of $\hat{S}_i := S_i U_i$ are orthonormal bases of the corresponding subspaces. Furthermore, $\hat{S}_1^* \hat{S}_2$ has the required form. It remains to show that the singular values σ_i are smaller or equal to 1. Note that σ_i is equal to the inner product of two orthonormal vectors (ith columns of $\hat{S}_1$ and $\hat{S}_2$). Thus, by the Schwarz inequality (Theorem 2.6), we have

$$\sigma_i = |\langle \hat{\mathbf{s}}_{1i}, \hat{\mathbf{s}}_{2i} \rangle| \leq \|\hat{\mathbf{s}}_{1i}\| \|\hat{\mathbf{s}}_{2i}\| = 1.$$

□

If the spaces $\mathcal{S}_i$ $(i = 1, 2)$, are of equal dimension k, then we will often include the singular values that are equal to zero in the notation:

$$\hat{S}_1^* \hat{S}_2 = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_k \end{bmatrix}, \qquad 1 \geq \sigma_1 \geq \ldots \geq \sigma_k \geq 0.$$

We define the *canonical angles* θ_i between $\mathcal{S}_1$ and $\mathcal{S}_2$ as

$$\cos(\theta_i) = \sigma_i.$$

If we put $\Theta = \operatorname{diag}\{\theta_1, \ldots, \theta_k\}$, we can write it simply as $\hat{S}_1^* \hat{S}_2 = \cos(\Theta)$. It is the matrix equivalent of the equation defining the angle between two vectors of norm 1.

We will now give the geometric interpretation of the canonical angles in the case of $\mathcal{S}_1$ and $\mathcal{S}_2$ being subspaces of dimension 2 in $\mathbb{R}^3$. We start with two orthonormal bases of $\mathcal{S}_1$ and $\mathcal{S}_2$ (Figure 3.2). After rotations of the axes, we have biorthogonal bases (Figure 3.3) and the matrix

$$S_1^* S_2 = \begin{bmatrix} 1 & 0 \\ 0 & \cos(\theta_2) \end{bmatrix}$$

is diagonal. In $\mathbb{R}^3$, biorthogonal bases of two subspaces of dimension 2 necessarily have a vector in common. Thus, θ_1 is equal to zero and its cosine is equal to 1. The other angle θ_2 is indeed the angle between the planes corresponding to $\mathcal{S}_1$ and $\mathcal{S}_2$.

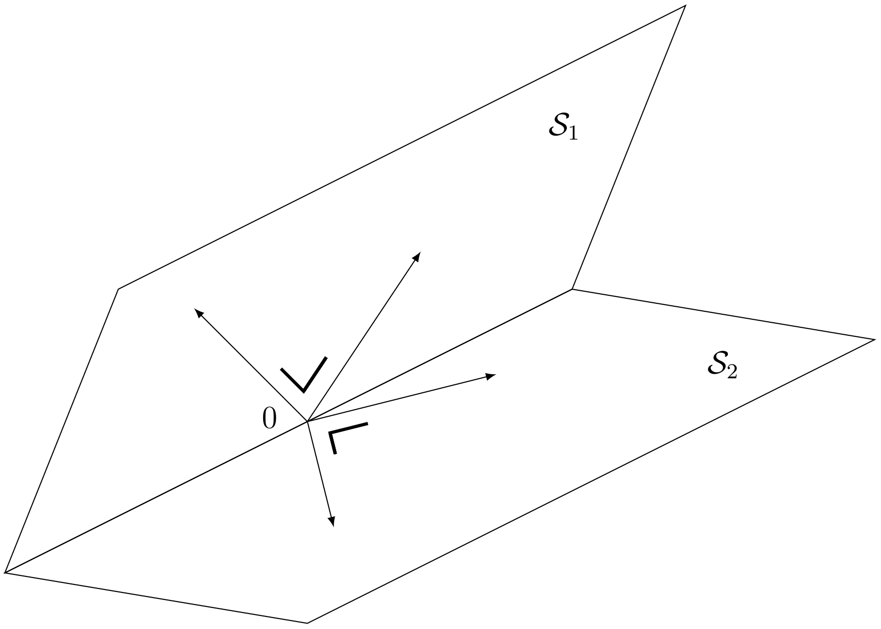

Figure 3.2: Arbitrary orthogonal bases of $\mathcal{S}_1$ and $\mathcal{S}_2$.

Exercise 3.3. *How can we define the notion of canonical angles between two spaces of different dimension?*

An application in signal processing that we will mention is based on the following lemma :

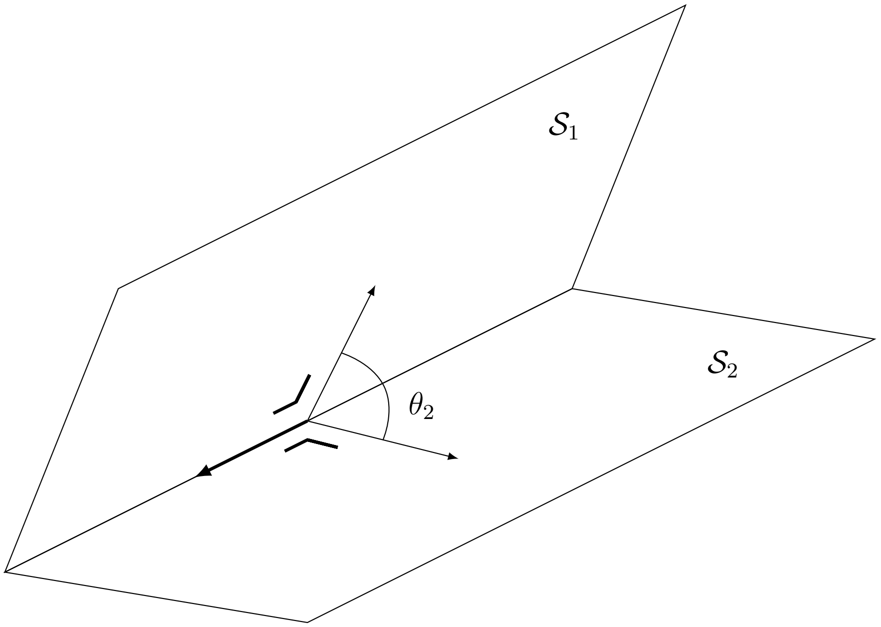

Figure 3.3: Biorthogonal bases of $\mathcal{S}_1$ and $\mathcal{S}_2$.

Lemma 3.20

Let $X \in \mathbb{C}^{m\times n}$ and $Y \in \mathbb{C}^{m\times n}$ be two matrices of rank n. There exist invertible transformations $T_x, T_y \in \mathbb{C}^{n\times n}$ such that $\hat{X} = XT_x$ and $\hat{Y} = YT_y$ satisfy

$$\left[\begin{array}{c} \hat{X}^* \\ \hline \hat{Y}^* \end{array}\right] \left[\begin{array}{c|c} \hat{X} & \hat{Y} \end{array}\right] = \left[\begin{array}{c|c} I_n & \Sigma \\ \hline \Sigma & I_n \end{array}\right],$$

where Σ is a diagonal real matrix with decreasing diagonal.

Proof. First, we perform the QR decomposition of the matrices X and Y:

$$X = Q_x R_x, \qquad Y = Q_y R_y.$$

Afterwards, we find the singular value decomposition of $Q_x^* Q_y$:

$$U_x^*(Q_x^* Q_y)U_y = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_k \end{bmatrix}.$$

It is not hard to verify that $T_x = R_x^{-1}U_x$ and $T_y = R_y^{-1}U_y$ provide the desired result. □

If the columns of X and Y are the samples of stochastic processes (or “signals”), then the QR factorization performs a “decorrelation” of all $\{\mathbf{x}_{:i}\}$ and all $\{\mathbf{y}_{:i}\}$. The $\{\sigma_i\}$ provided by the SVD measure the “principal correlations” between the signal spaces of $\{\mathbf{x}_{:i}\}$ and $\{\mathbf{y}_{:i}\}$.

3.2.7 Polar decomposition and the Procrustes problem

The singular value decomposition allows us to derive the *polar decomposition* of a square matrix.

Theorem 3.21

Every square matrix $A \in \mathbb{C}^{n\times n}$ admits a polar decomposition:

$$A = HQ,$$

where $H \in \mathbb{C}^{n\times n}$ is Hermitian positive semidefinite and $Q \in \mathbb{C}^{n\times n}$ is unitary ($Q^*Q = I_n$).

Proof. Starting from $A = U\Sigma V^*$, it suffices to take $H = U\Sigma U^*$ and $Q = UV^*$. □

We will see later that the matrix Q can be written as an imaginary exponential of a Hermitian matrix. So the polar decomposition of a matrix A can be written in the following form:

$$A = H_1 e^{iH_2} \qquad \text{where} \qquad H_i^* = H_i \in \mathbb{C}^{n\times n} \text{ are positive semidefinite.}$$

It can be seen as a natural extension of the polar decomposition of a scalar.

Exercise 3.4. *Show that for every positive semidefinite matrix H and every unitary matrix Q, we have*

$$|\mathrm{trace}(HQ)| \leq \mathrm{trace}(H).$$

Exercise 3.5. *How could you extend the polar decomposition to the case $m \neq n$?*

This decomposition arises in the *Procrustes problem.* This character from the Greek mythology used to force his victims to position themselves (in an optimal manner) on his bed and stretched them with a hammer if they were too small, or amputated the excess length with his axe if they were too large (he died after Theseus forced him to fit his own bed). The corresponding matrix problem is to find the optimal rotation Q that minimizes the error $\|AQ^\top - B\|_F^2$ (the rows of A represent the reference points of the victim, and the rows of B represent the reference points of the bed).

Exercise 3.6. *Show that the polar decomposition of $B^\top A$ leads to an optimal rotation Q that minimizes $\|AQ^\top - B\|_F^2$.*

3.2.8 Principal component analysis

In this application, usually known as PCA, we consider a set of points $\{\mathbf{x}_j\}_{j=1,\ldots,n}$ in $\mathbb{R}^m$ with components denoted by x_{ij} ($i = 1, \ldots, m$).

We consider the matrix $X_{m\times n}$ whose jth column contains the point $\mathbf{x}_j$, i.e., in the ith row and jth column we have x_{ij}. These points are assumed to be randomly generated according to a Gaussian distribution, and we would like to estimate the mean and covariance matrix of the process.

First, we will shift the values by means of

$$\hat{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{c}$$

in order to make the mean of $\hat{\mathbf{x}}_i$ equal to zero. Clearly, it suffices to set $\mathbf{c} = \frac{1}{n}X\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector of ones. Indeed, we have $\hat{X}\mathbf{1} = 0$ where

$$\hat{X} = X(I - n^{-1}\mathbf{1}\mathbf{1}^\top).$$

Afterwards, we perform a rotation

$$\tilde{\mathbf{x}}_i = U^\top \hat{\mathbf{x}}_i$$

in order to make the coordinates $\tilde{x}_{ij}$ and $\tilde{x}_{ik}$ of these new points mutually uncorrelated, that is, we would like to have $\sum_{i=1}^m \tilde{x}_{ij}\tilde{x}_{ik} = 0$ for every $j \neq k$. It immediately implies that $\tilde{X}\tilde{X}^\top$ is diagonal, where $\tilde{X} = U^\top \hat{X}$. We can find U from the SVD of $\hat{X} = U\Sigma V^\top$, since putting $\tilde{X} = U^\top \hat{X} = \Sigma V^\top$ leads to $\tilde{X}\tilde{X}^\top = \Sigma\Sigma^\top = D$ with $d_i = \sigma_i^2$ arranged in non-increasing order. This way, the autocorrelations $\sum_{i=1}^m \tilde{x}_{ij}^2 = d_j$ are ordered.

We will find more on this technique of extremely high practical importance in the seminar at the end of the semester…

3.3 Variational problems

In this section, we will show that the eigenvalues of a Hermitian matrix and the singular values of an arbitrary matrix can be seen as the stationary points of certain functions of $\mathbf{x}$. It will allow us to characterize the eigenvalues and the singular values as solutions to optimization problems.

For a Hermitian matrix $H \in \mathbb{C}^{n\times n}$, we define the *Rayleigh quotient* of a nonzero vector $\mathbf{x}$ as

$$R(\mathbf{x}) := \frac{\langle H\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\mathbf{x}^* H \mathbf{x}}{\mathbf{x}^*\mathbf{x}}, \qquad \mathbf{x} \neq 0 \in \mathbb{C}^n.$$

Theorem 3.22

The Rayleigh quotient of a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is real and satisfies

$$\lambda_{\min}(H) \leq R(\mathbf{x}) \leq \lambda_{\max}(H).$$

Proof. We start with the decomposition $H = U\Lambda U^*$ and put $\hat{\mathbf{x}} = U^*\mathbf{x}$. Substituting, we have

$$R(\mathbf{x}) = \frac{\sum_{i=1}^n \lambda_i |\hat{x}_i|^2}{\sum_{i=1}^n |\hat{x}_i|^2}.$$

The quotient is indeed real and it can be easily shown that the required inequalities are satisfied. □

By choosing $\hat{\mathbf{x}} = \mathbf{e}_1$ and $\hat{\mathbf{x}} = \mathbf{e}_n$, we observe that both bounds are attained. Thus, we have the following corollary (suppose $\lambda_1 \geq \ldots \geq \lambda_n$):

Corollary 3.23

$$\lambda_n = \min_{\mathbf{x}\neq 0} R(\mathbf{x}), \qquad \lambda_1 = \max_{\mathbf{x}\neq 0} R(\mathbf{x}).$$

In other words, the extremal eigenvalues of a Hermitian matrix are the stationary points of $R(\mathbf{x})$. The following theorem establishes this property for all eigenvalues of H.

Theorem 3.24

The stationary points of the Rayleigh quotient $R(\mathbf{x})$ are exactly the eigenvectors $\mathbf{x}_i$ of H. The corresponding values are the eigenvalues λ_i of H.

Proof. To simplify the presentation, we will present the proof only for the case of a real (and therefore symmetric) matrix H.

Since

$$\langle H\mathbf{x}, \mathbf{x}\rangle = \sum_{i,j} h_{ij}x_i x_j,$$

we have

$$\frac{\partial \langle H\mathbf{x}, \mathbf{x}\rangle}{\partial x_k} = 2\sum_{j=1}^{n} h_{kj}x_j = 2H_{k:}\mathbf{x}$$

where we have used $h_{kj} = h_{jk}$. By applying this formula to $H = I$, we obtain

$$\frac{\partial \langle \mathbf{x}, \mathbf{x}\rangle}{\partial x_k} = 2x_k.$$

Therefore,

$$\frac{\partial R(\mathbf{x})}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\frac{\langle H\mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}\right] = \frac{\langle \mathbf{x}, \mathbf{x}\rangle 2H_{k:}\mathbf{x} - \langle H\mathbf{x}, \mathbf{x}\rangle 2x_k}{\langle \mathbf{x}, \mathbf{x}\rangle^2}.$$

For $\mathbf{x} = \mathbf{x}_i$ where $\mathbf{x}_i$ is an eigenvector of H, we have

$$\frac{\partial R(\mathbf{x}_i)}{\partial x_k} = 2\frac{\langle \mathbf{x}_i, \mathbf{x}_i\rangle}{\langle \mathbf{x}_i, \mathbf{x}_i\rangle^2}[H_{k:}\mathbf{x}_i - \lambda_i x_k] = 0$$

and

$$R(\mathbf{x}_i) = \frac{\langle H\mathbf{x}_i, \mathbf{x}_i\rangle}{\langle \mathbf{x}_i, \mathbf{x}_i\rangle} = \lambda_i.$$

The gradient is given by

$$\nabla R(\mathbf{x}) = \left[\frac{\partial R(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial R(\mathbf{x})}{\partial x_n}\right]^\top = \frac{2[H\mathbf{x} - R(\mathbf{x})\mathbf{x}]}{\langle \mathbf{x}, \mathbf{x}\rangle}.$$

Clearly, it is equal to zero if and only if $\mathbf{x}$ is an eigenvector $\mathbf{x} = \mathbf{x}_i$ and $R(\mathbf{x}) = \lambda_i$. □

We are now able to derive a *variational definition of the eigenvalues of H.* Before we proceed, we require the following lemma:

Lemma 3.25

Let $\mathcal{S}_j \subseteq \mathbb{C}^n$ be a subspace of dimension j. Then, it holds that

$$\min_{\mathbf{x}\neq 0 \in \mathcal{S}_j} R(\mathbf{x}) \leq \lambda_j, \qquad \max_{\mathbf{x}\neq 0 \in \mathcal{S}_j} R(\mathbf{x}) \geq \lambda_{n-j+1}.$$

Proof. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the eigenvectors of H. Observe that

$$\hat{\mathcal{S}}_j = \operatorname{span}\{\mathbf{x}_j, \dots, \mathbf{x}_n\}$$

is a subspace of dimension $n-j+1$. Furthermore, it has a nonzero intersection with $\mathcal{S}_j$, otherwise $\hat{\mathcal{S}}_j + \mathcal{S}_j$ would be of dimension $(n-j+1)+j = n+1$. Let $\mathbf{x}_0 \in \mathcal{S}_j \cap \hat{\mathcal{S}}_j$. Since this vector belongs to $\hat{\mathcal{S}}_j$, we can represent it as

$$\mathbf{x}_0 = \sum_{i=j}^{n} \alpha_i \mathbf{x}_i \neq 0.$$

Observe now that its Rayleigh quotient is

$$R(\mathbf{x}_0) = \frac{\sum_{i=j}^{n} |\alpha_i|^2 \lambda_i}{\sum_{i=j}^{n} |\alpha_i|^2} \leq \lambda_j,$$

and thus

$$\min_{\mathbf{x} \neq 0 \in \mathcal{S}_j} R(\mathbf{x}) \leq \lambda_j.$$

The proof of the second inequality is dual. □

The following result known as the Courant–Fisher theorem is a direct consequence of this lemma.

Theorem 3.26: Courant–Fisher

For any Hermitian matrix $H \in \mathbb{C}^{n \times n}$, the Rayleigh quotient $R(\mathbf{x}) = \frac{\langle H\mathbf{x}, \mathbf{x} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$ satisfies

$$\begin{aligned} \lambda_j &= \max_{\mathcal{S}_j} \min_{\mathbf{x} \neq 0 \in \mathcal{S}_j} R(\mathbf{x}), \\ \lambda_{n-j+1} &= \min_{\mathcal{S}_j} \max_{\mathbf{x} \neq 0 \in \mathcal{S}_j} R(\mathbf{x}), \end{aligned}$$

where $\mathcal{S}_j \subseteq \mathbb{C}^n$ is a subspace of dimension j.

Proof. It suffices to show that the two bounds of the preceding lemma can be achieved. Indeed, by taking

$$\mathcal{S}_j = \operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_j\} \qquad [\text{resp. } \mathcal{S}_j = \operatorname{span}\{\mathbf{x}_{n-j+1}, \dots, \mathbf{x}_n\}],$$

we see that the equality is achieved due to $R(\mathbf{x}_i) = \lambda_i$. □

Now, we will switch to the case of an arbitrary matrix A. Recall that the matrices AA^* and A^*A are Hermitian, and thus, we can apply the preceding theorem to derive the following result about the singular values of A.

Theorem 3.27

The singular values of an arbitrary matrix $A \in \mathbb{C}^{m\times n}$ are given by

$$\begin{aligned}\sigma_j(A) &= \max_{\mathcal{S}_j} \min_{\mathbf{x}\neq 0\in\mathcal{S}_j} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2},\\ \sigma_{n-j+1}(A) &= \min_{\mathcal{S}_j} \max_{\mathbf{x}\neq 0\in\mathcal{S}_j} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2},\end{aligned}$$

where $\mathcal{S}_j \subseteq \mathbb{C}^n$ is a subspace of dimension j.

Proof. It is enough to apply the preceding theorem to the matrix A^*A or AA^*. □

This result finally leads us to a major application of the singular value decomposition.

Theorem 3.28: Low-rank approximation

Let $A \in \mathbb{C}^{m\times n}$ be a matrix of rank r. The best approximation (with respect to the matrix 2-norm) of A by a matrix $B \in \mathbb{C}^{m\times n}$ of rank $s < r$ satisfies

$$\min_{\operatorname{rank}(B)\leq s} \|A-B\|_2 = \sigma_{s+1}(A).$$

Proof. By Theorem 3.27, for a matrix B of rank s, we have

$$\sigma_{s+1}(A) \leq \max_{\substack{\mathbf{x}\neq 0\\ B\mathbf{x}=0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

since $\operatorname{Ker}(B)$ is a space of dimension $n-s$. Therefore,

$$\sigma_{s+1}(A) \leq \max_{\substack{\mathbf{x}\neq 0\\ B\mathbf{x}=0}} \frac{\|(A-B)\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \|A-B\|_2$$

and

$$\sigma_{s+1}(A) \leq \inf_{\operatorname{rank}(B)\leq s} \|A-B\|_2.$$

It is not hard to see that the infimum is reached by the following matrix

$$B = \sum_{i=1}^{s} \mathbf{u}_i\sigma_i\mathbf{v}_i^* \tag{3.15}$$

and thus

$$A-B = \sum_{i=s+1}^{r} \mathbf{u}_i\sigma_i\mathbf{v}_i^*.$$

Furthermore, it is clear that it is not possible to obtain a better bound by means of a matrix B of rank strictly smaller than s. □

The matrix (3.15) is always a solution of 3.28 and it is the unique solution only when $\sigma_{s+1} = 0$ (that is, if $s = r$).

Exercise 3.7. *Construct a matrix B, with* $\operatorname{rank}(B) \leq s < r$, *that is different from* (3.15) *and reaches the same bound on* $\|A - B\|_2$.

This exercise shows us that the solution of the theorem is far from being unique. The set of matrices satisfying the equality is quite complicated and in general poorly understood. By contrast, in the case of the Frobenius norm, we have more satisfying results:

Theorem 3.29: Eckart–Young

Let $A \in \mathbb{C}^{m\times n}$ be a matrix of rank r. The best approximation (with respect to the matrix Frobenius norm) of A by a matrix $B \in \mathbb{C}^{m\times n}$ of rank $s < r$ satisfies

$$\min_{\operatorname{rank}(B)\leq s} \|A - B\|_F^2 = \sigma_{s+1}^2 + \cdots + \sigma_r^2.$$

Proof. See [Wilkinson, 1965] or [Golub and Van Loan, 2012]. □

It turns out that the bound in the above theorem is attained by the same matrix (3.15), but this time, it is unique if and only if $\sigma_s > \sigma_{s+1}$.

These theorems help us to define the important concept of the “numerical rank” of a matrix. The problem of computing the rank of a matrix A is quite delicate, since the set of matrices of full rank is dense in the set of all matrices and every algorithm unavoidably introduces rounding errors, thus perturbing the matrix A. We will denote these perturbations by ΔA. Typically, a bound on ΔA is available:

$$\|\Delta A\|_2 \leq \epsilon c \|A\|_2.$$

Finally, the *numerical rank* of A is the minimal rank among the matrices $A + \Delta A$ such that

$$\|\Delta A\|_2 \leq \epsilon c \|A\|_2.$$

In order to find it, it suffices to discard the singular values $\sigma_i \leq \epsilon c \sigma_1$ and consider the corresponding approximation $A + \Delta A$. By Theorem 3.28, we can conclude that no matrix of rank smaller than $A + \Delta A$ can satisfy the condition. The matrix ΔA can be seen as the “noise” produced by the algorithm computing the rank. This discussion reveals the crucial importance of the singular value decomposition for the computation of the rank of a matrix.

Last but not least, as we will see in Chapter 4, the SVD can be computed robustly...

3.4 Recursive least squares

We here push further our analysis of the least squares problem discussed in Subsection 3.2.4.

Updating

Typically, when we seek a solution of the problem $\arg\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2$ where $A \in \mathbb{R}^{n\times k}$, we will not use the SVD as formalized in Subsection 3.2.4. Instead, we will rather first find a factorization

$$Q^\top [\, A \,|\, \mathbf{b} \,] = \left[\begin{array}{c|c} R_{k\times k} & \mathbf{r} \\ 0_{(n-k)\times k} & \boldsymbol{\rho} \end{array}\right],$$

where R is upper triangular, $\mathbf{r} \in \mathbb{R}^k$, $\boldsymbol{\rho} \in \mathbb{R}^{n-k}$, and $Q \in \mathbb{R}^{k\times k}$ is orthogonal. Afterwards, we reduce our problem to

$$\arg\min_{\mathbf{x}} \left\| \begin{bmatrix} R\mathbf{x} - \mathbf{r} \\ \boldsymbol{\rho} \end{bmatrix} \right\|_2 = \arg\min_{\mathbf{x}} \|R\mathbf{x} - \mathbf{r}\|_2 \tag{3.16}$$

Indeed, the latter problem is equivalent to the initial, since the 2-norm is unitarily invariant, and in this formulation, one does not need to compute the SVD.

If $\hat{Q} \in \mathbb{R}^{n\times k}$ is the matrix consisting of the first k columns of Q, then $\hat{Q}$ is an isometry (i.e., $\hat{Q}^\top \hat{Q} = I_k$) and $\hat{Q}R$ is the compact QR factorization of A (see Subsection 2.4.3). Then, assuming $\operatorname{rank}(A) = k$ (a hypothesis we will make all along this section), it follows from (3.16) that finding the least squares solution of $A\mathbf{x} = \mathbf{b}$ is equivalent to solving $R\mathbf{x} = \hat{Q}^\top \mathbf{b}$.

Now, in many situations, it happens that we need to resolve the original system *augmented with additional constraints* $\mathbf{a}_+\mathbf{x} = b_+$, which are added subsequently (after that a first least squares solution has been computed). Unfortunately, by combining this new equation, we lose the QR form that we had obtained. It turns out that one can recover a complete QR form without having to perform a completely new QR decomposition. Indeed, the new matrix A_{up} of the augmented system is almost triangularized by the following orthogonal transformation:

$$\begin{bmatrix} 1 & \\ & \hat{Q}^\top \end{bmatrix} A_{up} = \begin{bmatrix} 1 & \\ & \hat{Q}^\top \end{bmatrix} \begin{bmatrix} \mathbf{a}_+ \\ A \end{bmatrix} = \begin{bmatrix} \mathbf{a}_+ \\ R \end{bmatrix} = R_+. \tag{3.17}$$

The matrix R^+ is almost triangular since its structure is

$$R_+ = \begin{bmatrix} \times & \times & \cdots & \times \\ \times & \times & \cdots & \times \\ & \times & \cdots & \times \\ & & \ddots & \vdots \\ & & & \times \end{bmatrix}. \tag{3.18}$$

In order to triangularize R^+, it is enough to perform k Givens transformations of rows of R^+. Thus, the triangularization of R^+ requires at most

$$6\sum_{j=0}^{k-2}(k-j) = 6\sum_{j=2}^{k} j \cong 3k^2 \tag{3.19}$$

additions/multiplications. Indeed, at step j two rows of length $k-j$ are updated, and the update of one row takes 3 flops (two multiplications plus one addition).

Windowing

When we want new equations to have a larger influence on the system than the old ones, a common practice is to use exponential windowing. More precisely, the main idea is to solve the following problem:

$$\min_{\mathbf{x}} \|W(A\mathbf{x} - \mathbf{b})\|_2 = \min_{\mathbf{x}} \|(WA)\mathbf{x} - (W\mathbf{b})\|_2 \tag{3.20}$$

where

$$W = \operatorname{diag}\{\lambda^n, \lambda^{n-1}, \cdots, \lambda, 1\},$$

and $0 < \lambda < 1$ is the forgetting factor. For this new system, it is easy to see that if we have already obtained a triangularization

$$\hat{Q}^\top [\, WA \,|\, W\mathbf{b} \,] = [\, R \,|\, \mathbf{r} \,]$$

then we can obtain a QR factorization of the new system by finding Q_{up} such that

$$Q_{up}\left[\begin{array}{c|c} \lambda R & \lambda \mathbf{r} \\ \mathbf{a}_+ & b_+ \end{array}\right] = [\, R_+ \,|\, \mathbf{r}_+ \,] \tag{3.21}$$

using again k Givens transformations, just as before.

Downdating

We are now interested in a problem inverse to the updating. Let us consider a QR factorization of the matrix

$$A_{n\times k} = \begin{bmatrix} \mathbf{a}^\top \\ \hat{A} \end{bmatrix} = Q \begin{bmatrix} R_{k\times k} \\ 0_{(n-k)\times k} \end{bmatrix}$$

where $\mathbf{a} \in \mathbb{R}^k$, $Q \in \mathbb{R}^{n\times n}$ is orthogonal and R is upper triangular. We want to find a QR factorization of $\hat{A} \in \mathbb{R}^{(n-1)\times k}$ starting from Q and R.

Let $\mathbf{q}$ be the first row of Q and compute Givens rotations $G_1, \ldots, G_{n-1} \in \mathbb{R}^{n\times n}$ such that

$$G_1^\top \cdots G_{n-1}^\top \mathbf{q}^\top = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Note that $H = G_1^\top \cdots G_{n-1}^\top R$ is upper Hessenberg, i.e.,

$$H = G_1^\top \cdots G_{n-1}^\top R = \begin{bmatrix} \mathbf{v}^\top \\ \hat{R}_{k\times k} \\ 0_{(n-k-1)\times k} \end{bmatrix}$$

where $\hat{R}$ is upper triangular, since the Givens rotations G_i involve only consecutive rows from bottom to top. Also note that

$$QG_{n-1}\cdots G_1 = \begin{bmatrix} 1 & \\ & \hat{Q} \end{bmatrix}$$

where $\hat{Q} \in \mathbb{R}^{(n-1)\times(n-1)}$ is orthogonal. Finally,

$$A = \begin{bmatrix} \mathbf{a}^\top \\ \hat{A} \end{bmatrix} = (QG_{n-1}\cdots G_1)(G_1^\top \cdots G_{n-1}^\top R) = \left[\begin{array}{c|c} 1 & \\ \hline & \hat{Q} \end{array}\right] \left[\begin{array}{c} \mathbf{v}^\top \\ \hline \hat{R} \\ 0_{(n-k-1)\times k} \end{array}\right]$$

which gives the desired QR factorization

$$\hat{A} = \hat{Q} \begin{bmatrix} \hat{R} \\ 0_{(n-k-1)\times k} \end{bmatrix}.$$

It is not hard to check that the complexity of a “downdate” is $3k^2$ flops, coinciding with the complexity of an “update”.

Sliding window

We are now interested in a system whose solution $\mathbf{x} \in \mathbb{R}^k$ is slightly varying with time. Assume that at each moment of time, $\mathbf{x}$ satisfies a new equation $\mathbf{a}_i^\top \mathbf{x} = b_i$. As a first approximation, we can solve the system

$$\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}, \qquad \hat{A} \in \mathbb{R}^{\hat{n} \times k}, \quad \hat{\mathbf{a}}_{i:} = \mathbf{a}_i^\top, \quad \hat{b}_i = b_i \tag{3.22}$$

for the first $\hat{n}$ constraints. This way, we will find an average solution $\hat{\mathbf{x}}$ of the system for the first k steps. At the $(k+1)$st step, we can update the value of $\mathbf{x}$ by solving the new system

$$\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}, \qquad \tilde{A} \in \mathbb{R}^{\hat{n} \times k}, \quad \tilde{\mathbf{a}}_{i:} = \mathbf{a}_{i+1}^\top, \quad \tilde{b}_i = b_{i+1}.$$

Relying on the updating and downdating algorithms we can solve this system in $O(k^2)$: the complexity of the QR factorization of $\tilde{A}$ starting from the factorization of $\hat{A}$ is the complexity of an "update" followed by a "downdate", thus equal to $6k^2$ flops; the resolution of the modified triangular system requires around k^2 flops.

Chapter 4

Eigenvalues, eigenvectors and similarity transformations

In this chapter, we will treat the problem of eigenvalues of a matrix. We will address the similarity relations and the inherent canonical forms, as well as the computation and approximation of eigenvalues. We will also introduce the notion of eigenvector and invariant subspace.

The eigenvalue problem plays a central role in matrix theory. It arises in many ordinary or partial differential equations (with constant coefficients) problems, and helps to write the fundamental solutions for this type of equations.

4.1 Eigenvalues and eigenvectors of matrices

We define an eigenvalue $\lambda \in \mathbb{C}$ and the associated eigenvector $\mathbf{x} \in \mathbb{C}^n$ of a matrix $A \in \mathbb{C}^{n\times n}$ as the solutions of the equation

$$A\mathbf{x} = \lambda \mathbf{x}, \qquad \mathbf{x} \neq 0. \tag{4.1}$$

This is equivalent to

$$(\lambda I_n - A)\mathbf{x} = 0, \qquad \mathbf{x} \neq 0,$$

and allows us to define the eigenvalues λ as the roots of the polynomial

$$\det(\lambda I_n - A) = 0.$$

Now let us consider a transformation T of the space $\mathbb{C}^n$ (or $\mathbb{R}^n$) of vectors $\mathbf{x}$:

$$T\hat{\mathbf{x}} = \mathbf{x}, \qquad \det(T) \neq 0.$$

This transforms equation (4.1) into

$$T^{-1}AT\hat{\mathbf{x}} := A_T\hat{\mathbf{x}} = \lambda\hat{\mathbf{x}},$$

which leads to the following result:

Lemma 4.1

The eigenvalues of a matrix are invariant under similarity transformations.

Proof. For every invertible matrix T, we have

$$\det(I_n) = \det(T^{-1}T) = \det(T^{-1})\det(T) = 1.$$

Hence,

$$\det(\lambda I_n - A_T) = \det(T^{-1}(\lambda I_n - A)T) = \det(\lambda I_n - A).$$

This implies that A and A_T have the same eigenvalues. □

Since invertible matrices form a multiplicative group, the similarity transformations $A \mapsto TAT^{-1}$ define an equivalence class of matrices and every matrix A_T belonging to the similarity class of A has the same eigenvalues. We may ask whether the eigenvalues are the only invariants for this class. We will come back on this question later.

First, we restrict ourselves to unitary similarity transformations. This class of transformations already enables us to reduce every matrix $A \in \mathbb{C}^{n\times n}$ to a form revealing the eigenvalues of this matrix. This form is called the Schur form, and is upper triangular with the eigenvalues of A on the diagonal. We prove below the existence of the Schur form in a constructive way.

Since $A \in \mathbb{C}^{n\times n}$, its characteristic polynomial defined by

$$\chi(\lambda) = \det(\lambda I_n - A)$$

has at least one root λ_1 in the complex plane. Hence, there exists an eigenvector $\mathbf{u}_1$ such that

$$(\lambda_1 I_n - A)\mathbf{u}_1 = 0, \qquad \|\mathbf{u}_1\|_2 = 1.$$

It suffices to complete the vector $\mathbf{u}_1$ with an orthonormal basis (given by the columns of some matrix $U_1^\perp$) of its orthogonal complement to get an orthonormal basis of $\mathbb{C}^n$:

$$U_1 = \left[\, \mathbf{u}_1 \,\middle|\, U_1^\perp \,\right] \in \mathbb{C}^{n\times n}, \qquad U_1^* U_1 = U_1 U_1^* = I_n.$$

If we apply this transformation U_1 as a similarity transformation on A, we obtain

$$\hat{A} = U_1^* A U_1 = \left[\begin{array}{c|c} \lambda_1 & \mathbf{a}_1^\top \\ \hline 0 & \\ \vdots & A_2 \\ 0 & \end{array}\right], \tag{4.2}$$

because

$$U_1^* A \mathbf{u}_1 = \left[\begin{array}{c} \mathbf{u}_1^* \\ \hline (U_1^\perp)^* \end{array}\right] \mathbf{u}_1 \lambda_1 = \begin{bmatrix} \lambda_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We note here that

$$\det(\lambda I_n - A) = \det(\lambda I_n - \hat{A}) = (\lambda - \lambda_1)\det(\lambda I_{n-1} - A_2),$$

where the last identity corresponds to the expansion of $\lambda I_n - \hat{A}$ in cofactors. Thus, this unitary similarity transformation isolates a first eigenvalue of A on the diagonal $\hat{A}$, and defines a submatrix

A_2 having as eigenvalues all the other eigenvalues of A. We repeat the same construction with the matrix A_2. This gives

$$\hat{U}_2^* A_2 \hat{U}_2 = \left[\begin{array}{c|c} \lambda_2 & \mathbf{a}_2^\top \\ \hline 0 & \\ \vdots & A_3 \\ 0 & \end{array}\right]. \tag{4.3}$$

It remains to include (4.3) into (4.2) to obtain

$$U_2^* U_1^* A U_1 U_2 = \left[\begin{array}{cc|ccc} \lambda_1 & \times & \times & \cdots & \times \\ 0 & \lambda_2 & \times & \cdots & \times \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & A_3 & \\ 0 & 0 & & & \end{array}\right]$$

where

$$U_2 = \operatorname{diag}\{1, \hat{U}_2\}.$$

It is easy to see that a recursive application of this reasoning leads to the following theorem:

Theorem 4.2: Schur

Every matrix $A \in \mathbb{C}^{n\times n}$ can be upper triangularized under unitary similarity transformations:

$$U^* A U = \begin{bmatrix} \lambda_1 & \times & \cdots & \times \\ & \lambda_2 & \ddots & \vdots \\ & & \ddots & \times \\ 0 & & & \lambda_n \end{bmatrix} =: A_S,$$

where the diagonal of A_S consists of the eigenvalues of A.

Remark 4.1.

1. *If A is Hermitian, it is clear that A_S has the same property. Hence, A_S is diagonal and real. This form is canonical and shows that the only invariants of a Hermitian matrix under unitary similarity transformations are its eigenvalues.*

2. *The eigenvalues in the Schur form can be ordered. For example, we can order them in such a way that $|\lambda_i|$ is decreasing, and for eigenvalues with equal modulus, the phases are decreasing.*

3. *If $A \in \mathbb{R}^{n\times n}$, the eigenvalues and eigenvectors might nonetheless be complex. In this case, it is not hard to see that the complex conjugates of the eigenvalues and of the eigenvectors are also eigenvalues and eigenvectors of A.*

We would like to find out what is the most general class of matrices that can be diagonalized under unitary transformations.

Definition 4.3

A *normal* matrix is a square matrix $A \in \mathbb{C}^{n\times n}$ satisfying

$$AA^* = A^*A,$$

i.e., A commutes with its conjugate transpose.

Theorem 4.4

A matrix $A \in \mathbb{C}^{n\times n}$ is normal if and only if it is diagonalizable under unitary similarity transformations:

$$A = U\Lambda U^*, \qquad \Lambda = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\}.$$

Proof. If A is diagonalizable, then it is normal since

$$(U\Lambda U^*)\,(U\Lambda U^*)^* = (U\Lambda U^*)^*\,(U\Lambda U^*)\ .$$

If A is normal, then U^*AU is normal, hence it suffices to analyze the Schur form A_S of A: we want to show that

$$A_S A_S^* = A_S^* A_S \tag{4.4}$$

implies that A_S is diagonal. Therefore, we block-partition A_S in the following way:

$$A_S = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0_{n_2\times n_1} & A_{22} \end{array}\right]$$

where A_{ij} has dimensions $n_i \times n_j$ ($i = 1, 2,\ j = 1, 2$) and $n_1 + n_2 = n$. The block A_{21} is zero since A_S is upper triangular. Then (4.4) implies

$$A_{11}A_{11}^* + A_{12}A_{12}^* = A_{11}^*A_{11}.$$

Taking the trace, we obtain

$$\operatorname{trace}(A_{11}A_{11}^*) + \operatorname{trace}(A_{12}A_{12}^*) = \operatorname{trace}(A_{11}^*A_{11})$$

and since $\operatorname{trace}(X^*X) = \operatorname{trace}(XX^*) = \|X\|_F^2$, we have

$$\|A_{11}\|_F^2 + \|A_{12}\|_F^2 = \|A_{11}\|_F^2.$$

This implies $\|A_{12}\|_F = 0$, and thus $A_{12} = 0$. Since this is true for every $n_1 + n_2 = n$, we conclude that A_S is diagonal. □

Let us mention that, in general, the eigenvalues of a normal matrix are complex. If they are real, it is easy to see that the matrix is Hermitian. The unitary matrices are another subclass of normal matrices. They have their eigenvalues on the unit circle in the complex plane since $\Lambda\Lambda^* = \Lambda^*\Lambda = I_n$.

4.2 Invariant subspaces

An invariant subspace is a generalization of the concept of eigenvectors, for which an eigenvector is an invariant space of dimension one.

Definition 4.5

A subset $\mathcal{X} \subseteq \mathbb{C}^n$ is an *invariant subspace* under the operator $A \in \mathbb{C}^{n\times n}$ if

$$A\mathcal{X} \subseteq \mathcal{X}.$$

This implies that the vectors of the space $\mathcal{X}$ are mapped into $\mathcal{X}$ when we apply on them the linear transformation A. In other words, the application A leaves $\mathcal{X}$ invariant. Typical examples of invariant subspaces are the spaces generated by the eigenvectors:

$$\mathcal{X} = \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_i\} \quad \Longrightarrow \quad A\mathcal{X} = \operatorname{span}\{A\mathbf{x}_1, \ldots, A\mathbf{x}_i\} \subseteq \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_i\}$$

since $A\mathbf{x}_j = \lambda_j \mathbf{x}_j$.

However, this kind of invariant subspaces does not include the totality of invariant subspaces, as illustrated in the following example:

Example 4.1. *The matrix*

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

has only one eigenvalue $\lambda = 0$, and only one eigenvector satisfying

$$(A - 0I)\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 0.$$

However,

$$A \operatorname{Im} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \subseteq \operatorname{Im} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Exercise 4.1. *Show that the one-dimensional invariant subspaces of a matrix are those generated by its eigenvectors.*

We are now going to establish an important relation between the invariant subspaces and the triangular forms of a matrix:

Theorem 4.6

Let $\mathcal{X} \subseteq \mathbb{C}^n$ be a subspace of dimension k. Let $X \in \mathbb{C}^{n\times k}$ be such that the columns of X form a basis of $\mathcal{X}$, and let X_c be a completion of X such that $T := [\, X \,|\, X_c \,]$ is non-singular. Then the following three propositions are equivalent:

1. $A\mathcal{X} \subseteq \mathcal{X}$;
2. $AX = XA_{11}$;
3. $T^{-1}AT = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0_{(n-k)\times k} & A_{22} \end{array}\right]$;

(where $A_{11} \in \mathbb{C}^{k\times k}$, $A_{12} \in \mathbb{C}^{k\times(n-k)}$ and $A_{22} \in \mathbb{C}^{(n-k)\times(n-k)}$.)

Proof. $A\mathcal{X} \subseteq \mathcal{X}$ implies that AX contains only linear combinations of the columns of X, i.e., is equal to XA_{11}. Reversely, $AX = XA_{11}$ implies that $A\mathcal{X} = \mathrm{Im}(AX) = \mathrm{Im}(XA_{11}) \subseteq \mathrm{Im}(X) = \mathcal{X}$. Hence, the first two points are equivalent. We can rewrite the third point as

$$AT = T\begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix}, \qquad T = [\, X \,|\, X_c \,], \tag{4.5}$$

implying that $AX = XA_{11}$. Moreover, $AX = XA_{11}$ can be rewritten as

$$AX = [\, X \,|\, X_c \,]\begin{bmatrix} A_{11} \\ 0 \end{bmatrix}$$

for every matrix X_c. If we choose X_c such that T is invertible, we get equality (4.5) for appropriate matrices A_{12} and A_{22}. Hence, the last two points are equivalent. □

In the proof above, we have made no assumptions regarding the choice of the basis X and X_c. It is clear that if those bases are orthonormal, then T is unitary, and thus it is possible to draw a link with the Schur form. Indeed, if we partition the transformation U in the Schur form $U^*AU = A_S$ in a k-columns matrix U_1 and a $(n-k)$-columns matrix U_2, we get

$$[\, U_1 \,|\, U_2 \,]^* A \,[\, U_1 \,|\, U_2 \,] = \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix} = A_S, \qquad A_{11} \in \mathbb{C}^{k\times k}, \quad U_1 \in \mathbb{C}^{n\times k},$$

then the columns of U_1 provide an orthonormal basis of an invariant space of A.

This allows us to define a real-valued version of the Schur form of a matrix $A \in \mathbb{R}^{n\times n}$. For every *real* eigenvalue λ of A, it is clear that there exists at least one real eigenvector $\mathbf{u}$ associated to λ, since it is a solution to

$$(A - \lambda I_n)\mathbf{u} = 0. \tag{4.6}$$

On the other hand, for a *complex* eigenvalue $\lambda = \alpha + j\beta$ ($\beta \neq 0$) of A, any associated eigenvector $\mathbf{u} = \mathbf{x} + j\mathbf{y}$ will be complex. From the equation

$$[A - (\alpha + j\beta)I](\mathbf{x} + j\mathbf{y}) = 0,$$

it follows that

$$A[\,\mathbf{x} \,|\, \mathbf{y}\,] = [\,\mathbf{x} \,|\, \mathbf{y}\,]\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}. \tag{4.7}$$

Moreover, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ must be linearly independent: otherwise, there would exist an angle ϕ such that $\sin(\phi)\mathbf{x} + \cos(\phi)\mathbf{y} = 0$, and thus the eigenvector $(\mathbf{x} + j\mathbf{y})e^{j\phi}$ would be real, as well as the corresponding λ. This allows us to deduce the following lemma:

Lemma 4.7

A square real matrix always admits a *real* invariant subspace of dimension one or two.

Proof. If A has a real eigenvalue, then we refer to (4.6) to obtain $\mathcal{X} = \operatorname{span}\{\mathbf{u}\}$. If A has a complex eigenvalue $\alpha + j\beta$, we refer to (4.7) to obtain $\mathcal{X} = \operatorname{span}\{\mathbf{x}, \mathbf{y}\}$ which is an invariant subspace according to Theorem 4.6. □

For every invariant subspace, there exists, according to Theorem 4.6, a transformation U_1 which we choose real and orthogonal and such that

$$U_1^\top A U_1 = \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix}$$

where A_{11} is of size 1×1 or 2×2, depending on the case. By induction, we obtain the following theorem:

Theorem 4.8: Real Schur form

Every *real* matrix $A \in \mathbb{R}^{n\times n}$ can be almost triangularized under real orthogonal similarity transformations $U \in \mathbb{R}^{n\times n}$, and with blocks of dimensions 1×1 or 2×2 on its diagonal:

$$U^\top A U = \begin{bmatrix} A_{11} & \times & \cdots & \times \\ & A_{22} & \ddots & \vdots \\ & & \ddots & \times \\ & & & A_{kk} \end{bmatrix}, \qquad A_{ii} \in \mathbb{R}^{1\times 1} \cup \mathbb{R}^{2\times 2}.$$

Exercise 4.2. *Show that if $A \in \mathbb{R}^{n\times n}$ satisfies $A^\top A = AA^\top$ (i.e., A is normal), then its real Schur form is block-diagonal with blocks*

$$A_{ii} = \alpha_i \qquad \textit{or} \qquad A_{jj} = \begin{bmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{bmatrix}$$

for each real eigenvalue α_i and each complex eigenvalue $\alpha_j \pm j\beta_j$.

Exercise 4.3. *If $A \in \mathbb{R}^{n\times n}$ is anti-symmetric (i.e., $A = -A^\top$), then the real Schur form is block-diagonal with blocks*

$$A_{ii} = 0 \qquad \textit{or} \qquad A_{jj} = \begin{bmatrix} 0 & \beta_j \\ -\beta_j & 0 \end{bmatrix}.$$

4.3 Generalized eigenvalue problem

We now analyze the *generalized eigenvalue problem* and the related Schur decomposition. The motivation of this generalization is to study implicit differential equations. The simplest form of such equations is

$$B\dot{x}(t) = Ax(t) + f(t), \qquad x(0) := x_0.$$

We find this kind of differential equations for example in electrical circuit models obtained from kirchhoff's laws. The schema of the circuit represented in Figure 4.1 gives rise in the frequency domain (using Laplace transform) to the system of equations

$$(sB - A)x(s) = f(s)$$

where

$$A = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \qquad B = \begin{bmatrix} C_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$x(s) = [\, x_1(s) \,|\, \dots \,|\, x_k(s) \,]^\top, \qquad f(s) = [\, 0 \,|\, \dots \,|\, 0 \,|\, -e(s) \,]^\top.$$

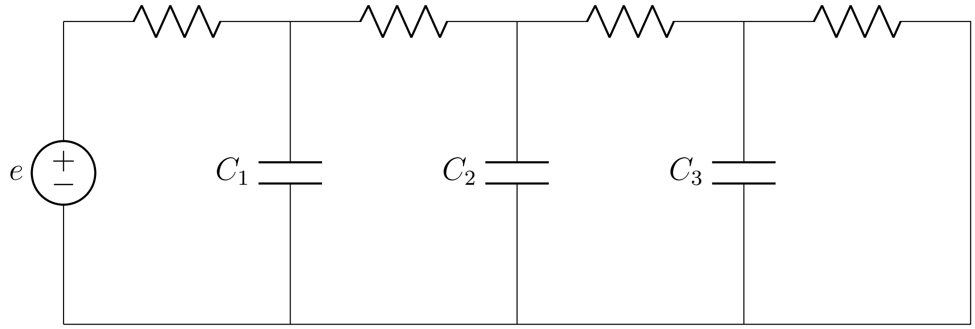

Figure 4.1: Electrical circuit with four resistances of 1 Ω and three capacitors.

When B is invertible, this implicit system can be reduced to an explicit system:

$$\dot{x}(t) = B^{-1}Ax(t) + B^{-1}f(t), \qquad x(0) := x_0.$$

The proper frequencies of this system are thus the eigenvalues of the matrix $B^{-1}A$, but they are also the roots of the following polynomial

$$\chi(s) := \det(sB - A)$$

since $\chi(s) = \det(B)\det(sI - B^{-1}A)$. Instead of computing the eigenvalues of $B^{-1}A$ by computing its Schur form, we derive the generalized Schur form as explained in the following theorem:

Theorem 4.9

Every pair of complex matrices $A, B \in \mathbb{C}^{n\times n}$ admits a triangularization under unitary transformations $Q, Z \in \mathbb{C}^{n\times n}$:

$$Q^*(sB - A)Z = sB_S - A_S$$

where $A_S = [\alpha_{ij}]$ and $B_S = [\beta_{ij}]$ are upper triangular. If $\chi(s)$ is not identically zero, then the roots of $\chi(s)$ are the quotients α_{ii}/β_{ii} for every $\beta_{ii} \neq 0$.

Proof. When B is invertible, the proof follows from the complex Schur form M_S of $M := B^{-1}A$, i.e., $M_S = U^*MU$ with M_S upper triangular. Let $Z := U$ and Q be such that $B_S := Q^*BZ$ is upper triangular (i.e., QB_S is a QR decomposition of BZ), and define $A_S := Q^*AZ$. Then $B_S^{-1}A_S = (Q^*BZ)^{-1}(Q^*AZ) = Z^*B^{-1}AZ = M_S$ is upper triangular. Hence, $A_S = B_S M_S$ is upper triangular as well.

When B is not invertible, we choose an infinitesimal perturbation B_ϵ and we compute the unitary matrices Q_ϵ and Z_ϵ. Taking the limit of $\epsilon \to 0$, the matrices Q_ϵ and Z_ϵ must converge toward unitary matrices since this is a compact set.

Moreover, the roots of $\det(sB - A)$ are the roots of $\det(sB_S - A_S)$ and thus the roots of the polynomials $s\beta_{ii} - \alpha_{ii}$, $i = 1, \ldots, n$. When $\beta_{ii} \neq 0$, this gives rise to roots equal to α_{ii}/β_{ii}. □

If $\det(B) = 0$, the polynomial $\chi(s) := \det(sB - A)$ has less than n finite roots. By a perturbation argument similar to the one used in the proof above, we observe that some roots are in fact *infinite*. For electrical circuits, this corresponds to short circuits.

Remark 4.2. *There also exists a real-valued version to the generalized Schur form, where B_S and A_S are block upper triangular with 1×1 or 2×2 diagonal blocks.*

4.4 Jordan canonical form

In this section, we show how to obtain the Jordan form of a matrix $A \in \mathbb{C}^{n\times n}$ under similarity transformations $T^{-1}AT$. This form is a canonical form, as we will see later. The proof of the theorem is constructive and starts with a block-diagonal form:

Theorem 4.10

Every matrix $A \in \mathbb{C}^{n\times n}$ admits a block-diagonal form under similarity transformations:

$$T^{-1}AT = A_S = \text{diag}\{A_{11}, \ldots, A_{kk}\} \tag{4.8}$$

where each block A_{ii} has only one eigenvalue (possibly with multiplicity larger than 1).

Proof. The proof is by induction on the dimension n of A. The case $n = 1$ is trivial. Assume that for every $m < n$ and $A \in \mathbb{C}^{m\times m}$, (4.8) holds.

Now let $A \in \mathbb{C}^{n\times n}$ and let λ be an eigenvalue of A. For each $k \in \{1, 2, \ldots\}$, let $\mathcal{X}_k = \text{Ker}((A - \lambda I)^k)$. Clearly, $\mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \mathcal{X}_3 \subseteq \ldots$ and $\dim(\mathcal{X}_1) \leq \dim(\mathcal{X}_2) \leq \dim(\mathcal{X}_3) \leq \ldots \leq n$. Hence, there is an ℓ such that $\dim(\mathcal{X}_\ell) = \dim(\mathcal{X}_k)$ for every $k \geq \ell$. Thus, $\text{Ker}((A - \lambda I)^\ell) = \text{Ker}((A - \lambda I)^k)$ for every $k \geq \ell$.

We will show that

$$\mathbb{C}^n = \operatorname{Ker}((A-\lambda I)^\ell) \oplus \operatorname{Im}((A-\lambda I)^\ell). \tag{4.9}$$

From Theorem 2.4, $\dim(\operatorname{Ker}((A-\lambda I)^\ell)) + \dim(\operatorname{Im}((A-\lambda I)^\ell)) = n$. Hence, it suffices to show that $\operatorname{Ker}((A-\lambda I)^\ell) \cap \operatorname{Im}((A-\lambda I)^\ell) = \{0\}$. Therefore, let $\mathbf{x}$ be in the intersection $\operatorname{Ker}((A-\lambda I)^\ell) \cap \operatorname{Im}((A-\lambda I)^\ell)$. Then $\mathbf{x} = (A-\lambda I)^\ell \mathbf{v}$ for some $\mathbf{v} \in \mathbb{C}^n$. Since $(A-\lambda I)^\ell \mathbf{x} = 0$, $\mathbf{v} \in \operatorname{Ker}((A-\lambda I)^{2\ell}) = \operatorname{Ker}((A-\lambda I)^\ell)$ by definition of ℓ. Hence, $\mathbf{x} = (A-\lambda I)^\ell \mathbf{v} = 0$. This proves (4.9).

Now, observe that $\operatorname{Ker}((A-\lambda I)^\ell)$ and $\operatorname{Im}((A-\lambda I)^\ell)$ are both invariant subspaces for A, since A commutes with $(A-\lambda I)^\ell$. Hence, if we let the columns of X be a basis of $\operatorname{Ker}((A-\lambda I)^\ell)$ and the columns of Y be a basis of $\operatorname{Im}((A-\lambda I)^\ell)$ and we define $T = [\,X \,|\, Y\,]$, we have

$$T^{-1}AT = \begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix}.$$

The block A_{11} cannot have another eigenvalue than λ: otherwise, there would exist an eigenvector of A: $\mathbf{x} \in \operatorname{Ker}((A-\lambda I)^\ell)$ associated to the eigenvalue $\mu \neq \lambda$. Then $(A-\lambda I)^\ell \mathbf{x} = (\mu-\lambda)^\ell \mathbf{x} \neq 0$, a contradiction with $\mathbf{x} \in \operatorname{Ker}((A-\lambda I)^\ell)$. Finally, we use the induction hypothesis to block-diagonalize A_{22} which has dimension strictly smaller than n. □

Lemma 4.11

Every matrix $A \in \mathbb{C}^{n\times n}$ satisfying $(A-\lambda I_n)^n = 0$ for some $\lambda \in \mathbb{C}$ can be transformed by similarity transformations into a block-diagonal form:

$$T^{-1}AT = \operatorname{diag}\{J_1(\lambda), \ldots, J_k(\lambda)\}$$

where each $J_i(\lambda) \in \mathbb{C}^{n_i \times n_i}$ is a *Jordan block*:

$$J_i(\lambda) = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix}. \tag{4.10}$$

Proof. The proof is constructive. By considering $A - \lambda I_n$ instead of A if necessary, we may assume without loss of generality that $\lambda = 0$. Let ℓ be the smallest positive integer such that $\operatorname{Ker}(A^\ell) = \mathbb{C}^n$. Choose vectors $\mathbf{x}_{0,1}, \ldots, \mathbf{x}_{0,n_0} \in \operatorname{Ker}(A^\ell)$ such that $\mathcal{V}_0 := \{\mathbf{x}_{0,1}, \ldots, \mathbf{x}_{0,n_0}\}$ is linearly independent and

$$\mathbb{C}^n = \operatorname{Ker}(A^{\ell-1}) \oplus \operatorname{span} \mathcal{V}_0 \tag{4.11}$$

is in *direct* sum.

We will show that $A\mathcal{V}_0$ is linearly independent and

$$\operatorname{Ker}(A^{\ell-2}) \cap \operatorname{span} A\mathcal{V}_0 = \{0\}.$$

Indeed, let $\mathbf{y} \in \operatorname{Ker}(A^{\ell-2})$ and $\alpha_1, \ldots, \alpha_{n_0} \in \mathbb{C}$ such that

$$\mathbf{y} + \sum_{k=1}^{n_0} \alpha_k A\mathbf{x}_{0,k} = 0.$$

We have to show that $\mathbf{y} = 0$ and $\alpha_1 = \ldots = \alpha_{n_0} = 0$. Therefore, apply $A^{\ell-2}$ on both sides of the above equation. This gives

$$\sum_{k=1}^{n_0} \alpha_k A^{\ell-1}\mathbf{x}_{0,k} = 0 \qquad \Longrightarrow \qquad \sum_{k=1}^{n_0} \alpha_k \mathbf{x}_{0,k} \in \operatorname{Ker}(A^{\ell-1}).$$

From (4.11) and $\mathcal{V}_0$ being linearly independent, this implies that $\alpha_1 = \ldots = \alpha_{n_0} = 0$, and thus $\mathbf{y} = 0$ as well.

Now, let $\mathcal{V}_1' = \{\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,n_1}\} \subseteq \operatorname{Ker}(A^{\ell-1})$ be a completion (possibly empty) of $A\mathcal{V}_0$ such that $\mathcal{V}_1 := A\mathcal{V}_0 \cup \mathcal{V}_1'$ is linearly independent and

$$\operatorname{Ker}(A^{\ell-1}) = \operatorname{Ker}(A^{\ell-2}) \oplus \operatorname{span} \mathcal{V}_1$$

is in *direct* sum. With an argument identical to the above one, we can show that $A\mathcal{V}_1$ is linearly independent and

$$\operatorname{Ker}(A^{\ell-3}) \cap \operatorname{span} A\mathcal{V}_1 = \{0\}.$$

Doing this recursively for $i = 1, \ldots, \ell - 1$, we may find sets $\mathcal{V}_i' \subseteq \operatorname{Ker}(A^{\ell-i})$ such that $\mathcal{V}_i := A\mathcal{V}_{i-1} \cup \mathcal{V}_i'$ is linearly independent and

$$\operatorname{Ker}(A^{\ell-i}) = \operatorname{Ker}(A^{\ell-i-1}) \oplus \operatorname{span} \mathcal{V}_i.$$

Obviously, for $i = \ell - 1$, we have $\operatorname{Ker}(A) = \operatorname{span} \mathcal{V}_{\ell-1}$.

The construction above shows that we may find a basis of $\mathbb{C}^n$ given by

$$\begin{aligned}
&\{A^{\ell-1}\mathbf{x}_{0,1}, A^{\ell-2}\mathbf{x}_{0,1}, \ldots, \mathbf{x}_{0,1}\} \cup \ldots \cup \{A^{\ell-1}\mathbf{x}_{0,n_0}, A^{\ell-2}\mathbf{x}_{0,n_0}, \ldots, \mathbf{x}_{0,n_0}\}\\
\cup &\{A^{\ell-2}\mathbf{x}_{1,1}, A^{\ell-3}\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,1}\} \cup \ldots \cup \{A^{\ell-2}\mathbf{x}_{1,n_1}, A^{\ell-3}\mathbf{x}_{1,n_1}, \ldots, \mathbf{x}_{1,n_1}\}\\
&\vdots\\
\cup &\{A\mathbf{x}_{\ell-2,1}, \mathbf{x}_{\ell-2,1}\} \cup \ldots \cup \{A\mathbf{x}_{\ell-2,n_{\ell-2}}, \mathbf{x}_{\ell-2,n_{\ell-2}}\}\\
\cup &\{\mathbf{x}_{\ell-1,1}\} \cup \ldots \cup \{\mathbf{x}_{\ell-1,n_{\ell-1}}\}.
\end{aligned}$$

If $T \in \mathbb{C}^{n\times n}$ is the matrix whose columns are given by the above basis, it is not hard to see that

$$AT = T \operatorname{diag}\Big\{ \underbrace{J_\ell, \ldots, J_\ell}_{n_0 \text{ times}}, \underbrace{J_{\ell-1}, \ldots, J_{\ell-1}}_{n_1 \text{ times}}, \ldots, \underbrace{J_2, \ldots, J_2}_{n_{\ell-2} \text{ times}}, \underbrace{0, \ldots, 0}_{n_{\ell-1} \text{ times}} \Big\}$$

where $J_i = J_i(0)$ is given by (4.10) with $\lambda = 0$. □

Theorem 4.12: Jordan canonical form

Every matrix $A \in \mathbb{C}^{n\times n}$ can be transformed by similarity transformations into a block-diagonal form, called the *Jordan form*:

$$T^{-1}AT = \operatorname{diag}\{J_1(\lambda_1), \ldots, J_k(\lambda_k)\}$$

where $J_i(\lambda)$ is given by (4.10).

Proof. Straightforward from Lemma 4.11 applied to the different blocks A_{ii} obtained in Theorem 4.10, which satisfy $(A_{ii} - \lambda_i I)^{\ell_i} = 0$ where λ_i and ℓ_i are as in (4.9). □

The following corollary expresses that the Jordan form is a *canonical form* under similarity transformations:

Corollary 4.13

Two matrices $A, B \in \mathbb{C}^{n \times n}$ are similar if and only if they have the same Jordan form.

Proof. If they have the same Jordan form, then

$$T_A^{-1} A T_A = J = T_B^{-1} B T_B$$

and clearly, A and B are similar:

$$A = T_A T_B^{-1} B T_B T_A^{-1}.$$

On the other hand, if $A = T^{-1} B T$, then for every $\lambda \in \mathbb{C}$,

$$A - \lambda I = T^{-1}(B - \lambda I) T$$

and more generally for every integer $k \geq 1$,

$$(A - \lambda I)^k = T^{-1}(B - \lambda I)^k T.$$

In particular, $\dim(\operatorname{Ker}((A - \lambda I)^k)) = \dim(\operatorname{Ker}((B - \lambda I)^k))$. Now observe that, for a given eigenvalue λ of A, the values of ℓ and $n_0, \ldots, n_{\ell-1}$ in the proof of Lemma 4.11 are uniquely determined by $\dim(\operatorname{Ker}((A - \lambda I)^k))$ for $k \geq 1$. This shows that A and B have the same Jordan form. □

Remark 4.3. *If we restrict ourselves to real similarity transformations for a matrix $A \in \mathbb{R}^{n \times n}$, then we cannot obtain a Jordan form if A has complex eigenvalues. In this case, we can however obtain a real-valued version of the Jordan form where the diagonal blocks have the following form:*

$$\left[\begin{array}{cc|cc|cc|cc|cc}
\alpha & \beta & 1 & 0 & & & & & & \\
-\beta & \alpha & 0 & 1 & & & & & & \\
\hline
& & \alpha & \beta & 1 & 0 & & & & \\
& & -\beta & \alpha & 0 & 1 & & & & \\
\hline
& & & & \multicolumn{2}{c|}{\ddots} & \multicolumn{2}{c|}{\ddots} & & \\
\hline
& & & & & & \alpha & \beta & 1 & 0 \\
& & & & & & -\beta & \alpha & 0 & 1 \\
\hline
& & & & & & & & \alpha & \beta \\
& & & & & & & & -\beta & \alpha
\end{array}\right]$$

for every pair of complex eigenvalues $\alpha \pm j\beta$.

4.5 Derivative of eigenvalues

In this section, we will analyze the eigenvalues of a matrix depending on a real variable t, i.e., of the matrix-valued function

$$A(t) \in \mathbb{C}^{n \times n}, \quad t \in \mathbb{R}.$$

We can show that the eigenvalues of $A(t)$ are differentiable with respect to t when the elements of $A(t)$ are also differentiable with respect to t, and when the eigenvalues of $A(0)$ are distinct. This result is based on the following theorem (for a proof, see, e.g., [Kato, 2013]):

Theorem 4.14

Let $A(t)$ be a complex $n \times n$ matrix whose elements $a_{ij}(t)$ are C^1 functions of $t \in \mathbb{R}$, and let λ_0 be an isolated eigenvalue of $A(0)$. Then there exist a neighborhood I_0 of $t = 0$ and a C^1 function $\lambda(t)$ on I_0 such that $\lambda(0) = \lambda_0$ and $\lambda(t)$ is an isolated eigenvalue of $A(t)$. The right-eigenvector $x(t) \in \mathbb{C}^{n\times 1}$ and left-eigenvector $y(t) \in \mathbb{C}^{1\times n}$ can be normalized in such a way that their elements are C^1 functions on I_0:

$$\begin{aligned} A(t)x(t) &= x(t)\lambda(t), \\ y(t)A(t) &= \lambda(t)y(t). \end{aligned}$$

If, in addition, every eigenvalue of $A(0)$ is isolated, then there exist matrices $\Lambda(t)$, $X(t)$ and $Y(t)$ whose elements are C^1 functions on a neighborhood I_0 of $t = 0$ such that

$$\begin{aligned} \Lambda(t) &= \operatorname{diag}\{\lambda_1(t), \ldots, \lambda_n(t)\}, \\ Y(t)X(t) &= I_n, \\ Y(t)A(t)X(t) &= \Lambda(t). \end{aligned} \tag{4.12}$$

This implies that the columns of $X(t)$ are the right-eigenvectors of $A(t)$, the rows of $Y(t)$ are the left-eigenvectors of $A(t)$ and those vectors are biorthogonal. Since these matrices are C^1 on I, we can develop them in Taylor series:

$$\begin{aligned} A(t) &= A_0 + tA_1 + O(t^2), \\ \Lambda(t) &= \Lambda_0 + t\Lambda_1 + O(t^2), \\ X(t) &= X_0 + tX_1 + O(t^2), \\ Y(t) &= Y_0 + tY_1 + O(t^2), \end{aligned}$$

and inject the constant and first-order terms in (4.12), which gives

$$\begin{aligned} &Y_0X_0 = I_n, \\ &Y_0A_0X_0 = \Lambda_0, \\ &Y_0X_1 + Y_1X_0 = 0, \\ &Y_1A_0X_0 + Y_0A_1X_0 + Y_0A_0X_1 = \Lambda_1. \end{aligned}$$

Using $A_0X_0 = X_0\Lambda_0$ and $Y_0A_0 = \Lambda_0Y_0$, we obtain

$$Y_1X_0\Lambda_0 + Y_0A_1X_0 + \Lambda_0Y_0X_1 = \Lambda_1,$$

but since the diagonal elements of $Y_0X_1 + Y_1X_0$ are zero, we have that

$$[Y_0A_1X_0]_{ii} = [\Lambda_1]_{ii}.$$

Since $[\Lambda_1]_{ii} = \frac{d}{dt}\lambda_i(t)\Big|_{t=0}$, we get a nice formula for this derivative:

$$\frac{d}{dt}\lambda_i(t)\Big|_{t=0} = y_iA_1x_i,$$

where x_i and y_i are the right- and left-eigenvectors of the eigenvalue $\lambda_i(0)$ of $A(0)$.

4.6 Computation of the eigenvalues

In this section, we describe the most popular method to compute the eigenvalues of an arbitrary matrix: the QR algorithm. Instead of presenting the method directly, we prefer to first introduce the power method, which is one of the oldest methods to compute the largest eigenvalue of a given matrix.

Let $A \in \mathbb{C}^{n\times n}$ and suppose that its eigenvalues are distinct. Thus A has n distinct eigenvectors uniquely defined up to a scalar coefficient (chosen such that the eigenvectors are normalized):

$$A\mathbf{x}_i = \lambda\mathbf{x}_i, \qquad \|\mathbf{x}_i\|_2 = 1.$$

Also suppose that the eigenvalues have different modulus ordered as follows:

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|.$$

The power method starts from an arbitrary nonzero vector $\mathbf{q}_{(0)}$ on which we apply the following iterative algorithm:

Algorithm 4.1: Power method

```
for $k = 1, 2, \ldots$ do
    $\mathbf{z}_{(k)} = A\mathbf{q}_{(k-1)}$;
    $\mathbf{q}_{(k)} = \mathbf{z}_{(k)}/\|\mathbf{z}_{(k)}\|_2$;
end
```

We prove now that the vectors $\mathbf{q}_{(k)}$ converge toward $\mathbf{x}_1$ up to the multiplication by some phase (i.e., by $e^{i\varphi}$ for some $\varphi \in \mathbb{R}$), and thus we also get the corresponding eigenvalue.

Theorem 4.15

If $A \in \mathbb{C}^{n\times n}$ has n eigenvalues with different modulus

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|,$$

and corresponding normalized eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, then Algorithm 4.1 produces vectors $\mathbf{q}_{(k)}$ converging to $\mathbf{x}_1$ (up to a phase multiplication):

$$\lim_{k\to\infty} \mathbf{q}_{(k)} e^{j\varphi_k} = \mathbf{x}_1, \tag{4.13}$$

as long as the vector $\mathbf{q}_{(0)}$ has a nonzero component in the direction of $\mathbf{x}_1$ (with respect to the basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$). Moreover,

$$\lim_{k\to\infty} \mathbf{q}_{(k)}^* A\mathbf{q}_{(k)} = \lambda_1.$$

Proof. If we decompose $\mathbf{q}_{(0)}$ in the basis of eigenvectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, i.e.,

$$\mathbf{q}_{(0)} = \sum_{i=1}^{n} c_i\mathbf{x}_i,$$

the assumption on $\mathbf{q}_{(0)}$ in the theorem implies that $c_1 \neq 0$. Hence, we have

$$A^k \mathbf{q}_{(0)} = \sum_{i=1}^{n} c_i \lambda_i^k \mathbf{x}_i = c_1 \lambda_1^k \left[\mathbf{x}_1 + \sum_{i=2}^{n} \left(\frac{\lambda_i}{\lambda_1} \right)^k \frac{c_i}{c_1} \mathbf{x}_i \right] = c_1 \lambda_1^k \left[\mathbf{x}_1 + O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \right]$$

On the other hand, Algorithm 4.1 provides

$$A^k \mathbf{q}_{(0)} = \mathbf{q}_{(k)} \prod_{i=1}^{k} \|\mathbf{z}_{(i)}\|_2$$

and thus the vectors $\mathbf{x}_1$ and $\mathbf{q}_{(k)}$ become more and more parallel when k grows. Since they are both normalized, it suffices to adjust the phase of $\mathbf{q}_{(k)}$ to obtain (4.13). This phase disappears in the expression $\mathbf{q}_{(k)}^* A \mathbf{q}_{(k)}$, and thus

$$\lim_{k \to \infty} \mathbf{q}_{(k)}^* A \mathbf{q}_{(k)} = \mathbf{x}_1^* A \mathbf{x}_1 = \lambda_1 .$$

□

Remark 4.4.

1. *A way to avoid the phase $e^{j\varphi_k}$ at step k is to choose a particular component of $\mathbf{q}_{(k)}$ and $\mathbf{x}_1$ to be real positive.*

2. *The theorem remains true if we impose the less restrictive condition*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n| \; ,$$

but the proof becomes more technical.

3. *The convergence of this algorithm is linear since the error behaves like*

$$\| \, \mathbf{q}_{(k)} e^{j\varphi_k} - \mathbf{x}_1 \| = O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) , \qquad \| \, \mathbf{q}_{(k)}^* A \mathbf{q}_{(k)} - \lambda_1 \| = O\left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) .$$

To be able to compute more than one eigenvalue of a matrix, we first present a simple generalization of the previous algorithm, which consists in using an orthonormal basis given by the columns of some matrix $Q_{(0)} \in \mathbb{C}^{n \times p}$ instead of a vector $\mathbf{q}_{(0)}$. If we start from an arbitrary matrix $Z_{(0)} \in \mathbb{C}^{n \times p}$, it suffices to compute a QR factorization of this matrix:

$$Z_{(0)} = Q_{(0)} R_{(0)}$$

to get $Q_{(0)}$ whose columns provide an orthonormal basis of $\mathrm{Im}(Z_{(0)})$. This procedure is repeated iteratively in the following algorithm:

Algorithm 4.2

```
for $k = 1, 2, \ldots$ do
        $Z_{(k)} = A Q_{(k-1)}$;
        $Q_{(k)} R_{(k)} = Z_{(k)}$;
end
```

The theorem related to this algorithm is the following:

Theorem 4.16

If the eigenvalues of $A \in \mathbb{C}^{n\times n}$ satisfy

$$|\lambda_1| \geq \ldots \geq |\lambda_p| > |\lambda_{p+1}| \geq \ldots \geq |\lambda_n|,$$

then for almost every isometry $Q_{(0)} \in \mathbb{C}^{n\times p}$ (i.e., $Q_{(0)}^* Q_{(0)} = I_p$), Algorithm 4.2 produces isometries $Q_{(k)} \in \mathbb{C}^{n\times p}$ that converge (up to some unitary transformation $U_{(k)} \in \mathbb{C}^{p\times p}$) to some isometry $X_1 \in \mathbb{C}^{n\times p}$, whose columns provide an orthonormal basis of the invariant space associated to the eigenvalues $\lambda_1, \ldots, \lambda_p$:

$$\lim_{k\to\infty} Q_{(k)} U_{(k)} = X_1. \tag{4.14}$$

As a consequence,

$$\lim_{k\to\infty} U_{(k)}^* Q_{(k)}^* A Q_{(k)} U_{(k)} = \hat{A}_{11} \tag{4.15}$$

where the eigenvalues of $\hat{A}_{11}$ are $\lambda_1, \ldots, \lambda_p$.

Proof. Let the columns of X_1 be an orthonormal basis of the invariant space associated to the eigenvalues $\lambda_1, \ldots, \lambda_p$, and let the columns of X_2 be a basis of the invariant space associated to the eigenvalues $\lambda_{p+1}, \ldots, \lambda_n$. Then

$$A = [\, X_1 \,|\, X_2 \,] \begin{bmatrix} A_{11} & \\ & A_{22} \end{bmatrix} [\, X_1 \,|\, X_2 \,]^{-1}$$

where the eigenvalues of A_{11} (resp. A_{22}) are $\lambda_1, \ldots, \lambda_p$ (resp. $\lambda_{p+1}, \ldots, \lambda_n$).

Decompose the columns of $Q_{(0)}$ in the basis defined by X_1 and X_2:

$$Q_{(0)} = X_1 C_1 + X_2 C_2, \qquad C_1 \in \mathbb{C}^{p\times p}, \quad C_2 \in \mathbb{C}^{(n-p)\times p}.$$

Then for almost every isometry $Q_{(0)} \in \mathbb{C}^{n\times p}$, $\det(C_1) \neq 0$ (i.e., C_1 is invertible). Hence, we get

$$A^k Q_{(0)} = X_1 A_{11}^k C_1 + X_2 A_{22}^k C_2.$$

On the other hand, Algorithm 4.2 produces the identity

$$A^k Q_{(0)} = Q_{(k)} R_{(k)} \cdots R_{(1)}. \tag{4.16}$$

Letting $\hat{R}_{(k)} = R_{(k)} \cdots R_{(1)}$, this gives

$$Q_{(k)} \hat{R}_{(k)} C_1^{-1} A_{11}^{-k} = X_1 + X_2 A_{22}^k C_2 C_1^{-1} A_{11}^{-k} = X_1 + E_{(k)}, \qquad \|E_{(k)}\| \in O\left(\left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k\right).$$

From the above equation, we can show that $\tilde{U}_{(k)} = \hat{R}_{(k)} C_1^{-1} A_{11}^{-k}$ "becomes more and more orthogonal", that is,

$$\lim_{k\to\infty} \tilde{U}_{(k)}^* \tilde{U}_{(k)} = \lim_{k\to\infty} (Q_{(k)} \tilde{U}_{(k)})^* (Q_{(k)} \tilde{U}_{(k)}) = \lim_{k\to\infty} (X_1 + E_{(k)})^* (X_1 + E_{(k)}) = X_1^* X_1 = I_p.$$

If we let $\tilde{U}_{(k)} = U_{(k)} T_{(k)}$ be the QR factorization of $\tilde{U}_{(k)}$ with $U_{(k)} \in \mathbb{C}^{n\times p}$ orthogonal and $T_{(k)} \in \mathbb{C}^{p\times p}$ upper triangular with positive diagonal, we find that

$$\lim_{k\to\infty} T^*_{(k)} T_{(k)} = \lim_{k\to\infty} \tilde{U}^*_{(k)} \tilde{U}_{(k)} = I_p,$$

implying that $\lim_{k\to\infty} T_{(k)} = I_p$ because of the upper triangular structure and positive diagonal of $T_{(k)}$ (can you prove this?). Thus $\tilde{U}_{(k)} - U_{(k)} \to 0$. This implies that

$$\lim_{k\to\infty} Q_{(k)} U_{(k)} = \lim_{k\to\infty} Q_{(k)} \tilde{U}_{(k)} = X_1,$$

proving thereby (4.14). □

Remark 4.5.

1. *We can show that the convergence is also linear in this case.*

2. *Because we converge to a matrix X_1 and a matrix A_{11}, we do not always have the eigenvalues and the eigenvectors themselves at the termination of the algorithm. Nevertheless, the reduction of the $n \times n$ problem to a $p \times p$ problem is already considered as a partial solution.*

We now have at our disposal all the necessary tools for describing the QR algorithm (aka. QR Francis' algorithm) [Francis, 1961]. We assume that the eigenvalues of A have different modulus and are ordered in the following way:

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|.$$

We start from a similarity transformation, given by the unitary matrix $Q_0 \in \mathbb{C}^{n\times n}$ (that we do not specify for the moment), and build the following algorithm:

Algorithm 4.3: QR algorithm

```
$A_0 = Q_0^* A Q_0$;
for $k = 1, 2, \ldots$ do
    $Q_k R_k = A_{k-1}$;        ($QR$ factorization of $A_{k-1}$)
    $A_k = R_k Q_k$;
end
```

At each step, we compute the QR factorization of an $n \times n$ matrix A_{k-1}, and we construct the new matrix A_k as the product of the factor matrices R and Q, in the reverse order of the factorization. To emphasize the link with Algorithm 4.2, we observe that each matrix A_k is similar to the previous matrix A_{k-1}:

$$A_k = Q_k^* (Q_k R_k) Q_k = Q_k^* A_{k-1} Q_k,$$

and thus they have the same eigenvalues.

By induction, we also have that

$$A_k = Q_k^* \cdots Q_0^* \, A \, Q_0 \cdots Q_k.$$

Moreover,

$$A_0^k = (Q_1 R_1)^k = Q_1 (R_1 Q_1)^{k-1} R_1 = Q_1 A_1^{k-1} R_1 = (Q_1 \cdots Q_k)\,(R_k \cdots R_1)$$

where the last identity is obtained by induction on k. Finally, from $A_0 = Q_0^* A Q_0$, we deduce that

$$A^k Q_0 = (Q_0 \cdots Q_k)\,(R_k \cdots R_1). \tag{4.17}$$

This identity reminds us (4.16). This link allows us to demonstrate the following theorem:

Theorem 4.17

The QR algorithm applied on a matrix $A \in \mathbb{C}^{n\times n}$ whose eigenvalues have different modulus converges to an upper triangular form:

$$\lim_{k\to\infty} A_k = \begin{bmatrix} \lambda_1 & \cdots & \times \\ & \ddots & \vdots \\ & & \lambda_n \end{bmatrix} = A_S$$

whose diagonal consists of the eigenvalues of A ordered with decreasing modulus.

Proof. We give only a sketch of the proof. For the details, we refer the reader to [Wilkinson, 1965]. We multiply equation (4.17) with the matrix $\left[\begin{smallmatrix} I_p \\ 0 \end{smallmatrix}\right]$ to obtain

$$A^k Q_{(0)} = Q_{(k)} R_{(k)} \tag{4.18}$$

where

$$Q_{(0)} = Q_0 \begin{bmatrix} I_p \\ 0 \end{bmatrix}, \qquad Q_{(k)} = Q_0 \cdots Q_k \begin{bmatrix} I_p \\ 0 \end{bmatrix}, \qquad R_{(k)} = \begin{bmatrix} I_p & 0 \end{bmatrix} R_k \cdots R_1 \begin{bmatrix} I_p \\ 0 \end{bmatrix}.$$

Indeed, the multiplication of the upper triangular matrices with $\left[\begin{smallmatrix} I_p \\ 0 \end{smallmatrix}\right]$ provides

$$R_k \cdots R_1 \begin{bmatrix} I_p \\ 0 \end{bmatrix} = \begin{bmatrix} I_p \\ 0 \end{bmatrix} \begin{bmatrix} I_p & 0 \end{bmatrix} R_k \cdots R_1 \begin{bmatrix} I_p \\ 0 \end{bmatrix}.$$

The equation (4.18) has now the same form as (4.16), and thus we get that $Q_{(k)}$ converges to an orthonormal basis of the invariant subspace corresponding to the first p eigenvalues of the matrix A. This is equivalent to saying that $\left[\begin{smallmatrix} I_p \\ 0 \end{smallmatrix}\right]$ is an invariant subspace of $A_k = Q_k^* \cdots Q_0^* \, A \, Q_0 \cdots Q_k$ for $k \to \infty$, or even that A_k converges to a matrix with the following form:

$$A_k \quad \longrightarrow \quad \begin{bmatrix} A_{11} & A_{12} \\ & A_{22} \end{bmatrix}$$

where the eigenvalues of $A_{11} \in \mathbb{C}^{p\times p}$ are the leading p eigenvalues of A. Since this is true for every p, we have proved the theorem. □

The convergence rate of this algorithm is linear (in the number of steps k), and each step requires $O(n^3)$ arithmetic operations, which is rather expensive. A way to significantly lower the cost of each step is to reduce the initial matrix A_0 to a condensed form that will be preserved throughout the QR algorithm.

Lemma 4.18: Hessenberg

It is always possible to compute a unitary matrix Q_0 such that $Q_0^*AQ_0$ is a Hessenberg matrix, i.e., $Q_0^*AQ_0$ has the form

$$Q_0^*AQ_0 = A_H = \begin{bmatrix} \times & \cdots & \cdots & \times \\ \times & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \times & \times \end{bmatrix}.$$

Proof. We build this Hessenberg matrix A_H with $n-2$ Householder transformations. Let $\hat{H}_1$ be a Householder transformation such that

$$\hat{H}_1^* \begin{bmatrix} a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix} = \begin{bmatrix} \times \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Then

$$H_1^*AH_1 = \left[\begin{array}{c|ccc} a_{11} & \times & \cdots & \times \\ \hline \times & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{array}\right] =: A_1$$

where $H_1 = \operatorname{diag}\{1, \hat{H}_1\}$. We then apply a similarity transformation $H_2 = \operatorname{diag}\{I_2, \hat{H}_2\}$ to the matrix A_1 that will preserve (because of the structure of H_2) the zeros in the first column of A_1. Moreover, we choose

$$\hat{H}_2^* \begin{bmatrix} a_{32}^{(1)} \\ a_{42}^{(1)} \\ \vdots \\ a_{n1}^{(1)} \end{bmatrix} = \begin{bmatrix} \times \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

implying that

$$H_2^*H_1^*AH_1H_2 = H_2^*A_1H_2 = \left[\begin{array}{cc|ccc} \times & \times & \times & \cdots & \times \\ \times & \times & \times & \cdots & \times \\ \hline 0 & \times & \times & \cdots & \times \\ 0 & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \cdots & \times \end{array}\right].$$

By induction, we finally obtain the desired form:

$$H_{n-2}^* \cdots H_1^* \, A \, H_1 \cdots H_{n-2} = Q_0^*AQ_0 = \begin{bmatrix} \times & \cdots & \cdots & \times \\ \times & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \times & \times \end{bmatrix}.$$

□

The complexity of this algorithm is in $O(n^3)$ [Golub and Van Loan, 2012], and this algorithm needs to be carried out only once. We now demonstrate that the Hessenberg form is preserved by the QR algorithm, lowering therefore significantly its complexity:

Lemma 4.19

The QR factorization $A_H = QR$ of a Hessenberg matrix $A_H \in \mathbb{C}^{n\times n}$ can be computed with $n-1$ Givens transformations. Moreover, the product RQ is again a Hessenberg matrix.

Proof. It suffices to use an appropriate Givens transformation between the first and second rows of A_H to get a zero at position $(2,1)$:

$$G_{1,2}\begin{bmatrix} \times & \times & \times & \cdots & \times \\ \times & \times & \times & \cdots & \times \\ & \times & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & \times & \times \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \cdots & \times \\ 0 & \times & \times & \cdots & \times \\ & \times & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & \times & \times \end{bmatrix}.$$

Then it suffices to transform the second and third rows to introduce a zero at position $(3,2)$, and so on. This is sketched in the following equation:

$$G_{n-1,n}\cdots G_{1,2}\,A_H = G_{n-1,n}\cdots G_{1,2}\begin{bmatrix} \times & \times & \times & \cdots & \times \\ \times & \times & \times & \cdots & \times \\ & \times & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & \times & \times \end{bmatrix} = \begin{bmatrix} \times & \times & \times & \cdots & \times \\ 0 & \times & \times & \cdots & \times \\ & 0 & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & 0 & \times \end{bmatrix},$$

Hence, we have

$$A_H = QR, \qquad Q = G_{1,2}^*\cdots G_{n-1,n}^*.$$

Moreover, the product RQ is again a Hessenberg matrix since

$$\begin{bmatrix} \times & \times & \times & \cdots & \times \\ 0_* & \times & \times & \cdots & \times \\ & 0_* & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & 0_* & \times \end{bmatrix} G_{1,2}^*\cdots G_{n-1,n}^* = \begin{bmatrix} \times & \times & \times & \cdots & \times \\ \times & \times & \times & \cdots & \times \\ & \times & \times & \cdots & \times \\ & & \ddots & \ddots & \vdots \\ & & & \times & \times \end{bmatrix}$$

"refills" the elements 0_*, as sketched above. □

At step i, applying the Givens transformation $G_{i,i+1}$ on the Hessenberg matrix updates two rows of length $n-i+1$, and updating one row takes 3 flops (two multiplications plus one addition). Then the complexity is $6(n-i+1)$ operations. The application of $G_{i,i+1}^*$ for the computation of Q takes the same number of flops $6(n-i+1)$. In total, this gives

$$2\sum_{i=1}^{n-1} 6(n-i+1) = 12\sum_{i=2}^{n} i \approx 6n^2 \quad \text{operations.}$$

This lowers thus the complexity of the base step of the QR algorithm from $O(n^3)$ to $O(n^2)$.

Another technique which brings a significant speed-up of the QR algorithm is the incorporation of shifts. Before explaining its effect, we give a formal description:

Algorithm 4.4: QR algorithm with shift

```
$A_0 = Q_0^* A Q_0$;    (Hessenberg)
for $k = 1, 2, \ldots$ do
    $Q_k R_k = A_{k-1} - \hat{\lambda}_k I$;    ($QR$ factorization of $A_{k-1} - \hat{\lambda}_k I$)
    $A_k = R_k Q_k + \hat{\lambda}_k I$;
end
```

where the shifts $\hat{\lambda}_k$ are computed at each step.

We first notice that the generated matrices A_k are still similar to each other:

$$Q_k^* A_{k-1} Q_k = R_k Q_k + \hat{\lambda}_k I = A_k,$$

and the shifts do not affect the Hessenberg form.

What is the purpose of these shifts? Suppose that $\hat{\lambda}$ is an estimation of order δ of λ_n, the smallest eigenvalue of A. Then the ratio between the shifted eigenvalues

$$|\lambda_1 - \hat{\lambda}| \geq \ldots \geq |\lambda_{n-1} - \hat{\lambda}| \gg |\lambda_n - \hat{\lambda}|$$

becomes more significant since $\lambda_n - \hat{\lambda} \approx \delta$.

We can show [Wilkinson, 1965] that this provides a quadratic rate of convergence to the shifted QR algorithm, as long as the shifts $\hat{\lambda}_k$ are adapted at each iteration in a specific way (not presented here). In very few steps, we will then converge to a matrix with the form

$$\left[\begin{array}{cccc|c} \times & \times & \cdots & \times & \times \\ \times & \times & \cdots & \times & \times \\ & \ddots & \ddots & \vdots & \vdots \\ & & \times & \times & \times \\ \hline & & & 0 & \lambda_n \end{array}\right]$$

and we restart the algorithm with the remaining (upper left) $(n-1) \times (n-1)$ submatrix, which is still Hessenberg. The total complexity of this algorithm is typically $15n^3$ operations to compute *all* the eigenvalues of an arbitrary $n \times n$ matrix. This result is quite surprising since the multiplication of two arbitrary matrices already requires $2n^3$ operations.

Remark 4.6.

1. *If a matrix is Hermitian, then the Hessenberg form is automatically tridiagonal.*

2. *A matrix of the type A^*A is always Hermitian, and its eigenvalues are the squared singular values of A. The tridiagonalization of A^*A can be computed implicitly by bidiagonalizing (see* Golub-Kahan-Lanczos bidiagonalization procedure*, not presented here) the matrix A:*

$$\begin{bmatrix} \times & \times & & & \\ & \times & \times & & \\ & & \ddots & \ddots & \\ & & & \times & \times \\ & & & & \times \end{bmatrix} = A_{Bi} = U_0^* A V_0.$$

Hence this implies that

$$A_{Bi}^* A_{Bi} = V_0^* A^* A V_0$$

is tridiagonal. This trick leads to an efficient QR-like algorithm for computing the singular values of an arbitrary matrix. The complexity of this algorithm is also in $O(n^3)$; see, e.g., [Golub and Van Loan, 2012].

3. *For a real matrix, it is possible to define a QR algorithm without dealing with any complex numbers. It suffices to implicitly compute a QR step of the following matrix product:*

$$(A - \lambda_i I)(A - \bar{\lambda}_i I) = A^2 - (\lambda_i + \bar{\lambda}_i)A + \lambda_i \bar{\lambda}_i I$$

which is clearly a real matrix. For the details, we refer to [Golub and Van Loan, 2012].

4.7 Estimation of the eigenvalues

In this section, we present different methods for estimating the eigenvalues of a given matrix. All these methods have the advantage of not requiring the computation of a form revealing the eigenvalues (for example the Schur form). Some of these estimation methods are useful in a framework wider than the simple evaluation of the eigenvalues, for example they are useful for the study of the convergence of recursive algorithms.

Definition 4.20

The *field of values* of a matrix $A \in \mathbb{C}^{n\times n}$ is the set

$$\mathcal{F}(A) := \left\{ \frac{\mathbf{x}^* A \mathbf{x}}{\mathbf{x}^* \mathbf{x}} \,\middle|\, \mathbf{x} \neq 0 \in \mathbb{C}^n \right\}.$$

Theorem 4.21: Hausdorff–Toeplitz

$\mathcal{F}(A)$ is a convex compact subset of $\mathbb{C}$ and it contains the eigenvalues of A.

Proof. If we rewrite the definition of $\mathcal{F}(A)$ as follows:

$$\mathcal{F}(A) := \{\mathbf{x}^* A \mathbf{x} \mid \|\mathbf{x}\|_2 = 1\},$$

we see that $\mathcal{F}(A)$ is the image of a compact set $\|\mathbf{x}\|_2 = 1$ by a continuous function $x \mapsto x^* A x$ and thus, $\mathcal{F}(A)$ is compact. It is also clear that $\lambda_i(A) \in \mathcal{F}(A)$ since for any eigenvector $\mathbf{x}_i$, $\|\mathbf{x}_i\| = 1$, associated to λ_i, we have $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$, and thus $\mathbf{x}_i^* A \mathbf{x}_i = \lambda_i$. Finally, for the proof of the convexity, we refer the reader to [Horn and Johnson, 1990]. □

Lemma 4.22

The field of values of a matrix is invariant under unitary similarity transformations:

$$\mathcal{F}(A) = \mathcal{F}(U^* A U), \qquad U^* U = U U^* = I.$$

Proof. If in the definition

$$\mathcal{F}(A) = \{\mathbf{x}^* A\mathbf{x} \mid \mathbf{x}^*\mathbf{x} = 1\},$$

we let $\mathbf{x} = U\mathbf{y}$, then

$$\mathcal{F}(A) = \{\mathbf{y}^* U^* AU\mathbf{y} \mid \mathbf{y}^* U^* U\mathbf{y} = \mathbf{y}^*\mathbf{y} = 1\} = \mathcal{F}(U^* AU).$$

□

Corollary 4.23

If A is normal, then $\mathcal{F}(A)$ is the convex hull of the eigenvalues of A.

Proof. Let $A = U^*\Lambda U$ be the (diagonal) Schur form of A. Then

$$\mathcal{F}(A) = \mathcal{F}(\Lambda) = \left\{\sum |y_i|^2\lambda_i \;\middle|\; \sum |y_i|^2 = 1\right\} = \left\{\sum \theta_i\lambda_i \;\middle|\; \theta_i \geq 0,\ \sum \theta_i = 1\right\}$$

where $\theta_i = |y_i|^2$. □

The Figure 4.2 illustrates the case of a real normal matrix.

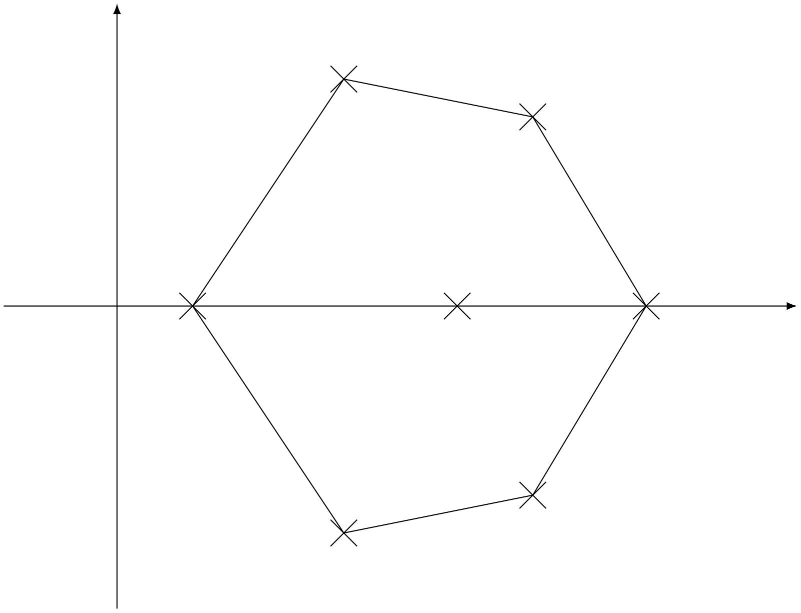

Figure 4.2: If A is normal, then $\mathcal{F}(A)$ is the convex hull of the eigenvalues of A.

Corollary 4.24

If A is Hermitian, then $\mathcal{F}(A)$ is the interval $[\lambda_{\min}(A), \lambda_{\max}(A)]$ on the real line.

For non-normal matrices, $\mathcal{F}(A)$ can be much larger than the convex hull of the eigenvalues of A (in a sense, $\mathcal{F}(A)$ can be seen as a measure of the normality of A). A way to evaluate the volume of $\mathcal{F}(A)$ consists in evaluating the extreme values of the real and imaginary parts of the points in $\mathcal{F}(A)$. This is stated formally in the following theorem:

Theorem 4.25

Let $\lambda \in \mathcal{F}(A)$. Then

$$\begin{aligned} \lambda_{\min}\left(\frac{A+A^*}{2}\right) \leq \Re(\lambda) \leq \lambda_{\max}\left(\frac{A+A^*}{2}\right), \\ \lambda_{\min}\left(\frac{A-A^*}{2j}\right) \leq \Im(\lambda) \leq \lambda_{\max}\left(\frac{A-A^*}{2j}\right). \end{aligned} \tag{4.19}$$

These intervals determine the smallest rectangle (in the complex plane) enclosing $\mathcal{F}(A)$.

Proof. Let $\lambda \in \mathcal{F}(A)$. We can always decompose a matrix A as follows:

$$A = \frac{A+A^*}{2} + j\frac{A-A^*}{2j} = H_1 + jH_2$$

where H_1 and H_2 are Hermitian. Hence,

$$\lambda = \mathbf{x}^* A \mathbf{x} = \mathbf{x}^* H_1 \mathbf{x} + j\mathbf{x}^* H_2 \mathbf{x} = \alpha + j\beta$$

where α and β are real and lie in (4.19). Since it is easy to find a vector $\mathbf{x}$ satisfying each extremum, the rectangle (4.19) is the smallest possible rectangle enclosing $\mathcal{F}(A)$. □

Corollary 4.26: Bendixon

The eigenvalues of a matrix $A \in \mathbb{C}^{n\times n}$ lie in the rectangle (4.19).

At first sight, the set $\mathcal{F}(A)$ is quite vague, and is thus not very useful to estimate the eigenvalues of A. However, it plays an important role in the convergence analysis of iterative methods to compute the eigenvalues of A. Suppose, for example, we have built a matrix

$$\hat{A} = Q^* A Q, \qquad Q \in \mathbb{C}^{n\times p}, \quad Q^* Q = I_p$$

where Q defines thus an orthonormal basis. What can we tell about the eigenvalues of $\hat{A}$ compared to the eigenvalues of A? It is easy to see that

$$\mathcal{F}(\hat{A}) = \{\mathbf{x}^* Q^* A Q \mathbf{x} \mid \mathbf{x}^*\mathbf{x} = 1\}.$$

If we set $\mathbf{y} = Q\mathbf{x}$, then

$$\mathcal{F}(\hat{A}) = \{\mathbf{y}^* A \mathbf{y} \mid \mathbf{y}^*\mathbf{y} = 1,\ \mathbf{y} \in \operatorname{Im}(Q)\} \subseteq \mathcal{F}(A).$$

Hence, we have that $\lambda_i(\hat{A}) \in \mathcal{F}(A)$. If we apply this to the matrices $\hat{A}_k := Q_{(k)}^* A Q_{(k)}$ of Algorithm 4.2 or to the matrices $\hat{A}_k := \begin{bmatrix} I_p & 0 \end{bmatrix} A_k \begin{bmatrix} I_p \\ 0 \end{bmatrix}$ of Algorithm 4.3, then we know that the values $\hat{A}_k$ converge to a subset of the values of A, and never leave the set $\mathcal{F}(A)$.

For a normal (or Hermitian) matrix, the convergence is monotone toward the exterior, that is, it converges to the vertices of $\mathcal{F}(A)$.

For arbitrary matrices, the convergence is not monotone anymore but the approximations $\lambda_i(\hat{A}_k)$ never leave the set $\mathcal{F}(A)$. The field of values $\mathcal{F}(A)$ plays a similar role in the convergence analysis of iterative methods for the solutions of linear systems, and in the stability analysis of dynamical systems.

Another method for locating the eigenvalues of a matrix is to use the Geršgorin disks. To introduce this method, we define the radii

$$r_p = \sum_{k \neq p} |a_{pk}|.$$

Theorem 4.27

Let $A \in \mathbb{C}^{n \times n}$. Then the eigenvalues of A lie in the union of the *Geršgorin disks*:

$$|z - a_{pp}| \leq r_p. \tag{4.20}$$

Proof. Let λ be an eigenvalue of A and let x_k be the components of a nonzero eigenvector $\mathbf{x}$ associated to λ. Then

$$\sum_{k=1}^{n} a_{jk} x_k = \lambda x_j, \qquad 1 \leq j \leq n.$$

Let p be the index of a maximal element of $\mathbf{x}$:

$$|x_p| = \max_j |x_j|,$$

Then we have the inequality

$$|\lambda - a_{pp}|\, |x_p| = \Big| \sum_{k \neq p} a_{pk} x_k \Big| \leq |x_p| \sum_{k \neq p} |a_{pk}|.$$

Hence we can write

$$|\lambda - a_{pp}| \leq r_p$$

because $x_p \neq 0$. This is valid for each eigenvalue. We can thus conclude that the eigenvalues will be in the union of the Geršgorin disks (4.20). □

Remark 4.7.

1. *In general, the Geršgorin disks are not disjoint. If they are, then we can show that each disk contains one and only one eigenvalue [Lancaster and Tismenetsky, 1985].*

2. *The same theorem applies to the matrix $A^\top$. The corresponding disks have the same centers but different radii. The choice of A or $A^\top$ can improve the localization of the eigenvalues.*

3. *A diagonal scaling DAD^{-1} is easy to compute and can also significantly reduce the radius of the disks, without changing their center.*

4.8 Estimation of the eigenvalues of a Hermitian matrix

For Hermitian matrices, we can expect more accurate bounds and properties than for arbitrary matrices. For example, we have already shown that the field of values reduces to an interval of the real line. But since the eigenvalues of a Hermitian matrix are stationary points of the Rayleigh quotient, we can probably go further. Therefore, we introduce the concept of constrained eigenvalues.

First of all, let us consider the constrained Rayleigh quotient:

$$R_A^{\mathcal{X}}(\mathbf{x}) = \frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}}, \qquad \mathbf{x} \neq 0 \in \mathcal{X} \tag{4.21}$$

where $\mathcal{X} \subseteq \mathbb{C}^n$ is a subspace of dimension $n-r$ (i.e., we impose r constraints). If $Q \in \mathbb{C}^{n\times(n-r)}$ provides an orthonormal basis of this space, i.e., $\mathcal{X} = \mathrm{Im}(Q)$ and $Q^*Q = I_{n-r}$, then

$$R_A^{Q}(\mathbf{x}) = \frac{\mathbf{x}^* A\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \frac{\mathbf{y}^* Q^* A Q\mathbf{y}}{\mathbf{y}^*\mathbf{y}} = R_{Q^*AQ}(\mathbf{y}).$$

This leads us to the following definition:

Definition 4.28

The *constrained eigenvalues* of a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ on a subspace $\mathcal{X} \subseteq \mathbb{C}^n$ are defined as the eigenvalues of $\hat{A} = Q^*AQ$ where the columns of Q provide an orthonormal basis of $\mathcal{X}$.

The matrix $\hat{A} = Q^*AQ$ is often called the *restriction* of A to the subspace $\mathrm{Im}(Q)$.

Theorem 4.29

Let $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_{n-r}\}$ be the (ordered with decreasing order) eigenvalues of the Hermitian matrices $A \in \mathbb{C}^{n\times n}$ and $\hat{A} = Q^*AQ$ respectively, and where $Q^*Q = I_{n-r}$. Then

$$\lambda_{i+r} \leq \mu_i \leq \lambda_i \ , \qquad i = 1, \ldots, n-r.$$

Proof. We use the variational properties of the eigenvalues of Hermitian matrices. Let $\mathcal{S}_i$ denote a subspace of dimension i. Then, due to Theorem 3.26, we have

$$\mu_i = \max_{\mathcal{S}_i \subseteq \mathcal{X}} \min_{\mathbf{x}\neq 0 \in \mathcal{S}_i} R(\mathbf{x}) \leq \max_{\mathcal{S}_i \subseteq \mathbb{C}^n} \min_{\mathbf{x}\neq 0 \in \mathcal{S}_i} R(\mathbf{x}) = \lambda_i.$$

To prove the second inequality, it suffices to consider the matrices $-A$ and $-\hat{A} = Q^*(-A)Q$. Since the order of the eigenvalues is reversed and becomes $\{-\lambda_n, \ldots, -\lambda_1\}$ and $\{-\mu_{n-r}, \ldots, -\mu_1\}$, we obtain in the same way: $-\lambda_{i+r} \geq -\mu_i$. □

Corollary 4.30

Let $\hat{A}$ be obtained from the Hermitian matrix $A \in \mathbb{C}^{n\times n}$ by removing one row and the corresponding column of A. Then the (ordered) eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and $\{\mu_1, \ldots, \mu_{n-1}\}$ of A and $\hat{A}$ are interlacing:

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \ldots \geq \mu_{n-1} \geq \lambda_n.$$

Proof. The proof (trivial) is left to the reader. □

An important application of this theorem arises in the case of tridiagonal matrices (see also Exercise 1.20):

Exercise 4.4. *The orthogonal polynomials defined by the recurrence*

$$\begin{cases} p_0(\lambda) & = & 1, \\ p_1(\lambda) & = & \lambda - \alpha_1, \\ p_i(\lambda) & = & (\lambda - \alpha_i)p_{i-1}(\lambda) - \beta_i^2 p_{i-2}(\lambda), \qquad i = 2, \ldots, n, \end{cases}$$

are the characteristic polynomials of the matrices

$$T_i = \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_i \\ & & \beta_i & \alpha_i \end{bmatrix}.$$

Show that the roots of two consecutive polynomials are interlacing. The interweaving is strict *if the β_j's are nonzero.*

These last results allow us to study the perturbation of Hermitian matrices.

Theorem 4.31

Let $\hat{A} = A + \Delta$ be a perturbed $n \times n$ Hermitian matrix. If the ordered eigenvalues of A, Δ and $\hat{A}$ are respectively $\{\lambda_1, \ldots, \lambda_n\}$, $\{\delta_1, \ldots, \delta_n\}$ and $\{\mu_1, \ldots, \mu_n\}$, then

$$\mu_{r+s-1} \le \lambda_r + \delta_s, \qquad r + s - 1 \le n.$$

Proof. Clearly, we have

$$R_{\hat{A}}(x) = R_A(x) + R_\Delta(x).$$

Let $\hat{\mathcal{S}}_j(H)$ be the subspace spanned by the j eigenvectors corresponding to the first j eigenvalues of a Hermitian matrix H. Then we define the subspace $\mathcal{X}$ as follows:

$$\mathcal{X} = [\hat{\mathcal{S}}_{r-1}(A) + \hat{\mathcal{S}}_{s-1}(\Delta)]^\perp = \hat{\mathcal{S}}_{r-1}^\perp(A) \cap \hat{\mathcal{S}}_{s-1}^\perp(\Delta).$$

The dimension of this space is at least equal to $n' = n - (r + s - 2)$. We can write

$$\begin{aligned} \min_{\mathcal{S}_{n'}} \max_{\mathbf{x} \neq 0 \in \mathcal{S}_{n'}} R_{\hat{A}}(\mathbf{x}) &\le \max_{\mathbf{x} \neq 0 \in \mathcal{X}} [R_A(\mathbf{x}) + R_\Delta(\mathbf{x})] \\ &\le \max_{\mathbf{x} \neq 0 \in \hat{\mathcal{S}}_{r-1}^\perp(A)} R_A(\mathbf{x}) + \max_{\mathbf{x} \neq 0 \in \hat{\mathcal{S}}_{s-1}^\perp(\Delta)} R_\Delta(\mathbf{x}) \end{aligned}$$

and thus, by Theorem 3.26, we get $\mu_{r+s-1} \le \lambda_r + \delta_s$. □

Corollary 4.32

If Δ is a small perturbation of the matrix $A \in \mathbb{C}^{n\times n}$, we have

$$\lambda_k + \delta_n \leq \mu_k \leq \lambda_k + \delta_1 \ , \qquad 1 \leq k \leq n.$$

Proof. It suffices to apply the previous theorem to the matrices $\hat{A} = A + \Delta$ and $A = \hat{A} - \Delta$ with $r = k$ and $s = 1$. □

This corollary is particularly useful for perturbations with small norm $\|\Delta\| = \delta$. Since $|\delta_i| \leq \|\Delta\|$, we have that

$$\lambda_k - \|\Delta\| \leq \mu_k \leq \lambda_k + \|\Delta\|,$$

which shows that the eigenvalues of a Hermitian matrix are shifted in the worst case with a shift equal the norm of the perturbation.

Exercise 4.5. *If $\hat{M}_{m\times n}$ is the matrix $M_{m\times n}$ in which we have replaced a row $\mathbf{m}_{i:}$ with a row of zeros, then the singular values $\{\sigma_1, \ldots, \sigma_n\}$ and $\{\hat{\sigma}_1, \ldots, \hat{\sigma}_n\}$ of M and $\hat{M}$ satisfy*

$$\sigma_k^2 - \|\mathbf{m}_{i:}\|_2^2 \leq \hat{\sigma}_k^2 \leq \sigma_k^2.$$

Hint. *It suffices to apply the previous theorem to the identity*

$$\hat{M}^\top \hat{M} = M^\top M - \mathbf{m}_{i:}^\top \mathbf{m}_{i:}.$$

4.9 Functions of matrices

Let

$$p(\lambda) = p_0 + p_1\lambda + \ldots + p_d\lambda^d$$

be a polynomial with real or complex coefficients (i.e., $p \in \mathbb{R}[\lambda]$ or $p \in \mathbb{C}[\lambda]$). Then for every matrix $A \in \mathbb{C}^{n\times n}$, we can define the *polynomial of this matrix* as follows

$$p(A) := p_0 I + p_1 A + \ldots + p_d A^d.$$

In this section, we analyze the eigenvalues of this kind of matrix functions.

Theorem 4.33

Let $A = TJT^{-1}$ be a Jordan decomposition of $A \in \mathbb{C}^{n\times n}$, then

$$p(A) = Tp(J)T^{-1}$$

and thus

$$\lambda_i\big(p(A)\big) = p\big(\lambda_i(A)\big).$$

Proof. It suffices to take the powers of A, $A^k = TJ^kT^{-1}$:

$$p(A) = p_0 I + p_1 TJT^{-1} + \ldots + p_d TJ^dT^{-1} = Tp(J)T^{-1}.$$

Since $p(J)$ is upper triangular, the diagonal elements give the eigenvalues $\lambda_i\Big(p(A)\Big)$. □

What could we tell further? Since $p(J)$ is block-diagonal, with the same structure as J, we can limit ourselves to the evaluation of the polynomial of a single Jordan block. We consider thus

$$A = J_k(\lambda_0) = \begin{bmatrix} \lambda_0 & 1 & & \\ & \lambda_0 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix}$$

with size $k \times k$, and we consider a simple example of polynomial: λ^d.

Lemma 4.34

The dth power of a Jordan block $J_k(\lambda_0)$ gives

$$\Big[J_k(\lambda_0)\Big]^d = \begin{bmatrix} \lambda_0^d & \binom{d}{1}\lambda_0^{d-1} & \binom{d}{2}\lambda_0^{d-2} & \cdots & \binom{d}{k-1}\lambda_0^{d-k+1} \\ & \lambda_0^d & \binom{d}{1}\lambda_0^{d-1} & \cdots & \binom{d}{k-2}\lambda_0^{d-k+2} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_0^d & \binom{d}{1}\lambda_0^{d-1} \\ & & & & \lambda_0^d \end{bmatrix}$$

where

$$\binom{d}{j} = \frac{d \cdot (d-1) \cdots (d-j+1)}{1 \cdot 2 \cdots j}.$$

Proof. The proof is by induction on the exponent d. The case $d = 0$ is trivial. For the induction step, observe that

$$\begin{bmatrix} \lambda_0^d & \binom{d}{1}\lambda_0^{d-1} & \cdots & \binom{d}{k-1}\lambda_0^{d-k+1} \\ & \lambda_0^d & \ddots & \vdots \\ & & \ddots & \binom{d}{1}\lambda_0^{d-1} \\ & & & \lambda_0^d \end{bmatrix} \begin{bmatrix} \lambda_0 & 1 & & \\ & \lambda_0 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix} = \begin{bmatrix} \lambda_0^{d+1} & \binom{d+1}{1}\lambda_0^{d} & \cdots & \binom{d+1}{k-1}\lambda_0^{d-k+2} \\ & \lambda_0^{d+1} & \ddots & \vdots \\ & & \ddots & \binom{d+1}{1}\lambda_0^{d} \\ & & & \lambda_0^{d+1} \end{bmatrix}$$

□

For general matrix functions, a result is given by the following theorem:

Theorem 4.35

Let $A \in \mathbb{C}^{n\times n}$. Then, for every analytic function $f(\lambda) = \sum_{i=0}^{\infty} a_i(\lambda - \lambda_0)^i$ whose radius of convergence R satisfies that $|\lambda_i - \lambda_0| < R$ for every eigenvalue λ_i of A, the quantity

$$f(A) = \sum_{i=0}^{\infty} a_i(A - \lambda_0 I_n)^i$$

is well defined.

Moreover, the derivatives $f^{(k)}(\lambda_i)$ $(1 \leq k \leq n-1)$ are defined at each eigenvalue λ_i of A, and we have that

$$f(A) = Tf(J)T^{-1}$$

where

$$f(J) = \operatorname{diag}\left\{f\left(J_{k_{i_j}}(\lambda_{i_j})\right)\right\}$$

and

$$f\Big(J_k(\lambda)\Big) = \begin{bmatrix} f(\lambda) & \frac{f'(\lambda)}{1!} & \frac{f^{(2)}(\lambda)}{2!} & \cdots & \frac{f^{(k-1)}(\lambda)}{(k-1)!} \\ & f(\lambda) & \frac{f'(\lambda)}{1!} & \cdots & \frac{f^{(k-2)}(\lambda)}{(k-2)!} \\ & & \ddots & \ddots & \vdots \\ & & & f(\lambda) & \frac{f'(\lambda)}{1!} \\ & & & & f(\lambda) \end{bmatrix}.$$

Proof. The proof is based on the Taylor expansion of $f(\lambda)$ around each eigenvalue λ_i of A, on some results concerning interpolation polynomials, and on the polynomial case considered above. For the details, we refer the reader to [Lancaster and Tismenetsky, 1985]. □

A straightforward corollary is that, for every analytic function $f(\lambda)$ with suitable domain of convergence, the eigenvalues of $f(A)$ are given by

$$\lambda_i\Big(f(A)\Big) = f\Big(\lambda_i(A)\Big).$$

This theorem also allows us to state that there exist polynomials $p(\lambda)$ for which $p(A)$ is identically zero. Indeed, consider the characteristic polynomial of A:

$$\chi(\lambda) = \det(\lambda I - A) = \det(\lambda I_n - J) = \prod(\lambda - \lambda_i)^{k_i}$$

where

$$f(J) = \operatorname{diag}\left\{f\left(J_{k_{i_j}}(\lambda_{i_j})\right)\right\}, \qquad k_i = \sum_{j=1}^{n_i} k_{i_j}$$

and n_i is the number of Jordan blocks with eigenvalue λ_i.

We also consider the minimal polynomial of A:

Definition 4.36

The *minimal polynomial* of $A \in \mathbb{C}^{n\times n}$ is the polynomial

$$m(\lambda) = \prod_i (\lambda - \lambda_i)^{k_i^*}$$

where

$$J = \operatorname{diag}\left\{J_{k_{i_j}}(\lambda_{i_j})\right\}, \qquad k_i^* = \max_{1\le j\le n_i} k_{i_j}$$

and n_i is the number of Jordan blocks with eigenvalue λ_i.

The following theorem is due to Cayley and Hamilton:

Theorem 4.37: Cayley–Hamilton

The characteristic polynomial of $A \in \mathbb{C}^{n\times n}$ satisfies $\chi(A) = 0$, and the monic polynomial $p(\lambda)$ with minimal degree satisfying $p(A) = 0$ is the minimal polynomial of A.

Proof. It is necessary and sufficient that these two polynomials satisfy $\chi^{(k)}(\lambda_i) = m^{(k)}(\lambda_i) = 0$ for every $0 \le k \le \max_j\{k_{i_j}\} - 1$. It is clear that these conditions are satisfied for $\chi(A)$ and $m(A)$, and also fix the minimal degree of $m(\lambda)$. □

An important matrix function is the exponential function:

$$\mathrm{e}^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$$

which converges for every $\lambda \in \mathbb{C}$. We can define e^A from the above Taylor expansion

$$\mathrm{e}^A = \sum_{i=0}^{\infty} \frac{A^i}{i!} = T\left(\sum_{i=0}^{\infty} \frac{J^i}{i!}\right)T^{-1}.$$

By Theorem 4.35, we know that e^A is well defined for every $A \in \mathbb{C}^{n\times n}$. Alternatively, we can see that this series converges for every A since

$$\left\|\sum_{i=0}^{\infty} \frac{A^i}{i!}\right\| \le \sum_{i=0}^{\infty} \frac{\|A\|^i}{i!} < \infty.$$

Proposition 4.38

If $Q \in \mathbb{C}^{n\times n}$ is unitary, then $Q = \mathrm{e}^{jH}$ for some Hermitian matrix $H \in \mathbb{C}^{n\times n}$.

Proof. A unitary matrix is normal, and thus has a diagonal Schur form:

$$Q = U\Lambda U^*.$$

Moreover $\Lambda^*\Lambda = I$ so that

$$\Lambda = \operatorname{diag}\{\mathrm{e}^{j\varphi_1}, \dots, \mathrm{e}^{j\varphi_n}\}.$$

Hence, $\Lambda = \mathrm{e}^{j\Phi}$

$$Q = U\mathrm{e}^{j\Phi}U^* = \mathrm{e}^{jU\Phi U^*} = \mathrm{e}^{jH}$$

where $H = U\Phi U^*$ is Hermitian. □

Proposition 4.39

If $A, B \in \mathbb{C}^{n\times n}$ satisfy $AB = BA$, then $\mathrm{e}^A\mathrm{e}^B = \mathrm{e}^{A+B}$.

Proof. The identity for A and B scalar

$$\left(\sum_{i=0}^{\infty}\frac{A^i}{i!}\right)\left(\sum_{i=0}^{\infty}\frac{B^i}{i!}\right) = \sum_{i=0}^{\infty}\frac{(A+B)^i}{i!}$$

is based on the relation

$$(A+B)^i = A^i + \binom{i}{1}A^{i-1}B + \binom{i}{2}A^{i-2}B^2\ldots + \binom{i}{i-1}AB^{i-1} + B^i.$$

If A and B are matrices, the identity remains valid as long as A and B commute for the multiplication. Hence the property follows. □

Exercise 4.6. *Show that*

$$\mathrm{e}^{J_k(\lambda_0)t} = \mathrm{e}^{(\lambda_0 I_k t + J_k(0)t)} = \mathrm{e}^{\lambda_0 t}\mathrm{e}^{J_k(0)t}.$$

Exercise 4.7. *From the Taylor expansion, show that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{At} = A\mathrm{e}^{At} = \mathrm{e}^{At}A.$$

This leads us finally to the application of differential equations:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + \mathbf{f}(t), \tag{4.22}$$

where the initial value $\mathbf{x}(0)$ is known.

If A admits a diagonal Jordan form

$$A = T\Lambda T^{-1},$$

then it is possible to uncouple equation (4.22) by introducing

$$\mathbf{x}(t) = T\hat{\mathbf{x}}(t), \qquad \mathbf{f}(t) = T\hat{\mathbf{f}}(t),$$

which gives

$$\dot{\hat{\mathbf{x}}}(t) = \Lambda\hat{\mathbf{x}}(t) + \hat{\mathbf{f}}(t)$$

where $\hat{\mathbf{x}}(0)$ is known. The solution of the uncoupled system is

$$\hat{x}_i(t) = \mathrm{e}^{\lambda_i t}\hat{x}_i(0) + \int_0^t \mathrm{e}^{\lambda_i(t-\tau)}\hat{f}_i(\tau)\,\mathrm{d}\tau$$

for each component $\hat{x}_i(t)$ of the vector $\hat{\mathbf{x}}(t)$ and each component $\hat{f}_i(t)$ of the vector $\hat{\mathbf{f}}(t)$.

Exercise 4.8. *Show that, in the general case, we have*

$$\mathbf{x}(t) = \mathrm{e}^{At}\mathbf{x}(0) + \int_0^t \mathrm{e}^{A(t-\tau)}\mathbf{f}(\tau)\,\mathrm{d}\tau.$$

Hint. *Use the differentiation of the exponential* e^{At} *to show that it satisfies equation* (4.22)*.*

This application clearly shows the importance of the Jordan blocks in the characterization of the type of solutions to (4.22). Indeed, take $\mathbf{f}(t) = 0$ (homogeneous case) and $A = J_k(\lambda_0) = \lambda_0 I_k + J_k(0)$. Then

$$e^{J_k(0)t} = I_k + \frac{J_k(0)t}{1!} + \frac{J_k(0)^2t^2}{2!} + \ldots + \frac{J_k(0)^{k-1}t^{k-1}}{(k-1)!}$$

since $J_k(0)^k = 0$ (equation of the characteristic polynomial). Thus, using Exercise 4.6,

$$\mathrm{e}^{At} = \mathrm{e}^{\lambda_0 t}\left[I_k + \frac{J_k(0)t}{1!} + \frac{J_k(0)^2t^2}{2!} + \ldots + \frac{J_k(0)^{k-1}t^{k-1}}{(k-1)!}\right].$$

The general solution is thus

$$\hat{\mathbf{x}}(t) = \mathrm{e}^{\lambda_0 t}\left[\hat{\mathbf{x}}(0) + \frac{J_k(0)t}{1!}\hat{\mathbf{x}}(0) + \frac{J_k(0)^2t^2}{2!}\hat{\mathbf{x}}(0) + \ldots + \frac{J_k(0)^{k-1}t^{k-1}}{(k-1)!}\hat{\mathbf{x}}(0)\right]$$

and we see that the Jordan blocks determine the "polynomial behavior" of the solutions. Those solutions usually present an overshoot behavior for λ_0 stable (see Figure 4.3).

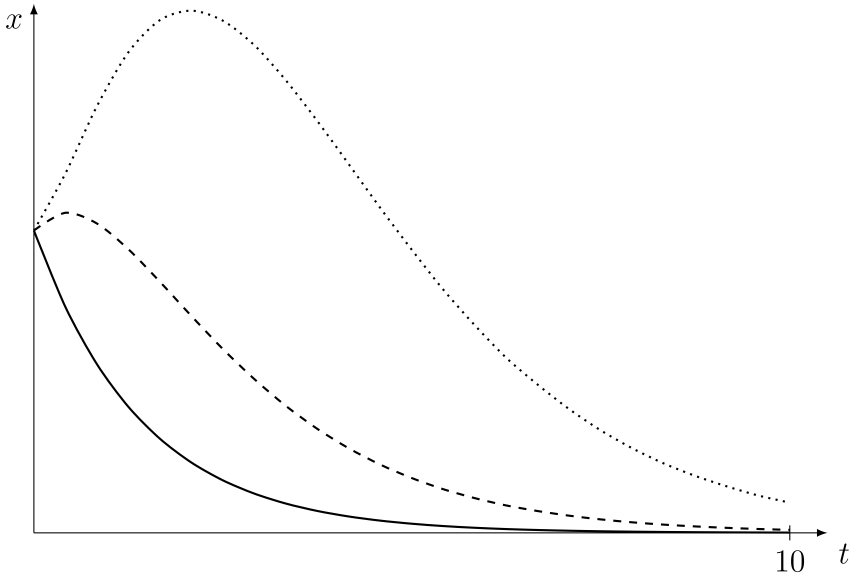

Figure 4.3: Curves: $\mathrm{e}^{\lambda_0 t}$ (plain), $\mathrm{e}^{\lambda_0 t}(1+t)$ (dashed) and $\mathrm{e}^{\lambda_0 t}(1+t+t^2)$ (dotted) for $\lambda_0 = -0.7$ and $0 \leq t \leq 10$. We observe an overshoot behavior.

Chapter 5

Inertia and stability of matrices

This chapter deals with the localization of the eigenvalues of a given matrix. First, we will study the sign of the eigenvalues of a Hermitian matrix; afterwards, we will consider the localization of the eigenvalues of an arbitrary matrix in the complex plane. More specifically, we will be able to count the number of eigenvalues having positive, zero or negative real part or the number of eigenvalues with the absolute value smaller, equal or greater than one. Problems of this type are fundamental for the analysis of stability of dynamical systems.

5.1 Congruences and inertia

In this section, we will introduce a group of transformations that acts on the set of Hermitian matrices and we will analyze its invariants.

Definition 5.1

Let $H_1, H_2 \in \mathbb{C}^{n\times n}$ be Hermitian matrices. We say that H_1 and H_2 are *congruent* if there is an invertible matrix $T \in \mathbb{C}^{n\times n}$ such that $H_1 = TH_2T^*$.

Observe that a matrix remains Hermitian under congruence transformations. Furthermore, these transformations form a multiplicative group, since T is invertible and the product of any two invertible matrices is invertible. Thus, we can ask what are the invariants of a Hermitian matrix under this group of transformations.

Definition 5.2

The *inertia* of a Hermitian matrix H is the triple

$$\mathrm{In}(H) = \{\pi_H, \nu_H, \delta_H\}$$

where

- π_H is the number of positive eigenvalues of H,
- ν_H is the number of negative eigenvalues of H,
- δ_H is the number of zero eigenvalues of H,

taking into account the multiplicities of the eigenvalues.

Typically, for a given transformation group, we aim at finding a *canonical form* that is as simple as possible. The following theorem brings us closer in this direction in the case of congruent transformations.

Theorem 5.3

Every Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is congruent to a diagonal matrix associated with its inertia: i.e., there is $T \in \mathbb{C}^{n\times n}$ invertible such that

$$THT^* = \operatorname{diag}\{I_{\pi_H}, -I_{\nu_H}, 0_{\delta_H}\}.$$

Proof. We start from the Schur form of the matrix H:

$$H = U\Lambda U^*$$

where Λ is diagonal and U is unitary. Since U is invertible, the matrices H and Λ are congruent. We can further assume that the diagonal values of Λ are ordered in the following way: first, all positive values appear, then all negative values, and finally all zero values, i.e.,

$$\Lambda = \operatorname{diag}\{\Lambda_+, \Lambda_-, \Lambda_0\}.$$

Now it remains to define

$$T = \operatorname{diag}\left\{\Lambda_+^{-\frac{1}{2}}, |\Lambda_-|^{-\frac{1}{2}}, I\right\} U^*$$

to get the desired result. □

Corollary 5.4

If two Hermitian matrices H_1 and H_2 have the same inertia, then they are congruent.

Proof. If $\operatorname{In}(H_1) = \operatorname{In}(H_2)$, then by Theorem 5.3, we conclude that there are invertible transformations T_1 and T_2 such that

$$T_1H_1T_1^* = T_2H_2T_2^*.$$

It immediately implies that H_1 and H_2 are congruent. □

The following theorem gives the converse of this corollary:

Theorem 5.5

If H_1 and H_2 are congruent Hermitian matrices, then they have the same inertia.

Proof. By Theorem 5.3, there are invertible transformations T_1 and T_2 such that

$$T_1H_1T_1^* = \operatorname{diag}\{I_{\pi_{H_1}}, -I_{\nu_{H_1}}, 0_{\delta_{H_1}}\} \tag{5.1}$$

and

$$T_2H_2T_2^* = \operatorname{diag}\{I_{\pi_{H_2}}, -I_{\nu_{H_2}}, 0_{\delta_{H_2}}\}\ . \tag{5.2}$$

The rank of a Hermitian matrix is equal to the number of its nonzero eigenvalues. Furthermore, the congruence transformations preserve the rank of a matrix, i.e., $\operatorname{rank}(H_1) = \operatorname{rank}(H_2)$. Thus, we have that $\delta_{H_1} = \delta_{H_2}$.

Hence, it suffices to show that $\pi_{H_1} = \pi_{H_2}$. Assume, for a contradiction, that $\pi_{H_1} > \pi_{H_2}$. Since H_1 and H_2 are congruent and due to the fact that their diagonal forms are (5.1) and (5.2) respectively, we obtain

$$\begin{bmatrix} I_{\pi_{H_1}} & & \\ & -I_{\nu_{H_1}} & \\ & & 0_\delta \end{bmatrix} = R^* \begin{bmatrix} I_{\pi_{H_2}} & & \\ & -I_{\nu_{H_2}} & \\ & & 0_\delta \end{bmatrix} R \tag{5.3}$$

for some invertible matrix R and $\delta = \delta_{H_1} = \delta_{H_2}$. Let us partition R as

$$R = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right],$$

where $R_{11} \in \mathbb{C}^{\pi_{H_2} \times \pi_{H_1}}$, and choose an appropriate vector

$$\mathbf{x} = \left[\begin{array}{c} \mathbf{x}_1 \\ \hline 0_{\nu_{H_1}+\delta} \end{array}\right]$$

such that $\mathbf{x}_1 \in \mathbb{C}^{\pi_{H_1}}$. Note that the product $R\mathbf{x}$ is equal to

$$\mathbf{y} = R\mathbf{x} = \left[\begin{array}{c} \mathbf{y}_1 \\ \hline \mathbf{y}_2 \end{array}\right], \qquad \mathbf{y}_1 = R_{11}\mathbf{x}_1 \in \mathbb{C}^{\pi_{H_2}}, \quad \mathbf{y}_2 = R_{21}\mathbf{x}_1 \in \mathbb{C}^{\nu_{H_2}+\delta}.$$

We define $\mathbf{y}_{21}, \mathbf{y}_{22}$ such that $\mathbf{y}_2 = \left[\begin{array}{c} \mathbf{y}_{21} \\ \hline \mathbf{y}_{22} \end{array}\right]$ with $\mathbf{y}_{21} \in \mathbb{C}^{\nu_{H_2}}$ and $\mathbf{y}_{22} \in \mathbb{C}^{\delta}$.
Since R_{11} has more columns than rows, we can further assume that $\mathbf{x}_1$ is a nonzero vector in $\operatorname{Ker}(R_{11})$, i.e., $\mathbf{y}_1 = 0$. Then, by multiplying (5.3) on the left by $\mathbf{x}^*$ and on the right by $\mathbf{x}$, we obtain $\mathbf{x}_1^*\mathbf{x}_1 = -\mathbf{y}_{21}^*\mathbf{y}_{21} \leq 0$, a contradiction with $\mathbf{x}_1 \neq 0$. If we assume that $\pi_{H_2} > \pi_{H_1}$, then we will get a similar contradiction with matrices H_1 and H_2 interchanged. □

Corollary 5.6

The diagonal matrix

$$\operatorname{diag}\{I_\pi, -I_\nu, 0_\delta\}, \qquad \pi + \nu + \delta = n,$$

is a *canonical form* of a Hermitian matrix $H \in \mathbb{C}^{n \times n}$ under congruence transformations, and its inertia $\operatorname{In} = \{\pi, \nu, \delta\}$ is the unique *invariant* under these transformations.

5.2 Cholesky and LDL factorization

To summarize, in the previous section, we have defined the invariants and the canonical form of a Hermitian matrix under congruence transformations. This canonical form also implies the existence of a well-known decomposition of positive definite matrices. Let us first recall their definition:

Definition 5.7

A Hermitian matrix H is *positive definite* if for every $\mathbf{x} \neq 0$,

$$\mathbf{x}^* H \mathbf{x} > 0.$$

We denote it by $H \succ 0$.

All eigenvalues of a positive definite matrix are positive, since

$$H\mathbf{x} = \lambda \mathbf{x}, \qquad \mathbf{x} \neq 0,$$

implies

$$\mathbf{x}^* H \mathbf{x} = \lambda \mathbf{x}^* \mathbf{x} > 0$$

and thus $\lambda > 0$. Therefore, we can write

$$H = TT^* \ .$$

This should be quite familiar, since every positive definite matrix has a *Cholesky factorization*

$$H = LL^*,$$

where L is a lower triangular matrix [Golub and Van Loan, 2012]. Furthermore, this decomposition can be computed in $O(n^3)$ operations and does not require the computation of the eigenvalues of H. In other words, it is possible to establish that all eigenvalues of a Hermitian matrix are positive without actually computing them. Immediately, the following natural question arises: is it possible to find the inertia of a Hermitian matrix without computing its eigenvalues? It will be the case if we define a symmetric Cholesky-type decomposition for an arbitrary Hermitian matrix; this is the so-called *LDL factorization* of a Hermitian matrix:

$$H = L\text{diag}\,\{D_{11}, \ldots, D_{kk}\} L^*,$$

where the diagonal blocks D_{ii} have dimensions 1×1 or 2×2. Moreover, the inertia of the 1×1 blocks is trivial, and the inertia of 2×2 blocks is $\{1, 1, 0\}$ by construction. Thus, this decomposition, which is merely a generalization of the Cholesky decomposition for arbitrary Hermitian matrices, indeed allows us to compute the inertia of a Hermitian matrix in a simple manner.

Only a sketch of the algorithm is presented (see Algorithm 5.1). Each time we move to step 2, we construct a 1×1 block. Thus, its sign contributes to the inertia. On the other hand, each time we move to step 3, we build a 2×2 block and its inertia is $\{1, 1, 0\}$. Indeed, each 2×2 block is Hermitian and its maximal elements are in positions $(1, 2)$ and $(2, 1)$, and thus, its determinant is negative which implies that this block has one positive eigenvalue and one negative eigenvalue (see also Proposition 5.9 below). At step 4, only a zero or empty matrix remains, finalizing the construction.

This algorithm has many variants that try to reduce the number of comparisons to find the maximum m^* of the submatrix M. Results of this type and a deeper analysis of this decomposition can be found, e.g., in [Golub and Van Loan, 2012].

In summary, we have shown that we can compute the inertia of a Hermitian matrix without computing its eigenvalues. The hardness of this computation is of the order $O(n^3)$.

Algorithm 5.1

```
Step 0: Set $k := 1$;
Step 1: Set $M := H[k:n, k:n]$;
        Find $(m^*, i^*, j^*) := \max_i \max_j |M_{ij}|$;
        – if $m^* = 0$: go to step 4;
        – if $i^* \neq j^*$: go to step 3;
Step 2: Permute row $i^* \leftrightarrow$ row 1; Permute column $i^* \leftrightarrow$ column 1;
        Eliminate $M[2:n, 1]$; Eliminate $M[1, 2:n]$;
        This gives
            $$LPMP^\top L^* = \left[\begin{array}{c|ccc} m^* & 0 & \cdots & 0 \\ \hline 0 & \times & \cdots & \times \\ \vdots & \vdots & & \vdots \\ 0 & \times & \cdots & \times \end{array}\right].$$
        Set $k := k + 1$;
        – if $k < n$: go to step 1;
        – else: go to step 4;
Step 3: Permute row $i^* \leftrightarrow$ row 1; Permute column $i^* \leftrightarrow$ column 1;
        Permute row $j^* \leftrightarrow$ row 2; Permute column $j^* \leftrightarrow$ column 2;
        Eliminate $M[3:n, 1:2]$; Eliminate $M[1:2, 3:n]$;
        This gives
            $$LPMP^\top L^* = \left[\begin{array}{cc|ccc} \times & m^* & 0 & \cdots & 0 \\ m^* & \times & 0 & \cdots & 0 \\ \hline 0 & 0 & \times & \cdots & \times \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \times & \cdots & \times \end{array}\right].$$
        Set $k := k + 2$;
        – if $k < n$: go to step 1;
        – else: go to step 4;
Step 4: End.
```

Theorem 5.8

Algorithm 5.1 computes a decomposition $P^\top HP = LDL^*$ of a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ in $\mathcal{O}(n^3)$ operations. The matrix D has diagonal blocks of dimension one or two. In the latter case, the inertia of these blocks is $(1,1,0)$.

The inertia of the 2×2 blocks can also be deduced from the following result:

Proposition 5.9

Let H be a Hermitian matrix. If the largest (in modulus) entry of H is not on the diagonal, then H has a negative eigenvalue.

Proof. Suppose that the (i,j)th element is the largest in modulus. Then the vector $\mathbf{x} = \mathbf{e}_i - \mathbf{e}_j$ satisfies $\mathbf{x}^* H\mathbf{x} < 0$, which forbids that H is positive definite. □

A different approach*

Another approach to count the number of positive, negative and zero eigenvalues of a Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is based on the tridiagonal form derived in Lemma 4.18 and Remark 4.6:

$$UHU^* = \begin{bmatrix} \alpha_1 & \beta_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_n \\ & & \beta_n & \alpha_n \end{bmatrix} =: T_n.$$

This tridiagonal form is a special case of the Hessenberg form for a Hermitian matrix. Since U corresponds to a similarity transformation, it is invertible and thus corresponds to a congruence. It remains to compute the inertia of T_n. The following recurrence allows us to compute the leading principal minors $d_i = \det(T[1:i,1:i])$ $(1 \le i \le n)$ of T_n:

Algorithm 5.2

```
$d_0 := 1;$  $d_1 := \alpha_1;$
for $i = 1,2,\dots,n-1$ do
    $d_{i+1} = \alpha_{i+1}d_i - \beta_{i+1}^2 d_{i-1};$
end
```

Note that, in contrast with Exercise 4.4, the constants $(-1)^i d_i$ are just the values of characteristic polynomials $p_i(\lambda)$ of $T[1:i,1:i]$ evaluated at 0:

$$(-1)^i d_i = p_i(0).$$

The following theorem follows from the properties of Sturm sequences.

Theorem 5.10

The number of sign changes in the sequence $\{d_0, d_1, \cdots, d_n\}$ is equal to the number of eigenvalues of H that are strictly smaller than 0.

An important asset of this method is that it can be applied to the matrix $H - \mu I_n$, since

$$U\left(H - \mu I_n\right)U^* = T_n - \mu I_n$$

is tridiagonal. By applying Theorem 5.10 to the matrix $T_n - \mu I_n$, we can compute the number of the eigenvalues of H that are strictly smaller than μ. As a consequence, we can use the tridiagonal form T_n to compute the number of eigenvalues in any interval of the real line.

5.3 Stability of dynamical systems

5.3.1 Continuous-time systems and the Lyapunov equation

In this section, we will consider the continuous-time linear system

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t), \qquad A \in \mathbb{C}^{n\times n}. \tag{5.4}$$

The system is *stable* if there exists a norm such that $\|\mathbf{x}(t)\|$ is strictly decreasing with time. Let

$$\|\mathbf{x}(t)\|^2 = p(\mathbf{x}(t)) := \mathbf{x}^*(t)P\mathbf{x}(t)\ , \qquad \text{with} \quad P \succ 0 \in \mathbb{C}^{n\times n}. \tag{5.5}$$

Our requirement that

$$\dot{p}(\mathbf{x}(t)) < 0 \qquad \forall t,\ \forall \mathbf{x}(0), \tag{5.6}$$

leads to the following condition on the matrix P:

$$\begin{aligned}\dot{p}(\mathbf{x}(t)) &= \dot{\mathbf{x}}(t)^*P\mathbf{x}(t) + \mathbf{x}(t)^*P\dot{\mathbf{x}}(t)\\ &= \mathbf{x}(t)^*(A^*P + PA)\mathbf{x}(t)\\ &= \mathbf{x}(t)^*(-Q)\mathbf{x}(t).\end{aligned}$$

Since $\dot{p}(\mathbf{x}(t))$ has to be strictly negative for all t, it implies that Q is a positive definite matrix. Thus, we arrive at the *Lyapunov equation*

$$A^*P + PA = -Q. \tag{5.7}$$

It implies that if P and Q are positive definite, then A is stable. We now present an algebraic version of this theorem:

Theorem 5.11

An arbitrary matrix $A \in \mathbb{C}^{n\times n}$ satisfies (5.7) with $P, Q \succ 0 \in \mathbb{C}^{n\times n}$ if and only if $\Re(\lambda_i) < 0$ for all eigenvalues λ_i of A.

Proof. First, we prove the "only if" direction. If $\mathbf{x}_i$ is an eigenvector corresponding to λ_i, then

$$-\mathbf{x}_i^*Q\mathbf{x}_i = \mathbf{x}_i^*(A^*P + PA)\mathbf{x}_i = \mathbf{x}_i^*P\mathbf{x}_i\,(\lambda_i + \bar{\lambda}_i).$$

Since $\mathbf{x}_i^*Q\mathbf{x}_i$ and $\mathbf{x}_i^*P\mathbf{x}_i$ are strictly positive, we have $\Re(\lambda_i) < 0$.

Now, we prove the "if" direction. For a given $Q \succ 0$, define

$$P = \int_0^\infty \mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau}\,\mathrm{d}\tau. \tag{5.8}$$

The integral is well defined since the eigenvalues of A have negative real parts, and thus the norm of $\mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau}$ converges exponentially to zero. Moreover, it is clear that P is Hermitian (as the integral of Hermitian matrices) and $P \succ 0$ since each $\mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau} \succ 0$ by Theorem 5.5 (remember that $\mathrm{e}^{A\tau}$ is invertible, with inverse $\mathrm{e}^{-A\tau}$).

It remains to show that (5.7) holds. Indeed, from Exercise 4.7, we have

$$\begin{aligned} A^*P + PA &= \int_0^\infty A^*\mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau} + \mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau}A \,\mathrm{d}\tau \\ &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}\tau}(\mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau})\,\mathrm{d}\tau = \left[\mathrm{e}^{A^*\tau}Q\mathrm{e}^{A\tau}\right]_{\tau=0}^{\tau\to\infty} = -Q. \end{aligned}$$

□

Remark 5.1. *It is possible to show [Horn and Johnson, 1990] that if $A, P, Q \in \mathbb{C}^{n\times n}$ satisfy (5.7) with P Hermitian and $Q \succ 0$, then*

$$\mathrm{In}(-P) = \mathrm{In}(A),$$

where $\mathrm{In}(A) = (i_+, i_-, i_0)$ are the numbers of eigenvalues of A with positive, negative and zero real part respectively.

5.3.2 Discrete-time systems and the Stein equation

The discrete-time equivalent of the dynamical system (5.4) is

$$\mathbf{x}_{k+1} = A\mathbf{x}_k, \qquad A \in \mathbb{C}^{n\times n}. \tag{5.9}$$

The system is called stable if the modulus of all eigenvalues of A are smaller than 1. The *Stein equation* is the equivalent of the Lyapunov equation:

$$P - A^*PA = Q, \qquad Q \succ 0. \tag{5.10}$$

The following theorem analyzes the discrete-time stability of A.

Theorem 5.12

An arbitrary matrix $A \in \mathbb{C}^{n\times n}$ satisfies (5.10) with $P, Q \succ 0 \in \mathbb{C}^{n\times n}$ if and only if $|\lambda_i| < 1$ for all eigenvalues λ_i of A.

Proof. To prove the "only if" direction, let $\mathbf{x}_i$ be an eigenvector corresponding to λ_i. Then

$$\mathbf{x}_i^*Q\mathbf{x}_i = \mathbf{x}_i^*(P - A^*PA)\mathbf{x}_i = \mathbf{x}_i^*P\mathbf{x}_i\,(1 - \lambda_i\bar{\lambda}_i).$$

Since $\mathbf{x}_i^*Q\mathbf{x}_i$ and $\mathbf{x}_i^*P\mathbf{x}_i$ are strictly positive, we have $|\lambda_i| < 1$.

Now, to prove the "if" direction, let $Q \succ 0$ be given and define

$$P = \sum_{i=0}^{\infty}(A^*)^iQA^i. \tag{5.11}$$

With a similar reasoning as in the proof of Theorem 5.11, we can show that the sum above is well defined and that A, P and Q satisfy (5.10) and $P \succ 0$. □

Remark 5.2. *We can show [Horn and Johnson, 1990] that if $A, P, Q \in \mathbb{C}^{n\times n}$ satisfy (5.10) with P Hermitian and $Q \succ 0$, then $\mathrm{In}(P)$ is equal to the number of eigenvalues of A whose absolute value is smaller than, larger than, or equal to 1 respectively.*

5.3.3 Computational aspects

The formulas (5.8) and (5.11) given in the proofs of Theorems 5.11 and 5.12 for the computation of P are not of great practical interest since they involve an infinite integral and an infinite sum. To conclude this section, we show how the matrix P can be efficiently computed for a given Q.

First, we note that Lyapunov (5.7) and Stein (5.10) equations are linear in the elements of P. It might not be immediately clear in their initial form, but if we define $\mathrm{vec}(P)$ as the "vectorization" of the matrix P, then we can show that $\mathrm{vec}(P)$ satisfies a system of linear equations with the right-hand side equal to $\mathrm{vec}(Q)$. If

$$\mathrm{vec}(P) \coloneqq \begin{bmatrix} P(:,1) \\ P(:,2) \\ \vdots \\ P(:,n) \end{bmatrix}, \qquad \mathrm{vec}(Q) \coloneqq \begin{bmatrix} Q(:,1) \\ Q(:,2) \\ \vdots \\ Q(:,n) \end{bmatrix},$$

then we can show that

$$\mathrm{vec}(BPA) = (A^\top \otimes B)\mathrm{vec}(P).$$

Actually, we just need to observe that

$$\begin{aligned} \mathrm{vec}(BPA) = \begin{bmatrix} BPA(:,1) \\ BPA(:,2) \\ \vdots \\ BPA(:,n) \end{bmatrix} &= \begin{bmatrix} B & & & \\ & B & & \\ & & \ddots & \\ & & & B \end{bmatrix} \begin{bmatrix} PA(:,1) \\ PA(:,2) \\ \vdots \\ PA(:,n) \end{bmatrix} \\ &= \begin{bmatrix} B & & & \\ & B & & \\ & & \ddots & \\ & & & B \end{bmatrix} \begin{bmatrix} a_{11}I & \cdots & a_{n1}I \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{1n}I & \cdots & a_{nn}I \end{bmatrix} \begin{bmatrix} P(:,1) \\ P(:,2) \\ \vdots \\ P(:,n) \end{bmatrix} \\ &= (A^\top \otimes B)\,\mathrm{vec}(P). \end{aligned}$$

If we apply this identity to the Lyapunov and Stein equations, we obtain respectively

$$(I_n \otimes A^* + A^\top \otimes I_n)\,\mathrm{vec}(P) = -\mathrm{vec}(Q) \tag{5.12}$$

$$(A^\top \otimes A^* - I_{n^2})\,\mathrm{vec}(P) = -\mathrm{vec}(Q) \tag{5.13}$$

that clearly demonstrate that P can be computed as the solution of a system of linear equations.

At first sight, it seems that we have found a simple way to count the number of stable and unstable eigenvalues of a dynamical system, without explicitly computing the eigenvalues themselves. However, the methods used in practice [Bartels and Stewart, 1972] to solve (5.12) and (5.13) boil down to the computation of the eigenvalues! Indeed, suppose that A is reduced by unitary similarity transformations to its Schur form:

$$U^*AU = \begin{bmatrix} \lambda_1 & & & \\ \times & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ \times & \cdots & \times & \lambda_n \end{bmatrix} \eqqcolon A_S, \qquad U^*U = UU^* = I_n.$$

Note that the lower triangular "Schur" form is essentially equivalent to the upper triangular Schur form, since it is enough to permute rows and columns to switch from one form to the other one:

$$\begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ \times & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ \times & \cdots & \times & \lambda_n \end{bmatrix} \begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{bmatrix} = \begin{bmatrix} \lambda_n & \times & \cdots & \times \\ & \ddots & \ddots & \vdots \\ & & \lambda_2 & \times \\ & & & \lambda_1 \end{bmatrix},$$

that are actually similarity transformations.

If we define the matrices

$$U^* P U \coloneqq P_S, \qquad U^* Q U \coloneqq Q_S,$$

then (5.7) and (5.10) become

$$\begin{aligned} A_S^* P_S + P_S A_S &= -Q_S, \\ A_S^* P_S A_S - P_S &= -Q_S. \end{aligned}$$

These equations can be solved in an efficient manner, since the equivalent systems

$$(I_n \otimes A_S^* + A_S^\top \otimes I_n)\,\mathrm{vec}(P_S) = -\mathrm{vec}(Q_S) \tag{5.14}$$

$$(A_S^\top \otimes A_S^* - I_{n^2})\,\mathrm{vec}(P_S) = -\mathrm{vec}(Q_S) \tag{5.15}$$

are upper triangular. It is enough to use substitutions to find the solution. This can be done in $O(n^3)$ operations. Thus, we have avoided the LU factorization of the original systems (5.12) and (5.13) that would have required a larger number of operations.

In fact, the reduction of the systems (5.12) and (5.13) into a triangular form (5.14) and (5.15) can be achieved with the unitary similarity transformation

$$V \coloneqq U^\top \otimes U.$$

This result is based on the property studied in the Exercise 1.6.

Exercise 5.1. *Show that*

$$V^* \left(I_n \otimes A^* + A^\top \otimes I_n\right) V = I_n \otimes A_S^* + A_S^\top \otimes I_n$$

whose eigenvalues are $\bar{\lambda}_j + \lambda_i$ *(*$i = 1, \ldots, n,\ j = 1, \ldots, n$*).*

Exercise 5.2. *Show that*

$$V^* \left(A^\top \otimes A^* - I_{n^2}\right) V = A_S^\top \otimes A_S^* - I_{n^2}$$

whose eigenvalues are $\bar{\lambda}_j \lambda_i - 1$ *(*$i = 1, \ldots, n,\ j = 1, \ldots, n$*).*

Exercise 5.3. *Apply these techniques to* Sylvester's equation

$$AX + XB = C.$$

5.4 Robustness of dynamical systems

In this section, we will study the influence of *perturbations* of dynamical systems on their stability. We will be mainly interested whether a perturbation $\Delta \in \mathbb{C}^{n\times n}$ can make a matrix $A \in \mathbb{C}^{n\times n}$ unstable. In the continuous-time case, it implies that $\Re(\lambda_i(A+\Delta)) \geq 0$; and in the discrete-time case, that $|\lambda_i(A+\Delta)| \geq 1$.

The following lemma provides a bound on such perturbations:

Lemma 5.13

The smallest norm of a perturbation Δ such that $A+\Delta$ has an eigenvalue λ^* satisfies

$$\|\Delta\|_2 = \sigma_{\min}(A - \lambda^* I).$$

Proof. Since λ^* is the eigenvalue of $A+\Delta$ we conclude that

$$\operatorname{rank}(A + \Delta - \lambda^* I) \leq n - 1.$$

Thus, we can interpret Δ as the smallest perturbation reducing the rank of $A - \lambda^* I$. In other words, $A + \Delta - \lambda^* I$ is a best approximation of $A - \lambda^* I$ with rank strictly smaller than n. The required bound on $\|\Delta\|_2$ follows then from Theorem 3.28. □

Note that the construction of a minimal-norm perturbation Δ is given by (3.15).

If we apply Lemma 5.13 to the stability of continuous-time systems, we obtain the following result:

Theorem 5.14

If $A \in \mathbb{C}^{n\times n}$ is stable ($\Re(\lambda_i) < 0$ for all λ_i) and $\|\Delta\|_2$ is bounded by

$$\|\Delta\|_2 < \min_{\lambda^* = j\omega} \sigma_{\min}(A - \lambda^* I),$$

then $A + \Delta$ remains stable.

Proof. Note first that the eigenvalues of a matrix depend continuously on its elements (since they are roots of the characteristic polynomial), so that the minimum (over $\lambda^* = j\omega$) exists. The matrix $A + \Delta$ can not become unstable unless it has an eigenvalue with the real part equal to zero for a suitable Δ: $\lambda(A+\Delta) = j\omega$. However, by Lemma 5.13, we conclude that the norm of such a destabilizing perturbation Δ is at least equal to $\min_{\lambda^*=j\omega} \sigma_{\min}(A - \lambda^* I)$. □

The equivalent result for discrete-time systems is the following theorem:

Theorem 5.15

If $A \in \mathbb{C}^{n\times n}$ is stable ($|\lambda_i| < 1$ for all λ_i) and $\|\Delta\|_2$ is bounded by

$$\|\Delta\|_2 < \min_{\lambda^* = e^{j\theta}} \sigma_{\min}(A - \lambda^* I),$$

then $A + \Delta$ remains stable.

The converses of Theorems 5.14 and 5.15 also hold: if we allow the norm of the perturbation Δ to be *equal* to the bounds presented in the above two theorems, then it is possible to find a Δ that *destabilizes* the system. Therefore, these bounds are often called the *stability radius* of the underlying systems.

Exercise 5.4. *Construct perturbations Δ with $\|\Delta\|_2$ equal to one of the bounds presented in Theorems 5.14 and 5.15 such that $A+\Delta$ is not stable.*

It is interesting to note that the function

$$\nu(\lambda) := \sigma_{\min}(A - \lambda I), \qquad \lambda \in \mathbb{C},$$

is a nonnegative real-valued function in the complex plane. Since the eigenvalues depend continuously on the elements of a matrix, we can conclude that $\nu(\lambda)$ is continuous with respect to λ. Furthermore, $\nu(\lambda^*) = 0$ if and only if λ^* is an eigenvalue of A. The level curves of $\nu(\lambda)$ are called the *pseudospectrum* of A. Lemma 5.13 implies that for each point of the complex plane such that $\nu(\lambda^*) = \delta$, there exists a perturbation Δ with $\|\Delta\|_2 \leq \delta$ such that λ^* is an eigenvalue of $A+\Delta$. It implies that the interior of the level curve δ is the set

$$\Lambda_\delta := \{\lambda_i(A+\Delta) \mid \|\Delta\|_2 \leq \delta\}.$$

In other words, it is the set of eigenvalues of perturbed matrices $A+\Delta$, where the perturbations Δ are bounded by δ. Therefore,

$$\delta_1 < \delta_2 \qquad \Longrightarrow \qquad \Lambda_{\delta_1} \subsetneq \Lambda_{\delta_2}.$$

The level curves are contained into each other.

Exercise 5.5. *Show that, when $\delta \to \infty$, Λ_δ tends to the disc of radius δ centered at the origin, i.e., the line curves for δ large become circles of radius δ.*

The pseudospectrum can be used to analyze the perturbations of the spectrum of A. For example, Theorems 5.14 and 5.15 can be reformulated in terms of the minimum of $\nu(\lambda)$ over $\lambda = j\omega$ or $\lambda = e^{j\theta}$ respectively. We can also study the stability of a system of non-linear differential equations using the pseudospectrum of its linearization, since nonlinearity can be seen as a perturbation of the linearization.

We will now present the pseudospectrum of several specific examples in order to demonstrate that the eigenvalues can change in a rather strange way. For each example, we will present the pseudospectrum computed with the help of $\nu(\lambda)$, and compute the spectrum of several randomly perturbed matrices for a fixed δ. Since we consider only real matrices for our examples, the graphs of the pseudospectra are symmetric with respect to the real axis.

Example 5.1. *Consider the Jordan matrix*

$$J_{32} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}.$$

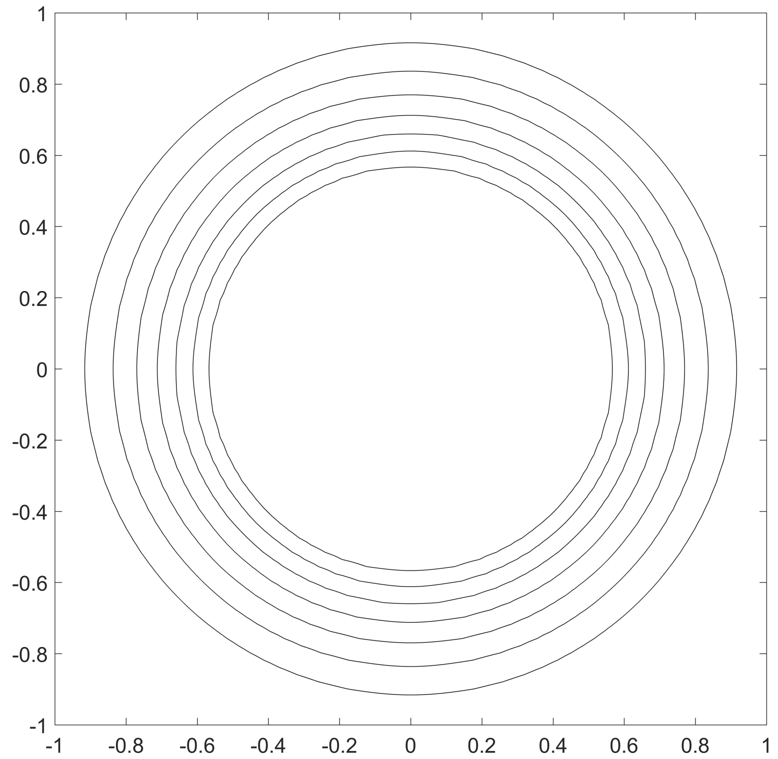

Figure 5.1: The pseudospectrum of the Jordan matrix

The only eigenvalue of this matrix is 0*. As we will see now, this matrix is very sensitive to perturbations. We have computed the line curves of the pseudospectrum (for the values* $10^{-2}, \ldots, 10^{-8}$*) of the Jordan matrix of size* 32 *(Figure 5.1). We have also computed the spectrum of* 100 *matrices of the form* $J_{32} + E$*, where* E *is a random matrix with* $\|E\|_2 = 10^{-2}$ *(Figure 5.2).*

By construction, all the points in Figure 5.2 are located in the largest disc of Figure 5.1. We can see that even a small perturbation of J_{32} *generates a totally different spectrum.*

Example 5.2. *Consider now a random* 16×16 *matrix* A*. The spectrum of* A *is given in Figure 5.3, and the level curves (for logarithmically spaced values between* 10^{-3} *and* 0.5*) of its pseudospectrum is given in Figure 5.4. It is clear from these pictures that the eigenvalues of this matrix are much less sensitive to perturbations.*

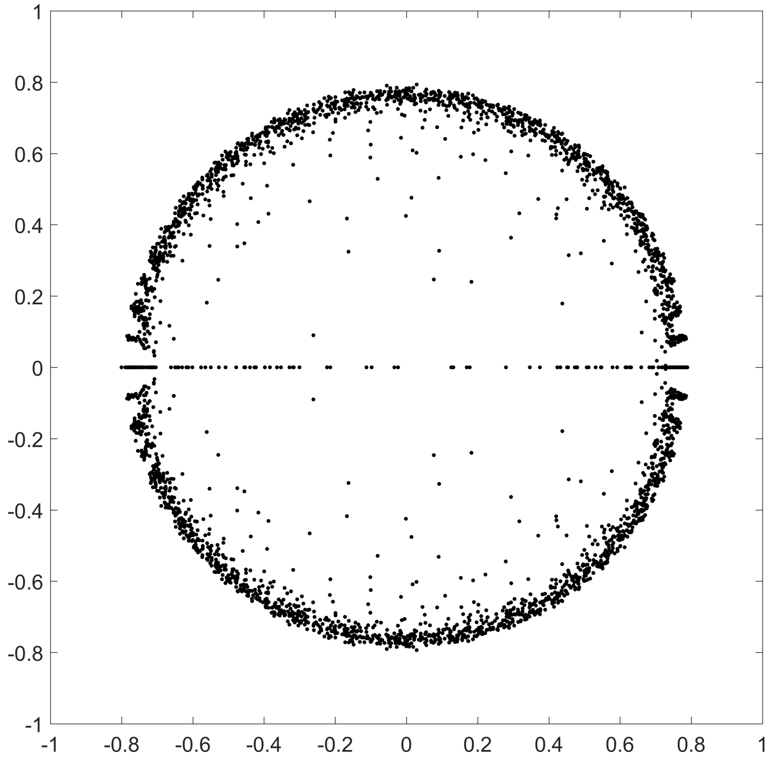

Figure 5.2: Spectra of the perturbations the Jordan matrix

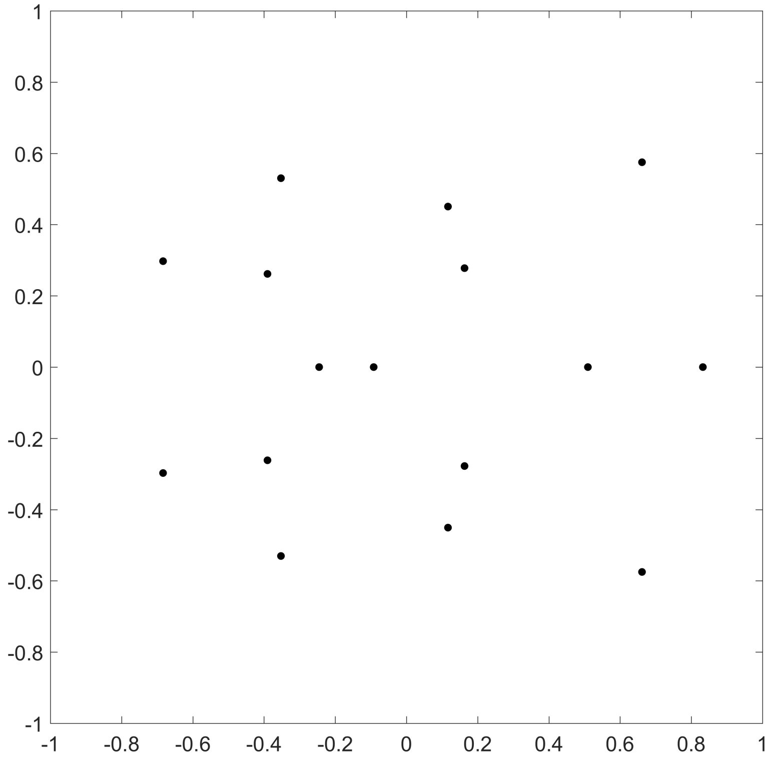

Figure 5.3: Spectrum of the matrix of Example 5.2.

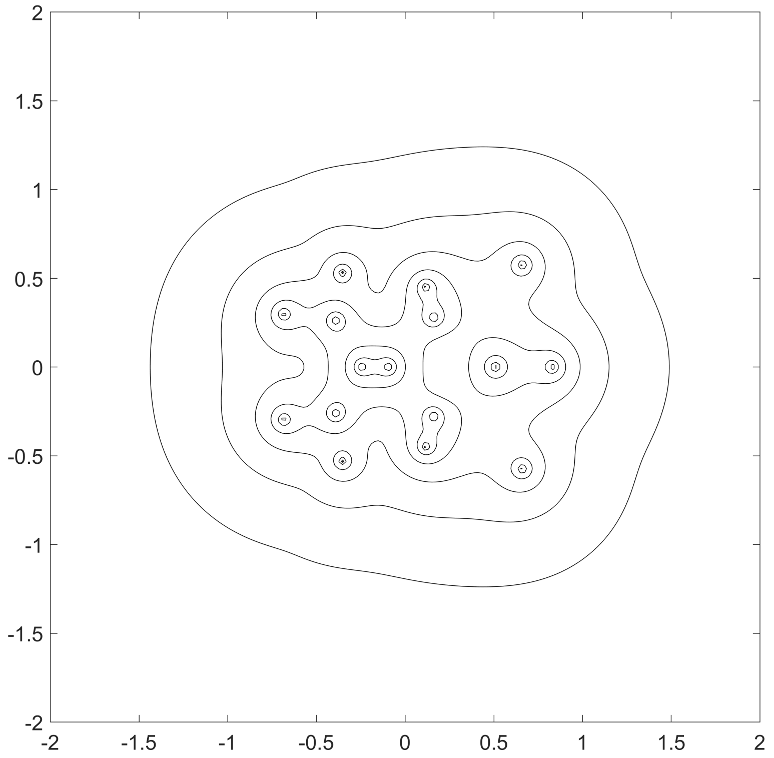

Figure 5.4: Pseudospectrum of the matrix of Example 5.2.

Chapter 6

Polynomial matrices

Polynomial matrices are mainly used for their applications in dynamical system theory. Indeed, if you compute the Laplace transformation of a system of differential equations with constant coefficients, you obtain an operator acting on a vector of functions. This operator has the form of a polynomial matrix. The same applies for the z-transform of a system of difference equations with constant coefficients. A third important application are convolutional codes which resort to difference equations on finite fields.

A polynomial matrix is a matrix whose entries are polynomials of a variable λ. For example,

$$P(\lambda) = \begin{bmatrix} \lambda^2 & 1+\lambda \\ 3\lambda & 3 \end{bmatrix} .$$

Equivalently, a polynomial matrix is a polynomial whose coefficients are matrices.

$$P(\lambda) = \bar{P}_0 + \bar{P}_1\lambda + \bar{P}_2\lambda^2 + \ldots + \bar{P}_d\lambda^d \ .$$

Note that, in this setting, $[P(\lambda)]_{i,j}$ are polynomials for all $1 \leq i \leq m$ and $1 \leq j \leq n$, and $\bar{P}_k$ are $m \times n$ matrices for all $0 \leq k \leq d$. We will denote by $\mathbb{R}^{m\times n}[\lambda]$ the set of $m \times n$ matrices whose elements are polynomials with real coefficients. We may, of course, also consider polynomials with coefficients in the field of complex numbers. In this case, the set of $m \times n$ "complex polynomial matrices" is denoted by $\mathbb{C}^{m\times n}[\lambda]$.

Because the entries of a polynomial matrix $P(\lambda)$ do not belong to a field, but rather belong to a ring (namely, to $\mathbb{R}[\lambda]$ or $\mathbb{C}[\lambda]$), we will expect to encounter other groups of transformations—and thus other canonical forms—than for the case of matrices with elements in $\mathbb{R}$ or $\mathbb{C}$. For instance, the key property that any non-singular square matrix is invertible is not true anymore. Indeed, if $P(\lambda)$ is a square polynomial matrix with nonzero determinant, the inverse of $P(\lambda)$ defined by

$$P^{-1}(\lambda) = \mathrm{adj}(P(\lambda))/\det(P(\lambda)) \tag{6.1}$$

is not necessarily a polynomial matrix (although the determinant is a polynomial and the adjugate matrix is a polynomial matrix). A first natural question is then to find the subclass of polynomial matrices which have an inverse in the form of a polynomial matrix.

Definition 6.1

A polynomial matrix $E(\lambda) \in \mathbb{C}^{n\times n}[\lambda]$ is *unimodular* if its determinant is a nonzero constant.

According to (6.1), we observe that every unimodular matrix has a polynomial inverse.

Exercise 6.1. *Show that the invertible polynomial matrices are exactly the unimodular matrices.*

Hint. *Suppose $E(\lambda)$ and $E^{-1}(\lambda)$ are both polynomial matrices. Then* $\det(E(\lambda)) \cdot \det(E^{-1}(\lambda)) = \det(I_n) = 1$. *The determinants are inverse of each other and thus are nonzero constants.*

Typical examples of unimodular matrices are

$$E_1(\lambda) = \begin{bmatrix} 1 & & p(\lambda) & \\ & 1 & q(\lambda) & \\ & & 1 & \\ & & & 1 \end{bmatrix}, \qquad E_2(\lambda) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 0 & 1 \\ & & 1 & 0 \end{bmatrix}, \tag{6.2}$$

which remind us the elementary transformations of Chapter 1. Note that the scaling of a matrix by means of a diagonal matrix $\operatorname{diag}\{1, \dots, 1, p(\lambda), 1, \dots, 1\}$ is not a unimodular transformation. The group of elementary transformations of polynomial matrices is thus smaller than the one for real or complex matrices.

Exercise 6.2. *Show that the elementary transformations of type 1 and 2 defined in* (6.2) *applied on the rows and on the columns of a polynomial matrix define a multiplicative group.*

We will show how to obtain a triangular or diagonal form using these transformations.

Theorem 6.2: Euclid–Stevin

For every two polynomials $a(\lambda), b(\lambda) \in \mathbb{C}[\lambda]$, there exists a unimodular transformation $U(\lambda) \in \mathbb{C}^{2\times 2}[\lambda]$ such that

$$\begin{bmatrix} a(\lambda) \\ b(\lambda) \end{bmatrix} = U(\lambda) \begin{bmatrix} d(\lambda) \\ 0 \end{bmatrix}$$

where $d(\lambda) = \gcd\{a(\lambda), b(\lambda)\}$.

Proof. We perform the polynomial division of $a(\lambda)$ by $b(\lambda)$. This gives the quotient $q_1(\lambda)$, and the residue $r_1(\lambda)$ whose degree is strictly smaller than the degree of $b(\lambda)$:

$$a(\lambda) = b(\lambda)q_1(\lambda) + r_1(\lambda).$$

This can be written in terms of a unimodular matrix:

$$\begin{bmatrix} a(\lambda) \\ b(\lambda) \end{bmatrix} = \begin{bmatrix} q_1(\lambda) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} b(\lambda) \\ r_1(\lambda) \end{bmatrix}, \tag{6.3}$$

(indeed observe that the determinant equals -1). We repeat the above reasoning for the division

$$b(\lambda) = r_1(\lambda)q_2(\lambda) + r_2(\lambda).$$

This gives

$$\begin{bmatrix} b(\lambda) \\ r_1(\lambda) \end{bmatrix} = \begin{bmatrix} q_2(\lambda) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} r_1(\lambda) \\ r_2(\lambda) \end{bmatrix} \tag{6.4}$$

which produces a polynomial $r_2(\lambda)$ of degree strictly smaller than the degree of $r_1(\lambda)$. Applying this procedure recursively, we finally obtain a residue $r_k(\lambda) = 0$. If we substitute (6.4) into (6.3) and so on, we get

$$\begin{bmatrix} a(\lambda) \\ b(\lambda) \end{bmatrix} = \begin{bmatrix} q_1(\lambda) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} q_2(\lambda) & 1 \\ 1 & 0 \end{bmatrix} \cdots \begin{bmatrix} q_k(\lambda) & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} d(\lambda) \\ 0 \end{bmatrix}.$$

We clearly recognize here the Euclid–Stevin algorithm for computing the greatest common divisor $d(\lambda)$ of two polynomials $a(\lambda)$ and $b(\lambda)$. Moreover, the product of the above 2×2 unimodular transformations is again unimodular. □

Corollary 6.3

For every n polynomials $p_1(\lambda), \ldots, p_n(\lambda) \in \mathbb{C}[\lambda]$, there exists a unimodular transformation $Q(\lambda) \in \mathbb{C}^{n\times n}[\lambda]$ such that

$$Q(\lambda) \begin{bmatrix} p_1(\lambda) \\ p_2(\lambda) \\ \vdots \\ p_n(\lambda) \end{bmatrix} = \begin{bmatrix} d(\lambda) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{6.5}$$

where $d(\lambda) = \gcd\{p_1(\lambda), \ldots, p_n(\lambda)\}$.

Proof. Using a construction similar to the one described in the proof of the previous theorem, we can recursively eliminate all the entries $p_i(\lambda)$ $(i = 2, \ldots, n)$ with a product of unimodular transformations. Hence, we have

$$\begin{bmatrix} p_1(\lambda) \\ p_2(\lambda) \\ \vdots \\ p_n(\lambda) \end{bmatrix} = U(\lambda) \begin{bmatrix} d(\lambda) \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{6.6}$$

By inverting $U(\lambda)$, we get (6.5) since $Q(\lambda) := U^{-1}(\lambda)$ is unimodular as well. Considering the first column of (6.6), we directly see that $d(\lambda)$ divides all the $p_i(\lambda)$:

$$p_i(\lambda) = u_{i1}(\lambda) d(\lambda).$$

Moreover, the quotients $u_{i1}(\lambda)$ have no nontrivial common divisor as otherwise this would be a nontrivial divisor of $\det(U(\lambda))$ as well, contradicting that $\det(U(\lambda))$ is constant. Hence, $d(\lambda)$ is the greatest common divisor of the polynomials $p_i(\lambda)$. □

The above corollary reminds us the Householder transformations for matrices in $\mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$, as it allows us to “compress” a vector into a single nonzero element. Henceforth, we can use this construction recursively to obtain the following result:

Theorem 6.4: Hermite

Every polynomial matrix $P(\lambda) \in \mathbb{C}^{m\times n}[\lambda]$ can be transformed into a quasi-triangular form

$$M(\lambda)P(\lambda)N = \left[\begin{array}{ccc|ccc} p_1(\lambda) & \cdots & \times & \times & \cdots & \times \\ & \ddots & \vdots & \vdots & & \vdots \\ & & p_r(\lambda) & \times & \cdots & \times \\ \hline \multicolumn{3}{c|}{0_{(m-r)\times r}} & \multicolumn{3}{c}{0_{(m-r)\times(n-r)}} \end{array}\right]$$

where $M(\lambda) \in \mathbb{C}^{m\times m}[\lambda]$ is unimodular and $N \in \mathbb{R}^{n\times n}$ is a permutation matrix.

Proof. The proof is constructive and similar to the proof of the QR factorization using Householder transformations. □

A more delicate question is the diagonalization by unimodular transformations. For this purpose, Algorithm 6.1 presented below provides a construction leading to the *Smith form* of an arbitrary polynomial matrix $P_{m\times n}(\lambda)$. This constructive algorithm provides a proof of the Smith theorem below.

Theorem 6.5: Smith

Every polynomial matrix $P(\lambda) \in \mathbb{C}^{m\times n}[\lambda]$ can be reduced by unimodular transformations $M(\lambda) \in \mathbb{C}^{m\times m}[\lambda]$ and $N(\lambda) \in \mathbb{C}^{n\times n}[\lambda]$ to a quasi-diagonal form

$$M(\lambda)P(\lambda)N(\lambda) = \left[\begin{array}{ccc|c} e_1(\lambda) & & 0 & \\ & \ddots & & 0_{r\times(n-r)} \\ 0 & & e_r(\lambda) & \\ \hline \multicolumn{3}{c|}{0_{(m-r)\times r}} & 0_{(m-r)\times(n-r)} \end{array}\right] \tag{6.7}$$

where $e_i(\lambda)$ divides $e_{i+1}(\lambda)$ $(i = 1, \ldots, r-1)$.

Proof. It suffices to observe that the transformations (of the rows and columns) described in Algorithm 6.1 are unimodular. □

Definition 6.6

The *normal rank* of a polynomial matrix $P(\lambda) \in \mathbb{C}^{m\times n}[\lambda]$ is the order of its largest nonzero minor.

Because the elementary (unimodular) operations do not affect the order of the largest nonzero minor of a matrix (cf. Theorem 1.9), it follows that the normal rank is equal to the number r of nonzero polynomials in the Smith decomposition (6.7).

Exercise 6.3. *The normal rank of $P(\lambda)$ is equal to the rank of $P(\lambda_0)$ for almost every (in the Lebesgue sense) $\lambda_0 \in \mathbb{C}$ (or $\mathbb{R}$). When they are not equal, the rank of $P(\lambda_0)$ is smaller than the normal rank.*

Algorithm 6.1

```
Step 1: If $P(\lambda) = 0$ go to step 5;
        Otherwise, permute the nonzero polynomial with minimal degree to position $(1,1)$;
Step 2: Apply one step of the Euclid–Stevin algorithm on the elements $p_{1j}(\lambda)$ for $j = 2, \dots, n$ and $p_{i1}(\lambda)$ for $i = 2, \dots, m$.
        Go back to step 1 unless you obtained
        $$M(\lambda)P(\lambda)N(\lambda) = \begin{bmatrix} p_{11}(\lambda) & 0 & \cdots & 0 \\ 0 & p_{22}(\lambda) & \cdots & p_{2n}(\lambda) \\ \vdots & \vdots & & \vdots \\ 0 & p_{m2}(\lambda) & \cdots & p_{mn}(\lambda) \end{bmatrix} \qquad (6.8)$$
        where the degree of $p_{11}(\lambda)$ is smaller or equal to the degree of the other polynomials;
Step 3: If $p_{11}(\lambda)$ does not divide all the polynomials $p_{ij}(\lambda)$ for $i = 2, \dots, m$ and $j = 2, \dots, n$, then add the $j$th column, that is not divisible by $p_{11}(\lambda)$, to the first column and go back to step 1;
Step 4: Because the degree of $p_{11}(\lambda)$ can only strictly decrease during steps 2 and 3, we finally obtain (6.8) with the property that $p_{11}(\lambda)$ divides all the polynomials of the submatrix $P[2:m, 2:n]$.
        Set $P \coloneqq P[2:m, 2:n]$, $m \coloneqq m - 1$ and $n \coloneqq n - 1$;
        If $m, n \neq 0$, then go back to step 1;
Step 5: End.
```

Hint. *It suffices to see that $M(\lambda_0)$ and $N(\lambda_0)$ in the Smith decomposition*

$$M(\lambda_0)P(\lambda_0)N(\lambda_0) = \left[\begin{array}{c|c} \operatorname{diag}\{e_i(\lambda_0)\} & 0 \\ \hline 0 & 0 \end{array}\right]$$

are invertible. Hence, $\operatorname{rank}(P(\lambda_0)) = \operatorname{rank}(\operatorname{diag}\{e_i(\lambda_0)\})$*, and the latter drops down if and only if λ_0 is a root of one of the polynomials $e_i(\lambda)$.*

Let us mention that if $P(\lambda)$ has full normal rank (i.e., $m = n = r$), then

$$\det(P(\lambda)) = c \cdot e_1(\lambda) \cdots e_r(\lambda).$$

The roots of the polynomials $e_i(\lambda)$ provide thus important information on the polynomial matrix $P(\lambda)$. For systems of differential equations, these roots are the characteristic/resonance frequencies of the solutions of the homogeneous system of differential equations

$$P\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)\mathbf{x}(t) = 0.$$

Remark 6.1.

1. *The Smith form* (6.7) *is a* canonical form *under the group of unimodular transformations on the left and on the right. These transformations are called* equivalences*. Two polynomial matrices are thus equivalent if and only if they have the same Smith form.*

2. *The Smith form was first introduced for matrices with integer coefficients. This is another example of matrices defined on a ring. The "unimodular" transformations in this case are the matrices whose determinant is invertible in the integers (why?), i.e., whose determinant is equal to ± 1.*

To conclude this chapter, we will establish the link between the Smith form and the Jordan form, which describes the characteristic solutions of a system of differential equations

$$\left(\frac{\mathrm{d}}{\mathrm{d}t}I - A\right)\mathbf{x}(t) = 0.$$

Because $\lambda I - A$ is a polynomial matrix, it admits a Smith form. What is the link between the Jordan form of A and the Smith form of $\lambda I - A$? Without going into the details, we show that the Smith form of a single Jordan block $\lambda I - J_n(\alpha)$ is $\operatorname{diag}\{1, 1, \ldots, 1, (\lambda-\alpha)^n\}$. Therefore, we give the unimodular transformations leading to this form

$$\begin{bmatrix} (\lambda-\alpha)^{n-1} & \cdots & (\lambda-\alpha) & 1 \\ -1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & & -1 & 0 \end{bmatrix} \begin{bmatrix} (\lambda-\alpha) & -1 & & \\ & (\lambda-\alpha) & \ddots & \\ & & \ddots & -1 \\ & & & (\lambda-\alpha) \end{bmatrix} \begin{bmatrix} 1 & & & \\ (\lambda-\alpha) & 1 & & \\ \vdots & \ddots & \ddots & \\ (\lambda-\alpha)^{n-1} & \cdots & (\lambda-\alpha) & 1 \end{bmatrix}$$

$$= \begin{bmatrix} (\lambda-\alpha)^n & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

It remains to apply a permutation to obtain the desired result. We can show that every Jordan block $J_d(\alpha)$ corresponds in fact to an elementary factor $(\lambda-\alpha)^d$ of the polynomials $e_i(\lambda)$:

$$e_i(\lambda) = (\lambda-\alpha_1)^{d_1^{(i)}} \cdots (\lambda-\alpha_k)^{d_k^{(i)}}.$$

For more details, we refer the reader to [Lancaster and Tismenetsky, 1985].

Exercise 6.4. *Verify that the Smith form of the matrix*

$$\lambda I_4 - \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 2 & 1 \\ & & & 2 \end{bmatrix}$$

is $\operatorname{diag}\{1, 1, \lambda-1, (\lambda-1)(\lambda-2)^2\}$.

Chapter 7

Positive matrices

In this chapter, we study matrices whose entries are positive or nonnegative real numbers. This type of matrices appears for example:

- in *graph theory*, where the elements of the matrix can represent the edges (possibly weighted) between the nodes of a graph (incidence matrix);
- in *statistics*, where the entries of the matrix can represent the probability of transition from one state to another (stochastic matrices);
- in *economy*, where the matrices can represent tables of demands and resources (Leontief model).

For the simplicity of notation, we will write

$$A > 0 \qquad \iff \qquad a_{ij} > 0 \quad \forall i,\ \forall j, \tag{7.1}$$

$$A \geq 0 \qquad \iff \qquad a_{ij} \geq 0 \quad \forall i,\ \forall j. \tag{7.2}$$

$$A \gneq 0 \qquad \iff \qquad A \geq 0 \text{ and } A \neq 0. \tag{7.3}$$

This should not be mistaken with the concepts of positive definite and positive semidefinite matrices introduced earlier in the notes, and for which we have used a typographically close notation. Note that (7.1) and (7.2) also apply to vectors and non-square matrices.

We will see in the following that the eigenvalues and eigenvectors of such matrices are in general not positive and can even be complex-valued. However, under certain assumptions, there will always exist a positive eigenvalue whose corresponding eigenvector is positive. This result, due to Perron and Frobenius, has important implications in many applications. Before stating the theorem, we need to introduce an important concept:

Definition 7.1

A nonnegative matrix $A \geq 0 \in \mathbb{R}^{n\times n}$ is *irreducible* if there exists no permutation $P \in \mathbb{R}^{n\times n}$ such that

$$PAP^\top = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0_{(n-n_1)\times n_1} & A_{22} \end{array}\right], \qquad A_{11} \in \mathbb{R}^{n_1\times n_1}, \quad 0 < n_1 < n.$$

Theorem 7.2

Let $A \geq 0 \in \mathbb{R}^{n\times n}$ be irreducible. Then

$$(I+A)^{n-1} > 0.$$

Proof. Let $\mathbf{y} \gneq 0$. Then

$$\mathbf{z} = (I+A)\mathbf{y}$$

is nonnegative and if $\mathbf{y}$ is not positive, then the number of positive elements of $\mathbf{z}$ is strictly larger than the number of positive elements of $\mathbf{y}$. Indeed, suppose that all the positive elements of $\mathbf{y}$ are in $\mathbf{y}_1$, the upper part of $\mathbf{y}$. Then

$$\begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{y}_1 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_1 > 0.$$

The vector $\mathbf{z}_1$ is clearly positive and $\mathbf{z}_2$ cannot be zero unless $A_{21} = 0$, a contradiction with A irreducible. If the nonzero elements of $\mathbf{y}$ are not gathered in $\mathbf{y}_1$, it suffices to perform a symmetric permutation of A.

If we repeat the above argument $n-1$ times, we obtain a vector

$$(I+A)^{n-1}\mathbf{y} > 0.$$

Since $\mathbf{y}$ is arbitrary, it suffices to choose $\mathbf{y} = \mathbf{e}_i = [0,\dots,0,1,0,\dots,0]^\top$, to prove that

$$(I+A)^{n-1} > 0.$$

□

Let $A \geq 0 \in \mathbb{R}^{n\times n}$ be irreducible. We define the following quotient

$$r(\mathbf{x}) := \min_{\substack{1\leq i\leq n \\ x_i \neq 0}} \frac{[A\mathbf{x}]_i}{x_i} \tag{7.4}$$

which is clearly nonnegative for every vector $\mathbf{x} \gneq 0 \in \mathbb{R}^n$.

Exercise 7.1. *Show that for all $\mathbf{x} \gneq 0 \in \mathbb{R}^n$, $\rho\mathbf{x} \leq A\mathbf{x}$ if and only if $\rho \leq r(\mathbf{x})$.*

For a given $\mathbf{x}$, we have thus that $r(\mathbf{x})$ is the supremum of all $\rho \in \mathbb{R}$ satisfying

$$\rho\mathbf{x} \leq A\mathbf{x}.$$

Now, consider the supremum of $r(\mathbf{x})$ for all $\mathbf{x}$. More precisely, define

$$r = \sup_{\mathbf{x}\gneq 0} r(\mathbf{x}) \ .$$

We would like to obtain a vector $\mathbf{x}$ for which the supremum is reached. This would be feasible if $r(\mathbf{x})$ were continuous and $\{\mathbf{x} \gneq 0 \in \mathbb{R}^n\}$ were bounded. We first note that (7.4) is invariant under scaling of $\mathbf{x}$ by a positive constant. Hence, we may consider the compact set

$$\mathcal{M} = \{\mathbf{x} \gneq 0 \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1\},$$

which satisfies

$$r = \sup_{\mathbf{x} \in \mathcal{M}} r(\mathbf{x}).$$

However, $r(\mathbf{x})$ is not continuous on $\mathcal{M}$ since (7.4) may present discontinuities when one of the entries x_i approaches zero. Therefore, we define the set

$$\mathcal{N} = \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} = (I + A)^{n-1}\mathbf{x},\ \mathbf{x} \in \mathcal{M}\}$$

which is compact and contains only positive vectors. On this set, $r(\mathbf{y})$ is continuous and thus

$$r_{\mathcal{N}} \coloneqq \sup_{\mathbf{y} \in \mathcal{N}} r(\mathbf{y}) = \max_{\mathbf{y} \in \mathcal{N}} r(\mathbf{y}).$$

Since $\mathcal{N} \subseteq \{\mathbf{x} \gneq 0\}$, we have that $r_{\mathcal{N}} \leq r$. On the other hand, observe that for any $\mathbf{x} \in \mathcal{M}$,

$$r(\mathbf{x})\mathbf{y} = r(\mathbf{x})(I + A)^{n-1}\mathbf{x} \leq (I + A)^{n-1}A\mathbf{x} \tag{7.5}$$

because $r(\mathbf{x})\mathbf{x} \leq A\mathbf{x}$. Hence, letting $\mathbf{y} = (I + A)^{n-1}\mathbf{x}$, (7.5) implies

$$r(\mathbf{x})\mathbf{y} \leq A\mathbf{y},$$

and thus Exercise 7.1 implies $r(\mathbf{x}) \leq r(\mathbf{y}) \leq r_{\mathcal{N}}$. In conclusion, we have

$$r = \max_{\mathbf{y} \in \mathcal{N}} r(\mathbf{y})$$

and r is reached by a positive vector. Finally, it is clear that $r > 0$. Indeed, it suffices to observe that $r \geq r(\mathbf{x}) > 0$ for any $\mathbf{x} > 0$ since A is irreducible.

Theorem 7.3

Let $A \geq 0 \in \mathbb{R}^{n \times n}$ be irreducible and r be as above. Then r is an eigenvalue of A and each extremal vector $\mathbf{x}$ (i.e., every vector $\mathbf{x} \gneq 0 \in \mathbb{R}^n$ satisfying $r(\mathbf{x}) = r$) is positive and is an eigenvector of A.

Proof. Let $\mathbf{x}$ be an extremal vector. Then $A\mathbf{x} \geq r\mathbf{x}$ (Exercise 7.1). If $A\mathbf{x} \neq r\mathbf{x}$, then $A\mathbf{x} - r\mathbf{x} \gneq 0$ so that

$$(I + A)^{n-1}(A\mathbf{x} - r\mathbf{x}) > 0.$$

Hence,

$$A(I + A)^{n-1}\mathbf{x} > r(I + A)^{n-1}\mathbf{x}$$

which implies, by letting $\mathbf{y} = (I + A)^{n-1}\mathbf{x}$,

$$A\mathbf{y} > r\mathbf{y},$$

contradicting $r = \sup_{\mathbf{x}} r(\mathbf{x})$. Thus, $A\mathbf{x} = r\mathbf{x}$. Finally, we have

$$\mathbf{y} = (I + A)^{n-1}\mathbf{x} = (1 + r)^{n-1}\mathbf{x} > 0$$

implying that

$$\mathbf{x} = (1 + r)^{1-n}\mathbf{y} > 0.$$

□

Theorem 7.4: Perron–Frobenius

Let $A \geq 0 \in \mathbb{R}^{n\times n}$ be irreducible and r be as above. Then r is the spectral radius of A. Moreover, r is a simple eigenvalue whose eigenspace is generated by a positive vector.

Proof. Firstly, let $\lambda \in \mathbb{C}$ be an eigenvalue of A and $\mathbf{z}$ be a corresponding eigenvector:

$$A\mathbf{z} = \lambda\mathbf{z}.$$

Then we have

$$|\lambda||\mathbf{z}| = |A\mathbf{z}| \leq A|\mathbf{z}|. \tag{7.6}$$

For $\mathbf{y} = |\mathbf{z}| \geq 0$, we can write

$$|\lambda| \leq r(\mathbf{y}) \leq r$$

so that the spectral radius $\rho(A) := \max_i |\lambda_i|$ is smaller than or equal to r.

Secondly, for every eigenvector $\mathbf{z}$ corresponding to r, we have

$$A\mathbf{z} = r\mathbf{z}$$

and thus

$$A|\mathbf{z}| \geq r|\mathbf{z}|$$

by a reasoning similar to (7.6). Moreover, $|\mathbf{z}| > 0$ because of the previous theorem, since it is an extremal vector. Hence, every eigenvector $\mathbf{z}$ associated to r cannot have a zero component. If there is more than one linearly independent eigenvector associated to r, then it is always possible to obtain, by linear combinations, another eigenvector with a zero component. This contradiction shows that r is a simple eigenvalue. □

The Perron–Frobenius theorem finds applications, e.g., in the study of stochastic matrices

$$S = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix}, \qquad \sum_{i=1}^{n} p_{ij} = 1 \tag{7.7}$$

whose entries p_{ij} are respectively the probability of transition from state j to state i. The normalization condition in (7.7) implies

$$[1, 1, \ldots, 1]\, S = [1, 1, \ldots, 1],$$

so that S has an eigenvalue equal to 1.

Theorem 7.5

If $S \geq 0 \in \mathbb{R}^{n\times n}$ is an irreducible stochastic matrix, then $\rho(S) = 1$ and 1 is a simple eigenvalue.

Proof. It suffices to show that 1 is the largest (in modulus) eigenvalue. Therefore, we note that

$$\|S\|_1 = \max_j \sum_i |p_{ij}| = 1,$$

and the spectral radius $\rho(S)$ is always smaller than or equal to any submultiplicative matrix norm. □

Exercise 7.2. *Show that if* $|\lambda_2(S)| < \rho(S)$*, then*

$$\lim_{n\to+\infty} S^n = \mathbf{x}\,[1, 1, \ldots, 1]$$

where $\mathbf{x}$ *is the Perron eigenvector of* S*, normalized such that* $\sum x_i = 1$*.*

This matrix is nonnegative and gives the "stationary" probabilities of transitions corresponding to the matrix S (see, e.g., [Lancaster and Tismenetsky, 1985]).

Remark 7.1. *An irreducible matrix can have other eigenvalues with modulus equal to* ρ*. A classical example is the* $n \times n$ *matrix*

$$S = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{bmatrix},$$

which is cyclic and stochastic. The eigenvalues and eigenvectors of S *are given by*

$$\lambda_i = \omega_i, \qquad \mathbf{x}_i = [1, \omega_i, \omega_i^2, \ldots, \omega_i^{n-1}]^\top$$

where ω_i *is an nth root of unity (i.e.,* $\omega_i^n = 1$*). We observe that*

$$|\lambda_i| = 1, \qquad |\mathbf{x}_i| = [1, 1, \ldots, 1]^\top.$$

Chapter 8

Semigroups of matrices

There are many situations in engineering where one has to build a product of different matrices (say, sampled from a finite set of given matrices), as for instance in the modelling and analysis of *switched systems* (see below). In this case, many of the problems that we have studied in this course become extremely difficult, and often impossible to solve algorithmically.

Nevertheless, given a particular set of matrices, one can define (and sometimes, actually compute, or approximate) numerical quantities that help to understand the structure of the generated *semigroup* (that is, the set of products that one can build from the given set of matrices). These quantities are called *joint spectral characteristics*: joint, because they characterize a set of matrices, and spectral, by analogy with the spectral radius, and more precisely its interpretation in terms of asymptotic behaviour of the powers of a matrix. This chapter constitutes a quick survey on the joint spectral characteristics. It mainly focuses on one of them: the *joint spectral radius*. Some of the results presented in this chapter require rather involved proofs. For this reason, this chapter is not self-contained. Its goal is to give a glimpse at modern tools in the mathematics of semigroups of matrices.

Perhaps the most natural way to introduce joint spectral characteristics is through switched systems. A switched linear system in discrete time is characterized by the equation

$$\mathbf{x}_{t+1} = A_t \mathbf{x}_t, \qquad \mathbf{x}_0 \in \mathbb{R}^n, \qquad A_t \in \Sigma, \tag{8.1}$$

where Σ is a set of real $n \times n$ matrices. We would like to estimate the evolution of the vector $\mathbf{x}$, and more particularly (if it exists) the asymptotic growth rate of its norm:

$$\lambda = \lim_{t \to \infty} \|\mathbf{x}_t\|^{1/t}.$$

Clearly, one cannot expect that this limit would exist in general. Indeed, even in dimension one, it is easy to design a dynamical system and a trajectory such that the limit above does not exist. Thus a typical relevant question for such a system is the extremal rate of growth: given a set of matrices Σ, what is the maximal value for λ over all initial vectors x_0 and all sequences of matrices A_t? In the case of dynamical systems for instance, such an analysis makes a lot of sense. Indeed, by computing the maximal growth rate, one can ensure the stability of the system, provided that this growth rate is less than one. We will see that the quantity characterizing this maximal rate of growth of a switched linear discrete time system is the *joint spectral radius*, introduced in 1960 by Rota and Strang [Rota and Strang, 1960]. Because of its interpretation in terms of dynamical systems, and for many other reasons that we will present later on, it has been widely studied during the last decades.

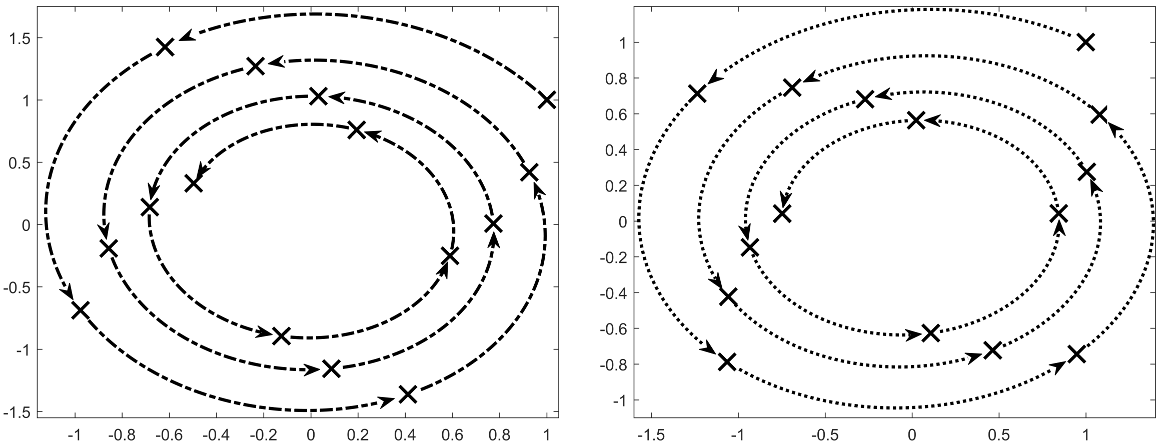

Figure 8.1: Trajectories of two stable matrices

When the set of matrices consists of a single matrix A, the problem is simple: the maximal growth rate is the largest magnitude of the eigenvalues of A. As a consequence, a matrix is stable if and only if the magnitudes of its eigenvalues are less than one. However, if the set of matrices consists in more than just one matrix, the problem is far more complex: the matrices could well all be stable, while the system itself could be unstable! This phenomenon, which motivates the study of the joint spectral radius, is illustrated by the next example.

Example 8.1. *Consider the set of matrices*

$$\Sigma = \left\{ A_0 = \frac{2}{3}\begin{bmatrix} \cos(1.5) & \sin(1.5) \\ -2\sin(1.5) & 2\cos(1.5) \end{bmatrix}, \; A_1 = \frac{2}{3}\begin{bmatrix} 2\cos(1.5) & 2\sin(1.5) \\ -\sin(1.5) & \cos(1.5) \end{bmatrix} \right\}.$$

The dynamics of A_0 (resp. A_1) are illustrated in Figure 8.1 (left) (resp. right), with the initial point $\mathbf{x}_0 = [1,1]^\top$. Since both matrices are stable ($\rho(A_0) = \rho(A_1) = 0.9428$, where $\rho(A)$ is the spectral radius *of A, i.e., is the largest magnitude of its eigenvalues), the trajectories go to the origin. But if one combines the action of A_0 and A_1 alternatively, a diverging behavior occurs (Figure 8.2). The explanation is straightforward: the spectral radius of A_0A_1 is equal to* $1.751 > 1$.

In practical applications, some other quantities can be of importance, as for instance the *minimal* rate of growth. This concept corresponds to the notion of *joint spectral subradius*. In this chapter we first present precise definitions of the main concepts (Section 8.1). In Section 8.2, we show that these definitions are well posed, and we present some basic properties on the joint spectral radius and the joint spectral subradius. Then, we show that these notions are "useful", in the sense that they actually characterize the maximal and minimal growth rates of a switched dynamical system of the type (8.1). As the reader will discover, this is not so obvious.

Some of the results presented in this chapter require rather involved proofs, which we skip.

8.1 Definitions

The joint spectral radius characterizes the maximal asymptotic growth rate of the norms of long products of matrices taken in a set Σ. All the considered matrix norms in this chapter are assumed to be submultiplicative, i.e., $\|AB\| \leq \|A\| \, \|B\|$ (see also Appendix B).

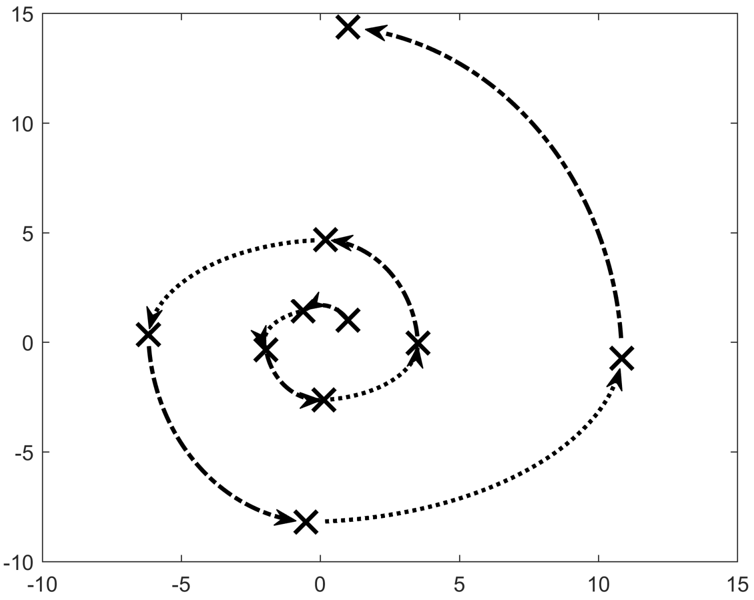

Figure 8.2: Unstable behavior by combining two stable matrices

So, let $\|\cdot\|$ be a matrix norm, and $A \in \mathbb{R}^{n\times n}$ be a real matrix. As shown by Gelfand in 1941, the spectral radius of A, that is, the maximal modulus of its eigenvalues, represents the asymptotic growth rate of the norm of the successive powers of A:

$$\rho(A) = \lim_{t\to\infty} \|A^t\|^{1/t}. \tag{8.2}$$

This quantity does provably not depend on the norm used, and one can see that it characterizes the maximal rate of growth for the norm of a point $\mathbf{x}_t$ subject to a linear time-invariant dynamical system.

In order to generalize Gelfand's formula (8.2) to a set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$, let us introduce the following notation:

$$\Sigma^t := \{A_1 \cdots A_t \mid A_i \in \Sigma\}.$$

We define the two following quantities that are good candidates to quantify the maximal norm of products of length t:

$$\hat{\rho}_t(\Sigma) := \sup\{\|A\|^{1/t} \mid A \in \Sigma^t\},$$

$$\rho_t(\Sigma) := \sup\{\rho(A)^{1/t} \mid A \in \Sigma^t\}.$$

For a matrix $A \in \Sigma^t$, we call $\|A\|^{1/t}$ and $\rho(A)^{1/t}$ respectively the *averaged norm* and the *averaged spectral radius* of the matrix, in the sense that it is averaged with respect to the length of the product. Rota and Strang introduced the *joint spectral radius* as the limit [Rota and Strang, 1960]:

$$\hat{\rho}(\Sigma) := \lim_{t\to\infty} \hat{\rho}_t(\Sigma).$$

Using the equivalence of norms in finite-dimensional spaces, it can be shown that this definition is independent of the norm used in the definition of $\hat{\rho}_t$. Daubechies and Lagarias introduced the *generalized spectral radius* as [Daubechies and Lagarias, 1992]:

$$\rho(\Sigma) := \limsup_{t\to\infty} \rho_t(\Sigma).$$

It turns out that for *bounded* sets of matrices these two quantities are equal (see Theorem 8.3 below). Based on this equivalence, we use the following definition:

Definition 8.1

The *joint spectral radius* of a bounded set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$ is defined by

$$\rho(\Sigma) = \limsup_{t\to\infty} \rho_t(\Sigma) = \lim_{t\to\infty} \hat{\rho}_t(\Sigma).$$

Example 8.2. *Let us consider the following set of matrices:*

$$\Sigma = \left\{ \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \right\}.$$

The spectral radius of both matrices is one. However, by multiplying them, one can obtain the matrix

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

whose spectral radius is equal to two. Hence, $\rho(\Sigma) \geq \sqrt{2}$, since

$$\lim_{t\to\infty} \hat{\rho}_t(\Sigma) \geq \lim_{t\to\infty} \|A^{t/2}\|^{1/t} = \sqrt{2}.$$

Now, $\hat{\rho}_2 = \sqrt{2}$ (where we have used the spectral norm) and, as we will see below, $\hat{\rho}_t$ is an upper bound on ρ for any t. So we get $\rho(\Sigma) = \sqrt{2}$.

Let us now interest ourself to the minimal rate of growth. We can still define similar quantities, describing the minimal rate of growth of the spectral radius and of the norms of products in Σ^t. These notions were introduced later than the joint spectral radius [Gurvits, 1995]:

$$\check{\rho}_t(\Sigma) := \inf\{\|A\|^{1/t} \mid A \in \Sigma^t\},$$

$$\underline{\rho}_t(\Sigma) := \inf\{\rho(A)^{1/t} \mid A \in \Sigma^t\}.$$

Then the *joint spectral subradius* is defined as the limit:

$$\check{\rho}(\Sigma) := \lim_{t\to\infty} \check{\rho}_t(\Sigma)$$

which is still independent of the norm used in the definition of $\check{\rho}_t$. We define the *generalized spectral subradius* as

$$\underline{\rho}(\Sigma) := \lim_{t\to\infty} \underline{\rho}_t(\Sigma)$$

It turns out that for any sets of matrices (not necessarily bounded) these two quantities are equal (see Theorem 8.5 below), and we use the following definition:

Definition 8.2

The *joint spectral subradius* of a set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$ is defined by

$$\check{\rho}(\Sigma) = \lim_{t\to\infty} \check{\rho}_t = \lim_{t\to\infty} \underline{\rho}_t.$$

Example 8.3. *Let us consider the following set of matrices:*

$$\Sigma = \left\{ \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 3 \end{bmatrix} \right\}.$$

The spectral radius of both matrices is greater than one. However, by multiplying them, one can obtain the zero matrix, and thus the joint spectral subradius is zero.

The above examples are simple but, as the reader will see, the situation is sometimes much more complex.

8.2 Fundamental theorems

8.2.1 The joint spectral radius theorem

It is well known that the spectral radius of a matrix satisfies $\rho(A^t) = \rho(A)^t$ and satisfies Gelfand's formula (8.2). One would like to generalize these relations to "inhomogeneous" products of matrices, that is, products where factors are not all equal to a same matrix A. This is possible, as proved in 1992 by Berger and Wang [Berger and Wang, 1992] in the so-called *Joint Spectral Radius Theorem*:

Theorem 8.3

For any bounded set of matrices $\Sigma \subseteq \mathbb{R}^{n \times n}$, the values $\hat{\rho}(\Sigma)$ and $\rho(\Sigma)$ are equal

Proof. We refer the reader to [Jungers, 2009, Theorem 2.3]. □

Observe that the joint spectral radius theorem cannot be generalized to unbounded sets of matrices, as can be seen with the following example:

Example 8.4. *Let us consider the following set of matrices:*

$$\Sigma = \left\{ \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \dots \right\}.$$

For this set, we have $\rho(\Sigma) = 1$, while $\hat{\rho}(\Sigma) = \infty$.

8.2.2 The joint spectral subradius theorem

Let us now consider the joint spectral subradius. It appears that now both $\underline{\rho}_t$ and $\check{\rho}_t$ converge:

Proposition 8.4

For any set $\Sigma \subseteq \mathbb{R}^{n \times n}$, the sequence $\check{\rho}_t(\Sigma)$ converges when $t \to \infty$, and

$$\lim_{t \to \infty} \check{\rho}_t(\Sigma) = \inf_{t > 0} \check{\rho}_t(\Sigma).$$

Moreover, the sequence $\underline{\rho}_t(\Sigma)$ converges when $t \to \infty$, and

$$\lim_{t \to \infty} \underline{\rho}_t(\Sigma) = \inf_{t > 0} \underline{\rho}_t(\Sigma).$$

Proof. We refer the reader to [Jungers, 2009, Proposition 1.1]. □

We also have the equality between $\check{\rho}$ and $\underline{\rho}$. Moreover in this case the set need not be bounded:

Theorem 8.5

For any set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$,

$$\lim_{t\to\infty} \inf\{\rho(A)^{1/t} \mid A \in \Sigma^t\} = \lim_{t\to\infty} \inf\{\|A\|^{1/t} \mid A \in \Sigma^t\} =: \check{\rho}(\Sigma).$$

Proof. Clearly,

$$\lim_{t\to\infty} \inf\{\rho(A)^{1/t} \mid A \in \Sigma^t\} \le \lim_{t\to\infty} \inf\{\|A\|^{1/t} \mid A \in \Sigma^t\}$$

because for any matrix A, $\rho(A) \le \|A\|$.

Now, for $\varepsilon > 0$, let $A \in \Sigma^t$ with averaged spectral radius $r \le \underline{\rho}(\Sigma) + \varepsilon$. Then the product $A^k \in \Sigma^{kt}$ is such that $\|A^k\|^{1/kt} \to r$ as $k \to \infty$ (Gelfand's formula), so that

$$\lim_{t\to\infty} \inf\{\|A\|^{1/kt} \mid A \in \Sigma^{kt}\} \le r \le \underline{\rho}(\Sigma) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, this concludes the proof. □

For the sake of completeness, and because it is worth to keep in mind that many problems become very hard when dealing with semigroups of matrices, we briefly cite one classical infeasibility result on the joint spectral subradius. It is based on a famous result by Paterson [Paterson, 1970] on the *mortality problem*. In this problem, one is given a set of matrices Σ, and it is asked whether there exists a product of matrices in $\Sigma^* = \bigcup_{t\ge 1} \Sigma^t$ that is equal to zero.

Theorem 8.6

The mortality problem is undecidable. This is true even for sets of $2(n_p + 1)$ matrices with dimensions 3×3, where n_p is any number for which Post's correspondence problem is undecidable.

Corollary 8.7

The mortality problem is undecidable for sets of 16 matrices with dimensions 3×3.

Proof. Matiyasevitch and Sénizergues have shown that Post's correspondence problem is undecidable even for 7 pairs of words [Matiyasevich and Senizergues, 1996]. □

From this, one can easily show that computing the joint spectral subradius, or more precisely, *deciding whether the joint spectral subradius of a set of matrices is zero* is undecidable.

8.2.3 Other basic properties

The proofs in this subsection are elementary applications of concepts previously seen in this course, and we leave them as an exercise.

Invariance properties

Proposition 8.8: Scaling invariance

For any set $\Sigma \subseteq \mathbb{R}^{n\times n}$ and any $\alpha \in \mathbb{R}$,

$$\begin{aligned}\hat{\rho}(\alpha\Sigma) &= |\alpha|\hat{\rho}(\Sigma),\\ \rho(\alpha\Sigma) &= |\alpha|\rho(\Sigma),\\ \check{\rho}(\alpha\Sigma) &= |\alpha|\check{\rho}(\Sigma),\\ \underline{\rho}(\alpha\Sigma) &= |\alpha|\underline{\rho}(\Sigma).\end{aligned}$$

Proposition 8.9: Invariance under similarity

For any set $\Sigma \subseteq \mathbb{R}^{n\times n}$ and any invertible matrix $T \in \mathbb{R}^{n\times n}$,

$$\begin{aligned}\hat{\rho}(\Sigma) &= \hat{\rho}(T\Sigma T^{-1}),\\ \rho(\Sigma) &= \rho(T\Sigma T^{-1}),\\ \check{\rho}(\Sigma) &= \check{\rho}(T\Sigma T^{-1}),\\ \underline{\rho}(\Sigma) &= \underline{\rho}(T\Sigma T^{-1}).\end{aligned}$$

Common reducibility

We will say that a set of matrices is *commonly reducible* (or simply *reducible*) if there is a nontrivial linear subspace (i.e., different from $\{0\}$ and $\mathbb{R}^n$) that is invariant under all matrices in Σ. This property is equivalent to the existence of an invertible matrix T that block-triangularizes simultaneously all matrices in Σ:

Definition 8.10

The set $\Sigma \subseteq \mathbb{R}^{n\times n}$ is reducible if and only if there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ and an integer $0 < n' < n$ such that for every $A_i \in \Sigma$,

$$TA_iT^{-1} = \begin{bmatrix} B_i & C_i \\ 0 & D_i \end{bmatrix}$$

where $D_i \in \mathbb{R}^{n'\times n'}$.

We will say that a set of matrices is *commonly irreducible* (or simply *irreducible*), if it is not commonly reducible.

Proposition 8.11

With the notations of Definition 8.10, if $\Sigma \subseteq \mathbb{R}^{n\times n}$ is bounded and reducible, then

$$\rho(\Sigma) = \max\{\rho(\{B_i\}), \rho(\{D_i\})\},$$
$$\check{\rho}(\Sigma) \geq \max\{\check{\rho}(\{B_i\}), \check{\rho}(\{D_i\})\}.$$

Three members inequalities

Proposition 8.12

For any set $\Sigma \subseteq \mathbb{R}^{n\times n}$ and any $t \in \mathbb{Z}_{>0}$,

$$\rho_t(\Sigma) \leq \rho(\Sigma) \leq \hat{\rho}(\Sigma) \leq \hat{\rho}_t(\Sigma).$$

If Σ is bounded, then the central inequality is an equality, $\rho(\Sigma) = \hat{\rho}(\Sigma)$ (Theorem 8.3).

For the joint spectral subradius, it appears that both quantities $\underline{\rho}_t$ and $\check{\rho}_t$ are in fact upper bounds:

Proposition 8.13

For any set $\Sigma \subseteq \mathbb{R}^{n\times n}$ and any $t \in \mathbb{Z}_{>0}$,

$$\check{\rho}(\Sigma) \leq \underline{\rho}_t(\Sigma) \leq \check{\rho}_t(\Sigma).$$

8.3 Stability of dynamical systems

As explained in the introduction, one possible use of the joint spectral radius is to characterize the maximal asymptotic behavior of a dynamical system. But is this exactly what we are doing, when we compute a joint spectral radius? The notion of stability of a dynamical system, like the system defined in (8.1), is somewhat fuzzy in the literature, and many different (and not equivalent) definitions appear. According to the natural intuition, and to the most commonly-used definition, we introduce the next definition:

Definition 8.14

A switched dynamical system (8.1) is *stable* if for any initial condition $\mathbf{x}_0 \in \mathbb{R}^n$, and any sequence of matrices $\{A_t\} \subseteq \Sigma$, $\lim_{t\to\infty} \mathbf{x}_t = 0$.

Clearly, if $\hat{\rho}(\Sigma) < 1$, then the dynamical system is stable, because $\mathbf{x}_t = A\mathbf{x}_0$, with $A \in \Sigma^t$, and so $|\mathbf{x}_t| \leq \|A\|\,|\mathbf{x}_0| \leq \hat{\rho}_t(\Sigma)^t|\mathbf{x}_0| \to 0$.

But the converse statement is less obvious: could the condition $\hat{\rho} < 1$ be too strong for stability? Could it be that for any length, one is able to provide a product of this length that is not too small, but yet that any *actual trajectory*, defined by an infinite sequence of matrices, is bound to tend to zero? The next example shows that such a case appears with unbounded sets:

Example 8.5. *Let*

$$\Sigma = \left\{ A = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \cup \left\{ B_k = \begin{bmatrix} 0 & k \\ 0 & 0 \end{bmatrix},\ k \in \mathbb{N} \right\}.$$

For any length t, $\hat{\rho}_t = \infty$, but one can check easily that every infinite product tends to zero. To see this, observe that a left-infinite product has one of these forms, each of which tends to zero

$$\begin{aligned} \|\cdots AA\| &\approx (1/2)^t, \\ \|\cdots A \cdots AB_kA\| &\approx k(1/2)^{t-1}, \\ \|\cdots A \cdots AB_kA \cdots AB_\ell A\| &= 0. \end{aligned}$$

The following theorem [Berger and Wang, 1992] ensures that such a pathological situation does not appear with bounded sets:

Theorem 8.15

For any bounded set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$, there exists a left-infinite product $\cdots A_2A_1$ that does not converge to zero if and only if $\rho(\Sigma) \geq 1$.

Proof. The proof of this theorem is not trivial. The reader will find a proof of this important result in [Jungers, 2009, Section 2.1]. □

This proves that the joint spectral radius rules the stability of dynamical systems:

Corollary 8.16

For any bounded set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$, the corresponding switched dynamical system (8.1) is stable if and only if $\rho(\Sigma) < 1$.

In Theorem 8.15 and in Corollary 8.16, the boundedness assumption cannot be removed, as shown by Example 8.5.

The equivalent problem for the joint spectral subradius is obvious: for any bounded set of matrices Σ, the corresponding switched dynamical system (8.1) is *stabilizable*, i.e., there exists an infinite product of matrices whose norm tends to zero, if and only if $\check{\rho}(\Sigma) < 1$. Indeed, if $\check{\rho} < 1$, there exists a real γ, and a finite product $A \in \Sigma^t$ such that $\|A\| \leq \gamma < 1$, and thus $\lim_{k\to\infty} A^k = 0$. On the other hand, if $\check{\rho} \geq 1$, then for all $A \in \Sigma^t$, $\|A\| \geq 1$ because $\check{\rho}_t(\Sigma) \geq \check{\rho}(\Sigma)$ (Proposition 8.13), and so no long product of matrices tends to zero. There is however a nontrivial counterpart to Corollary 8.16. To see this, let us rephrase Theorem 8.15 in the following corollary:

Corollary 8.17

For any bounded set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$, there is an infinite product of these matrices reaching the joint spectral radius. More precisely, there is a sequence of matrices $A_0, A_1, \ldots$ in Σ such that

$$\lim_{t\to\infty} \|A_t \cdots A_1\|^{1/t} = \rho(\Sigma).$$

The counterpart for the joint spectral subradius is the following result:

Corollary 8.18

For any (even unbounded) set of matrices $\Sigma \subseteq \mathbb{R}^{n\times n}$, there is an infinite product of these matrices reaching the joint spectral subradius. More precisely, there is a sequence of matrices $A_0, A_1, \ldots$ in Σ such that

$$\lim_{t\to\infty} \|A_t \cdots A_1\|^{1/t} = \check{\rho}(\Sigma).$$

8.4 Other joint spectral characteristics

While the joint spectral radius and subradius can easily be understood as a natural control theoretical quantity, other quantities can be defined to further describe semigroups of matrices. Such a quantity, often called the *p-radius*, has motivations in functional analysis. See, e.g., [Jia, 1995] for early work on the topic. It considers the average norm over *all* the products of length t:

$$\rho_p(\Sigma) = \lim_{t\to\infty} \left[\frac{1}{m^t} \sum_{A\in\Sigma^t} \|A\|^p\right]^{1/(pt)}.$$

The next quantity also considers the asymptotic evolution of some average norm among all the products of length t, but here, the geometric average is taken:

$$\bar{\rho}(\Sigma) = \lim_{t\to\infty} \left[\prod_{A\in\Sigma^t} \|A\|\right]^{1/(tm^t)}.$$

In control, we often call it the *Lyapunov exponent* of the system (8.1) referring implicitly to a system where (equal) probabilities are appended to each matrix in the set, so that at each time step, one matrix is sampled from the set according to the probabilities. In this context, the Lyapunov exponent provides the rate of growth of the switching system with probability one. See [Pollicott, 2010] for a more formal statement of this result and recent computational approaches.

Finally, the last quantity that is also concerned with the smallest possible rate of growth, but now it is assumed that at every step t, one can choose the matrix depending on the present value of $\mathbf{x}_t$. (This last joint spectral quantity is thus smaller than the subradius.) It has only been introduced formally recently [Jungers and Mason, 2017], but the reader can find earlier implicit studies of it in the literature. The *stabilizability radius* is defined as follows:

$$\tilde{\rho}(\Sigma) = \sup_{\mathbf{x}_0\in\mathbb{R}^n} \left[\inf\{\lambda\in\mathbb{R} \mid \exists\{t_0, t_1, \ldots\}, \exists M > 0 : \forall t \geq 0,\ |\mathbf{x}_t| \leq M\lambda^t |\mathbf{x}(t)|\}\right].$$

8.5 Conclusion

The goal of this chapter was to understand properly the notions of joint spectral characteristics in a glance, and provide a window on problems arising in the theory of matrix semigroups. Needless to say, this theory of matrix semigroups goes way beyond joint spectral characteristics.

We have limited ourself to joint spectral characteristics for which, as the reader has seen, even some basic facts, such as the equivalence between the joint and generalized spectral radii, require some advanced results. The study of matrix semigroups, and in particular joint spectral characteristics, is still the subject of active research in the mathematics and control community.

Appendix A

Algebraic structures

▷ A *semigroup* is a set together with an associative binary operation.

▷ A *monoid* is a semigroup with a neutral element.

▷ A *group* is a monoid in which every element has an inverse.

▷ A *commutative (or abelian)* group is a group whose binary operation is commutative.

When a set has two associative binary operations, they are commonly denoted by $+$ (addition) and $\cdot$ (multiplication).

▷ A *ring* is a triple $(E, +, \cdot)$ such that

- $(E, +)$ is a commutative group;
- $(E, \cdot)$ is a monoid;
- $\cdot$ is distributive with respect to $+$.

Examples: $\mathbb{Z}$, $\mathbb{R}[z]$ (polynomials of the variable z with real coefficients).

In a ring, the neutral for the addition is denoted by 0 and the neutral for the multiplication is denoted by 1.

▷ A ring $(E, +, \cdot)$ is *commutative* if $(E, \cdot)$ is commutative.

▷ An *integral domain* is a commutative ring in which the product of any two nonzero elements is nonzero. This implies that the equation $ax = b$ with $a \neq 0$ has at most one solution.

▷ A *Euclidean domain* is an integral domain such that for every two elements in the domain, we can perform the Euclidean division:

$$\forall a_1, a_2, \quad \exists q, r \qquad \text{s.t.} \qquad a_1 = a_2 q + r \quad \text{with} \quad r < a_2.$$

▷ A *field* is a commutative ring $(E, +, \cdot)$ such that every $a \in E \setminus \{0\}$ has a multiplicative inverse.

Examples: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$.

We may also defined structures on pairs of sets K and E equipped with an *external composition operation* $K \times E \to E$, also denoted by $\cdot$.

- ▷ We say that $(K, E, +)$ is a module over the ring $(K, +, \cdot)$ if
  - $(E, +)$ is a commutative group;
  - the external composition operation $\cdot : K \times E \to E$ satisfies

$$\begin{aligned} &(a+b)\cdot x = a\cdot x + b\cdot x && \text{(mixed distributivity)},\\ &a\cdot(x+y) = a\cdot x + a\cdot y && \text{(distributivity)},\\ &a\cdot(b\cdot x) = (a\cdot b)\cdot x && \text{(mixed associativity)},\\ &1\cdot x = x && \text{(common neutral element)}. \end{aligned} \tag{A.1}$$

  Examples: $\mathbb{R}^n[z]$, $\mathbb{C}^n[z]$.

- ▷ If, on top of this, $(K, +, \cdot)$ is a field, we say that $(K, E, +)$ is a *vector space* over $(K, +, \cdot)$.

  Examples: $\mathbb{R}^n$, $\mathbb{C}^n$.

We may define an *internal composition operation* $E \times E \to E$ that we denote by $\cdot$ again.

- ▷ We say that $(K, E, +, \cdot)$ is an *algebra* if
  - $(K, E, +)$ is a module or a vector space;
  - the internal composition operation $\cdot : E \times E \to E$ is bilinear.

  Examples: the square matrices with elements in a field or a ring, e.g., $\mathbb{R}^{n\times n}$, $\mathbb{C}^{n\times n}$, $\mathbb{R}^{n\times n}[z]$, $\mathbb{C}^{n\times n}[z]$.

Appendix B

Norms

A *vector norm* is a function $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ satisfying the following properties:

- $\|\mathbf{x}\| \geq 0$,
- $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$,
- $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\| \qquad \forall\alpha \in \mathbb{C},\ \forall\mathbf{x} \in \mathbb{C}^n$,
- $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad \forall\mathbf{x},\mathbf{y} \in \mathbb{C}^n$.

The most commonly used norms are

$$\begin{aligned}
\|\mathbf{x}\|_1 &= \sum_{i=1}^{n} |x_i|, \\
\|\mathbf{x}\|_2 &= \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2} = \left(\mathbf{x}^*\mathbf{x}\right)^{1/2}, \\
\|\mathbf{x}\|_\infty &= \max_{1\leq i\leq n} |x_i|,
\end{aligned}$$

which are particular cases of the p-norm:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \qquad p \geq 1.$$

Among the vector norms, we often consider the norm $\|\cdot\|_2$ because it is derivable (its gradient is equal to $\mathbf{x}/\|\mathbf{x}\|_2$) and is unitarily invariant: $\|U\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for every unitary matrix U (i.e., every matrix satisfying $U^*U = UU^* = I_n$).

A *matrix norm* is a function $\|\cdot\| : \mathbb{C}^{m\times n} \to \mathbb{R}$ which satisfies the same properties as a vector norm. The most frequently used matrix norms are the following:

- The *Frobenius* norm:

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{i,j}|^2\right)^{1/2} = \Big[\mathrm{trace}(A^*A)\Big]^{1/2},$$

- The matrix norms induced by a vector norm (aka. *operator norms*):

$$\|A\|_p = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}, \qquad p = 1, 2, \infty.$$

We can show [Horn and Johnson, 1990] that

$$\begin{aligned}
\|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{i,j}|, \\
\|A\|_\infty &= \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{i,j}|, \\
\|A\|_2 &= \Big[\lambda_{\max}(A^*A)\Big]^{1/2} = \sigma_{\max}(A), \qquad \text{(spectral norm)}.
\end{aligned}$$

A matrix norm $\|\cdot\|$ is *submultiplicative* if $\|AB\| \leq \|A\|\|B\|$ for every matrices A and B for which the product makes sense and the norms of A, B and AB are well defined. The above norms are all submultiplicative.

A matrix norm is *unitarily invariant* if $\|M\| = \|UMV\|$ for every unitary matrices U (i.e., $U^*U = UU^* = I_m$) and V (i.e., $V^*V = VV^* = I_n$). Among the above norms, only $\|\cdot\|_2$ and $\|\cdot\|_F$ are unitarily invariant. This is why they are used so often.

Bibliography

[Bartels and Stewart, 1972] Bartels, R. H. and Stewart, G. W. (1972). Solution of the matrix equation $ax + xb = c$ [f4]. *Communications of the ACM*, 15(9):820–826.

[Berger and Wang, 1992] Berger, M. A. and Wang, Y. (1992). Bounded semigroups of matrices. *Linear Algebra and its Applications*, 166:21–27.

[Daubechies and Lagarias, 1992] Daubechies, I. and Lagarias, J. C. (1992). Sets of matrices all infinite products of which converge. *Linear algebra and its applications*, 161:227–263.

[Francis, 1961] Francis, J. G. (1961). The qr transformation a unitary analogue to the lr transformation—part 1. *The Computer Journal*, 4(3):265–271.

[Golub and Van Loan, 2012] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

[Gurvits, 1995] Gurvits, L. (1995). Stability of discrete linear inclusion. *Linear algebra and its applications*, 231:47–85.

[Horn and Johnson, 1990] Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis.* Cambridge university press.

[Jia, 1995] Jia, R.-Q. (1995). Subdivision schemes in l^p spaces. *Advances in Computational Mathematics*, 3(4):309–341.

[Jungers, 2009] Jungers, R. M. (2009). *The joint spectral radius: theory and applications*, volume 385. Springer Science & Business Media.

[Jungers and Mason, 2017] Jungers, R. M. and Mason, P. (2017). On feedback stabilization of linear switched systems via switching signal control. *SIAM Journal on Control and Optimization*, 55(2):1179–1198.

[Kato, 2013] Kato, T. (2013). *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media.

[Lancaster and Tismenetsky, 1985] Lancaster, P. and Tismenetsky, M. (1985). *The theory of matrices: with applications.* Elsevier.

[Matiyasevich and Senizergues, 1996] Matiyasevich, Y. and Senizergues, G. (1996). Decision problems for semi-thue systems with a few rules. In *Logic in Computer Science, 1996. LICS'96. Proceedings., Eleventh Annual IEEE Symposium on*, pages 523–531. IEEE.

[Paterson, 1970] Paterson, M. S. (1970). Unsolvability in 3 × 3 matrices. *Studies in Applied Mathematics*, 49(1):105–107.

[Pollicott, 2010] Pollicott, M. (2010). Maximal Lyapunov exponents for random matrix products. *Inventiones mathematicae*, 181(1):209–226.

[Rota and Strang, 1960] Rota, G.-C. and Strang, G. W. (1960). A note on the joint spectral radius.

[Stewart, 1973] Stewart, G. W. (1973). Introduction to matrix computations.

[Wilkinson, 1965] Wilkinson, J. H. (1965). *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford.