LINEAR ALGEBRA

with DIFFERENTIAL EQUATIONS

M. Erol Sezer
Yaşar University, İzmir

Preface to the e-Edition

This book was first printed by Bilkent University (Ankara, Turkey) to be used as a textbook in the course Linear Algebra and Differential Equations that I had been teaching to sophomore engineering students of Bilkent for some years. An electronic version was also available for public use on my personal web page. After I left Bilkent for a new academic position, it was largely forgotten until a research assistant at Yaşar University (where I hold a position currently) told me she had managed to compile the chapters of the book from my previous course pages to prepare for her qualification exam. I gave her one of the printed copies that I happen to carry around with me, and decided to make it available to public again. Thus is born this e-edition, now with an ISBN.

I am grateful to both Bilkent University for printing the original book and for giving me the opportunity to try it with very special students, and to Yaşar University for publishing the electronic version on its institutional website.

M. Erol Sezer
July 2018
Yaşar University
erol.sezer@yasar.edu.tr

Preface to the Print Edition

Linear algebra is not only a powerful mathematical theory; it is also a useful computational tool supported by abundant computer software. It provides the theoretical foundations of the concept of linearity as well as efficient methods for solving problems formulated within this framework. For these reasons it has found applications in many diverse fields outside mathematics, ranging over various engineering disciplines, computer science, economics, statistics, and more.

This book is intended to introduce the theory and applications of linear algebra to engineering students, especially those in electrical engineering. Its main objective is to provide engineering students with a firm understanding of the concept of linearity at an early stage of their program. It attempts to achieve this objective by integrating basic topics of linear algebra with linear differential equations and linear systems at a suitable level.

The book is primarily a text on linear algebra supplemented with linear differential equations. Although merging linear differential equations in a text on linear algebra is observed to be pedagogically useful in connecting different concepts, the book is not intended to serve as a text on differential equations. This is evident from its contents, as many important topics covered in an introductory course on differential equations, such as series solutions and introductory partial differential equations are left out; other topics such as first order nonlinear differential equations, numerical solution techniques, and Laplace transforms are mentioned only briefly. In contrast, it contains all essential material of Linear Algebra that engineering students might need throughout their undergraduate and early graduate curriculum so that they can use it as a reference book even if not everything in the text is taught in a specific course. This feature also makes it suitable for self-study.

The main difficulty I faced in preparing the manuscript was to make a choice for the content and presentation between conflicting alternatives: Amount of material to be included versus suitability for a one-semester or two-quarter course, emphasis on theory versus operational aspects, mathematical formalism versus explanation of implications of the results, and finally examples versus exercises.

- I preferred a self-contained text to one tailored for a specific course. To help the instructor/reader in selecting the material to be covered in a quick treatment, I indicated more advanced and/or abstract topics (including examples) with an asterisk (*). However, this does not mean that all the marked sections or examples can be omitted without destroying the completeness of the book.

- I endeavored to develop the material from the concrete to the abstract, in most instances explaining the need to introduce a new definition or result. Although I gave priority to a logical development of the theory, I explained, whenever possible, what the theory is good for. For example, the basic definitions and properties of matrices are followed by systems of linear equations to help the reader appreciate the convenience and power

inherent in the matrix formulation of linear equations. I encouraged the student to use MATLAB to solve matrix-related problems starting immediately in the first chapter without substituting it for the underlying theory. I avoided integrating MATLAB with the text, leaving that to the instructor. However, I provided sufficient exercises to demonstrate MATLAB's power and, to a lesser degree, its limitations.

- I provided sufficient examples to explain the theoretical development. Rather than leaving the often difficult task of establishing a connection between a new result and a previous one to the reader, I attempted to explain the logic behind the development, often by referring to previous examples. However, I avoided multiple similar examples except when they emphasize different aspects of the same concept. Instead of filling the pages with redundant examples, I preferred to include many exercises with hints to formulation and/or solution. Again, I avoided similar exercises that differ only in the numerical values involved.

- Although I avoided a Definition-Theorem-Proof structure to make the reading easier and less dry, I did not state any result without a proof except when:

  – I thought the reader could provide a fairly straightforward proof independently by simply imitating the steps of similar results for which a formal proof is given, or

  – In few instances, the proof of the stated result is beyond the scope of the book, but its implications are too significant to omit. In such cases I tried to provide insight into the meaning and the implications of the stated result.

The book is built upon a rigorous treatment of vector spaces and linear transformations, which are motivated by linear systems of algebraic equations and first and second order linear differential equations. The first three chapters contain the most essential material for a first course.

Chapter 1 is a self-contained treatment of simple matrix algebra, where basic definitions and operations on matrices are introduced and systems of linear algebraic equations are studied. Properties of matrix addition and scalar multiplication are stated in a manner consistent with the corresponding properties of vector addition and scalar multiplication to prepare the student for the more abstract concepts to follow. The Gaussian elimination is introduced not only as a systematic approach to solving linear equations, but also as a theoretical tool leading to the concepts of rank and particular and complementary solutions, which in turn pave the road to a more abstract treatment of linear equations in terms of the kernel and image of the associated linear transformation. Through simple examples and exercise problems the student is urged to use MATLAB to check the results of their hand calculations and to digest the idea that a matrix is a data unit (like a number) on which they can perform some operations.

In Chapter 2 first and second order linear differential equations are studied with emphasis on the constant coefficient case. The three objectives of the chapter are (i) to provide the students with solution techniques for simple differential equations that they can immediately start utilizing in concurrent courses on circuits or dynamical systems, (ii) to further prepare the student for linear transformations by repeating the concepts of particular and complementary solutions in a different context and by introducing linear differential operators, and (iii) to introduce the basics of numerical solution techniques so that the student can begin to use MATLAB or other software, and at the same time, to give an idea of linear difference equations, which involve yet another type of linear transformation.

Chapter 3 contains an abstract treatment of vector spaces and linear transformations based on the ideas introduced in the preceding two chapters. The concept of a vector space is extended beyond the familiar n-spaces with the aim of unifying linear algebraic and differential equations under a common framework. By interpreting an n-vector as a function defined over a finite domain, the student is prepared for function spaces. The concept of basis is given special emphasis to establish the link between abstract vectors and the more familiar n-vectors, as well as between abstract linear transformations and matrices. Discrete Fourier series are introduced as an example of representation of vectors of a finite-dimensional vector space with respect to a fixed basis. This chapter also contains some more advanced topics such as inverse transformations, direct sum decompositions, and projections.

Chapter 4 introduces rank and inverse of matrices. Rank is defined in terms of the row and column spaces. Left, right, two-sided and generalized inverses are based on elementary matrices without reference to determinants. Concepts of equivalence and similarity are related to change of basis. The LU decomposition is studied as a natural and useful application of elementary operations. Determinants are considered mainly for traditional reasons to mention the role they play in the solution of linear equations with square coefficient matrices.

Chapter 5 deals with the eigenvalues, eigenvectors and diagonalization of square matrices from a geometric perspective. The diagonalization problem is related to the decomposition of the underlying vector space into invariant subspaces to motivate the much more advanced Jordan form. The chapter concludes with a treatment of functions of a matrix, the main objective being to define the exponential matrix function that will be needed in the study of systems of linear differential equations.

In Chapter 6 we return to linear differential equations. As opposed to the traditional approach of treating nth order linear differential equations and systems of first order linear differential equations separately, and in that order, systems of differential equations are studied first and the results developed in that context are used to derive the corresponding results for nth order differential equations. This is consistent with the matrix theoretic approach of the text to the treatment of linear problems, which relates the abstract concepts of bases, direct sum decomposition of vector spaces, and the Jordan form to solutions of a homogeneous system of linear differential equations and their modal decomposition. The method of undetermined coefficients is included as a practical way of solving linear differential equations with special forcing functions that are common in engineering applications.

Chapter 7 treats normed and inner product spaces with emphasis on the concepts of orthogonality and orthogonal projections, where the Gram-Schmidt orthogonalization process, the least-squares problem and the Fourier series are formulated as applications of the projection theorem.

Chapter 8 deals with unitary and Hermitian matrices for the purpose of presenting such useful applications as quadratic forms and the singular-value decomposition, which is related to the least-squares problem and matrix norms.

I was able to cover most of the material in a 56-class-hour one-semester course I taught to a class of select students at the Electrical Engineering Department of Bilkent University. However, an average class of second or third year students would need about 60 class hours to cover the essential material. For such a course, Sections 1.6, 2.6, 3.4.3, 3.6 5.4, 6.2 and Chapter 8 may be omitted.

The text was developed over some years of my experience with teaching linear algebra at the Middle East Technical University and Bilkent University. My long time colleague Özay Oral and I prepared some lecture notes to meet the demand for a text for a combined

course on Linear Algebra and Differential Equations. Although the present text is completely different from those lecture notes, both in its approach and in contents, it would not have come to fruition without those initial efforts. I am indebted to Özay for his motivation and encouragement that led first to the lecture notes and eventually to the present version of the text. Thanks are also due to my colleagues for their suggestions and constructive criticisms.

M. Erol Sezer
Bilkent University

Contents

Chapter 1
Matrices and
Systems of Linear Equations

1.1 Basic Matrix Definitions

An $m \times n$ (read "*m-by-n*") ***matrix*** is an array of mn elements of a field $\mathbb{F}$ arranged in m rows and n columns as

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

A matrix with m rows and n columns is said to be of ***order*** (size, dimension) $m \times n$. We denote matrices with uppercase letters and their elements with corresponding lowercase letters, and use the notation $A = [\, a_{ij} \,]_{m \times n}$ to describe an $m \times n$ matrix where a_{ij} typifies the element in the ith row and the jth column. When the order of A need not be specified we simply write $A = [\, a_{ij} \,]$. The set of all $m \times n$ matrices with elements from $\mathbb{F}$ is denoted by $\mathbb{F}^{m \times n}$. Throughout the book we will assume that the underlying field $\mathbb{F}$ is either $\mathbb{R}$ (in which case A is a real matrix) or $\mathbb{C}$ (in which case A is a complex matrix).[1]

A $1 \times n$ matrix is called a ***row matrix*** or a ***row vector***, and an $m \times 1$ matrix is called a ***column matrix*** or a ***column vector***.[2] We denote row and column vectors with boldface lowercase letters. Thus

$$\boldsymbol{x} = [\, x_1 \;\; x_2 \;\; \cdots \;\; x_n \,]$$

is a $1 \times n$ row vector, and

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

is an $m \times 1$ column vector, which we also denote as

$$\boldsymbol{y} = \text{col}\,[\, y_1, y_2, \ldots, y_m \,]$$

[1] Definition of a field and a brief review of complex numbers are given in Appendix A. Since the field of real numbers is a subfield of complex numbers, every real matrix can also be viewed as a complex matrix.

[2] The use of the term "vector" for a column or a row matrix is justified in Chapter 3.

to save space. Note that a single column or row index suffices to denote elements of a row or a column vector.

An $n \times n$ matrix is called a ***square matrix*** of order n. The sum of the diagonal elements $a_{11}, \ldots, a_{nn}$ of a square matrix $A = [\, a_{ij} \,]_{n\times n}$ is the ***trace*** of A:

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

A square matrix $D = [\, d_{ij} \,]_{n\times n}$ in which $d_{ij} = 0$ for all $i \neq j$ is called a ***diagonal*** matrix. As for a row or column vector, a single index suffices to denote elements of a diagonal matrix:

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} = \text{diag}\,[\, d_1, d_2, \ldots, d_n \,]$$

A square matrix $L = [\, l_{ij} \,]_{n\times n}$ in which $l_{ij} = 0$ whenever $i < j$ is a ***lower triangular*** matrix for the obvious reason that all the elements located above its diagonal are zero. Similarly, a square matrix $U = [\, u_{ij} \,]_{n\times n}$ with $u_{ij} = 0$ whenever $i > j$ is an ***upper triangular*** matrix.

Example 1.1

The array

$$A = \begin{bmatrix} 0 & \sqrt{2} & -1 \\ 3 & e & \ln 5 \end{bmatrix}$$

is a 2×3 real matrix with elements $a_{11} = 0, a_{12} = \sqrt{2}, \ldots, a_{23} = \ln 5$.

The array

$$B = \begin{bmatrix} 1+2i & -3 & -1+i \\ 0 & -3i & 5 \\ -1 & -2 & 3+2i \end{bmatrix}$$

is a complex square matrix of order 3 with

$$\text{tr}\,(B) = (1+2i) + (-3i) + (3+2i) = 4+i$$

The matrices

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 0 & 4 & 5 \end{bmatrix}, \quad D = \begin{bmatrix} 3+i & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

are lower triangular and diagonal, respectively.

1.2 Basic Matrix Operations

1.2.1 Matrix Addition and Scalar Multiplication

Let $A = [\,a_{ij}\,]$ and $B = [\,b_{ij}\,]$ be two matrices in $\mathbb{F}^{m\times n}$. A and B are said to be ***equal***, denoted $A = B$, if $a_{ij} = b_{ij}$ for all (i, j).[3] Note that it would be meaningless to talk about equality of two matrices unless they are of the same size and their elements are comparable, that is, they are also of the same type.

If $a_{ij} = 0$ for all (i, j), then A is called an $m \times n$ ***zero matrix*** (null matrix), denoted $O_{m\times n}$, or simply O if the order is known. That is, $O_{m\times n} = [\,0\,]_{m\times n}$.

The ***sum*** of A and B, denoted $A + B$, is defined in terms of their elements as

$$A + B = [\,a_{ij} + b_{ij}\,]_{m\times n}$$

That is, the (i, j)th element of $A + B$ is the sum of the corresponding elements of A and B. The subtraction operation is defined in terms of addition as

$$A - B = A + (-B)$$

where

$$-B = [\,-b_{ij}\,]$$

Note that, like equality, addition and subtraction operations are defined only for matrices belonging to the same class, and that if $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{m\times n}$, then $A + B \in \mathbb{F}^{m\times n}$, $-B \in \mathbb{F}^{m\times n}$, and therefore, $A - B \in \mathbb{F}^{m\times n}$.

Any element of the field $\mathbb{F}$ over which the matrices of concern are defined is called a ***scalar***. The ***scalar product*** of a matrix A with a scalar c, denoted cA, is also defined element-by-element as

$$cA = [\,ca_{ij}\,]_{m\times n}$$

Thus the (i, j)th element of cA is c times the corresponding element of A. It follows from the definition that $cA \in \mathbb{F}^{m\times n}$. Clearly,

$$0A = O \quad \text{and} \quad (-1)A = -A$$

Example 1.2

$$\begin{bmatrix} -1 & 1-i \\ 1+2i & 0 \\ -i & 3+2i \end{bmatrix} + \begin{bmatrix} 2 & 1 \\ -1 & -2 \\ 3 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 2-i \\ 2i & -2 \\ 3-i & 2+2i \end{bmatrix}$$

and

$$(1+i)\begin{bmatrix} 1-i & 0 \\ 1 & -1+2i \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1+i & -3+i \end{bmatrix}$$

The first example above shows that we can add a real matrix to a complex matrix by treating it as a complex matrix. Similarly, a real matrix can be multiplied with a complex scalar, and a complex matrix with a real scalar.

[3] We will use the phrase "for all (i, j)" to mean "for all $1 \le i \le m$ and $1 \le j \le n$" if the ranges of the indices i (1 to m) and j (1 to n) are known.

We list below some properties of matrix addition and scalar multiplication.

A1. $A + B = B + A$

A2. $(A + B) + C = A + (B + C)$

A3. $A + O = A$

A4. $A + (-A) = O$

S1. $c(dA) = (cd)A$

S2. $(c + d)A = cA + dA$

S3. $c(A + B) = cA + cB$

S4. $1A = A$, where 1 is the multiplicative identity of $\mathbb{F}$.

Note that in property S2 the same symbol "$+$" is used to denote both the addition of the scalars c and d, and also the addition of the matrices cA and dA. This should cause no confusion as which operation is meant is clear from the operands. Similarly, in property S1, cd represents the product of the scalars c and d, whereas dA represents the scalar multiplication of the matrix A with the scalar d, although no symbol is used to denote either of these two different types of multiplication.[4]

The properties above follow from the properties of addition and multiplication of the scalars involved. For example, property S2 can be proved as

$$\begin{aligned}(c+d)A &= [\,(c+d)a_{ij}\,] &&= [\,ca_{ij} + da_{ij}\,] \\ &= [\,ca_{ij}\,] + [\,da_{ij}\,] &&= c\,[\,a_{ij}\,] + d\,[\,a_{ij}\,] &&= cA + dA\end{aligned}$$

Proofs of the other properties are left to the reader.

From the basic properties above we can derive further useful properties of matrix addition and scalar multiplication. For example,

$$\begin{aligned}A + B = O &\implies B = -A \\ A + B = A + C &\implies B = C \\ cA = O &\implies c = 0 \quad \text{or} \quad A = O\end{aligned}$$

We finally note that if $A_1, A_2, \ldots, A_k \in \mathbb{F}^{m\times n}$ then because of property A2, an expression of the form $A_1 + A_2 + \cdots + A_k$ unambiguously defines a matrix in $\mathbb{F}^{m\times n}$. For example,

$$A + B - A = B$$

and

$$\underbrace{A + \cdots + A}_{k} = kA$$

[4] If we used a symbol for scalar multiplication, say "$\cdot$", then property S1 would be stated as

$$c \cdot (d \cdot A) = (cd) \cdot A$$

1.2.2 Matrix Multiplication

Let $A = [\, a_{ij} \,]_{m\times n}$ and $B = [\, b_{ij} \,]_{p\times q}$. If A has exactly as many columns as B has rows, that is, if $n = p$, then A and B are said to be ***compatible*** for the product AB. The product is then defined to be an $m \times q$ matrix $AB = C = [\, c_{ij} \,]_{m\times q}$ whose elements are

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

That is, the (i, j)th element of the product is the sum of the ordered products of the ith row elements of A with the jth column elements of B.[5] Thus it takes n multiplications to compute a single element of the product, and therefore, mnq multiplications to compute C.

Example 1.3

Let

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 3 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & -1 & 1 & 3 \\ 1 & 0 & -1 & 2 \end{bmatrix}$$

Since A is 2×3 and B is 3×4, the product $C = AB$ is defined, and is a 2×4 matrix. First column elements of C are found as

$$\begin{aligned} c_{11} &= 1 \cdot 1 + (-1) \cdot 0 + 2 \cdot 1 &&= 3 \\ c_{21} &= 3 \cdot 1 + 0 \cdot 0 + 1 \cdot 1 &&= 4 \end{aligned}$$

Computing other elements of C similarly, we obtain

$$C = \begin{bmatrix} 3 & 3 & -3 & 0 \\ 4 & 6 & -1 & -1 \end{bmatrix}$$

On the other hand, the product BA is not defined.

Example 1.4

Let

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 3 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} -1 & 1 \end{bmatrix}$$

Then

$$\begin{aligned} A\boldsymbol{x} &= \begin{bmatrix} -1 \\ 5 \end{bmatrix} \\ \boldsymbol{y}A &= \begin{bmatrix} 2 & 1 & -1 \end{bmatrix} \\ \boldsymbol{x}\boldsymbol{y} &= \begin{bmatrix} -2 & 2 \\ -1 & 1 \\ 1 & -1 \end{bmatrix} \end{aligned}$$

Other pairwise products are not defined.

[5] Implicit in the definition of matrix multiplication is the assumption that elements of A can be multiplied with those of B, which requires that they belong to the same field. However, we can multiply a complex matrix with a real one.

Examples 1.3 and 1.4 illustrate that matrix multiplication is not commutative. If A is $m \times n$ and B is $n \times q$, then AB is an $m \times q$ matrix, but BA is not defined unless $q = m$. If $q = m$, that is, when B is $n \times m$, then both AB and BA are defined, but AB is $m \times m$, while BA is $n \times n$. Even when A and B are both $n \times n$ square matrices, so that both AB and BA are defined and are $n \times n$ matrices, in general $AB \neq BA$. For example, if

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{1.1}$$

then $AB = O$, but $BA = A$. Therefore, the order in which two matrices are multiplied is important. In the product AB, B is said to be premultiplied (or multiplied from left) with A, and A is said to be postmultiplied (or multiplied from right) with B.

Two square matrices A and B of the same order are said to ***commute*** if $AB = BA$. For example, if

$$C = \text{diag}\,[\,c_1, \ldots, c_n\,] \quad \text{and} \quad D = \text{diag}\,[\,d_1, \ldots, d_n\,]$$

then

$$CD = DC = \text{diag}\,[\,c_1 d_1, \ldots, c_n d_n\,]$$

Obviously, every square matrix commutes with itself. If A is a square matrix then we denote the product AA by A^2. Higher order powers of A are defined recursively as $A^{k+1} = AA^k\,, \quad k = 1, 2, \ldots$.

We now state several properties of matrix multiplication.

M1. $(AB)C = A(BC)$

M2. $A(B + C) = AB + AC \quad$ and $\quad (A + B)C = AC + BC$

M3. $OA = O \quad$ and $\quad AO = O$

In stating these properties, we implicitly assume that the products involved are defined. Note that the two properties in item M2, as well as those in item M3, are different (that is, one does not follow from the other), as matrix multiplication is not commutative.

The properties above follow directly from definitions. For example, letting

$$A = [\,a_{ij}\,]_{m \times n}\,, \quad B = [\,b_{ij}\,]_{n \times q}\,, \quad C = [\,c_{ij}\,]_{n \times q}$$

we have

$$\begin{aligned} A(B + C) &= [\,a_{ij}\,]\,[\,b_{ij} + c_{ij}\,] &&= [\,\sum_{k=1}^{n} a_{ik}(b_{kj} + c_{kj})\,] \\ &= [\,\sum_{k=1}^{n} a_{ik} b_{kj} + \sum_{k=1}^{n} a_{ik} c_{kj}\,] &&= [\,\sum_{k=1}^{n} a_{ik} b_{kj}\,] + [\,\sum_{k=1}^{n} a_{ik} c_{kj}\,] \\ &= AB + AC \end{aligned}$$

Some usual properties of multiplication of scalars (like commutativity) do not hold for matrices. For example, $AB = O$ does not necessarily imply that $A = O$ or $B = O$ as for the matrices in (1.1). Similarly, $AB = AC$ does not necessarily imply that $B = C$.

Finally, as in the case of matrix addition, an expression of the form $A_1 A_2 \cdots A_k$ is unambiguous due to property M1, and can be evaluated by computing pairwise products of

adjacent matrices in any sequence without changing the original order of the matrices. For example, the product $ABCD$ can be evaluated as

$$((AB)C)D) \quad \text{or} \quad (AB)(CD) \quad \text{or} \quad (A(BC))D \quad \text{or} \quad A((BC)D) \quad \text{or} \quad A(B(CD))$$

However, a careful reader might observe that one of these equivalent expressions might be easier to compute depending on the order of the matrices (see Exercise 1.9).

Let

$$D = [\, d_{ij} \,] = \text{diag}\,[\, d_1, d_2, \ldots, d_n \,]$$

Also, let $A = [\, a_{ij} \,]_{m\times n}$, and consider the product $C = AD = [\, c_{ij} \,]_{m\times n}$. Since $d_{kj} = 0$ for $k \neq j$ and $d_{jj} = d_j$,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}d_{kj} = a_{ij}d_{jj} = a_{ij}d_j$$

Thus

$$AD = \begin{bmatrix} a_{11}d_1 & a_{12}d_2 & \cdots & a_{1n}d_n \\ a_{21}d_1 & a_{22}d_2 & \cdots & a_{2n}d_n \\ \vdots & \vdots & & \vdots \\ a_{m1}d_1 & a_{m2}d_2 & \cdots & a_{mn}d_n \end{bmatrix}$$

that is, the product AD is obtained simply by scaling the columns of A with the diagonal elements of D.

Similarly, if $B = [\, b_{ij} \,]_{n\times p}$ then

$$DB = \begin{bmatrix} d_1b_{11} & d_1b_{12} & \cdots & d_1b_{1p} \\ d_2b_{21} & d_2b_{22} & \cdots & d_2b_{2p} \\ \vdots & \vdots & & \vdots \\ d_nb_{n1} & d_nb_{n2} & \cdots & d_nb_{np} \end{bmatrix}$$

that is, the product DB is obtained by scaling the rows of B with the diagonal elements of D.

Now consider a diagonal matrix having all 1's on its diagonal. Such a matrix is called an ***identity*** matrix, denoted I or I_n if the order needs to be specified. That is,

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n\times n} = \text{diag}\,[\, 1, 1, \ldots, 1 \,]$$

Applying the above results about the product of a matrix with a diagonal matrix to $D = I$ we obtain

M5. $IA = A \quad$ and $\quad AI = A$

as further properties of matrix multiplication. In fact, it is because of these properties that I is called an identity matrix: It acts like a multiplicative identity.

1.3 Transpose and Hermitian Adjoint

Let A be an $m \times n$ matrix. The $n \times m$ matrix obtained by interchanging the rows and columns of A is called the ***transpose*** of A, denoted A^t. Thus if $A = [\,a_{ij}\,]_{m\times n}$ then $A^t = B = [\,b_{ij}\,]_{n\times m}$ where $b_{ij} = a_{ji}$ for all (i, j). From the definition it follows that $(A^t)^t = A$.

If A is a complex matrix then its conjugate transpose (obtained by transposing A and replacing every element with its complex conjugate, or vice versa) is called the ***Hermitian adjoint*** of A, denoted A^h. Thus if $A = [\,a_{ij}\,]_{m\times n}$ then $A^h = C = [\,c_{ij}\,]_{n\times m}$ where $c_{ij} = a^*_{ji}$ for all (i, j). Again, the definition implies that $(A^h)^h = A$.

Note that if A is real then $A^h = A^t$. Hence all properties concerning the Hermitian adjoint of a complex matrix are valid for the transpose of a real matrix. For this reason, from now on we will state and prove such properties only for the Hermitian adjoint. For example, the properties

$$\begin{aligned}(AB)^h &= B^h A^h\\ (A+B)^h &= A^h + B^h\\ (cA)^h &= c^* A^h\end{aligned}$$

involving complex matrices A and B and a complex scalar c can be shown in one or two steps. We can then safely state without proof that

$$\begin{aligned}(AB)^t &= B^t A^t\\ (A+B)^t &= A^t + B^t\\ (cA)^t &= cA^t\end{aligned}$$

for real matrices A and B and a real scalar c.

Clearly, the transpose or Hermitian adjoint of a row vector is a column vector, and vice versa. Also, $D^t = D$ and $D^h = D^*$ for any diagonal matrix D. Finally, $O^h_{m\times n} = O^t_{m\times n} = O_{n\times m}$ whether O is treated as a real or as a complex matrix.

A square matrix A is called ***symmetric*** if $A^t = A$, and ***skew-symmetric*** if $A^t = -A$. Thus $A = [\,a_{ij}\,]_{n\times n}$ is symmetric if and only if $a_{ij} = a_{ji}$ for all (i, j), and skew-symmetric if and only if $a_{ij} = -a_{ji}$ for all (i, j). Note that if A is skew-symmetric then $a_{ii} = -a_{ii}$, which requires that the diagonal elements should be zero.

A complex square matrix is called ***Hermitian*** if $A^h = A$, and ***skew-Hermitian*** if $A^h = -A$. Clearly, a real Hermitian matrix is symmetric, and a real skew-Hermitian matrix is skew-symmetric. Further properties of Hermitian matrices are dealt with in Exercises 1.14-1.17.

Example 1.5

The transpose and the Hermitian adjoint of the matrix B in Example 1.1 are

$$B^t = \begin{bmatrix} 1+2i & 0 & -1\\ -3 & -3i & -2\\ -1+i & 5 & 3+2i \end{bmatrix}, \quad B^h = \begin{bmatrix} 1-2i & 0 & -1\\ -3 & 3i & -2\\ -1-i & 5 & 3-2i \end{bmatrix}$$

The matrices

$$\begin{bmatrix} 1 & 3 & -1\\ 3 & 0 & 2\\ -1 & 2 & 4 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & -2\\ -1 & 0 & 0\\ 2 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1+i\\ 1-i & 2 \end{bmatrix}$$

are symmetric, skew-symmetric, and Hermitian, respectively.

1.4 Partitioned Matrices

Let A be any matrix. By deleting some of the rows and some of the columns of A we obtain a smaller matrix called a ***submatrix*** of A.

Let us partition the rows of an $m \times n$ matrix A into p groups of size $m_1, \ldots, m_p$, and the columns into q groups of size $n_1, \ldots, n_q$, where

$$\sum_{i=1}^{p} m_i = m\,, \quad \sum_{j=1}^{q} n_j = n$$

Representing the submatrix of A consisting of the ith group of m_i rows and the jth group of n_j columns as A_{ij}, we write

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2q} \\ \vdots & \vdots & & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pq} \end{bmatrix} = [\, A_{ij} \,]$$

Each submatrix A_{ij} in the above representation is called a ***block*** of A.[6] In general, the blocks A_{ij} are of different order; however, all blocks in the same row block have the same number of rows, and all blocks in the same column block have the same number of columns. Note also that only the rows (but not the columns) or only the columns (but not the rows) of a matrix may be partitioned.

Partitioning may be useful to display structure of some special matrices. For example, an identity matrix may be expressed in terms of its columns or in terms of its rows as

$$I_n = \left[\begin{array}{c|c|c|c} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & & 1 \end{array}\right] = [\, \boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\cdots\; \boldsymbol{e}_n \,]$$

$$= \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & \cdots & 0 \\ \hline & & \vdots & \\ \hline 0 & 0 & \cdots & 1 \end{array}\right] = \begin{bmatrix} \boldsymbol{e}_1^t \\ \boldsymbol{e}_2^t \\ \vdots \\ \boldsymbol{e}_n^t \end{bmatrix}$$

As another example, partitioning allows us to define a ***block diagonal*** matrix as

$$A = \left[\begin{array}{cc|ccc} a_{11} & a_{12} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ \hline 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & a_{43} & a_{44} & a_{45} \\ 0 & 0 & a_{53} & a_{54} & a_{55} \end{array}\right] = \text{diag}\,[\, A_1, A_2 \,]$$

[6] Representing a partitioned matrix A in terms of its blocks does not mean that A is a matrix with elements themselves being matrices. However, the blocks of a partitioned matrix may be treated just like its elements in some matrix operations as explained later in this section.

where

$$A_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} a_{33} & a_{34} & a_{35} \\ a_{43} & a_{44} & a_{45} \\ a_{53} & a_{54} & a_{55} \end{bmatrix}$$

We can similarly define upper or lower ***block triangular*** matrices.

If A and B are matrices of the same order and are partitioned in exactly the same way so that their corresponding blocks are of the same order, then the sum $A + B$ can be obtained by adding the corresponding blocks. That is, if

$$A = [A_{ij}] \quad \text{and} \quad B = [B_{ij}]$$

where the blocks A_{ij} and B_{ij} are of the same order for all (i, j), then

$$A + B = [A_{ij} + B_{ij}]$$

If A and B are partitionqed matrices compatible for the product AB, then the product can be obtained by treating the blocks of A and B as if they were their elements. That is, if $A = [\, A_{ij} \,]$ and $B = [\, B_{ij} \,]$ then $AB = C = [\, C_{ij} \,]$, where

$$C_{ij} = \sum_k A_{ik} B_{kj}$$

Of course, this requires that the blocks A_{ik} and B_{kj} be compatible for the products $A_{ik}B_{kj}$ for all (i, k, j). In other words, columns of A must be partitioned in exactly the same way as the rows of B are partitioned.

Block multiplication is useful in expressing matrix products in a compact form. For example, if A is an $m \times n$ matrix and B is an $n \times q$ matrix partitioned into its columns as

$$B = [\, \boldsymbol{b}_1 \;\; \boldsymbol{b}_2 \;\; \cdots \;\; \boldsymbol{b}_q \,]$$

where each $\boldsymbol{b}_j$ is an $n \times 1$ column vector, then the product AB can be obtained by block multiplication as

$$AB = A\,[\, \boldsymbol{b}_1 \;\; \boldsymbol{b}_2 \;\; \cdots \;\; \boldsymbol{b}_q \,] = [\, A\boldsymbol{b}_1 \;\; A\boldsymbol{b}_2 \;\; \cdots \;\; A\boldsymbol{b}_q \,]$$

Observe that the product is also partitioned into its columns, the jth column being an $m \times 1$ vector obtained by premultiplying the jth column of B with A.

Similarly, if A is partitioned into its rows as

$$A = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{bmatrix}$$

where each $\boldsymbol{\alpha}_i$ is a $1 \times n$ row vector, then the product AB can be expressed as

$$AB = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{bmatrix} B = \begin{bmatrix} \boldsymbol{\alpha}_1 B \\ \boldsymbol{\alpha}_2 B \\ \vdots \\ \boldsymbol{\alpha}_m B \end{bmatrix}$$

Now the product is partitioned into its rows, the ith row being a $1 \times q$ vector obtained by postmultiplying the ith row of A with B.

If both A and B are partitioned as above (A into its rows and B into its columns), then $AB = C = [\, C_{ij} \,]_{m \times q}$, where each block $C_{ij} = \boldsymbol{\alpha}_i \boldsymbol{b}_j$ is a scalar. In fact, $C_{ij} = c_{ij}$, the (i, j)th element of C, as expected. This is actually how matrix multiplication is defined. The (i, j)th element of the product AB is the product of the ith row of A with the jth column of B.

Alternatively, we may choose to partition A into its columns and B into its rows as

$$A = [\, \boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \cdots \;\; \boldsymbol{a}_n \,] \,, \quad B = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_n \end{bmatrix}$$

where now each $\boldsymbol{a}_i$ is an $m \times 1$ column vector and each $\boldsymbol{\beta}_j$ is a $1 \times q$ row vector. Then the product AB consists of only one block given as

$$AB = \boldsymbol{a}_1 \boldsymbol{\beta}_1 + \boldsymbol{a}_2 \boldsymbol{\beta}_2 + \ldots + \boldsymbol{a}_n \boldsymbol{\beta}_n$$

where each product term $\boldsymbol{a}_i \, \boldsymbol{\beta}_i$ is an $m \times q$ matrix.

As a final property of partitioned matrices we note that if $A = [\, A_{ij} \,]$ then $A^h = [\, A^h_{ji} \,]$ as the reader can verify by examples.

1.5 Systems of Linear Equations

An $m \times n$ system of linear equations, or an $m \times n$ ***linear system***, is a set of m equations in n unknown variables $x_1, x_2, \ldots, x_n$, written as

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ & & & & & & & \vdots & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array} \tag{1.2}$$

where the coefficients a_{ij} and the constants b_i are fixed scalars. Letting

$$A = [\, a_{ij} \,] \,, \quad \boldsymbol{x} = \text{col}\,[\, x_1, x_2, \ldots, x_n \,] \,, \quad \boldsymbol{b} = \text{col}\,[\, b_1, b_2, \ldots, b_n \,]$$

the system in (1.2) can be written in matrix form as

$$A\boldsymbol{x} = \boldsymbol{b} \tag{1.3}$$

A is called the ***coefficient matrix*** of (1.3). If $\boldsymbol{b} = \boldsymbol{0}$, then the system (1.3) is said to be ***homogeneous***.

An $n \times 1$ column vector $\boldsymbol{x} = \boldsymbol{\phi}$ is called a ***solution*** of (1.3) if $A\boldsymbol{\phi} = \boldsymbol{b}$. A system may have no solution, a unique solution, or more than one solution. If it has at least one solution, it is said to be ***consistent***, otherwise, ***inconsistent***. A homogeneous system is always consistent as it has at least the ***trivial*** (null) solution $\boldsymbol{x} = \boldsymbol{0}$.

We are interested in the following problems associated with a linear system:

a) Determine whether the system is consistent.

b) If it is consistent

   i. determine if it has a unique solution or many solutions,

   ii. if it has a unique solution find it,

   iii. if it has many solutions, find a solution or all solutions.

c) If it is inconsistent find $\boldsymbol{x} = \boldsymbol{\phi}$ that is closest to being a solution.

In this section we will deal with problems (a) and (b), leaving (c) to Chapter 7.

Example 1.6

The system

$$\begin{array}{rcrcl} x_1 & - & x_2 & = & 1 \\ x_1 & + & x_2 & = & 5 \end{array} \tag{1.4}$$

has a unique solution

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \tag{1.5}$$

which can be obtained graphically as illustrated in Figure 1.1(a). Each of the equations describes a straight line in the x_1x_2 plane, and the solution is their intersection point.

The equations of the system

$$\begin{array}{rcrcr} x_1 & - & x_2 & = & 1 \\ 2x_1 & - & 2x_2 & = & -6 \end{array}$$

describe parallel lines in the x_1x_2 plane as illustrated in Figure 1.1(b). Since there is no point common to both lines, the system has no solution.

On the other hand, the equations of the system

$$\begin{array}{rcrcl} x_1 & - & x_2 & = & 1 \\ 2x_1 & - & 2x_2 & = & 2 \end{array}$$

are proportional, and describe the same line shown in Figure 1.1(c). Since any point on this line satisfies both equations, the system has infinitely many solutions. To characterize these solutions we choose one of the variables, say x_2, arbitrarily as $x_2 = c$, and determine the other variable from either of the equations as $x_1 = 1+c$. Thus we obtain a one-parameter family of solutions described as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1+c \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The geometric interpretation of equations and their solution(s) can easily be generalized to systems containing three variables x_1, x_2 and x_3, where each equation defines a plane in the $x_1x_2x_3$ space, and any point common to all planes, if it exists, defines a solution (see Exercise 1.29). However, when there are more than three variables, the geometric interpretation loses much of its appeal (for we can not visualize what an equation in four or more variables describes), and we resort to pure algebraic methods. One such method is the elimination method, which we recall by reconsidering the example above.

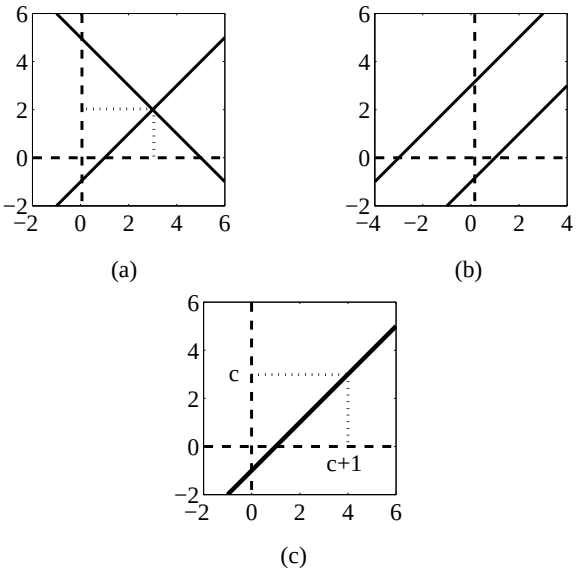

Figure 1.1: Geometric representation of systems in Example 1.6

Example 1.7

Let us consider the system in (1.4). One way of finding the solution is to express one of the variables in terms of the other by using one of the equations, and then to substitute this expression into the other equation. For example, by expressing x_1 in terms of x_2 using the first equation we get

$$x_1 = 1 + x_2$$

Substituting this expression for x_1 into the second equation and rearranging the terms we obtain

$$2x_2 = 4 \tag{1.6}$$

The last equation gives $x_2 = 2$, and from the expression for x_1 we get

$$x_1 = 1 + x_2 = 1 + 2 = 3$$

thus obtaining the solution in (1.5).

What we did by expressing x_1 in terms of x_2 and substituting this expression into the second equation was to eliminate x_1 from the second equation. We could do this simply by subtracting the first equation from the second, which would give (1.6) directly.

Alternatively, we could eliminate x_2 from the second equation by adding the first equation to it. This would give

$$2x_1 = 6$$

from which we would obtain $x_1 = 3$. Then from the first equation we would get

$$x_2 = x_1 - 1 = 3 - 1 = 2$$

reaching the same solution.

For the simple example considered above, it makes no difference whether we eliminated x_1 or x_2. However, to solve larger equations we need to be more systematic in the elimination process, especially if we are using a computer program to do the job. A systematic procedure is based on transforming the given system into a simpler equivalent system in which variables can be solved one after the other by successive substitutions as we explain by the following example.

Example 1.8

Let us solve the system

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & - & x_3 & = & 1 \\ -2x_1 & + & 8x_2 & - & x_3 & = & 5 \\ 2x_1 & - & 6x_2 & + & 2x_3 & = & -4 \end{array} \tag{1.7}$$

by using the elimination method.

We first eliminate the variable x_1 from all equations except one. Since it appears in all three equations, we can associate it with any one of them and eliminate from the other two. Associating x_1 arbitrarily with the first equation, we eliminate it from the second and third equations by adding 2 times the first equation to the second and -2 times the first equation to the third. After these manipulations the equations become

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & - & x_3 & = & 1 \\ & & 4x_2 & - & 3x_3 & = & 7 \\ & - & 2x_2 & + & 4x_3 & = & -6 \end{array} \tag{1.8}$$

Next we eliminate x_2 from one of the last two equations. This can be done by associating x_2 with the second equation and eliminating it from the third (by adding $1/2$ times the second equation to the third). Alternatively, we may associate x_2 with the third equation and eliminate it from the second (by adding 2 times the third equation to the second). To avoid dealing with fractions we choose the latter. However, before the elimination we first interchange the second and third equations so that the equation with which x_2 is associated comes before those from which it is to be eliminated. This gives

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & - & x_3 & = & 1 \\ & - & 2x_2 & + & 4x_3 & = & -6 \\ & & 4x_2 & - & 3x_3 & = & 7 \end{array} \tag{1.9}$$

Now we add 2 times the second equation to the third and obtain

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & - & x_3 & = & 1 \\ & - & 2x_2 & + & 4x_3 & = & -6 \\ & & & & 5x_3 & = & -5 \end{array} \tag{1.10}$$

The system in (1.10) has a triangular shape which allows us to solve the unknown variables starting from the last equation and working backwards. From the last equation we obtain

$$x_3 = -5/5 = -1$$

Substituting the value of x_3 into the second equation we find

$$x_2 = (-1/2)(-6 - 4x_3) = (-1/2)(-6 + 4) = 1$$

Finally, substituting the values of x_2 and x_3 into the first equation we get

$$x_1 = 1 + 2x_2 + x_3 = 1 + 2 - 1 = 2$$

Thus we obtain the solution of the system (1.7) as

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} \tag{1.11}$$

Instead of finding x_2 and x_1 by successive backward substitutions, we can continue elimination of the variables in (1.10) in the reverse order. Starting with the system in (1.10), we first multiply the last equation with $1/5$ to normalize the coefficient of x_3 to 1, and then add 1 and -4 times the resulting equation to the first and the second equations to eliminate x_3 from the first two equations, respectively. After these operations the system becomes

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & & & = & 0 \\ & - & 2x_2 & & & = & -2 \\ & & & & x_3 & = & -1 \end{array}$$

Now we scale the second equation with $-1/2$, add 2 times the resulting equation to the first to eliminate x_2 from the first equation, and thus obtain

$$\begin{array}{rcrcrcr} x_1 & & & & & = & 2 \\ & & x_2 & & & = & 1 \\ & & & & x_3 & = & -1 \end{array} \tag{1.12}$$

Note that in the last two steps we not only eliminated the variables x_2 and x_3 from the first two equations but also normalized their coefficients to 1. (Coefficient of x_1 in the first equation was already 1 at the start, so we need not do anything about it.) The final system in (1.12) is so simple that it displays the solution.

Backward elimination need not wait until forward elimination is completed; they can be performed simultaneously. Consider the system in (1.9) in which x_1 is already eliminated from the second and third equations. Scaling the second equation with $-1/2$, and adding 2 and -4 times the resulting equation to the first and the third equations, we eliminate x_2 not only from the third equation but also from the first equations, and get

$$\begin{array}{rcrcrcr} x_1 & & & - & 5x_3 & = & 7 \\ & & x_2 & - & 2x_3 & = & 3 \\ & & & & 5x_3 & = & -5 \end{array}$$

Now, multiplying the third equation with $1/5$ and adding 5 and 2 times the resulting equation to the first and second equations, we eliminate x_3 from these equations and end up with (1.12). However, a careful reader may observe that it is not smart to perform forward and backward eliminations simultaneously (see Exercise 1.38).

Example 1.8 illustrates the recursive nature of the elimination method. After the elimination of x_1, the last two equations in (1.8) form a 2×2 system

$$\begin{array}{rcrcr} 4x_2 & - & 3x_3 & = & 7 \\ -2x_2 & + & 4x_3 & = & -6 \end{array} \tag{1.13}$$

which is obviously easier to solve than the original 3×3 system. Once x_2 and x_3 are solved from (1.13), x_1 can easily be found from the first equation in (1.8) by substitution. Now the process can be repeated for (1.13) to further reduce it to a simpler system. This is exactly what we do when we eliminate x_2 from one of the equations in (1.13) to reach the triangular

system in (1.10). Thus at every step of the elimination process, both the number of equations and the number of unknowns are reduced by at least one.[7]

Let us consider the system (1.7) in matrix form:

$$\begin{bmatrix} 1 & -2 & -1 \\ -2 & 8 & -1 \\ 2 & -6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ -4 \end{bmatrix} \tag{1.14}$$

After the first two operations the equations take the form in (1.8), which has the matrix representation

$$\begin{bmatrix} 1 & -2 & -1 \\ 0 & 4 & -3 \\ 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 7 \\ -6 \end{bmatrix}$$

Apparently, the operation of adding 2 times the first equation to the second equation is reflected as adding 2 times the first row of the coefficient matrix in (1.14) to its second row, and at the same time, adding 2 times the first element of the column matrix on the right-hand side of (1.14) to its second element. Since the same operations are involved in the coefficient matrix and the column on the right-hand side, we conveniently form an ***augmented matrix***

$$[\,A \;\; \boldsymbol{b}\,] = \left[\begin{array}{ccc|c} 1 & -2 & -1 & 1 \\ -2 & 8 & -1 & 5 \\ 2 & -6 & 2 & -4 \end{array}\right]$$

associated with the system in (1.7), and represent the operations leading to (1.12) as row operations on the augmented matrix as

$$\left[\begin{array}{ccc|c} 1 & -2 & -1 & 1 \\ -2 & 8 & -1 & 5 \\ 2 & -6 & 2 & -4 \end{array}\right] \begin{array}{c} 2\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ -2\mathrm{R}_1 + \mathrm{R}_3 \to \mathrm{R}_3 \\ \longrightarrow \end{array} \left[\begin{array}{ccc|c} 1 & -2 & -1 & 1 \\ 0 & 4 & -3 & 7 \\ 0 & -2 & 4 & -6 \end{array}\right]$$

$$\begin{array}{c} \mathrm{R}_2 \leftrightarrow \mathrm{R}_3 \\ 2\mathrm{R}_2 + \mathrm{R}_3 \to \mathrm{R}_3 \\ \longrightarrow \end{array} \left[\begin{array}{ccc|c} 1 & -2 & -1 & 1 \\ 0 & -2 & 4 & -6 \\ 0 & 0 & 5 & -5 \end{array}\right]$$

$$\begin{array}{c} (1/5)\mathrm{R}_3 \to \mathrm{R}_3 \\ \mathrm{R}_3 + \mathrm{R}_1 \to \mathrm{R}_1 \\ -4\mathrm{R}_3 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \longrightarrow \end{array} \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & -2 & 0 & -2 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

$$\begin{array}{c} -(1/2)\mathrm{R}_2 \to \mathrm{R}_2 \\ 2\mathrm{R}_2 + \mathrm{R}_1 \to \mathrm{R}_1 \\ \longrightarrow \end{array} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{array}\right]$$

The first two steps of the above operations correspond to forward elimination of the variables, and the last two steps correspond to backward elimination. The notation "$\mathrm{R}_i \leftrightarrow \mathrm{R}_j$"

[7] Although in Example 1.8 the number of equations and and the number of unknowns are reduced by exactly one, we will later see examples where either or both are reduced by more than one.

denotes interchange of the ith and the jth rows, "$c\mathrm{R}_i \to \mathrm{R}_i$" denotes multiplication of the ith row by a nonzero scalar c, and "$c\mathrm{R}_i + \mathrm{R}_j \to \mathrm{R}_j$" denotes addition of c times the ith row to the jth row. Note that after the operation $\mathrm{R}_2 \leftrightarrow \mathrm{R}_3$ at the second step above, R_2 and R_3 denote the current second and third rows.

The following three types of operations on the rows of the augmented matrix are involved in the elimination process.

I: Interchange any two rows

II: Multiply any row by a nonzero scalar

III: Add any scalar multiple of a row to another row

These operations performed on the rows of a matrix are called ***elementary row operations***. An $m \times n$ matrix B is said to be ***row equivalent*** to an $m \times n$ matrix A if it can be obtained from A by a finite number of elementary row operations. Clearly, to every elementary row operation there corresponds an inverse elementary row operation of the same kind. For example, the inverse of adding c times the ith row to the jth row is to add $-c$ times the ith row to the jth row. If B is obtained from A by a single elementary row operation, then A can be restored from B by performing the inverse operation. Thus if B is row equivalent to A, then A is row equivalent to B, and we say that A and B are row equivalent.[8] Two $m \times n$ systems of linear equations are said to be ***equivalent*** if their augmented matrices are row equivalent. Two equivalent systems either have the same solution(s), or are both inconsistent. We have already used this fact in Example 1.8 to find the solution of a system. Below we consider another example.

Example 1.9

Find the value of the parameter q such that the system

$$\begin{bmatrix} 1 & 1 & -1 & 2 & 0 \\ 2 & 2 & -2 & 3 & 2 \\ -1 & -1 & 3 & -4 & 2 \\ 1 & 1 & 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1 \\ -3 \\ 3 \\ q \end{bmatrix} \tag{1.15}$$

is consistent, and then find all solutions.

We first associate x_1 with the first equation and eliminate it from the last three equations. The operations involved in the elimination of x_1 are represented by elementary row operations on the

[8] An equivalence relation defined on a set $\boldsymbol{\mathcal{S}}$, denoted $\equiv$, is a reflexive, symmetric and transitive relation among the elements of $\boldsymbol{\mathcal{S}}$. That is, for all $a, b, c \in \boldsymbol{\mathcal{S}}$, $a \equiv a$, if $a \equiv b$ then $b \equiv a$, and if $a \equiv b$ and $b \equiv c$ then $a \equiv c$. In this sense, row equivalence is indeed an equivalence relation on $\mathbb{F}^{m\times n}$. An equivalence relation partitions $\boldsymbol{\mathcal{S}}$ into disjoint subsets, called equivalence classes, such that every element of the set belongs to one and only one equivalence class and any two equivalent elements belong to the same equivalence class. Hence row equivalence partitions $\mathbb{F}^{m\times n}$ into equivalence classes such that any two matrix in the same equivalence class can be obtained from each other by a finite sequence of elementary row operations.

augmented matrix as

$$\left[\begin{array}{rrrrr|r} 1 & 1 & -1 & 2 & 0 & -1 \\ 2 & 2 & -2 & 3 & 2 & -3 \\ -1 & -1 & 3 & -4 & 2 & 3 \\ 1 & 1 & 2 & 1 & -1 & q \end{array}\right]$$

$$\begin{array}{r} -2\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \mathrm{R}_1 + \mathrm{R}_3 \to \mathrm{R}_3 \\ -\mathrm{R}_1 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \left[\begin{array}{rrrrr|r} 1 & 1 & -1 & 2 & 0 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 2 & -2 & 2 & 2 \\ 0 & 0 & 3 & -1 & -1 & q+1 \end{array}\right]$$

At this point we observe that incidentally x_2 is also eliminated from the last three equations. We then continue with the elimination of the next variable, x_3, which appears in the third and fourth equations, and must be associated with one of them. We associate x_3 with the third equation and interchange the second and third equations. The rest of the elimination process is straightforward, and is summarized below.

$$\begin{array}{r} \mathrm{R}_2 \leftrightarrow \mathrm{R}_3 \\ -(3/2)\mathrm{R}_2 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \left[\begin{array}{rrrrr|r} 1 & 1 & -1 & 2 & 0 & -1 \\ 0 & 0 & 2 & -2 & 2 & 2 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 2 & -4 & q-2 \end{array}\right]$$

$$\begin{array}{r} \\ 2\mathrm{R}_3 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \begin{array}{c} \downarrow \quad\quad \downarrow \quad \downarrow \\ \left[\begin{array}{rrrrr|r} 1 & 1 & -1 & 2 & 0 & -1 \\ 0 & 0 & 2 & -2 & 2 & 2 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & q-4 \end{array}\right] \end{array} \tag{1.16}$$

$$\begin{array}{r} -\mathrm{R}_3 \to \mathrm{R}_3 \\ -2\mathrm{R}_3 + \mathrm{R}_1 \to \mathrm{R}_1 \\ 2\mathrm{R}_3 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \longrightarrow \end{array} \quad \left[\begin{array}{rrrrr|r} 1 & 1 & -1 & 0 & 4 & -3 \\ 0 & 0 & 2 & 0 & -2 & 4 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & q-4 \end{array}\right]$$

$$\begin{array}{r} \\ (1/2)\mathrm{R}_2 \to \mathrm{R}_2 \\ \mathrm{R}_2 + \mathrm{R}_1 \to \mathrm{R}_1 \\ \longrightarrow \end{array} \quad \begin{array}{c} \downarrow \quad\quad \downarrow \quad \downarrow \\ \left[\begin{array}{rrrrr|r} 1 & 1 & 0 & 0 & 3 & -1 \\ 0 & 0 & 1 & 0 & -1 & 2 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & q-4 \end{array}\right] \end{array} \tag{1.17}$$

In the above sequence of elementary row operations, steps leading to (1.16) correspond to forward elimination of the variables associated with the columns marked by arrows, and those leading to (1.17) correspond to scaling of the equations and backward elimination of the same variables.

The last equation associated with the augmented matrix in (1.17) is

$$0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 + 0 \cdot x_5 = q - 4$$

from which we observe that if $q \neq 4$ then the system is inconsistent as this equation is never satisfied. On the other hand, if $q = 4$ then this equation is a trivial identity $0 = 0$ for any choice of the variables, and can be discarded. Then using the first three equations we express the variables associated with the marked columns of the augmented matrix in terms of the others as

$$\begin{aligned} x_1 &= -1 - x_2 - 3x_5 \\ x_3 &= 2 + x_5 \\ x_4 &= 1 + 2x_5 \end{aligned} \tag{1.18}$$

From (1.18) we see that we can choose the variables x_2 and x_5 arbitrarily, and calculate x_1, x_3 and x_4 from these relations to obtain a solution. Letting $x_2 = c_1$ and $x_5 = c_2$, where $c_1, c_2 \in \mathbb{R}$ are arbitrary, and calculating x_1, x_3 and x_4 from (1.18), we obtain the solution in parametric form as

$$\begin{aligned} x_1 &= -1 - c_1 - 3c_2 \\ x_2 &= \quad c_1 \\ x_3 &= 2 + \quad c_2 \\ x_4 &= 1 + \quad 2c_2 \\ x_5 &= \quad c_2 \end{aligned}$$

or equivalently, as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 2 \\ 1 \\ 0 \end{bmatrix} + c_1 \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} -3 \\ 0 \\ 1 \\ 2 \\ 1 \end{bmatrix} \tag{1.19}$$

Note that we could obtain (1.18) and hence the solution in (1.19) from the augmented matrix in (1.16) by back substitution of the marked variables.

Unlike the system in (1.7), which has the unique solution given in (1.11), the system in (1.15) is either inconsistent (if $q \neq 4$) or has infinitely many solutions as given in (1.19). For example,

$$\boldsymbol{x} = \begin{bmatrix} -1 \\ 0 \\ 2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

are two solutions corresponding to the choice of the arbitrary constants as $(c_1, c_2) = (0, 0)$ and $(c_1, c_2) = (1, -1)$, respectively.

Example 1.9 outlines a general procedure to determine whether a given system is consistent and to find solutions when it is consistent. It is based on transforming the coefficient matrix A into a simple form by performing elementary row operations on the augmented matrix. We now give a precise definition of what we mean by a simple form.

An $m \times n$ matrix R is said to be in ***row echelon form*** if it has the following characteristics.

i) First r rows of R are nonzero, and the remaining $m - r$ rows are zero for some $0 \leq r \leq m$.

ii) The first nonzero element in each of the first $r - 1$ rows lies to the left of the first nonzero element in the subsequent row. (If $r = 0$ or $r = 1$ this item does not apply.)

The number of nonzero rows, r, is called the ***row rank*** of R. The first nonzero element of each of the first r rows is called the ***leading entry*** of its row, and the column which contains the leading entry of the ith row is called the ith ***basic column***. Thus, if $1 \leq i < p \leq r$, then the ith basic column lies to the left of the pth basic column. Note that this requirement is equivalent to R having all zero elements below a jagged diagonal defined by the leading entries.

A matrix R in row echelon form is said to be in ***reduced row echelon form*** if it satisfies the following additional conditions.

iii) Each leading entry is a 1.

iv) The ith leading entry is the only nonzero element in the ith basic column.

For example, the matrices

$$R_1 = \begin{bmatrix} 0 & 1 & -1 & 2 & -4 \\ 0 & 0 & 1 & -3 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad R_2 = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

are in row echelon form with $r(R_1) = 3$ and $r(R_2) = 2$, and the matrix

$$R_3 = \begin{bmatrix} 1 & -3 & 0 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is in reduced row echelon form. As further examples, the reader can check that the coefficient matrix associated with the augmented matrix in (1.16) is in row echelon form, and the coefficient matrix associated with the augmented matrix in (1.17) is in reduced row echelon form.

An algorithm to bring a given $m \times n$ matrix into reduced row echelon form by means of elementary row operations is given in Table 1.1. The algorithm, which is known as ***Gaussian elimination***, simply imitates the steps involved in Example 1.9.

The algorithm returns the reduced row echelon form of A written over A, the rank of the reduced row echelon form (r), and the column indices of the basic columns ($j_1, \ldots, j_r$). In steps 1-11, forward elimination is performed and A is brought to a row echelon form. The nonzero element a_{pj} found in Step 3 of the algorithm is called a ***pivot element*** of the jth column. After incrementing r in Step 4, the pivot element is brought to the (r, j)th position by a row interchange in Step 6 (if it is not already there), and it becomes the leading entry of its row. In steps 7-10, the pivot element is used to nullify the elements below it. In steps 12-18, leading entries are normalized to unity and A is brought into reduced row echelon form by means of backward elimination.[9]

Whether a system of linear equations is consistent can be determined from the reduced row echelon form of the augmented matrix of the system. If it is consistent, all solutions can be obtained in parametric form by choosing the non-basic variables arbitrarily and expressing the basic variables in terms of the non-basic variables. (The basic and non-basic variables are the unknowns corresponding to the basic and non-basic columns of the coefficient matrix.)

As illustrated in Example 1.9, the general form of the solution is

$$\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu = \boldsymbol{\phi}_p + \boldsymbol{\phi}_c$$

where $\nu = n - r$ is the number of non-basic variables. $\boldsymbol{\phi}_p$ and $\boldsymbol{\phi}_c$ are called a ***particular solution*** and the ***complementary solution***, respectively. Obviously, when $r = n$, i.e., when there are no non-basic variables that can be chosen arbitrarily, then the complementary solution does not exist. The significance of particular and complementary solutions is studied in the next section.

[9] Some books use the term "Gaussian elimination" to refer to the forward elimination process only, and call the complete algorithm in Table 1.1 the ***Gauss-Jordan algorithm***.

Table 1.1: Gaussian Elimination Algorithm

1. Set $r = 0$

   [Forward Elimination]

2. For $j = 1 : n$
3. Find $r < p \leq m$ such that $a_{pj} \neq 0$. If none, increment j
4. $r \leftarrow r + 1$
5. $j_r \leftarrow j$
6. If $p > r$, $\mathrm{R}_p \leftrightarrow \mathrm{R}_r$
7. For $i = r + 1 : m$
8. $\mu_{ij} \leftarrow a_{ij}/a_{rj}$
9. $\mathrm{R}_i \leftarrow \mathrm{R}_i - \mu_{ij}\mathrm{R}_r$
10. End
11. End

    [Backward Elimination]

12. For $p = r : 2$
13. $\mathrm{R}_p \leftarrow (1/a_{pj_p})\mathrm{R}_p$
14. For $i = 1 : p - 1$
15. $\mathrm{R}_i \leftarrow \mathrm{R}_i - a_{ij_p}\mathrm{R}_p$
16. End
17. End
18. $\mathrm{R}_1 \leftarrow (1/a_{1j_1})\mathrm{R}_1$

Depending on the choice of the pivot element, Gaussian Elimination may produce different matrices in row echelon form at the end of step 11. However, upon completion of the algorithm, we end up with a unique matrix in reduced row echelon form. In fact, uniqueness of the reduced row echelon form is not specific to Gaussian elimination, but is a result of the fact that any given matrix A is row equivalent to a unique reduced row echelon matrix, which we define as the reduced row echelon form of A. In other words, independent of the algorithm used, if a finite sequence of elementary row operations on A results in a reduced row echelon matrix R, then R is the unique reduced row echelon form of A.[10] Consequently, all matrices which are row equivalent to A have the same reduced row echelon form, which can be interpreted as a convenient representative of its equivalence class.

The ***row rank*** of a matrix, denoted $r(A)$, is defined to be the row rank of its unique reduced row echelon form. Thus all row equivalent matrices have the same row rank. From the reduced row echelon form we deduce that if A is $m \times n$, then not only $r(A) \leq m$ but also $r(A) \leq n$.

[10] We shall prove this fact in Chapter 4.

Example 1.10

In Example 1.9 we obtained a row echelon form of the coefficient matrix in (1.15) as in (1.16). A different row echelon form of the same matrix can be obtained by choosing different pivots as

$$\begin{bmatrix} 1 & 1 & -1 & 2 & 0 \\ 2 & 2 & -2 & 3 & 2 \\ -1 & -1 & 3 & -4 & 2 \\ 1 & 1 & 2 & 1 & -1 \end{bmatrix}$$

$$\begin{array}{r} \mathrm{R}_3 \leftrightarrow \mathrm{R}_1 \\ 2\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \mathrm{R}_1 + \mathrm{R}_3 \to \mathrm{R}_3 \\ \mathrm{R}_1 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \begin{bmatrix} -1 & -1 & 3 & -4 & 2 \\ 0 & 0 & 4 & -5 & 6 \\ 0 & 0 & 2 & -2 & 2 \\ 0 & 0 & 5 & -3 & 1 \end{bmatrix}$$

$$\begin{array}{r} \mathrm{R}_3 \leftrightarrow \mathrm{R}_2 \\ -2\mathrm{R}_2 + \mathrm{R}_3 \to \mathrm{R}_3 \\ -(5/2)\mathrm{R}_2 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \begin{bmatrix} -1 & -1 & 3 & -4 & 2 \\ 0 & 0 & 2 & -2 & 2 \\ 0 & 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 2 & -4 \end{bmatrix}$$

$$\begin{array}{r} 2\mathrm{R}_3 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \quad \begin{bmatrix} -1 & -1 & 3 & -4 & 2 \\ 0 & 0 & 2 & -2 & 2 \\ 0 & 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{1.20}$$

However, if we perform backward elimination on the matrix in (1.20) we end up with the same matrix in (1.17). This illustrates the uniqueness of the reduced row echelon form. Observe that both row echelon forms in (1.16) and (1.20) as well as the reduced row echelon form in (1.17) have the same row rank, $r = 3$, which is the row rank of the coefficient matrix in (1.15).

Example 1.11

While computing the reduced row echelon form of a matrix by hand, simplifying the matrix as much as possible before selecting the pivot element may help avoid dealing with complex numbers or fractions. Consider the matrix

$$\begin{bmatrix} 1+i & i & i \\ 2+i & 1+i & 1+2i \end{bmatrix}$$

A straightforward application of the Gaussian elimination algorithm requires operations with complex numbers. However, much of the complex arithmetic can be avoided at the expense of more row operations as

$$\begin{bmatrix} 1+i & i & i \\ 2+i & 1+i & 1+2i \end{bmatrix} \quad \begin{array}{c} -\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1+i & i & i \\ 1 & 1 & 1+i \end{bmatrix}$$

$$\begin{array}{r} \mathrm{R}_1 \leftrightarrow \mathrm{R}_2 \\ -\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1 & 1 & 1+i \\ i & -1+i & -1 \end{bmatrix}$$

$$\begin{array}{r} -i\mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1 & 1 & 1+i \\ 0 & -1 & -i \end{bmatrix}$$

$$\begin{array}{c} -\mathrm{R}_2 \leftrightarrow \mathrm{R}_2 \\ \mathrm{R}_1 - \mathrm{R}_2 \ \to \ \mathrm{R}_1 \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & i \end{bmatrix}$$

* 1.6 Solution Properties of Linear Equations

Suppose that the $m \times n$ system in (1.3) is transformed into

$$R\boldsymbol{x} = \boldsymbol{d} \tag{1.21}$$

by elementary row operations, where R is the reduced row echelon form of A. Let $r(A) = r(R) = r$.

We first consider the general case where $r < \min\{m, n\}$.

Partitioning the rows of R into two groups consisting of the nonzero and zero rows, (1.21) can be written as

$$\begin{bmatrix} F \\ O \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix} \tag{1.22}$$

where F is an $r \times n$ matrix consisting of the nonzero rows of R, O is the $(m - r) \times n$ zero matrix, and $\boldsymbol{p}$ and $\boldsymbol{q}$ are the corresponding blocks of $\boldsymbol{d}$.

Observe that if $\tilde{\boldsymbol{x}}$ is obtained from $\boldsymbol{x}$ by reordering its components and $\tilde{F}$ is obtained from F by the same reordering of its columns, then (1.22) is equivalent to

$$\begin{bmatrix} \tilde{F} \\ O \end{bmatrix} \tilde{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix} \tag{1.23}$$

In particular, let the reordering of the components of $\boldsymbol{x}$ be such that the basic variables (corresponding to the basic columns of F) occupy the first r positions, and the non-basic variables occupy the last $n - r$ positions. That is,

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix}$$

where $\boldsymbol{u}$ contains the basic variables, and $\boldsymbol{v}$ contains the non-basic variables. Since the jth basic column of F contains all 0's except a 1 in the jth position, the basic columns of F make up the matrix I_r. Hence, $\tilde{F}$ is of the form

$$\tilde{F} = [\, I_r \;\; H \,]$$

where H consists of the non-basic columns of F. With $\tilde{F}$ and $\tilde{\boldsymbol{x}}$ partitioned this way, (1.23) can be written as

$$\begin{bmatrix} I & H \\ O & O \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix} \tag{1.24}$$

Referring to the system in Example 1.9, (1.22) corresponds to

$$\left[\begin{array}{ccccc} 1 & 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -2 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \left[\begin{array}{c} -1 \\ 2 \\ 1 \\ \hline q - 4 \end{array}\right] \tag{1.25}$$

which can be written down from the augmented matrix in (1.17). The basic variables are x_1, x_3 and x_4, and the non-basic variables are x_2 and x_5. Reordering the variables so that the

basic variables appear before the non-basic variables, and performing the same reordering on the columns of the coefficient matrix, (1.25) becomes

$$\left[\begin{array}{ccc|cc} 1 & 0 & 0 & 1 & 3 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -2 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_3 \\ x_4 \\ \hline x_2 \\ x_5 \end{array}\right] = \left[\begin{array}{c} -1 \\ 2 \\ 1 \\ \hline q-4 \end{array}\right] \tag{1.26}$$

which is in the form of (1.24).

From (1.24) we immediately observe that if $\boldsymbol{q} \neq \boldsymbol{0}$, then the system (1.21), and therefore, the system (1.3) are inconsistent. This happens if and only if the row rank of the augmented matrix $[\,A \;\; \boldsymbol{b}\,]$ is larger than the row rank of the coefficient matrix A (see Exercise 1.35). Thus we obtain our first result: If $r[\,A \;\; \boldsymbol{b}\,] > r$, then the system (1.3) is inconsistent.

On the other hand, if $\boldsymbol{q} = \boldsymbol{0}$, that is, if $r[\,A \;\; \boldsymbol{b}\,] = r$, then discarding the last $m - r$ trivial equations, (1.24) is reduced to

$$[\,I \;\; H\,] \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{u} + H\boldsymbol{v} = \boldsymbol{p} \tag{1.27}$$

which can be rewritten as

$$\boldsymbol{u} = \boldsymbol{p} - H\boldsymbol{v} = \boldsymbol{p} - v_1\boldsymbol{h}_1 - \cdots - v_\nu \boldsymbol{h}_\nu \tag{1.28}$$

where $\nu = n - r$ and $\boldsymbol{h}_i$ are the columns of H. Equation (1.28) specifies the basic variables in terms of the non-basic variables, which can be chosen arbitrarily. Letting $v_1 = c_1, \ldots, v_\nu = c_\nu$, where $c_1, \ldots, c_\nu$ are arbitrary constants, we express $\boldsymbol{v}$ as

$$\boldsymbol{v} = \begin{bmatrix} c_1 \\ \vdots \\ c_\nu \end{bmatrix} = c_1\boldsymbol{e}_1 + \cdots + c_\nu \boldsymbol{e}_\nu$$

where $\boldsymbol{e}_i$ is the ith column of I_ν. Then $\boldsymbol{u}$ is obtained from (1.28) as

$$\boldsymbol{u} = \boldsymbol{p} + c_1(-\boldsymbol{h}_1) + \cdots + c_\nu(-\boldsymbol{h}_\nu)$$

Hence we obtain the following general expression for the solution in terms of the renamed variables.

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{0} \end{bmatrix} + c_1 \begin{bmatrix} -\boldsymbol{h}_1 \\ \boldsymbol{e}_1 \end{bmatrix} + \cdots + c_\nu \begin{bmatrix} -\boldsymbol{h}_\nu \\ \boldsymbol{e}_\nu \end{bmatrix} \tag{1.29}$$

or in more compact form as

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{\phi}}_p + c_1\tilde{\boldsymbol{\phi}}_1 + \cdots + c_\nu\tilde{\boldsymbol{\phi}}_\nu = \tilde{\boldsymbol{\phi}}_p + \tilde{\boldsymbol{\phi}}_c \tag{1.30}$$

where

$$\tilde{\boldsymbol{\phi}}_p = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{0} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\phi}}_i = \begin{bmatrix} -\boldsymbol{h}_i \\ \boldsymbol{e}_i \end{bmatrix}, \quad i = 1, \ldots, \nu$$

Reversing the reordering and using the original names for the variables, the solution above is expressed as

$$\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu = \boldsymbol{\phi}_p + \boldsymbol{\phi}_c \tag{1.31}$$

where $c_1, \ldots, c_\nu$ are arbitrary constants.

To illustrate expressions (1.30) and (1.31) we refer to the system in Example 1.9 again, which has been simplified to (1.26). We first observe that the system is consistent if and only if $q - 4 = 0$, in which case $r[\,A \;\; \boldsymbol{b}\,] = r(A) = 3$. Assuming so and discarding the last equation, we rewrite (1.26) in the form of (1.28) as

$$\begin{bmatrix} x_1 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} - x_2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - x_5 \begin{bmatrix} 3 \\ -1 \\ -2 \end{bmatrix} \tag{1.32}$$

Now letting $x_2 = c_1, x_5 = c_2$, (1.32) gives the solution in the form of (1.30) as

$$\left[\begin{array}{c} x_1 \\ x_3 \\ \hline x_4 \\ \hline x_2 \\ x_5 \end{array}\right] = \left[\begin{array}{r} -1 \\ 2 \\ \hline 1 \\ \hline 0 \\ 0 \end{array}\right] + c_1 \left[\begin{array}{r} -1 \\ 0 \\ \hline 0 \\ \hline 1 \\ 0 \end{array}\right] + c_1 \left[\begin{array}{r} -3 \\ 1 \\ \hline 2 \\ \hline 0 \\ 1 \end{array}\right] \tag{1.33}$$

where

$$\tilde{\boldsymbol{\phi}}_p = \left[\begin{array}{r} -1 \\ 2 \\ \hline 1 \\ \hline 0 \\ 0 \end{array}\right], \quad \tilde{\boldsymbol{\phi}}_1 = \left[\begin{array}{r} -1 \\ 0 \\ \hline 0 \\ \hline 1 \\ 0 \end{array}\right], \quad \tilde{\boldsymbol{\phi}}_2 = \left[\begin{array}{r} -3 \\ 1 \\ \hline 2 \\ \hline 0 \\ 1 \end{array}\right] \tag{1.34}$$

Finally restoring the original order of the variables we get the solution in (1.19), which has the form of (1.31).

Having studied the general case, we now consider other possible cases.

If $r = m < n$, then (1.24) is already in the form of (1.27) with $r[\,A \;\; \boldsymbol{b}\,] = r(A) = m$. Thus the system is consistent, and the solution is given by (1.31).

If $r = n < m$, then (1.24) is reduced to

$$\begin{bmatrix} I \\ O \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix} \tag{1.35}$$

Again, the system is consistent if and only if $\boldsymbol{q} = \boldsymbol{0}$, or equivalently, if and only if $r[\,A \;\; \boldsymbol{b}\,] = r(A) = n$. In this case, (1.35) is further reduced to

$$\boldsymbol{x} = \boldsymbol{p} \tag{1.36}$$

which displays a unique solution.

Finally, if $r = m = n$, then the system is consistent, and has a unique solution given in (1.36).

The solution expression in (1.31) consists of two parts: The first part, $\boldsymbol{\phi}_p$, is fixed and is due to the right-hand side of the system. If $\boldsymbol{b} = \boldsymbol{0}$, that is, if the system is homogeneous, then

$\boldsymbol{p} = \mathbf{0}$, and therefore, $\boldsymbol{\phi}_p = \tilde{\boldsymbol{\phi}}_p = \mathbf{0}$. The second part, $\boldsymbol{\phi}_c$, exists if and only if $r < n$. If $r = n$, then $\boldsymbol{\phi}_c = \mathbf{0}$, and $\boldsymbol{x} = \boldsymbol{\phi}_p$ would be the unique solution. (As a consequence, if the system is homogeneous and if $r = n$, then the only solution is the null solution $\boldsymbol{x} = \mathbf{0}$.)

Since $\boldsymbol{\phi}_p = \mathbf{0}$ when $\boldsymbol{b} = \mathbf{0}$, we conclude that when $r < n$

$$\boldsymbol{\phi}_c = c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu$$

is a nontrivial solution of the associated homogeneous system

$$A\boldsymbol{x} = \mathbf{0} \tag{1.37}$$

Since it contains ν arbitrary constants, $\boldsymbol{\phi}_c$ defines a family of solutions of (1.37). For every fixed choice of the arbitrary constants we get a member of this family. In particular, each $\boldsymbol{\phi}_i$ is a member of this family, $\boldsymbol{\phi}_1$ corresponding to the choice $c_1 = 1, c_2 = \cdots = c_\nu = 0$, $\boldsymbol{\phi}_2$ corresponding to the choice $c_1 = 0, c_2 = 1, c_3 = \cdots = c_\nu = 0$, etc. These solutions have the property that none of them can be expressed in terms of the others, and are said to be ***linearly independent***. This follows from the fact that each $\boldsymbol{e}_i$ contains a single 1 at a different position, so that no $\boldsymbol{e}_i$ can be expressed in terms of the others. Then the same must be true for $\tilde{\boldsymbol{\phi}}_i$, and therefore, for $\boldsymbol{\phi}_i$. The reader should examine the expressions for $\tilde{\boldsymbol{\phi}}_1$ and $\tilde{\boldsymbol{\phi}}_2$ in (1.34) to verify this fact. The significance of the solutions $\boldsymbol{\phi}_i$ being linearly independent is that they can not be combined to simplify $\boldsymbol{\phi}_c$, which implies that for every choice of the arbitrary constants $c_1, \ldots, c_\nu$ we get a different solution of the homogeneous equation (see Exercise 1.40).[11]

Like $\boldsymbol{\phi}_c$, the expression in (1.31) also defines a family of solutions of the non-homogeneous system in (1.3). Any member of this family, obtained by assigning fixed values to the arbitrary constants $c_1, \ldots, c_\nu$ is called a ***particular solution***. A simple particular solution is obtained by choosing $c_1 = \cdots = c_\nu = 0$ to be

$$\boldsymbol{x} = \boldsymbol{\phi}_p$$

The second part of the expression in (1.31), $\boldsymbol{\phi}_c$, is called a ***complementary solution*** of (1.3), because by adding to $\boldsymbol{\phi}_p$ any member of the family defined by $\boldsymbol{\phi}_c$, we obtain another particular solution of (1.3).

Finally, we note that since any solution of (1.3) must satisfy (1.28), it must be of the form in (1.31). In other words, the family characterized by the expression in (1.31) contains all solutions of (1.3). For this reason, we call this expression a ***general solution*** of (1.3).

Thus we not only showed that the system is consistent if $r[\,A\;\;\boldsymbol{b}\,] = r(A)$, but also gave a systematic procedure to find a general form of the solution. We summarize these results as a theorem.

---

[11] The concept of linear independence will be discussed in Chapter 3. For the time being it suffices to know that if $\boldsymbol{\phi}_i$ were not linearly independent, then one of them, say the kth one, would be expressed in terms of the others as

$$\boldsymbol{\phi}_k = \sum_{i \neq k} \alpha_i \boldsymbol{\phi}_i$$

Then $\boldsymbol{\phi}_c$ would reduce to

$$\boldsymbol{\phi} = \sum_{i \neq k} (c_i + c_k\alpha_i)\boldsymbol{\phi}_i = \sum_{i \neq k} c_i'\boldsymbol{\phi}_i$$

containing only $\nu - 1$ arbitrary constants, and one degree of freedom would be lost.

Theorem 1.1 *Let A be an $m \times n$ matrix with $r(A) = r$.*

a) The homogeneous linear system $A\boldsymbol{x} = \mathbf{0}$ is consistent, and

i. if $r = n$, then the only solution is the trivial solution $\boldsymbol{x} = \mathbf{0}$,

ii. if $r < n$, then there exist $\nu = n - r$ linearly independent solutions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu$, and $\boldsymbol{x} = c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu$ is a solution for every choice of the arbitrary constants $c_1, \ldots, c_\nu \in \mathbb{R}$.

b) The non-homogeneous system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent if and only if $r[\,A \quad \boldsymbol{b}\,] = r$, in which case

i. if $r = n$, then there exists a unique solution $\boldsymbol{x} = \boldsymbol{\phi}_p$,

ii. if $r < n$, then $\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \ldots + c_\nu\boldsymbol{\phi}_\nu$ is a solution for every choice of the arbitrary constants $c_1, \ldots, c_\nu \in \mathbb{R}$, where $\nu = n - r$, $\boldsymbol{\phi}_p$ is any particular solution, and $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu$ are the linearly independent solutions of the associated homogeneous system.

The analysis above shows that existence and uniqueness of solution of a given system of linear equations cannot be deduced from the number of equations (m) and the number of unknowns (n) alone. Indeed, Example 1.9 illustrates that if a system contains more unknowns than equations that does not necessarily imply that the system has a solution. The converse is not true either: A system that contains more equations than unknowns may still have a solution as we illustrate by the following example.

Example 1.12

Check if the system

$$\begin{array}{rcrcrcr} x_1 & - & 2x_2 & + & 3x_3 & = & 11 \\ -x_1 & + & 3x_2 & - & 2x_3 & = & -11 \\ 2x_1 & - & 3x_2 & + & 5x_3 & = & 18 \\ -x_1 & + & x_2 & - & 2x_3 & = & -7 \end{array}$$

is consistent, and if so, find the solution.

We form the augmented matrix and simplify it as described below.

$$\left[\begin{array}{rrr|r} 1 & -2 & 3 & 11 \\ -1 & 3 & -2 & -11 \\ 2 & -3 & 5 & 18 \\ -1 & 1 & -2 & -7 \end{array}\right] \begin{array}{r} \\ \mathrm{R}_1 + \mathrm{R}_2 \to \mathrm{R}_2 \\ -2\mathrm{R}_1 + \mathrm{R}_3 \to \mathrm{R}_3 \\ \mathrm{R}_1 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \left[\begin{array}{rrr|r} 1 & -2 & 3 & 11 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & -4 \\ 0 & -1 & 1 & 4 \end{array}\right]$$

$$\begin{array}{r} 2\mathrm{R}_2 + \mathrm{R}_1 \to \mathrm{R}_1 \\ -\mathrm{R}_2 + \mathrm{R}_3 \to \mathrm{R}_3 \\ \mathrm{R}_2 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \left[\begin{array}{rrr|r} 1 & 0 & 5 & 11 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & 2 & 4 \end{array}\right]$$

$$\begin{array}{r} -(1/2)\mathrm{R}_3 \to \mathrm{R}_3 \\ -5\mathrm{R}_3 + \mathrm{R}_1 \to \mathrm{R}_1 \\ -\mathrm{R}_3 + \mathrm{R}_2 \to \mathrm{R}_2 \\ -\mathrm{R}_3 + \mathrm{R}_4 \to \mathrm{R}_4 \\ \longrightarrow \end{array} \left[\begin{array}{rrr|r} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

From the last augmented matrix we find that the given system is consistent, and has a unique solution

$$x_1 = 1\,, \quad x_2 = -2\,, \quad x_3 = 2$$

Although a system that contains more equations than unknowns may be consistent, the reader should not expect to come across such systems often. Indeed, unless the elements of A and $\boldsymbol{b}$ are somehow related, such a system will almost always be inconsistent (see Exercise 1.39). However, certain problems lead to systems with more equations than unknowns, which, by the nature of the problem, are consistent (see Exercise 1.47).

1.7 Numerical Considerations

Computers use limited space to represent numbers, that is, they have finite precision. This may lead to numerical errors in evaluating expressions that involve operations with several numbers. Anyone who tries to evaluate the expression

$$(1/3) \cdot 3 - 1$$

with a hand calculator would probably get an answer like 1×10^{-6} or 1×10^{-12} depending on the precision of the calculator, instead of the exact answer 0. The reason is that the number $1/3$ cannot be represented exactly by the calculator (no matter how many digits are used), so that when it is multiplied with 3 the result will be slightly different from 1, hence the difference from 1 slightly different from zero. If the result of these operations is later used in other expressions, the errors accumulate, possibly reaching unacceptable levels.

Since operations with matrices involve many successive operations with scalars (as in the case of Gaussian elimination), one should be alert about numerical errors that might result from a sequence of such operations, and try to avoid them as much as possible.

Example 1.13

Consider the linear system

$$\begin{array}{rcrcl} -0.002\,x_1 & + & 2.000\,x_2 & = & 2.000 \\ 2.000\,x_1 & + & 1.000\,x_2 & = & 3.000 \end{array}$$

Eliminating x_1 from the second equation we obtain

$$\begin{array}{rcrcl} -0.002\,x_1 & + & 2.000\,x_2 & = & 2.000 \\ & & 2001.\,x_2 & = & 2003. \end{array}$$

from which we obtain the exact solution

$$x_2 = \frac{2003}{2001} = 1 + \frac{2}{2001}\,, \quad x_1 = \frac{2000}{2001} = 1 - \frac{1}{2001}$$

Now suppose we try to solve the same system using a calculator that can represent numbers with three significant digits only. If we proceed with eliminating x_1 as above, we obtain the system

$$\begin{array}{rcrcl} -0.002\,x_1 & + & 2.000\,x_2 & = & 2.000 \\ & & 2000.\,x_2 & = & 2000. \end{array}$$

because both 2001 and 2003 are represented as $2000.$ in our calculator. We then obtain

$$x_2 = 1.000\,, \quad x_1 = 0.000$$

which is far from being a solution.

What happened is that the information in the second equation was lost when the first equation is multiplied by 1000 and added to the second. That is why the "computed" solution above satisfies the first equation but not the second.

Fortunately, however, the problem can be overcome simply by interchanging the equations. Gaussian elimination applied to the reordered system

$$\begin{array}{rcrcl} 2.000\,x_1 & + & 1.000\,x_2 & = & 3.000 \\ -0.002\,x_1 & + & 2.000\,x_2 & = & 2.000 \end{array}$$

produces

$$\begin{array}{rcrcl} 2.000\,x_1 & + & 1.000\,x_2 & = & 3.000 \\ & & 2.000\,x_2 & = & 2.000 \end{array}$$

in the calculator, from which the solution is computed as

$$x_2 = 1.000\,, \quad x_1 = 1.000$$

The "computed" solution is acceptable now.

The situation in the above example is similar in nature to computation of

$$(1/3) \cdot 3 - 1$$

Just like rephrasing this expression as

$$(1 \cdot 3)/3 - 1$$

eliminates the error, reordering the equations before applying Gaussian elimination reduces error to an acceptable level. The error in the first attempt originates from choosing a very small pivot at the first step, which results in a very large multiplier that erases the information in the second equation. Reordering the equations leads to a choice of a larger pivot that does not cause a significant loss of information. This strategy (of picking as large a pivot as possible among the candidates) is called ***partial pivoting***.

Although pivoting can handle many difficult situations, there are problems that are inherently "bad", and error is unavoidable.

Example 1.14

The system

$$\begin{bmatrix} 0.9900 & 0.9800 \\ 0.9800 & 0.9700 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.970 \\ 1.950 \end{bmatrix} \tag{1.38}$$

has the exact solution

$$\boldsymbol{x}_e = \begin{bmatrix} 1.000 \\ 1.000 \end{bmatrix}$$

Gaussian elimination with four-digit floating-point arithmetic reduces the augmented matrix to

$$\left[\begin{array}{cc|c} 0.9900 & 0.9800 & 1.970 \\ 0.0000 & -0.0001 & .0000 \end{array}\right]$$

The reduced system has the solution

$$\boldsymbol{x}_c = \begin{bmatrix} 1.990 \\ .0000 \end{bmatrix}$$

which is nowhere near the exact solution. Interchanging the equations as in the previous example is of no use; we end up with a similar erroneous result.

Suppose we did not know the exact solution, and wanted to check the "computed" solution $\boldsymbol{x}_c$ by substituting it into the original system. We would get

$$\begin{bmatrix} 0.9900 & 0.9800 \\ 0.9800 & 0.9700 \end{bmatrix} \begin{bmatrix} 1.990 \\ .0000 \end{bmatrix} = \begin{bmatrix} 1.970 \\ 1.950 \end{bmatrix}$$

which is the same as the right-hand side of (1.38) up to the fourth significant digits. What is more interesting is that when we evaluate $A\boldsymbol{x}$ for

$$\boldsymbol{x}_1 = \begin{bmatrix} 6.910 \\ -4.970 \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}_2 = \begin{bmatrix} -4.870 \\ 6.930 \end{bmatrix}$$

which are totally unrelated to each other and to the exact solution, we get

$$A\boldsymbol{x}_1 = \begin{bmatrix} 1.970 \\ 1.949 \end{bmatrix} \quad \text{and} \quad A\boldsymbol{x}_2 = \begin{bmatrix} 1.970 \\ 1.951 \end{bmatrix}$$

which differ from the right-hand side of (1.38) only in the fourth significant digit. Apparently, our check is not reliable.

The numerical difficulties encountered in the solution of the system in (1.38) stem from the fact that the lines described by the equations of the system are almost parallel. Although they intersect at a point whose coordinates are specified by the exact solution $\boldsymbol{x}_e$, the points with coordinates defined by $\boldsymbol{x}_c$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not far from these lines either. Unfortunately, the problem is inherent in the system, and no cure (other than increasing the precision) is available. Such systems are said to be ***ill-conditioned***.

1.8 Exercises

1. Study the tutorial in Appendix D. Experiment with MATLAB, and learn
   (a) how to input a real and a complex matrix,
   (b) basic matrix operations (addition, multiplication, transposition),
   (c) how to create special matrices (identity matrix, zero matrix, diagonal matrix, etc.),
   (d) how to construct a partitioned matrix from given blocks and how to extract a submatrix from a given matrix,
   (e) how to create and execute an M-file.
2. Let
$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$
   (a) Compute $\boldsymbol{x}^t\boldsymbol{x}$, $\boldsymbol{y}^t\boldsymbol{y}$, $\boldsymbol{x}\boldsymbol{x}^t$, $\boldsymbol{y}\boldsymbol{y}^t$, $\boldsymbol{x}^t\boldsymbol{y}$, $\boldsymbol{y}^t\boldsymbol{x}$, $\boldsymbol{x}\boldsymbol{y}^t$, $\boldsymbol{y}\boldsymbol{x}^t$, $A\boldsymbol{x}$, $\boldsymbol{x}^tA$, $A\boldsymbol{y}$, $\boldsymbol{y}^tA$, $\boldsymbol{x}^tA\boldsymbol{x}$, $\boldsymbol{x}^tA\boldsymbol{y}$, $\boldsymbol{y}^tA\boldsymbol{x}$ and $\boldsymbol{y}^tA\boldsymbol{y}$.
   (b) Use MATLAB to find the products in part (a).
   (c) Compute tr (A), tr $(\boldsymbol{x}\boldsymbol{x}^t)$, tr $(\boldsymbol{x}\boldsymbol{y}^t)$, tr $(\boldsymbol{y}\boldsymbol{x}^t)$, tr $(\boldsymbol{y}\boldsymbol{y}^t)$.
3. Show, by an example, that $AB = AC$ does not imply that $B = C$.

4. (a) Show that if $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times m}$ then $\text{tr}\,(AB) = \text{tr}\,(BA)$.

   (b) Show that if $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{F}^{n\times 1}$ then

   $$\text{tr}\,(\boldsymbol{x}\boldsymbol{y}^t) = \sum_{i=1}^{n} x_i y_i$$

5. (a) Show that $A^p A^q = A^{p+q} = A^q A^p$.

   (b) Use MATLAB to verify the result in part (a) for

   $$A = \begin{bmatrix} 1 & -2 \\ 2 & 3 \end{bmatrix}$$

   and several p and q.

6. Let $A, B \in \mathbb{R}^{n\times n}$. Determine under what condition

   $$(A+B)^2 = A^2 + 2AB + B^2$$

7. Show that if $A = \text{diag}\,[\,d_1, \ldots, d_n\,]$, then $A^k = \text{diag}\,[\,d_1^k, \ldots, d_n^k\,]$.

8. (a) Prove that the product of two lower (upper) triangular matrices is also lower (upper) triangular.

   (b) Verify the result in part (a) by computing the product of two arbitrarily chosen 3×3 upper triangular matrices using MATLAB .

9. If A, B and C are 100×2, 2×100 and 100×10 matrices, would you compute the product ABC as $(AB)C$ or as $A(BC)$? Why?

10. Show that the matrices

    $$A = \begin{bmatrix} I_n & O \\ C & I_n \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} I_n & O \\ D & I_n \end{bmatrix}$$

    commute.

11. Find a general expression for A^n for the A matrix in Exercise 1.10.

12. Let

    $$A = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$$

    (a) Obtain a general formula for A^n.

    (b) Verify your formula by calculating A^n for $n = 2, 3, 4, 5$ using MATLAB.

13. Let

    $$A = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$$

    (a) Use MATLAB to compute A^5 and A^{10}. Would you expect A^n to blow up or converge to a finite limit matrix as $n \to \infty$?

    (b) Use MATLAB to compute A^n for $n = 1, 2, \ldots$ until the maximum element in absolute value of $A^n - A^{n-1}$ is smaller than a sufficiently small number, say 10^{-6}.

14. (a) Show that $K \in \mathbb{C}^{n\times n}$ is skew-Hermitian if and only if $K = iH$ where H is Hermitian.

    (b) Let $A = B + iC$, where $B, C \in \mathbb{R}^{n\times n}$. Show that A is Hermitian if and only if B is symmetric and C is skew-symmetric. State and prove a corresponding result for A to be skew-Hermitian.

15. (a) Show that every $A \in \mathbb{R}^{n\times n}$ can be decomposed as $A = S + Q$ where S is symmetric and Q is skew-symmetric.

    (b) Show that every $A \in \mathbb{C}^{n\times n}$ can be decomposed as $A = H + K$ where H is Hermitian and K is skew-Hermitian. (Thus, combined with the result of Exercise 1.14(a), we have $A = H_1 + iH_2$, where H_1 and H_2 are both Hermitian, similar to the decomposition of a complex number into its real and imaginary components.)

16. What can you say about the diagonal elements of a Hermitian and a skew-Hermitian matrix?

17. Show that A^hA is a Hermitian matrix for any $A \in \mathbb{C}^{m\times n}$. State and prove a corresponding result for $A \in \mathbb{R}^{m\times n}$.

18. Let $\boldsymbol{e}_i$ denote the ith column of I_n.

    (a) Interpret the products

    $$\boldsymbol{e}_i^t\boldsymbol{e}_i\,, \quad \boldsymbol{e}_i^t\boldsymbol{e}_j\,, \quad \boldsymbol{e}_i\boldsymbol{e}_j^t\,, \quad \boldsymbol{e}_i^tA\,, \quad A\boldsymbol{e}_j\,, \quad \boldsymbol{e}_i^tA\boldsymbol{e}_j$$

    (b) Use MATLAB to verify your interpretation by calculating the above products for the A matrix in Exercise 1.2 and for $i, j = 1, 2, 3$.

19. Let

    $$Q_n = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n\times n}$$

    (a) Obtain a general expression for $Q_n^k,\ k = 1, \ldots, m \geq n$.

    (b) Use MATLAB to verify your result for $n = 4$ and $k = 1, \ldots, 5$.

20. Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$ and Q_n be as in Exercise 1.14.

    (a) Express AQ_n^k in terms of the columns of A.

    (b) Express Q_n^kB in terms of the rows of B.

    (c) Construct the matrices Q_5, $A = [\,10i + j\,]_{3\times 5}$ and $B = [\,10i + j\,]_{5\times 4}$ in MATLAB, and calculate AQ_5^k and Q_5^kB for $k = 1, \ldots, 5$.

21. Let A be a 10×4 matrix partitioned into its columns as

    $$A = [\,\boldsymbol{a}_1\ \ \boldsymbol{a}_2\ \ \boldsymbol{a}_3\ \ \boldsymbol{a}_4\,]$$

    and let Q be such that

    $$AQ = [\,\boldsymbol{a}_4\ \ \boldsymbol{a}_3\ \ \boldsymbol{a}_1\ \ \boldsymbol{a}_2\,]$$

    (a) Find Q

    (b) Find AQ^{25} in terms of the columns of A

22. For each of the following $(A, \boldsymbol{b})$ pairs find the reduced row echelon form of the augmented matrix by hand, and check your result using the MATLAB command `rref`. Using the reduced row echelon form of the augmented matrix determine if the system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent, and if so, find the general solution.

    (a) $$A = \begin{bmatrix} 2 & -1 & 1 \\ -1 & -1 & -6 \\ 5 & -3 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

(b) $A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & -2 & -4 \\ 2 & 4 & 7 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$

(c) $A = \begin{bmatrix} 5 & -6 & 1 \\ 2 & -3 & 1 \\ 4 & -3 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix}$

(d) $A = \begin{bmatrix} 1 & 0 & 2 & -1 \\ 1 & 4 & 2 & 7 \\ 2 & -2 & 4 & -6 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 3 \\ 7 \\ 4 \end{bmatrix}$

(e) $A = \begin{bmatrix} 2 & -4 & -3 & -4 \\ -1 & 2 & 2 & 3 \\ 1 & -2 & -1 & -1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$

(f) $A = \begin{bmatrix} 1 & -1 & 2 & 1 \\ 2 & 1 & -3 & -1 \\ 4 & -1 & 1 & 1 \\ 1 & 2 & -5 & -2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}$

(g) $A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \\ -1 \end{bmatrix}$

(h) $A = \begin{bmatrix} 1+i & 0 & -i & 1 \\ 0 & 1 & 0 & -1 \\ -i & -1+i & 1 & 1 \\ 1 & i & 1-i & 2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ i \end{bmatrix}$

23. Attempt to solve the linear systems in Exercise 1.22 by using the MATLAB command `x=A\b`, and interpret the results.

24. Find the value of the scalar p such that the system

$$\begin{bmatrix} 1 & -2 & 3 \\ 2 & -4 & 7 \\ 1 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ p \end{bmatrix}$$

is consistent, and then find the general solution.

25. Repeat Exercise 1.24 for the system

$$\begin{bmatrix} 2 & -1 & 1 \\ -1 & -1 & -6 \\ 5 & -3 & 1 \\ 1 & -2 & p \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

26. For the following pair, determine all values of the constants p and q such that the system $A\boldsymbol{x} = \boldsymbol{b}$ has (a) no solution, (b) infinitely many solutions, (c) a unique solution.

$$A = \begin{bmatrix} 1 & -1 & 2 \\ 4 & -3 & 2 \\ -2 & -1 & p \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ q \end{bmatrix}$$

27. Determine geometrically if the following systems are consistent, and if so, find their solutions.

(a)
$$\begin{array}{rcrcl} x_1 & - & x_2 & = & 1 \\ x_1 & + & x_2 & = & 5 \\ x_1 & - & 3x_2 & = & 3 \end{array}$$

(b)
$$\begin{array}{rcrcr} x_1 & - & x_2 & = & 1 \\ x_1 & + & x_2 & = & 5 \\ x_1 & - & 3x_2 & = & -3 \end{array}$$

28. Write the equation of a straight line passing through the points $(x_1, y_1) \neq (x_2, y_2)$ in the xy plane. Hint: A straight line in the xy plane is described by an equation of the form $ax+by+c = 0$ in the most general case. Translate the problem into solving a system of two linear equations in the unknowns a, b and c.

29. Consider a square pyramid with base vertices at $\boldsymbol{v}_1 = (0, 0, 0)$, $\boldsymbol{v}_2 = (2, 0, 0)$, $\boldsymbol{v}_3 = (2, 2, 0)$ and $\boldsymbol{v}_4 = (0, 2, 0)$, and the tip at $\boldsymbol{v}_0 = (1, 1, 1)$ in the $x_1x_2x_3$ space. Write equations of the four planar faces of the pyramid and obtain the solution of the 4×3 system consisting of these equations. Is the answer what you expect?

30. Write a MATLAB program to implement the Gaussian Elimination Algorithm in Table 1.1, and save it for your future use. Use your program to obtain the row echelon and reduced row echelon forms of the matrices in Exercise 1.22.

31. Find the reduced row echelon form of the augmented matrix in (1.16) if $q \neq 4$.

32. Let the $m \times n$ matrix A have rank $r < m$ and the reduced row echelon form

$$\begin{bmatrix} R \\ O \end{bmatrix}$$

where R is $r \times n$ and O is $(m - r) \times n$. Find the rank and the reduced row echelon form of

$$B = \begin{bmatrix} A \\ A \end{bmatrix}$$

33. (a) Define ***elementary column operations*** on a matrix by imitating the definition of elementary row operations.

    (b) Give a precise definition of ***column equivalence*** of matrices.

    (c) Define ***column echelon form*** and ***reduced column echelon form*** of a matrix.

    (d) Define column rank of a matrix.

34. (a) Explain how the Gaussian Elimination algorithm of Section 1.4 can be modified to obtain the reduced column echelon form of a matrix.

    (b) Show that the reduced column echelon form of a matrix A is the transpose of the reduced row echelon form of A^t.

35. Let $r(A) = r$ and

$$[\,A \;\; \boldsymbol{b}\,] \longrightarrow [\,R \;\; \boldsymbol{d}\,] = \begin{bmatrix} F & \boldsymbol{p} \\ O & \boldsymbol{q} \end{bmatrix}$$

where R is the reduced row echelon form of A. Clearly, $r[\,A \;\; \boldsymbol{b}\,] = r$ if and only if $\boldsymbol{q} = \boldsymbol{0}$. Find the reduced row echelon form of the augmented matrix and its rank when $r[\,A \;\; \boldsymbol{b}\,] \neq r$.

36. (a) Construct the matrices

```
A=eye(3)+ones(3,3);b=[2;0;2]
```

in MATLAB, and solve the equation $A\boldsymbol{x} = \boldsymbol{b}$ by using your Gaussian elimination algorithm and also by the MATLAB command `x=A\b`.

(b) Repeat (a) for

```
A=eye(3)+i*ones(3,3);b=[0;-2;-1+i]
```

37. Calculate the total number of multiplication/division operations required to solve an $n \times n$ system by Gaussian elimination, assuming that a pivot can be chosen at every step. Include in your calculation divisions by 1, but exclude multiplications with 0. Hint: Consider the loop consisting of steps 7-10 of the Gaussian elimination algorithm applied to the augmented matrix. Elimination of x_r from each of the remaining $n - r$ equations requires 1 division to find the multiplier μ_{ir} at Step 8, and $n - r + 1$ multiplications to modify the ith row at Step 9. Thus the loop requires $(n-r)(n-r+2)$ multiplications/divisions, and the whole forward elimination process requires

$$\sum_{r=1}^{n-1}(n-r)(n-r+2)$$

such operations. Calculate the operations required by backward elimination similarly, and then find closed form expressions for the sums.

38. Repeat Exercise 1.37 if forward and backward substitutions are performed simultaneously, and explain why forward elimination followed by backward elimination is more efficient (in terms of the number of multiplication/division operations) than simultaneous elimination.

39. Use MATLAB command `M=rand(5,4)` to generate a 5×4 augmented matrix $M = [\, A \;\; \boldsymbol{b}\,]$ with random elements. Use either the MATLAB code written in Exercise 1.30 or MATLAB's build-in function `rref` to compute the reduced row echelon form of M, and determine if the associated system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent. Repeat several times.

40. Let $\{\, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu \,\}$ be given set of column vectors. Show that if

$$a_1\boldsymbol{\phi}_1 + \cdots + a_\nu\boldsymbol{\phi}_\nu = b_1\boldsymbol{\phi}_1 + \cdots + b_\nu\boldsymbol{\phi}_\nu$$

for two different ordered sets of scalars $(a_1, \ldots, a_\nu)$ and $(b_1, \ldots, b_\nu)$, then at least one of $\boldsymbol{\phi}_i$ can be expressed in terms of the others. This shows that if $\{\, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu \,\}$ is linearly independent then different choices of the arbitrary constants $c_1, \ldots, c_\nu$ in the expression

$$\boldsymbol{\phi}_c = c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu$$

yield different vectors.

41. Show that if a linear system has two distinct solutions then it has infinitely many solutions. Hint: Let $\boldsymbol{x} = \boldsymbol{\phi}_1$ and $\boldsymbol{x} = \boldsymbol{\phi}_2$ be two distinct solutions of $A\boldsymbol{x} = \boldsymbol{b}$, and consider $A(\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2)$.

42. Suppose that

$$\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu$$

is the general solution of $A\boldsymbol{x} = \boldsymbol{b}$ and that $\boldsymbol{x} = \boldsymbol{\psi}$ is a particular solution of $A\boldsymbol{x} = \boldsymbol{c}$. Find the general solution of

$$A\boldsymbol{x} = 2\boldsymbol{b} - \boldsymbol{c}$$

43. Consider the system

$$\begin{bmatrix} 0.19 & 0.18 \\ 0.18 & 0.17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.70 \end{bmatrix}$$

(a) Find the exact solution.

(b) Show that Gaussian Elimination with 3-digit floating point arithmetic results in an inconsistent system.

(c) Solve the system by using 4-digit floating point arithmetic.

(d) Solve the system by using 5-digit floating point arithmetic.

44. Repeat Exercise 1.43 for the system

$$\begin{bmatrix} 0.820 & 0.528 \\ 0.730 & 0.470 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.340 \\ 0.300 \end{bmatrix}$$

45. Consider the system

$$\begin{array}{rcrcrcr} x_1 & + & 2\,x_2 & - & x_3 & = & 1 \\ x_1 & + & x_2 & + & \epsilon\, x_3 & = & -1 \\ x_1 & + & \epsilon\, x_2 & + & x_3 & = & -1 \end{array}$$

where ϵ is smaller than the precision of a calculator (that is, the calculator can represent ϵ alone, but rounds $1 + \epsilon$ to 1).

(a) Find the exact solution of the system.

(b) Show that, independent of the choice of the pivot elements, Gaussian elimination implemented on the calculator fails to produce a solution.

(c) Rewrite the equations in terms of new variables $z_1 = x_1$, $z_2 = x_1 + x_2$, $z_3 = x_1 + x_3$. Can you solve the resulting system with the same calculator?

46. Find all possible values of s such that the system

$$\begin{bmatrix} s & 1 & 0 \\ 0 & s & 1 \\ 0 & 1 & s \end{bmatrix} \boldsymbol{x} = \mathbf{0}$$

has a nontrival solution.

47. (Application) Consider the resistive electrical network shown in Figure 1.2(a), where v_1 and v_2 are the voltages supplied by external sources. The problem is to determine the voltages across and currents through all components of the network using Kirchoff's voltage and current laws and the voltage/current relations of the resistors. Kirchoff's voltage law states that the algebraic sum of the voltages across components that form a closed circuit is zero, and the current law states that the algebraic sum of currents through components that form a cut-set (a hypothetical line that separates the network into two disjoint parts) is zero. For convenience in identifying the circuits and cut-sets of the circuit, we associate with it a directed graph as shown in Figure 1.2(b), where each edge corresponds to a component. Direction of the edges are assigned arbitrarily with the convention that if a current flows in the assigned direction then it has a positive value, and that if the voltage across the component drops in the assigned direction then it has a positive value. The voltage/current relation of a resistor is $v = Ri$.

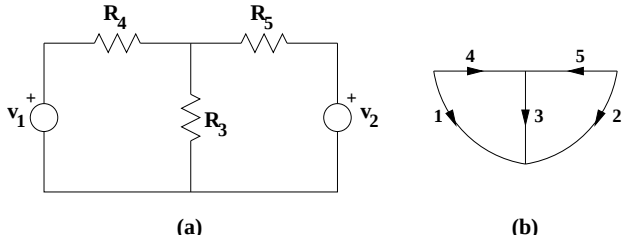

Figure 1.2: A resistive network

(a) Identify all circuits in the network, and write the circuit equations (relating the voltages of the components in the circuit) using Kirchoff's voltage law. Hint: There are three circuits.

(b) Identify all cut-sets in the network, and write the cut-set equations (relating the currents of the components in the circuit) using Kirchoff's current law. Hint: There are seven cut-sets.

(c) Use circuit and cut-set equations together with the voltage/current equations of the resistors to obtain a linear system in which all voltages and currents except v_1 and v_2 appear as unknowns to be solved in terms of v_1 and v_2.

(d) Solve the linear system you obtained above for the specific values $v_1 = 30V$, $v_2 = 60V$, $R_3 = 6K\Omega$, $R_4 = 2K\Omega$, $R_5 = 6K\Omega$. Show that although there are more equations than unknowns, the system is consistent and has a unique solution.

(e) Show that the solution above is independent of the value of R_4.

(f) Apparently not all the circuit and cut-set equations are independent (i.e., some of them can be obtained from the others, and are, therefore, redundant). Show that only two of the three circuit equations and only three of the seven cut-set equations are independent. (Thus, together with the three voltage/current equations of the resistors, there are eight equations in eight unknowns.)

48. (Application) The diagram in Figure 1.3 shows the major pipelines of the water distribution network of a town, where q_1 and q_2 denote the supply flow rates (in thousand m^3/sec) into the network from two reservoirs, and q_3 and q_4 the outflow rates from the network to town. It is assumed that no water is stored anywhere in the network, so that $q_1 + q_2 = q_3 + q_4$. The variables f_i associated with each pipe denote the flow rate in the direction arbitrarily assigned to the pipe. (Thus a negative value indicates that the flow is in the reverse direction.)

(a) Obtain a linear system in the variables f_i by equating the inflow rate at each node to the outflow rate.

(b) Obtain a general solution of the system obtained in part (a). How many variables can be chosen arbitrarily?

(c) Suppose that the pipes have a limited capacity so that $-F \leq f_i \leq F$, where $F = 18$. Obtain a region in the parameter space of arbitrary parameters that appear in the general solution in which every combination of the parameters gives a solution that satisfy the capacity constraints.

(d) Find F such that the system has a unique solution. Find also the corresponding solution.

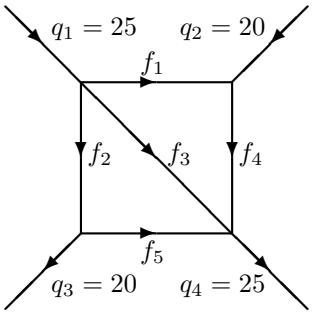

Figure 1.3: Water distribution network

49. (Application) Figure 1.4 shows a planar structure consisting of two rigid pieces pinned to each other and to two fixed supports on the ground. Let α_1 and α_2 denote the interior angles of the pieces with the horizontal, and let the weights W_1 and W_2 of the pieces be represented as downward forces acting at their midpoints. The problem is to find the forces on the supports. Let F_{ix} and F_{iy}, $i = 1, 2, 3$, denote the horizontal and vertical components of the reaction forces on the pin joints. Since each rigid piece is in equilibrium, the net horizontal and vertical force as well as the net torque acting on each piece is zero.

    (a) Write three equations for each piece to describe the equilibrium conditions to obtain a linear system of a total of six equations in the six unknowns F_{ix} and F_{iy}, $i = 1, 2, 3$. Hint: Use the geometry of the structure.

    (b) Solve the system formulated in part (a) to find the forces on the supports in terms of W_1, W_2 and α_1, α_2.

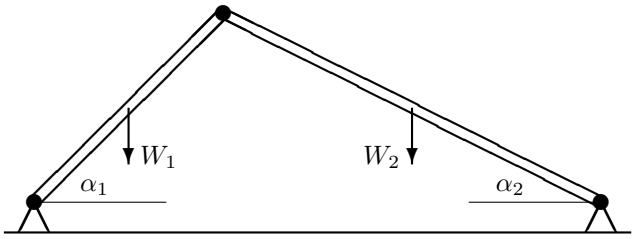

Figure 1.4: Rigid planar structure

Chapter 2
Introduction to Differential Equations

2.1 Basic Definitions

An equation involving a real-valued function of one or more real independent variables and its derivatives (with respect to these variables) is called a ***differential equation***. Some examples are

$$\begin{aligned} y' + (\ln x)y^2 &= 0\,, & y = y(x) \\ \frac{d^2y}{dt^2} + 3\frac{dy}{dt} + 2y &= \cos t\,, & y = y(t) \\ \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= 0\,, & u = u(x,y) \\ u_t - \alpha^2 u_{xx} &= 0\,, & u = u(t,x) \end{aligned}$$

A differential equation involving derivatives of a function of a single independent variable is called an ***ordinary*** differential equation, and one involving partial derivatives of a function of two or more independent variables is called a ***partial*** differential equation. First two equations above are ordinary differential equations, and the last two are partial differential equations. We will deal only with ordinary differential equations.

The ***order*** of a differential equation is the order of the highest derivative appearing in the equation. The first equation above is a first order differential equation, the others are second order.

An nth order ordinary differential equation in a function y of an independent variable t is of the form

$$F(t, y, y', \ldots, y^{(n)}) = 0 \tag{2.1}$$

where F is a given real-valued function, and $y', y'', \ldots, y^{(n)}$ denote the first, second, and the nth derivative of y.[1] If $y^{(n)}$ can be written explicitly in terms of the remaining variables, then (2.1) becomes

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}) \tag{2.2}$$

[1] If the independent variable t appears explicitly in a differential equation, then it is understood that the dependent variable y is a function of t, and y', y'', etc. refer to the derivatives of y with respect to t. However, if the independent variable does not appear explicitly in the differential equation, then it is better to denote the derivatives of the dependent variable by $\frac{dy}{dt}$, $\frac{d^2y}{dt^2}$, etc., to indicate that the independent variable is t and y is a function of t.

A real-valued function $\phi(t)$ defined on an open interval $\boldsymbol{I} = (t_i, t_f)$ is called a ***solution*** of the differential equation in (2.1) if

$$F(t, \phi(t), \phi'(t), \ldots, \phi^{(n)}(t)) = 0 \quad \text{for all} \;\; t \in \boldsymbol{I}$$

Obviously, this requires that $\phi'(t), \ldots, \phi^{(n)}(t)$ and $F(t, \phi(t), \phi'(t), \ldots, \phi^{(n)}(t))$ exist for all $t \in \boldsymbol{I}$. The graph of $y = \phi(t)$ is called a ***solution curve***.

Example 2.1

The function

$$\phi(t) = 1 + \sin t$$

is a solution of the differential equation

$$y'' + y = 1$$

on the interval $-\infty < t < \infty$, because $\phi'(t) = \cos t$ and $\phi''(t) = -\sin t$ exist and

$$\phi''(t) + \phi(t) = -\sin t + 1 + \sin t = 1$$

for all $-\infty < t < \infty$. The function

$$\psi(t) = 1 - 2\cos t$$

is also a solution of the same differential equation on the interval $-\infty < t < \infty$ as can be verified similarly.

Example 2.2

Any function of the form

$$\phi(t) = c/t$$

where c is a real number, is a solution of the differential equation

$$ty' + y = 0$$

on each of the intervals $-\infty < t < 0$ and $0 < t < \infty$. The solutions curves corresponding to different choices of c are shown in Figure 2.1.

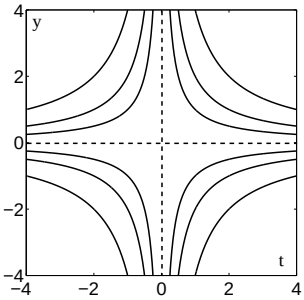

Figure 2.1: Solution curves of the differential equation in Example 2.2

If the differential equation in (2.1) can be written as

$$y^{(n)} + a_1(t)y^{(n-1)} + \cdots + a_{n-1}(t)y' + a_n(t)y = u(t) \tag{2.3}$$

where $a_1(t), \ldots, a_n(t)$ and $u(t)$ are given real-valued functions, then it is called a ***linear*** differential equation (LDE). If $u = 0$ in (2.3), then it is called ***homogeneous***. We will deal mostly with linear differential equations having constant coefficients: $a_i(t) = a_i \in \mathbb{R}, i = 1, 2, \ldots, n$.

2.2 First Order LDE with Constant Coefficients

A first order linear differential equation with a constant coefficient is of the form

$$y' + ay = u(t)$$

We deal with the problem of solving the above differential equation in two steps: We first find a solution of a homogeneous equation, and then generate from it a solution of the non-homogeneous equation.

2.2.1 Homogeneous Equations

Consider the homogeneous equation

$$y' + ay = 0 \tag{2.4}$$

Clearly, $y = 0$ is a trivial solution of (2.4) for all t. In search of a nontrivial solution we rewrite the equation as $y' = -ay$, from which observe that the derivative of the solution must be a multiple of the solution itself. One such function is the exponential function. Motivated with this observation, we seek a solution of the form $y = e^{st}$ where s is a real number. Substituting y and $y' = se^{st}$ into (2.4) we get

$$se^{st} + ae^{st} = (s + a)e^{st} = 0$$

Since $e^{st} \neq 0$ for all t, we must have

$$s + a = 0$$

which is called the ***characteristic equation*** of (2.4). The characteristic equation has the root $s = -a$, which implies that $y = e^{-at}$ is a solution. But then it is easy to see that any multiple of e^{-at}, that is, any function of the form

$$y = ce^{-at}, \quad c \in \mathbb{R} \tag{2.5}$$

is also a solution for all t. The constant c in expression (2.5) can be chosen arbitrarily, and for each choice of c we get a different solution as shown in Figure 2.2 for $a = 1$. Thus the expression in (2.5) defines a one-parameter family of solutions.

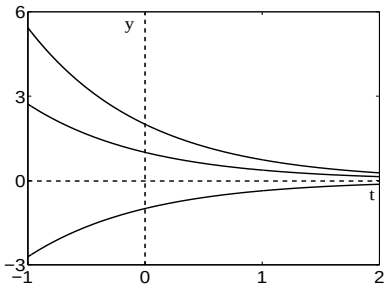

Figure 2.2: Solutions of (2.4) for $a = 1$.

2.2.2 Non-homogeneous Equations

Now consider the non-homogeneous equation

$$y' + ay = u(t) \tag{2.6}$$

We replace the constant c in (2.5) with a function $v(t)$, and look for a solution of the form $y = e^{-at}v(t)$.[2] Substituting y and $y' = e^{-at}v'(t) - ae^{-at}v(t)$ into (2.6), we get

$$e^{-at}v'(t) - ae^{-at}v(t) + ae^{-at}v(t) = e^{-at}v'(t) = u(t)$$

or equivalently,

$$v'(t) = e^{at}u(t)$$

Thus

$$v(t) = \int e^{at}u(t)dt = V(t) + c$$

where $V(t)$ is any antiderivative of $v'(t) = e^{at}u(t)$, and $c \in \mathbb{R}$ is an arbitrary constant. Hence any function of the form

$$y = e^{-at}[V(t) + c], \quad c \in \mathbb{R} \tag{2.7}$$

is a solution of the non-homogeneous equation in (2.6).[3]

As in the homogeneous case, the expression in (2.7) contains an arbitrary constant c, and thus defines a family of solutions. By analogy to the solution of linear systems considered in Chapter 1, any member of this family obtained by assigning a fixed value to the arbitrary constant c is called a ***particular solution***, denoted ϕ_p. A simple particular solution is obtained by choosing $c = 0$ as

$$\phi_p(t) = e^{-at}V(t)$$

[2]This method of finding a solution of a non-homogeneous linear differential equation from a solution of the associated homogeneous equation is known as the method of ***variation of parameters***, and is also applicable to higher order differential equations.

[3]From now on, we will omit the phrase "$c \in \mathbb{R}$" from a solution expression for simplicity in notation.

Then the family of solutions in (2.7) can be written as

$$y = e^{-at}V(t) + ce^{-at} = \phi_p(t) + \phi_c(t) \tag{2.8}$$

where

$$\phi_c(t) = ce^{-at}$$

characterizes a family of solutions of the homogeneous equation (2.4) associated with (2.6). ϕ_c is called a ***complementary solution*** of (2.6) because by adding to ϕ_p any member of the family defined by ϕ_c we obtain another particular solution.

Example 2.3

A complementary solution of the differential equation

$$y' + 2y = 5\cos t \tag{2.9}$$

is

$$\phi_c(t) = ce^{-2t}$$

To find a particular solution, we substitute $y = e^{-2t}v(t)$ and its derivative into the given equation, and obtain

$$e^{-2t}v'(t) - 2e^{-2t}v(t) + 2e^{-2t}v(t) = 5\cos t$$

or equivalently,

$$v'(t) = 5e^{2t}\cos t$$

Taking the antiderivative of both sides, we get

$$v(t) = \int 5e^{2t}\cos t\, dt = e^{2t}(2\cos t + \sin t) + c$$

Thus a family of solutions is obtained as

$$y = e^{-2t}v(t) = 2\cos t + \sin t + ce^{-2t}$$

where

$$\phi_p(t) = 2\cos t + \sin t \tag{2.10}$$

is a particular solution.

Now consider the same differential equation with a different right-hand side:

$$y' + 2y = 4t \tag{2.11}$$

Complementary solution is still $\phi_c(t) = ce^{-2t}$. Following the same steps as above, a particular solution can be found as

$$\phi_p(t) = 2t - 1 \tag{2.12}$$

What if the right-hand side of the differential equation is the sum of the right-hand sides of (2.9) and (2.11)? The reader might suspect that a particular solution would be the sum of the particular solutions in (2.10) and (2.12). Indeed, it is easy to verify that

$$y = 2\cos t + \sin t + 2t - 1$$

is a particular solution of

$$y' + 2y = 5\cos t + 4t$$

So far we have shown that any member of the family defined by (2.8) is a solution of the differential equation (2.6), but left the question whether there may be other solutions that do not belong to this family unanswered. We will consider this issue in the next section.

2.3 Initial Conditions

Suppose that we are interested in finding among the family of solutions given in (2.8) a particular one which has the value $y = y_0$ when $t = t_0$. In other words, we look for a particular solution whose graph passes through the point (t_0, y_0) in the ty plane. Such a condition is called an ***initial condition***, and a differential equation with an initial condition attached to it is called an ***initial-value problem***. We describe an initial-value problem involving a first order differential equation as

$$y' + ay = u(t)\,, \quad y(t_0) = y_0 \tag{2.13}$$

A function $\phi(t)$ is called a solution of the initial-value problem in (2.13) on an interval $\boldsymbol{I}$ that includes t_0 if it is a solution of the differential equation on $\boldsymbol{I}$ and $\phi(t_0) = y_0$.

If we assume that the solution of the initial-value problem in (2.13) is included in the family of solutions given by (2.8), then to find it all we have to do is to fix the arbitrary constant c in expression (2.8) to satisfy the initial condition as we illustrate by the following example.

Example 2.4

Let us solve the initial-value problem

$$y' + 2y = 5\cos t\,, \quad y(0) = 1$$

A family of solutions of the differential equation has already been obtained in Example 2.3 as

$$y = 2\cos t + \sin t + ce^{-2t}$$

To evaluate the arbitrary constant, we substitute $t_0 = 0$ for t and $y_0 = 1$ for y, and get

$$1 = 2\cos 0 + \sin 0 + ce^{-2.0} = 2 + c$$

which gives $c = -1$. Thus the solution of the initial-value problem is obtained as

$$y = 2\cos t + \sin t - e^{-2t}$$

There are two questions concerning the initial-value problem in (2.13).

a) Under what conditions does there exist a solution?

b) If a solution exists, is it included in the family of solutions given by (2.8)?

The first question is answered by the following theorem, whose proof is given in Appendix B for a more general case.

Theorem 2.1 *Suppose that the function $u(t)$ is piece-wise continuous on an interval $\boldsymbol{I}$ which includes t_0.*[4] *Then there exists a unique continuous function $\phi(t)$ defined on $\boldsymbol{I}$ such that $\phi(t_0) = y_0$ and $y = \phi(t)$ is a solution of the differential equation in (2.13) on every subinterval of $\boldsymbol{I}$ that does not contain a discontinuity point of $u(t)$.*

[4] A function $f(t)$ defined on a finite interval is said to be piece-wise continuous if it is continuous everywhere except for a *finite* number of discontinuity points, and left and right limits of f exist at the discontinuity points. A function defined on an infinite or semi-infinite interval is piece-wise continuous if it is piece-wise continuous on every finite subinterval.

The function $\phi(t)$ in the statement of Theorem 2.1 satisfies the differential equation (2.13) for all $t \in \boldsymbol{I}$ except the discontinuity points of $u(t)$, where $\phi(t)$ is well defined (as it is continuous) but $\phi'(t)$ fails to exist. However, since there are only a finite number of such points in every finite subinterval of $\boldsymbol{I}$, we can extend the definition of solution to include such piecewise differentiable functions. With this extended definition of solution, Theorem 2.1 states that the initial-value problem in (2.13) has a unique continuous, piece-wise differentiable solution $y = \phi(t)$ on $\boldsymbol{I}$.

Note that the theorem tells more than the existence of a solution. It also states that the solution is unique. It is the uniqueness of the solution that allows us to answer the second question.

Consider the function

$$\phi(t) = e^{-at}[V(t) - V(t_0) + e^{at_0}y_0] \tag{2.14}$$

which is a particular solution of the differential equation (2.13) obtained from (2.8) by choosing $c = e^{at_0}y_0 - V(t_0)$. Evaluating this function at $t = t_0$, we get

$$\phi(t_0) = e^{-at_0}[V(t_0) - V(t_0) + e^{at_0}y_0] = y_0$$

that is, ϕ also satisfies the initial condition. Then it must be the unique solution of (2.13). This shows that the solution of the initial-value problem (2.13) is indeed included in the family of solutions given by (2.8).

The reader may wonder how ϕ can be unique while V can be chosen to be any antiderivative of v'. The answer is that although $V(t)$ is not unique, $V(t) - V(t_0)$ is, because all antiderivatives differ only by a constant. If $\tilde{V}$ is any other antiderivative, then $\tilde{V}(t) = V(t) + C$ and $\tilde{V}(t_0) = V(t_0) + C$, so that $\tilde{V}(t) - \tilde{V}(t_0) = V(t) - V(t_0)$. A convenient choice for V is given by the definite integral

$$V(t) = \int_{t_0}^{t} e^{a\tau}u(\tau)d\tau$$

for which $V(t_0) = 0$. With this choice of $V(t)$, the unique solution of (2.13) is obtained from (2.14) as

$$\phi(t) = e^{-a(t-t_0)}y_0 + e^{-at}\int_{t_0}^{t} e^{a\tau}u(\tau)\,d\tau = \phi_o(t) + \phi_u(t) \tag{2.15}$$

This expression gives the solution as the sum of two parts; one part, $\phi_o(t)$, due to the initial condition y_0, and the other part, $\phi_u(t)$, due to the forcing function $u(t)$.[5]

Let us go back to the non-homogeneous differential equation (2.6). Let $y = \psi(t)$ be a solution of (2.6) on an interval $\boldsymbol{I}$, and let $\psi(t_0) = \psi_0$ at some arbitrary $t_0 \in \boldsymbol{I}$. Then obviously $y = \psi(t)$ is a solution of the initial-value problem

$$y' + ay = u(t)\,, \quad y(t_0) = \psi_0$$

and, by the above discussion, it must be included in the family of solutions given by (2.8). This shows that the expression (2.8) includes all possible solutions of (2.6). Because of this reason it is called a ***general solution*** of (2.6). Note that since $V(t)$ in (2.8) is not unique, a general solution may be expressed in many different ways.

[5] Unlike the decomposition of a solution into particular and complementary solutions as in (2.8), the parts $\phi_o(t)$ and $\phi_u(t)$ in (2.15) are not themselves solutions of (2.13).

Example 2.5

Let us find the solution of the initial-value problem

$$y' + ay = au(t)\,, \quad y(t_0) = y_0 \tag{2.16}$$

where $a \neq 0$, and

$$u(t) = \begin{cases} 1\,, & t > 0 \\ 0\,, & t < 0 \end{cases}$$

Such a function u is called a ***unit step*** function, and is common in many engineering applications. Note that unit step function is continuous everywhere except $t = 0$, where it has a jump as shown in Figure 2.3.

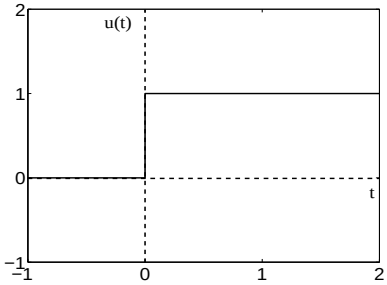

Figure 2.3: Unit step function

Since $u(t)$ has a discontinuity at $t = 0$, it is reasonable to look for separate solutions on the intervals $-\infty < t < 0$ and $0 < t < \infty$.

On the interval $-\infty < t < 0$, $u(t) = 0$, and a general solution of the resulting homogeneous differential equation

$$y' + ay = 0$$

is given as

$$y = c_1 e^{-at}\,, \quad t < 0 \tag{2.17}$$

On the interval $0 < t < \infty$, $u(t) = 1$, and the differential equation becomes

$$y' + ay = a$$

Now a general solution is

$$y = 1 + c_2 e^{-at}\,, \quad t > 0 \tag{2.18}$$

The solution curves defined by (2.17) and (2.18) are shown in Figure 2.4 for $a = 1$. Note that these solutions are not defined at $t = 0$. However, for any given y_0, there exist a particular solution ϕ_1 in the family defined by (2.17) and a particular solution ϕ_2 in the family defined by (2.18) such that

$$\lim_{t \to 0^-} \phi_1(t) = y_0 = \lim_{t \to 0^+} \phi_2(t) \tag{2.19}$$

The first condition in (2.19) requires that $c_1 = y_0$. Hence

$$\phi_1(t) = e^{-at} y_0\,, \quad t < 0$$

Similarly, using the second condition in (2.19) to evaluate c_2 in (2.18), we get

$$\phi_2(t) = e^{-at}y_0 + 1 - e^{-at}\,, \quad t > 0$$

Combining ϕ_1 and ϕ_2 and extending their domains of definition to include $t = 0$, we obtain a single solution for all $-\infty < t < \infty$ as

$$y \quad = \quad \begin{cases} e^{-at}y_0 \qquad\quad , & t \le 0 \\ e^{-at}y_0 + 1 - e^{-at}\,, & t \ge 0 \end{cases} \tag{2.20}$$

Note that the solution is continuous at all t, and it satisfies the given differential equation at all t except $t = 0$. This is what we mean by a solution in the extended sense of Theorem 2.1.

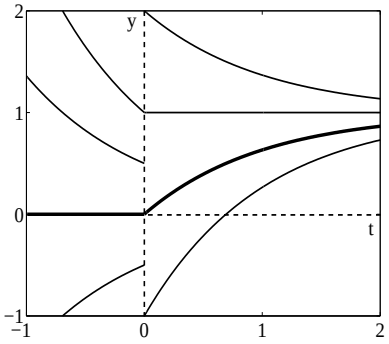

Figure 2.4: Solution curves of (2.16) for $a = 1$

The separate treatment of the cases $t < 0$ and $t > 0$ can be avoided by using the solution expression given in (2.15). If $t < 0$ then $u(\tau) = 0 \quad$ for all $\; t < \tau < 0$, and therefore, the integral term is zero. The solution is then given by

$$y = e^{-at}y_0\,, \quad t \le 0$$

If $t > 0$, $u(\tau) = 1 \quad$ for all $\; 0 < \tau < t$, and the solution is found as

$$y = e^{-at}y_0 + e^{-at}\int_0^t ae^{a\tau}d\tau = e^{-at}y_0 + 1 - e^{-at}\,, \quad t \ge 0$$

Of particular interest is the case when $y_0 = 0$. In this case, the solution becomes

$$y = \begin{cases} 0 \quad\; , & t \le 0 \\ 1 - e^{-at}\,, & t \ge 0 \end{cases}$$

which is called the ***step response*** of the differential equation, and is indicated by the thick curve in Figure 2.4.

Example 2.6

Let us consider the initial-value problem

$$y' + y = u_T(t)\,, \quad y(0) = 0 \tag{2.21}$$

where

$$u_T(t) = \begin{cases} 0 \;\; , & t < 0 \quad \text{or} \quad t > T \\ 1/T\,, & 0 < t < T \end{cases} \tag{2.22}$$

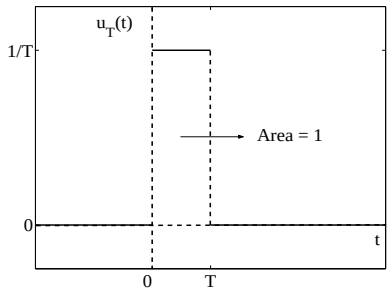

Figure 2.5: Unit pulse function.

Note that the area under the graph of $u_T(t)$ is 1 as shown in Figure 2.5. Such a function is called a ***unit pulse***.

As in the previous example, $y = 0$ for $t \leq 0$.

If $0 < t \leq T$, then $u_T(\tau) = 1/T \quad$ for all $\; 0 < \tau < t$, and

$$y = e^{-t} \int_0^t (1/T) e^{\tau} d\tau = \frac{1 - e^{-t}}{T}$$

If $t \geq T$, $u_T(\tau)$ contributes to the integral only for $0 < \tau < T$, so that

$$y = e^{-t} \int_0^T (1/T) e^{\tau} d\tau = \frac{e^T - 1}{T}\, e^{-t}$$

Combining these solutions, and extending their domains to include the discontinuity points $t = 0$ and $t = T$, we obtain

$$y = \begin{cases} 0 \quad , & t \leq 0 \\ \frac{1-e^{-t}}{T} \quad , & 0 \leq t \leq T \\ \frac{e^T - 1}{T}\, e^{-t} \, , & t \geq T \end{cases}$$

The solution is shown in Figure 2.6 for several values of T.

It is interesting to examine the behavior of $u_T(t)$ and the solution as $T \to 0$. As T gets smaller, the height of the pulse $u_T(t)$ tends to ∞ at $t = 0$ and to 0 everywhere else, while the area under the pulse remains unchanged. The limit of u_T is called a ***unit impulse***, denoted $\delta(t)$.[6] Now the corresponding solution tends to

$$y = \frac{e^T - 1}{T}\, e^{-t} \quad \to \quad e^{-t} \, , \quad t > 0$$

We formally say that the initial-value problem

$$y' + y = \delta(t) \, , \quad y(0^-) = 0$$

[6]Unit impulse is not a function in the ordinary sense, because it is not defined at $t = 0$, it is zero everywhere except $t = 0$, and yet

$$\int_{-\epsilon}^{\epsilon} \delta(t) dt = 1$$

for any $\epsilon > 0$.

has the solution

$$y = \begin{cases} 0\,, & t < 0 \\ e^{-t}\,, & t > 0 \end{cases}$$

which we call the ***impulse response*** of the differential equation. Note that the impulse response is not continuous, but has a jump at $t = 0$, as indicated by the thick solution curve in Figure 2.6. Since impulse is not a piece-wise continuous function, we should not expect to get a continuous solution. Theorem 2.1 is not applicable to this case. Note also that the initial condition is specified not exactly at $t = 0$, but at some $t = -\epsilon$ where $\epsilon > 0$ is arbitrarily small (which is denoted by 0^- for convenience). The reason is that we do not know what is going on at $t = 0$. Looking at the solution for $t > 0$, we observe that it is actually the same as the solution of the homogeneous initial-value problem

$$y' + y = 0\,, \quad y(0) = 1$$

It looks as if the impulse has changed the initial condition from 0 to 1 instantaneously at $t = 0$.

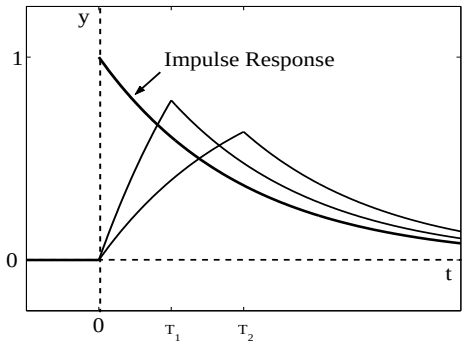

Figure 2.6: Solution of (2.21) for several T.

Example 2.7

In the previous two examples the point at which the initial condition was specified was also a discontinuity point of the forcing function $u(t)$, but these were just coincidence and irrelevant for the solution formula in (2.15). To illustrate this point consider the initial-value problem

$$y' + y = u(t)\,, \quad y(0) = y_0 \tag{2.23}$$

where

$$u(t) = \begin{cases} 1\,, & t > 1 \\ 0\,, & t < 1 \end{cases}$$

is a shifted unit step function.

For $t < 1$

$$y \;=\; \phi_1(t) \;=\; e^{-t}y_0 + e^{-t}\int_0^t e^{\tau}u(\tau)\,d\tau \;=\; e^{-t}y_0$$

where the second equality follows from the fact that $u(\tau) = 0$ on the interval of integration. For $t > 1$

$$\begin{aligned} y \;=\; \phi_2(t) \;&=\; e^{-t}y_0 + e^{-t}\int_0^t e^{\tau}u(\tau)\,d\tau \\ &=\; e^{-t}y_0 + e^{-t}\int_1^t e^{\tau}\,d\tau \;=\; e^{-t}y_0 + 1 - e^{-(t-1)} \end{aligned}$$

Note that

$$\lim_{t\,\to\,1^-}\phi_1(t) = y_0/e = \lim_{t\,\to\,1^+}\phi_2(t)$$

Combining ϕ_1 and ϕ_2 after extending their domains to include the discontinuity point $t = 1$ of u, we obtain a continuous solution whose graph is shown in Figure 2.7 for $y_0 = 0.4$.

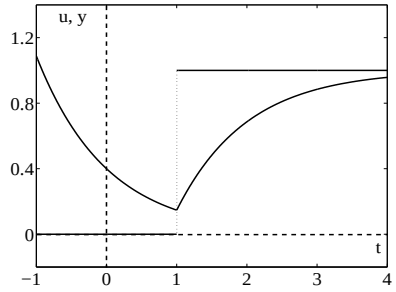

Figure 2.7: Solution of (2.23)

The formula in (2.15) provides us with a nice property of the solution of an initial-value problem involving a first order linear differential equation with a constant coefficient: If the initial-value problem

$$y' + ay = u(t)\,, \quad y(0) = y_0$$

has the solution $y = \phi(t)$, then the initial-value problem

$$y' + ay = u(t - t_0)\,, \quad y(t_0) = y_0$$

has the solution $y = \psi(t) = \phi(t-t_0)$ (see Exercise 2.6). This property allows us to derive the solution of a differential equation with an initial condition specified at some arbitrary t_0 from the solution of a modified differential equation with an initial condition specified at $t_0 = 0$. (Modification involves an appropriate shift of the forcing function $u(t)$.) It should however be emphasized that this property does not hold for a differential equation with a non-constant coefficient.

Example 2.8

The initial-value problem

$$y' + y = 2\,\cos t\,, \quad y(0) = 1$$

has the solution

$$y = \phi(t) = \sin t + \cos t$$

as can be verified by observing that

$$\phi'(t) + \phi(t) = (\cos t - \sin t) + (\sin t + \cos t) = 2\,\cos t$$

and

$$\phi(0) = \sin 0 + \cos 0 = 1$$

Then the initial-value problem

$$y' + y = 2\,\cos(t - \pi/2) = 2\,\sin t\,, \quad y(\pi/2) = 1$$

must have the solution

$$y = \psi(t) = \phi(t - \pi/2) = \sin(t - \pi/2) + \cos(t - \pi/2) = \sin t - \cos t$$

Indeed,

$$\psi'(t) + \psi(t) = (\cos t + \sin t) + (\sin t - \cos t) = 2\,\sin t$$

and

$$\psi(\pi/2) = \sin \pi/2 - \cos \pi/2 = 1$$

2.4 Second Order LDE with Constant Coefficients

2.4.1 Homogeneous Second Order Equations

Again we start with the homogeneous equation

$$y'' + a_1 y' + a_2 y = 0 \tag{2.24}$$

where $a_1, a_2 \in \mathbb{R}$ are given constants. As in the first order equations, $y = 0$ is a trivial solution, and we look for nontrivial solutions of the form $y = e^{st}$. Substituting y and its derivatives into the equation, we get the characteristic equation

$$s^2 + a_1 s + a_2 = 0$$

The characteristic equation is a second order equation with real coefficients. According to the fundamental theorem of algebra, it has two roots which may be real or complex. We investigate three possible cases separately.

Characteristic equation has distinct real roots

If $a_1^2 - 4a_2 > 0$, then the characteristic equation has two distinct real roots, $s = \sigma_1$ and $s = \sigma_2$. Then each of the functions

$$\phi_1(t) = e^{\sigma_1 t} \quad \text{and} \quad \phi_2(t) = e^{\sigma_2 t}$$

is a solution of the differential equation (2.24). Moreover, any function of the form

$$y = c_1\phi_1(t) + c_2\phi_2(t) = c_1 e^{\sigma_1 t} + c_2 e^{\sigma_2 t}$$

where $c_1, c_2 \in \mathbb{R}$ are arbitrary constants, is also a solution as can easily be verified by substitution.

The solutions ϕ_1 and ϕ_2 have the property that neither of them can be expressed as a multiple of the other, and are said to be ***linearly independent***. The importance of linear independence of ϕ_1 and ϕ_2 is that the solution expression above cannot be simplified by combining the two terms, which means that the two arbitrary constants c_1 and c_2 can be assigned arbitrary values independently, and we get a different solution for every different choice of the pair (c_1, c_2).[7]

Characteristic equation has complex roots

If $a_1^2 - 4a_2 < 0$, then the characteristic equation has a pair of complex conjugate roots $s = \lambda_1 = \sigma + i\omega$ and $s = \lambda_1^* = \sigma - i\omega$, where $\sigma, \omega \in \mathbb{R}$. Then each of the complex-valued functions

$$\psi_1(t) = e^{(\sigma+i\omega)t} = e^{\sigma t}(\cos\omega t + i\sin\omega t)$$

and

$$\psi_2(t) = \psi_1^*(t) = e^{(\sigma-i\omega)t} = e^{\sigma t}(\cos\omega t - i\sin\omega t)$$

satisfies the differential equation (2.24), and is called a ***complex solution***. To find real solutions we write $\psi_1(t) = \phi_1(t) + i\phi_2(t)$ where the real and imaginary parts

$$\phi_1(t) = e^{\sigma t}\cos\omega t \quad \text{and} \quad \phi_2(t) = e^{\sigma t}\sin\omega t$$

of ψ_1 are real-valued functions. Since ψ_1 satisfies the differential equation, we have

$$0 = \psi_1'' + a_1\psi_1' + a_2\psi_1 = (\phi_1'' + a_1\phi_1' + a_2\phi_1) + i(\phi_2'' + a_1\phi_2' + a_2\phi_2)$$

which implies

$$\phi_1'' + a_1\phi_1' + a_2\phi_1 = 0$$

and

$$\phi_2'' + a_1\phi_2' + a_2\phi_2 = 0$$

where the argument t of the functions are dropped for convenience. This shows that both the real part ϕ_1 and the imaginary part ϕ_2 of ψ_1 are solutions of (2.24). The reader can also verify this by substituting ϕ_1 and its derivatives and ϕ_2 and its derivatives into (2.24) (see Exercise 2.8). We would reach the same result if we considered ψ_2 instead of ψ_1, because their real parts are the same and imaginary parts differ only in sign.

The functions $\phi_1(t) = e^{\sigma t}\cos\omega t$ and $\phi_2(t) = e^{\sigma t}\sin\omega t$ are linearly independent, and any function of the form

$$y = c_1\phi_1(t) + c_2\phi_2(t) = c_1e^{\sigma t}\cos\omega t + c_2e^{\sigma t}\sin\omega t$$

is also a solution.

[7] Recall that we talked about linear independence of column vectors in Chapter 1 in connection with solution of linear systems. Here the same concept is used for functions. A precise definition of linear independence of functions will be given in Chapter 3.

Characteristic equation has a double real root

If $a_1^2 - 4a_2 = 0$, then the characteristic equation is of the form

$$s^2 - 2\sigma s + \sigma^2 = (s - \sigma)^2 = 0$$

and it has a double root at $s = \sigma$. In this case the function $\phi_1(t) = e^{\sigma t}$ is a solution of (2.24). Since there is no reason to think that this case is any different from the previous two cases in an essential way, we look for a second solution which is linearly independent of ϕ_1. To find a second solution, we define a new dependent variable as $x = y' - \sigma y$. Then the differential equation becomes

$$y'' - 2\sigma y' + \sigma^2 y = (y'' - \sigma y') - \sigma(y' - \sigma y) = x' - \sigma x = 0$$

Thus the original second order equation in y is reduced to a first order equation in the new variable x. The solution of this equation is

$$x = c_2 e^{\sigma t}$$

Substituting this expression in the equation defining x in terms of y, we obtain

$$y' - \sigma y = c_2 e^{\sigma t}$$

which is another first order differential equation in y, but now it is a non-homogeneous one. Solving this equation we obtain

$$y = c_1 e^{\sigma t} + c_2 t e^{\sigma t}$$

This expression is already in the form $y = c_1\phi_1(t) + c_2\phi_2(t)$, where $\phi_1(t) = e^{\sigma t}$ and $\phi_2(t) = te^{\sigma t}$. Thus we not only recover ϕ_1, but also obtain a second solution ϕ_2. As in the previous two cases, ϕ_1 and ϕ_2 are linearly independent.

In summary, the second order homogeneous linear differential equation (2.24) has a solution of the form

$$y = c_1\phi_1(t) + c_2\phi_2(t) \tag{2.25}$$

where ϕ_1 and ϕ_2 are linearly independent solutions that are defined by the roots of the characteristic equation.

2.4.2 Non-homogeneous Second Order Equations

We now consider the non-homogeneous equation

$$y'' + a_1 y' + a_2 y = u(t) \tag{2.26}$$

As in the first order equation, we use the method of variation of parameters, and assume a solution of the form

$$y = \phi_1(t)v_1(t) + \phi_2(t)v_2(t)$$

where ϕ_1 and ϕ_2 are linearly independent solutions of the associated homogeneous equation, and the functions v_1 and v_2 are to be determined. The derivative of y is obtained as

$$y' = \phi_1' v_1 + \phi_1 v_1' + \phi_2' v_2 + \phi_2 v_2'$$

where the argument t is dropped for simplicity. Let us impose the condition

$$\phi_1(t)v_1'(t) + \phi_2(t)v_2'(t) = 0 \tag{2.27}$$

on v_1 and v_2.[8] Then the expression for y' reduces to

$$y' = \phi_1' v_1 + \phi_2' v_2$$

and differentiating once more we get

$$y'' = \phi_1'' v_1 + \phi_1' v_1' + \phi_2'' v_2 + \phi_2' v_2'$$

Substituting y, y', and y'' into the differential equation in (2.26) and grouping the terms, we obtain

$$\begin{aligned} & y'' + a_1 y' + a_2 y \\ &= (\phi_1'' + a_1\phi_1' + a_2\phi_1)v_1 + (\phi_2'' + a_1\phi_2' + a_2\phi_2)v_2 + \phi_1' v_1' + \phi_2' v_2' \\ &= \phi_1' v_1' + \phi_2' v_2' \;=\; u \end{aligned}$$

where the second equation follows from the fact that ϕ_1 and ϕ_2 are solutions of the homogeneous part of (2.26) so that the expressions in the parentheses are zero. Thus we obtain a second equation in v_1' and v_2'

$$\phi_1'(t)v_1'(t) + \phi_2'(t)v_2'(t) = u(t) \tag{2.28}$$

Linear independence of ϕ_1 and ϕ_2 guarantee that equations (2.27) and (2.28) can be solved simultaneously for v_1' and v_2'.[9] The reader can verify that

$$v_1'(t) = \frac{\phi_2(t)u(t)}{\phi_1'(t)\phi_2(t) - \phi_1(t)\phi_2'(t)}$$

and

$$v_2'(t) = \frac{-\phi_1(t)u(t)}{\phi_1'(t)\phi_2(t) - \phi_1(t)\phi_2'(t)}$$

satisfy (2.27) and (2.28) simultaneously. Integrating these expressions we obtain

$$v_1(t) = V_1(t) + c_1 \quad \text{and} \quad v_2(t) = V_2(t) + c_2$$

where V_1 and V_2 are fixed antiderivatives of v_1' and v_2', and $c_1, c_2 \in \mathbb{R}$ are arbitrary constants. A solution of the non-homogeneous differential equation (2.26) is thus obtained as

$$y = \phi_1(t)V_1(t) + \phi_2(t)V_2(t) + c_1\phi_1(t) + c_2\phi_2(t) \tag{2.29}$$

We note that, as in the case of first order differential equations, the solution in (2.29) is also of the form

$$y = \phi_p(t) + \phi_c(t)$$

[8] The significance of this condition is explained in Chapter 6 for the general case of nth order linear differential equations.

[9] This will be proved in Chapter 6 for a general nth order linear differential equation.

where

$$\phi_p(t) = \phi_1(t)V_1(t) + \phi_2(t)V_2(t)$$

is a particular solution, and

$$\phi_c(t) = c_1\phi_1(t) + c_2\phi_2(t)$$

is a complementary solution (solution of the homogeneous equation (2.24) associated with (2.26)). We will show in Chapter 6 that the expression in (2.29) includes all solutions of (2.26), and therefore, it is a general solution of (2.26).

As in the case of first order equations, we might be interested in finding among the family of solutions given by (2.29) a particular one that satisfies additional conditions. Since a general solution contains two arbitrary constants, we need two conditions to determine the values of the arbitrary constants . If these conditions involve the values of the solution and its derivative at some t_0, then the problem becomes an initial-value problem specified as

$$y'' + a_1y' + a_2y = u(t)\,, \quad y(t_0) = y_0\,, \quad y'(t_0) = y_1$$

Equating the value of the general solution in (2.26) at t_0 to y_0 and the value of its derivative at t_0 to y_1, we get

$$\begin{aligned} \phi_p(t_0) + c_1\phi_1(t_0) + c_2\phi_2(t_0) &= y_0 \\ \phi_p'(t_0) + c_1\phi_1'(t_0) + c_2\phi_2'(t_0) &= y_1 \end{aligned}$$

Again linear independence of ϕ_1 and ϕ_2 guarantees that these equations can be solved for c_1 and c_2 to obtain the required solution of the initial-value problem.

Example 2.9

Solve the initial-value problem

$$y'' = 0\,, \quad y(0) = 1\,, \quad y'(0) = -1$$

The characteristic equation $s^2 = 0$ has a double root $s = 0$. Consequently,

$$\phi_1(t) = e^{0.t} = 1 \quad \text{and} \quad \phi_2(t) = te^{0.t} = t$$

are two linearly independent solutions, and a general solution is

$$y = c_1 + c_2t$$

The initial conditions require

$$y(0) = c_1 = 1\,, \quad y'(0) = c_2 = -1$$

Thus the solution of the initial-value problem is obtained as

$$y = 1 - t$$

This problem is so simple that the solution can be found without going into the systematic procedure of finding the roots of the characteristic equation. All we have to do is to integrate y twice:

$$y'(t) = y'(0) + \int_0^t y''(\tau)\,d\tau = -1 + \int_0^t 0\,dt = -1$$

and

$$y(t) = y(0) + \int_0^t y'(\tau)\,d\tau = 1 + \int_0^t (-1)\,d\tau = 1 - t$$

Example 2.10

Let us find the unit step response of

$$y'' + py' + y = u(t)\,, \quad y(0) = y'(0) = 0 \tag{2.30}$$

for several different values of the parameter p.

We note that whatever p is, the solution for $t \leq 0$ is given by $y = 0$. We are more interested in the solution for $t > 0$, for which $u(t) = 1$.

For $p = 5.2$, the characteristic equation

$$s^2 + 5.2s + 1 = 0$$

has two real roots $s = -0.2$ and $s = -5$, and a complementary solution is

$$y = c_1 e^{-0.2t} + c_2 e^{-5t}$$

To find a general solution we let

$$y = e^{-0.2t} v_1(t) + e^{-5t} v_2(t)$$

With the restriction

$$e^{-0.2t} v_1'(t) + e^{-5t} v_2'(t) = 0 \tag{2.31}$$

the derivatives of y are calculated as

$$y' = -0.2e^{-0.2t} v_1(t) - 5e^{-5t} v_2(t)$$

and

$$y'' = e^{-0.2t}(0.04v_1(t) - 0.2v_1'(t)) + e^{-5t}(25v_2(t) - 5v_2'(t))$$

Substituting y and its derivatives into the equation, we get after simplification

$$-0.2e^{-0.2t} v_1'(t) - 5e^{-5t} v_2'(t) = u(t) = 1 \tag{2.32}$$

Solving v_1' and v_2' from (2.31) and (2.32) we obtain

$$v_1'(t) = (1/4.8)e^{0.2t}\,, \quad v_2'(t) = (-1/4.8)e^{5t}$$

Hence

$$v_1(t) = (5/4.8)e^{0.2t} + c_1\,, \quad v_2(t) = -(0.2/4.8)e^{5t} + c_2$$

and a general solution for $t > 0$ is

$$\begin{aligned} y &= e^{-0.2t}((5/4.8)e^{0.2t} + c_1) + e^{-5t}(-(0.2/4.8)e^{5t} + c_2) \\ &= 1 + c_1 e^{-0.2t} + c_2 e^{-5t} \end{aligned}$$

The initial conditions

$$\begin{aligned} y(0) &= 1 + c_1 + c_2 \quad = 0 \\ y'(0) &= -0.2c_1 - 5c_2 = 0 \end{aligned}$$

give $c_1 = -25/24$ and $c_2 = 1/24$. Thus the solution of the initial-value problem is obtained as

$$y = 1 - (25/24)e^{-0.2t} + (1/24)e^{-5t}\,, \quad t \geq 0$$

For $p = 2$, the characteristic equation

$$s^2 + 2s + 1 = (s + 1)^2 = 0$$

has a double root $s = -1$, and a complementary solution is

$$y = c_1 e^{-t} + c_2 t e^{-t}$$

A general solution can be found by following the same steps as above. However, a simple observation allows us to avoid the burden of lengthy manipulations. We note that, for this particular problem, whatever p is, the function $\phi_p(t) = 1$ is a particular solution because

$$\phi_p''(t) + p\phi_p'(t) + \phi_p(t) = 0 + p \cdot 0 + 1 \cdot 1 = 1$$

Based on this observation we immediately write a general solution as

$$y = 1 + c_1 e^{-t} + c_2 t e^{-t}, \quad t \geq 0$$

Evaluating the arbitrary constants using the initial conditions, we get $c_1 = c_2 = -1$, and the solution of the initial-value problem is obtained as

$$y = 1 - (1 + t)e^{-t}, \quad t \geq 0$$

For $p = 0.6$ the characteristic equation

$$s^2 + 0.6s + 1 = 0$$

has the complex conjugate roots $s = -\sigma \mp i\omega$, where $\sigma = 0.3$ and $\omega = \sqrt{1 - \sigma^2}$. Consequently, a general solution is

$$y = 1 + c_1 e^{-\sigma t} \cos \omega t + c_2 e^{-\sigma t} \sin \omega t, \quad t \geq 0$$

Evaluating the arbitrary constants from the initial conditions, we obtain the solution of the initial-value problem as

$$y = 1 - e^{-\sigma t} \cos \omega t - \frac{\sigma}{\omega} e^{-\sigma t} \sin \omega t, \quad t \geq 0$$

It is left to the reader as an exercise to show that the solution for $p = 0$ is

$$y = 1 - \cos t, \quad t \geq 0$$

for $p = -0.6$ is

$$y = 1 - e^{\sigma t} \cos \omega t + \frac{\sigma}{\omega} e^{\sigma t} \sin \omega t, \quad t \geq 0$$

and for $p = -2.5$ is

$$y = 1 - (4/3)e^{0.5t} + (1/3)e^{2t}, \quad t \geq 0$$

The graphs of the solutions corresponding to the different values of p are shown in Figure 2.8. The reader is urged to interpret the results.

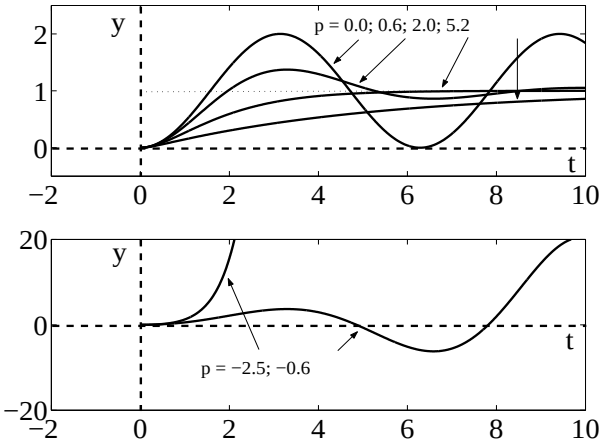

Figure 2.8: Solutions of (2.30) for several p.

2.5 Differential Operators

For a differentiable function f we use the notation f' to denote its derivative. If f' is also differentiable, then f'' denotes its second derivative, etc. An alternative notation is to denote the derivative of f by $\mathcal{D}(f)$, where $\mathcal{D}$ stands for the differentiation operation. We call $\mathcal{D}$ the ***differential operator***. Then the higher order derivatives of f can be expressed as

$$\begin{aligned} f'' &= \mathcal{D}(f') = \mathcal{D}(\mathcal{D}(f)) = \mathcal{D}^2(f) \\ f^{(3)} &= \mathcal{D}(f'') = \mathcal{D}(\mathcal{D}^2(f)) = \mathcal{D}^3(f) \end{aligned}$$

and so on, where $\mathcal{D}^2$ is a short notation for the compound operator $\mathcal{D} \circ \mathcal{D}$, $\mathcal{D}^3$ for $\mathcal{D} \circ \mathcal{D}^2$, etc. Each of the operators $\mathcal{D}$, $\mathcal{D}^2$, $\mathcal{D}^3$, etc. can be viewed as a mapping from a set of functions into another such that the image of a function f under $\mathcal{D}$ is f', under $\mathcal{D}^2$, f'', under $\mathcal{D}^3$, $f^{(3)}$, etc.

Using the operator notation, an nth order linear differential equation with constant coefficients can be written as

$$(\mathcal{D}^n + a_1\mathcal{D}^{n-1} + \cdots + a_{n-1}\mathcal{D} + a_n\mathcal{I})(y) = u(t) \tag{2.33}$$

where $\mathcal{I}$ stands for the identity operator, $\mathcal{I}(f) = f$. Letting

$$L(\mathcal{D}) = \mathcal{D}^n + a_1\mathcal{D}^{n-1} + \cdots + a_{n-1}\mathcal{D} + a_n\mathcal{I}$$

(2.33) can be written in a compact way as

$$L(\mathcal{D})(y) = u(t)$$

$L(\mathcal{D})$ is called a ***linear differential operator*** with constant coefficients. Like $\mathcal{D}$, $\mathcal{D}^2$, etc., $L(\mathcal{D})$ can be viewed as a mapping that maps a function f into a combination of itself and various order derivatives.

We know from calculus that

$$\mathcal{D}(af + bg) = a\mathcal{D}(f) + b\mathcal{D}(g) \tag{2.34}$$

for arbitrary functions f and g and scalars a and b. Then it is easy to show by induction on k that

$$\mathcal{D}^k(af + bg) = a\mathcal{D}^k(f) + b\mathcal{D}^k(g)\,, \quad k = 1, \ldots, n$$

which in turn implies

$$L(\mathcal{D})(af + bg) = aL(\mathcal{D})(f) + bL(\mathcal{D})(g) \tag{2.35}$$

The significance of the property in (2.35) is that if $y = \phi(t)$ and $y = \psi(t)$ are any two solutions of the linear differential equations

$$L(\mathcal{D})(y) = u(t)$$

and

$$L(\mathcal{D})(y) = v(t)$$

then $y = a\phi(t) + b\psi(t)$ is a solution of [10]

$$L(\mathcal{D})(y) = au(t) + bv(t)$$

Linear differential operators with constant coefficients provide great notational and conceptual simplification because they can be treated like polynomials. For example, if $L_1(\mathcal{D})$ and $L_2(\mathcal{D})$ are two such operators then we can define a product operator $L = L_1L_2$ such that

$$L(\mathcal{D})(y) = L_1(\mathcal{D})[L_2(\mathcal{D})(y)]$$

This definition allows us to factor a given linear differential operator with constant coefficients as if it were a polynomial in $\mathcal{D}$.

Example 2.11

Consider a second order linear differential equation with constant coefficients

$$y'' - 3y' + 2y = 0 \tag{2.36}$$

which can be written using the operator notation as

$$(\mathcal{D}^2 - 3\mathcal{D} + 2\mathcal{I})(y) = 0$$

Treating the linear differential operator $\mathcal{D}^2 - 3\mathcal{D} + 2\mathcal{I}$ like a polynomial, we can factor it out as

$$\mathcal{D}^2 - 3\mathcal{D} + 2\mathcal{I} = (\mathcal{D} - \mathcal{I})(\mathcal{D} - 2\mathcal{I})$$

and rewrite the differential equation as

$$(\mathcal{D} - \mathcal{I})[(\mathcal{D} - 2\mathcal{I})(y)] = 0 \tag{2.37}$$

We can do this because

$$\begin{aligned}(\mathcal{D} - \mathcal{I})[(\mathcal{D} - 2\mathcal{I})(y)] &= (\mathcal{D} - \mathcal{I})(y' - 2y)\\ &= \mathcal{D}(y' - 2y) - (y' - 2y)\\ &= (y'' - 2y') - (y' - 2y)\\ &= y'' - 3y' + 2y\\ &= (\mathcal{D}^2 - 3\mathcal{D} + 2\mathcal{I})(y)\end{aligned}$$

[10] This property of linear differential equations has already been mentioned in Example 2.3.

Now letting $(\mathcal{D} - 2\mathcal{I})(y) = z$, (2.37) is transformed into a first order equation

$$(\mathcal{D} - \mathcal{I})(z) = z' - z = 0$$

whose solution can easily be obtained as

$$z = c_1 e^t$$

Substituting the solution back into the definition of z, we get another first order differential equation

$$(\mathcal{D} - 2\mathcal{I})(y) = y' - 2y = c_1 e^t$$

Solving this final equation we obtain

$$y = c_1 e^t + c_2 e^{2t}$$

which is a general solution of (2.36).

The reader might have noticed that this is exactly what we did in finding the general solution of a second order equation whose characteristic equation has a double real root.

We will discuss the significance of linear differential operators in Chapter 3 in connection with linear transformations.

* 2.6 Further Topics on Differential Equations

2.6.1 First Order LDE with Non-Constant Coefficients

Consider a first order linear homogeneous differential equation with a non-constant coefficient

$$y' + a(t)y = 0 \tag{2.38}$$

where $a(t)$ is a given function. Writing (2.38) as

$$y'/y = -a(t)$$

and integrating both sides, we obtain

$$\ln |\, y \,| = c_1 - A(t)$$

or equivalently,

$$|\, y \,| = e^{c_1} e^{-A(t)}$$

where $A(t)$ is any antiderivative of $a(t)$. Noting that $e^{-A(t)} > 0$, we can remove the absolute value in the above expression by defining $e^{c_1} = |\, c \,|$, and thus obtain

$$y = ce^{-A(t)} \tag{2.39}$$

The expression in (2.39), which contains an arbitrary constant, is a general solution of (2.38).

Now consider the non-homogeneous equation

$$y' + a(t)y = u(t) \tag{2.40}$$

Following the method of variation of parameters, we look for a solution of the form $y = e^{-A(t)}v(t)$. Substituting y and y' into (2.40) and simplifying the equation, we get

$$v'(t) = e^{A(t)}u(t)$$

If $V(t)$ is any antiderivative of $v'(t)$ above, we have $v(t) = V(t) + c$, and a general solution of (2.40) is obtained as

$$y = e^{-A(t)}V(t) + ce^{-A(t)} = \phi_p(t) + \phi_c(t) \tag{2.41}$$

where ϕ_p is a particular solution and ϕ_c is a complementary solution.

If an initial condition $y(t_0) = y_0$ is specified, useful choices for $A(t)$ and $V(t)$ are

$$A(t) = \int_{t_0}^{t} a(\tau)\, d\tau$$

and

$$V(t) = \int_{t_0}^{t} e^{A(\tau)}u(\tau)\, d\tau = \int_{t_0}^{t} e^{\int_{t_0}^{\tau} a(\gamma)\, d\gamma}u(\tau)\, d\tau$$

With these choices, the solution of the initial-value problem

$$y' + a(t)y = u(t)\,, \quad y(t_0) = y_0 \tag{2.42}$$

is obtained as

$$y = e^{-\int_{t_0}^{t} a(\tau)\, d\tau}(y_0 + \int_{t_0}^{t} e^{\int_{t_0}^{\tau} a(\gamma)\, d\gamma}u(\tau)\, d\tau)$$

Defining

$$\phi(t, \tau) = e^{-\int_{\tau}^{t} a(\gamma)\, d\gamma} \tag{2.43}$$

and noting that

$$e^{\int_{t_0}^{\tau} a(\gamma)\, d\gamma} = e^{-\int_{\tau}^{t_0} a(\gamma)\, d\gamma} = \phi(t_0, \tau)$$

the solution above can be expressed in more compact form as

$$y = \phi(t, t_0)y_0 + \phi(t, t_0) \int_{t_0}^{t} \phi(t_0, \tau)u(\tau)\, d\tau) \tag{2.44}$$

Taking $\phi(t, t_0)$ inside the integral (it is independent of the variable of integration), and noting that (see Exercise 2.14)

$$\phi(t, t_0)\phi(t_0, \tau) = \phi(t, \tau)$$

an alternative expression for the solution is obtained as

$$y = \phi(t, t_0)y_0 + \int_{t_0}^{t} \phi(t, \tau)u(\tau)\, d\tau = \phi_o(t) + \phi_u(t) \tag{2.45}$$

Note that (2.15) is a special case of (2.45) corresponding to $\phi(t, t_0) = e^{-a(t-t_0)}$. Like (2.15), the expression in (2.45) gives the solution as the sum of two parts, one due to the initial condition y_0, and the other due to the forcing function $u(t)$.

Example 2.12

Let us solve the initial-value problem

$$y' - (\cos t)y = \cos t\,, \quad y(t_0) = y_0$$

Writing the associated homogeneous equation as

$$y'/y = \cos t$$

and integrating both sides, we obtain

$$\ln|\,y\,| = \sin t + c_1$$

or equivalently,

$$y = ce^{\sin t}$$

To find a solution of the non-homogeneous equation, we substitute $y = e^{\sin t}v(t)$ and its derivative, and obtain

$$e^{\sin t}v'(t) + (\cos t)e^{\sin t}v(t) - (\cos t)e^{\sin t}v(t) = e^{\sin t}v'(t) = \cos t$$

Thus

$$v'(t) = (\cos t)e^{-\sin t}\,, \quad v(t) = c - e^{-\sin t}$$

and a general solution is

$$y = e^{\sin t}v(t) = ce^{\sin t} - 1$$

Note that the particular solution $\phi_p(t) = -1$ could have been found by inspection.

The arbitrary constant in the general solution is evaluated using the initial condition

$$y_0 = ce^{\sin t_0} - 1 \implies c = e^{-\sin t_0}(y_0 + 1)$$

Thus the solution of the given initial-value problem is found as

$$y = e^{\sin t}e^{-\sin t_0}(y_0 + 1) - 1 = e^{\sin t - \sin t_0}y_0 + e^{\sin t - \sin t_0} - 1$$

Alternatively, we can use the formula in (2.45). Calculating

$$\phi(t,\tau) = e^{-\int_\tau^t(-\cos\delta)\,d\delta} = e^{\sin t - \sin\tau}$$

(2.45) gives the solution as

$$y = e^{\sin t - \sin t_0}y_0 + \int_{t_0}^{t} e^{\sin t - \sin\tau}\cos\tau\,d\tau = e^{\sin t - \sin t_0}y_0 + e^{\sin t - \sin t_0} - 1$$

2.6.2 Exact Equations

Nonlinear differential equations are difficult to solve even when they are first order. We now consider some special types of first order nonlinear equations for which we can find a solution.

A first order differential equation expressed in differential form

$$M(t,y)\,dt + N(t,y)\,dy = 0 \tag{2.46}$$

where M and N are given functions of two real variables t and y, defined in some rectangular region $\boldsymbol{R}$ of the ty plane, is said to be ***exact*** if there exists a function $F(t,y)$ defined and having continuous first partial derivatives in $\boldsymbol{R}$ such that

$$\frac{\partial F(t,y)}{\partial t} = M(t,y)\,, \quad \frac{\partial F(t,y)}{\partial y} = N(t,y) \tag{2.47}$$

Recall that the differential of a function $F(t,y)$ of two variables is

$$dF(t,y) = \frac{\partial F(t,y)}{\partial t}\,dt + \frac{\partial F(t,y)}{\partial y}\,dy$$

Thus if F satisfies the conditions in (2.47), then the equation (2.46) can be expressed as

$$M(t,y)\,dt + N(t,y)\,dy = dF(t,y) = 0$$

from which we obtain

$$F(t,y) = c$$

If this relation between t and y defines y as a differentiable function of t as $y = \phi(t)$, then ϕ is a solution of (2.46). Conversely, it can be shown that any solution $y = \phi(t)$ of the exact equation (2.46) must satisfy $F(t,\phi(t)) = c$. Such a relation between t and y is called an ***implicit solution***.

Determining whether (2.46) is exact using the definition is equivalent to finding an implicit solution. Fortunately, there is a much simpler way of checking (2.46) for exactness, which also provides a constructive method to find a solution.

Suppose that equation (2.46) is exact so that there exists a function F satisfying conditions (2.47) in some rectangular region $\boldsymbol{R}$. If the functions M and N in (2.46) have continuous first partial derivatives in $\boldsymbol{R}$, then

$$\frac{\partial M(t,y)}{\partial y} = \frac{\partial^2 F(t,y)}{\partial y\,\partial t} = \frac{\partial^2 F(t,y)}{\partial t\,\partial y} = \frac{\partial N(t,y)}{\partial t}$$

On the other hand, if M and N have continuous first partial derivatives and satisfy

$$\frac{\partial M(t,y)}{\partial y} = \frac{\partial N(t,y)}{\partial t} \tag{2.48}$$

in some region $\boldsymbol{R}$, then the function

$$F(t,y) = \int_{t_0}^{t} M(\tau, y_0)\,d\tau + \int_{y_0}^{y} N(t,z)\,dz \tag{2.49}$$

where (t_0, y_0) is an arbitrary point in $\boldsymbol{R}$, satisfies (2.47) (see Exercise 2.16). Thus (2.48) gives necessary and sufficient conditions for (2.46) to be exact when the functions M and N have continuous first partial derivatives, and (2.49) provides a formula to obtain the function F when these conditions are satisfied.

Example 2.13

Show that the equation

$$(y+1)\,dt + (t-y)\,dy = 0$$

is exact, and then find a solution satisfying the initial condition $y(0) = 1$.

Since

$$\frac{\partial M(t,y)}{\partial y} = 1 = \frac{\partial N(t,y)}{\partial t}$$

everywhere, the equation is exact. Then there exists F that satisfies

$$\frac{\partial F(t,y)}{\partial t} = M(t,y) = y + 1$$

Integrating both sides of this expression with respect to t, we get

$$F(t,y) = (y+1)t + f(y)$$

where f is a function to be determined. Using

$$\frac{\partial F(t,y)}{\partial y} = t + f'(y) = N(t,y) = t - y$$

we obtain

$$f'(y) = -y \implies f(y) = -y^2/2 + c_1$$

Thus

$$F(t,y) = (y+1)t - y^2/2 + c_1$$

and an implicit solution is obtained as

$$(y+1)t - y^2/2 + c_1 = c_2$$

or equivalently as

$$y^2 - 2ty - 2t = c$$

Alternatively, $F(t,y)$ can be obtained from the formula in (2.49) (see Exercise 2.17).

The implicit solution above defines two families of solutions

$$y = t - (t^2 + 2t + c)^{1/2}$$

and

$$y = t + (t^2 + 2t + c)^{1/2}$$

No member of the first family satisfies the initial condition. Substituting the initial conditions in the expression for the second family we get $c = 1$, and the required solution is obtained as

$$y = t + (t^2 + 2t + 1)^{1/2} = 2t + 1$$

Note that, unlike linear differential equations, we cannot say that the solution above is the only solution of the initial-value problem considered.

When an equation of the form (2.46) is not exact, it is sometimes possible to find a function $I(t, y)$ such that

$$I(t, y)M(t, y)\,dt + I(t, y)N(t, y)\,dy = 0 \tag{2.50}$$

is exact (see Exercises 2.18 and 2.19). Such a function is called an ***integrating factor***. If $I(t, y) \neq 0$ in a rectangular region in which equation (2.50) is exact, then (2.46) and (2.50) have the same solutions.

Example 2.14

The linear equation $y' + 2y = 0$ written in differential form as

$$2y\,dt + dy = 0$$

is not exact. Multiplying both sides by e^{2t} (which is nonzero everywhere), we get an exact equation

$$2e^{2t}y\,dt + e^{2t}\,dy = 0$$

for which F can be obtained as

$$F(t, y) = e^{2t}y + c_1$$

Thus an implicit solution is

$$e^{2t}y + c_1 = c_2$$

Letting $c = c_2 - c_1$ we get the expected explicit solution

$$y = ce^{-2t}$$

2.6.3 Separable Equations

A differential equation of the form

$$M_1(t)M_2(y)\,dt + N_1(t)N_2(y)\,dy = 0 \tag{2.51}$$

is said to be ***separable***.

In a region where $N_1 \neq 0$ and $M_2 \neq 0$, (2.51) is equivalent to

$$p(t)\,dt + q(y)\,dy = 0$$

where $p = M_1/N_1$ and $q = N_2/M_2$. Integrating both sides we get an implicit solution

$$P(t) + Q(y) = c$$

where P and Q are arbitrary antiderivatives of p and q. Any function $y = \phi(t)$ that satisfies the implicit solution is an explicit solution of (2.51). In addition, if the equation $M_2(y) = 0$ has a real root

$$y = r\,, \quad r \in \mathbb{R}$$

then it is also a solution (which is lost when dividing (2.51) by $N_1(t)M_2(y)$ to obtain the equivalent equation).

Example 2.15

The equation

$$2ty^2\,dt + dy = 0$$

is separable as it can be written as

$$2t\,dt + (1/y^2)\,dy = 0$$

Integrating the last equation, we obtain an implicit solution as $t^2 - y^{-1} = c$, which defines a family of solutions

$$y = \frac{1}{t^2 - c}$$

In addition, $y = 0$ is also a solution not included in this family.

2.6.4 Reduction of Order

Consider a second order linear differential equation with non-constant coefficients described as

$$y'' + a_1(t)y' + a_2(t)y = u(t) \tag{2.52}$$

If two linearly independent solutions of the associated homogeneous equation are known (that is, if a complementary solution is known), then a general solution can be obtained by the method of variation of parameters. Unfortunately, except in special cases, there is no general method of finding a complementary solution. However, if a solution $y = \phi(t)$ of the associated homogeneous equation

$$y'' + a_1(t)y' + a_2(t)y = 0$$

is known, then a general solution of the form $y = v(t)\phi(t)$ can be obtained as follows. By substituting y and its derivatives

$$\begin{aligned} y' &= v'\phi + v\phi' \\ y'' &= v''\phi + 2v'\phi' + v\phi'' \end{aligned}$$

into (2.52), we get

$$(\phi'' + a_1\phi' + a_2\phi) + \phi v'' + (2\phi' + a_1\phi)v' = u$$

where we dropped the argument t for simplicity. Since ϕ is a solution of the associated homogeneous equation, the term in the first parenthesis above vanishes. Letting $w = v'(t)$ the equation reduces to

$$\phi(t)w' + [2\phi'(t) + a_1(t)\phi(t)]w = u(t)$$

which is a first order equation in w that can be solved by known methods. Then v is obtained by integrating w.

Example 2.16

Solve the second order differential equation

$$y'' - (3/t)y' + (4/t^2)y = 1/t\,, \quad t > 0$$

if it is given that $y = t^2$ is a solution of the associated homogeneous equation.

Letting $y = t^2 v(t)$ and substituting

$$\begin{aligned} y' &= 2tv(t) + t^2 v'(t) \\ y'' &= 2v(t) + 4tv'(t) + t^2 v''(t) \end{aligned}$$

into the given equation, we obtain after simplification

$$t^2 v''(t) + tv'(t) = 1/t$$

Defining $w = v'(t)$, the last equation reduces to a first order equation

$$w' + (1/t)w = 1/t^3$$

A solution of the last equation is found as

$$w = c_1/t - 1/t^2$$

Integrating, we get

$$v = c_1 \ln t + c_2 - 1/t$$

Thus a solution of the original problem is obtained as

$$y = c_1 t^2 \ln t + c_2 t^2 - t$$

2.7 Systems of Differential Equations

Consider an nth order differential equation of the form (2.2)

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}) \tag{2.53}$$

together with a set of n initial conditions

$$y(t_0) = y_0\,,\ y'(t_0) = y_1\,,\ \ldots\,,\ y^{(n-1)}(t_0) = y_{n-1} \tag{2.54}$$

Let us define a set of new dependent variables

$$x_1 = y\,,\ x_2 = y'\,,\ \ldots\,,\ x_n = y^{(n-1)}$$

Their derivatives can easily be obtained using the definition and (2.53) as

$$\begin{aligned} x_1' &= y' &&= x_2 \\ x_2' &= y'' &&= x_3 \\ &\vdots \\ x_{n-1}' &= y^{(n-1)} &&= x_n \\ x_n' &= y^{(n)} &&= f(t, y, y', \ldots, y^{(n-1)}) &&= f(t, x_1, x_2, \ldots, x_n) \end{aligned}$$

(2.55)

The equations in (2.55) form a system of n first order differential equations which can be written in matrix form as

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}), \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{2.56}$$

where

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} y_0 \\ \vdots \\ y_{n-2} \\ y_{n-1} \end{bmatrix}, \quad \boldsymbol{f}(t, \boldsymbol{x}) = \begin{bmatrix} x_2 \\ \vdots \\ x_n \\ f(t, x_1, x_2, \ldots, x_n) \end{bmatrix}$$

and $\boldsymbol{x}'$ denotes element-by-element derivative of $\boldsymbol{x}$.

The nth order differential equation (2.53) and the system of first order differential equations in (2.56) are equivalent in the sense that there is a one-to-one correspondence between their solutions: If $y = \phi(t)$ is the solution of (2.53) corresponding to the initial conditions in (2.54), then

$$\boldsymbol{x} = \boldsymbol{\phi}(t) = \text{col}\,[\,\phi(t), \phi'(t), \ldots, \phi^{(n-1)}(t)\,]$$

is the solution of (2.56). Conversely, if

$$\boldsymbol{x} = \boldsymbol{\phi}(t) = \text{col}\,[\,\phi_1(t), \phi_2(t), \ldots, \phi_n(t)\,]$$

is the solution of (2.56), then $y = \phi_1(t)$ is the solution of (2.53) that satisfies the initial conditions in (2.54). Furthermore, $\phi_1'(t) = \phi_2(t), \ldots, \phi_{n-1}'(t) = \phi_n(t)$.[11]

Note that if the differential equation in (2.53) is a linear one as in (2.3), then the last equation in (2.55) becomes

$$x_n' = -a_n(t)x_1 - \cdots - a_2(t)x_{n-1} - a_1(t)x_n + u(t)$$

and accordingly, the system in (2.56) takes the form

$$\boldsymbol{x}' = A(t)\boldsymbol{x} + \boldsymbol{u}(t) \tag{2.57}$$

where

$$A(t) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ -a_n(t) & -a_{n-1}(t) & \cdots & -a_1(t) \end{bmatrix}, \quad \boldsymbol{u}(t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ u(t) \end{bmatrix}$$

If, in addition, the linear differential equation has constant coefficients, then $A(t)$ becomes a constant matrix. (2.57) is called a system of linear differential equations. We will study systems of linear differential equations in Chapter 6.

[11] Here we assumed that (2.53) and (2.56) have unique solutions that satisfy the given initial conditions. This is indeed the case under certain assumptions concerning the function f in (2.53). The reader is referred to Appendix B for details.

Example 2.17

The second order differential equation

$$y'' + a_1 y' + a_2 y = u(t)\,, \quad y(t_0) = y_0\,, \quad y'(t_0) = y_1$$

is equivalent to the system

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} x_2 \\ -a_2 x_1 - a_1 x_2 + u(t) \end{bmatrix}, \quad \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

where $x_1 = y$ and $x_2 = y'$.

Note that since

$$\begin{aligned} \boldsymbol{f}(t, \boldsymbol{x}) &= \begin{bmatrix} x_2 \\ -a_2 x_1 - a_1 x_2 + u(t) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ u(t) \end{bmatrix} = A\boldsymbol{x} + \boldsymbol{u}(t) \end{aligned}$$

the system of differential equations is linear.

Rewriting an nth order differential equation as an equivalent system of first order differential equations allows us to use some well-established numerical solution techniques as we consider in the next section.

2.8 Numerical Solution of Differential Equations

As we mentioned earlier, nonlinear differential equations and even linear equations of order higher than one with non-constant coefficients are difficult, and most of the time impossible to solve analytically. However, solutions of such equations, if they are known to exist, can be approximated with a desired degree of accuracy by using numerical techniques.

Consider an initial-value problem involving a first-order differential equation, not necessarily linear:

$$y' = f(t, y)\,, \quad y(t_0) = y_0 \tag{2.58}$$

Suppose that (2.58) has unique solution $y = \phi(t)$ on some interval $\mathcal{I}$ that includes t_0 so that

$$\phi'(t) = f(t, \phi(t))\,, \quad \phi(t_0) = y_0 \tag{2.59}$$

Consider the Taylor series expansion of ϕ about a point $t \geq t_0$ in $\mathcal{I}$.

$$\phi(t + h) = \phi(t) + h\phi'(t) + R(t, h)$$

where $R(t, h)$ denotes the remainder and is proportional to h^2. For sufficiently small h, $\phi(t + h)$ can be approximated as

$$\phi(t + h) \approx \phi(t) + hf(t, \phi(t))$$

where $\phi'(t)$ is substituted from (2.59). Let $t_k = t_0 + kh, k = 0, 1, \ldots$, and let w_k denote the approximate value of $\phi(t_k)$. Then the approximate expression for $\phi(t + h)$ above evaluated at $t = t_k$ becomes

$$w_{k+1} = w_k + hf(t_k, w_k)\,, \quad k = 0, 1, \ldots \tag{2.60}$$

which allows us to obtain the approximate values of the solution recursively, starting with $w_0 = \phi(t_0) = y_0$. This technique of obtaining an approximate solution to an initial-value problem is known as the ***Euler method***, and is suitable for computer implementation.

Note that the recursion relation in (2.60) runs forward and gives the approximate values of the solution for $t \geq t_0$. To obtain an approximate solution for $t \leq t_0$ all we have to do is to replace h with $-h$. Then with $t_k = t_0 + kh, k = 0, -1, \ldots$, the backward recursion becomes

$$w_{k-1} = w_k - hf(t_k, w_k)\,, \quad k = 0, -1, \ldots \tag{2.61}$$

Example 2.18

The initial-value problem

$$y' = -y + 1\,, \quad y(0) = 0 \tag{2.62}$$

has the exact solution

$$y = 1 - e^{-t}\,, \quad t \geq 0$$

The Euler method gives the recursion relation

$$w_{k+1} = w_k + h(1 - w_k)\,, \quad w_0 = 0$$

for the approximate solution. The following MATLAB code runs the recursion relation for $0 \leq t_k \leq 5$ with a step size of $h = 0.5$, and plots the approximate solution together with the exact solution.

```
t=0:0.01:5;                % range of t for solution
y=1-exp(-t);               % exact solution
h=0.5;                     % step size
k=1; tk(1)=0; wk(1)=0;     % initialization
while tk(k)<5              % recursion
   tk(k+1)=tk(k)+h;
   wk(k+1)=(1-h)*wk(k)+h;
   k=k+1;
end
plot(t,y,tk,wk,'.')
```

Plots of the exact and approximate solutions are shown in Figure 2.9. The reader may try a different step size to observe its effect on the quality of approximation.

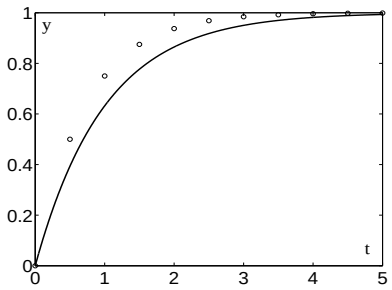

Figure 2.9: Exact and approximate solutions of (2.62).

Note that the recursion formula in (2.60) can be obtained directly from (2.58) by the following replacement of the variables.

$$\begin{aligned} t &\leftarrow t_k \\ y &\leftarrow w_k \\ y' &\leftarrow \frac{1}{h}\,(w_{k+1} - w_k) \end{aligned} \tag{2.63}$$

This observation suggests that the Euler method can be generalized to higher order differential equations provided that we derive approximate expressions for higher derivatives of the solution. For this purpose, let us define $\psi(t) = \phi'(t)$ and denote the approximate value of $\psi(t_k) = \phi'(t_k)$ by v_k. Then

$$\begin{aligned} \phi''(t_k) = \psi'(t_k) &\approx \frac{1}{h}\,(v_{k+1} - v_k) \\ &\approx \frac{1}{h}\,[\phi'(t_{k+1}) - \phi'(t_k)] \\ &\approx \frac{1}{h}\,(\frac{w_{k+2} - w_{k+1}}{h} - \frac{w_{k+1} - w_k}{h}) \\ &= \frac{1}{h^2}\,(w_{k+2} - 2w_{k+1} + w_k) \end{aligned}$$

Thus the replacement

$$y'' \leftarrow \frac{1}{h^2}\,(w_{k+2} - 2w_{k+1} + w_k) \tag{2.64}$$

in addition to those in (2.63), in a second order differential equation

$$y'' = f(t, y, y')\,, \quad y(t_0) = y_0\,, \quad y'(t_0) = y_1$$

yields the recursion relation

$$w_{k+2} = 2w_{k+1} - w_k + h^2 f(t_k, w_k, \frac{w_{k+1} - w_k}{h})$$

The initial values w_0 and w_1 needed to start the recursion are obtained from the initial conditions as

$$w_0 = y_0\,, \quad w_1 \approx h\phi'(t_0) + w_0 = hy_1 + y_0$$

Example 2.19

The second order initial-value problem

$$y'' + 0.6y' + y = 0\,, \quad y(0) = 0\,, \quad y'(0) = 1 \tag{2.65}$$

is similar to the one considered in Example 2.10, and has the exact solution

$$y = \frac{1}{\omega}\, e^{-\sigma t} \sin \omega t$$

where $\sigma = 0.3$ and $\omega = \sqrt{1 - \sigma^2} \approx 0.9539$.

The substitutions in (2.63) and (2.64) yield

$$w_{k+2} = (2 - 0.6\,h)w_{k+1} + (-1 + 0.6\,h - h^2)w_k,\ w_0 = 0,\ w_1 = h \tag{2.66}$$

The approximate solution obtained from the recursion relation above with $h = 0.1$ is shown in Figure 2.10 together with the exact solution.

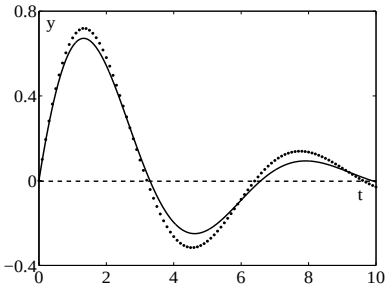

Figure 2.10: Exact and approximate solutions of (2.65).

An alternative approach to solving a high order differential equation numerically is to transform it into a system of first order differential equations as explained in the previous section. Note that a system of differential equations as in (2.56) is no different than a first order differential equation except that $\boldsymbol{x}$ and $\boldsymbol{f}$ are now column vectors rather than scalars. However, this does not make any difference in the application of the Euler method. Following the same argument leading to (2.60), a recursion relation

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + h\,\boldsymbol{f}(t_k, \boldsymbol{w}_k)\,, \quad \boldsymbol{w}_0 = \boldsymbol{x}_0 \tag{2.67}$$

can be obtained for the approximate solution.

Example 2.20

With $x_1 = y, x_2 = y'$, the second order differential equation in (2.65) is transformed into

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 - 0.6\,x_2 \end{bmatrix}, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.68}$$

and the Euler method yields

$$\begin{bmatrix} w_{1,k+1} \\ w_{2,k+1} \end{bmatrix} = \begin{bmatrix} w_{1k} + hw_{2k} \\ (1 - 0.6\,h)w_{2k} - hw_{1k} \end{bmatrix}, \quad \begin{bmatrix} w_{10} \\ w_{20} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.69}$$

Running the recursion relation with $h = 0.1$ and plotting w_{1k} we obtain the same approximate solution as the one in Figure 2.10. Of course, this is not a coincidence, because the recursion relations in (2.66) and (2.69) are equivalent (see Exercise 2.32).

There are other numerical solution methods that are more sophisticated and more accurate than Euler method. MATLAB provides several built-in functions that use a variable step size to solve systems of first-order differential equations of the form (2.56). The reader is referred to Appendix D for a brief summary of the use of these functions.

2.9 Exercises

1. Find a general solution of the following first order linear differential equations
   (a) $y' + y = 3e^{2t} + 5\sin 2t$
   (b) $y' + 2y = 2t - 1$

(c) $y' - y = e^t$

(d) $y' + (1/t)y = 0$

(e) $y' - (\cos t)y = \cos t$

2. Solve the following initial-value problems

(a) $2y' + y = t\,, \quad y(0) = 0$

(b) $y' + y = t^2\,, \quad y(0) = 2$

(c) $y' + y = \begin{cases} 0\,, & t < 0 \quad \text{or} \quad t > 1 \\ 1\,, & -1 < t < 1 \end{cases}\,, \quad y(0) = 0$

(d) $x^2 \dfrac{dy}{dx} + 2xy = 2\sin x\,, \quad y(2\pi) = 0$

3. Suppose that $\sigma \neq -a$.

(a) Show that the first order differential equation

$$y' + ay = e^{\sigma t}$$

has a particular solution of the form $\phi_p(t) = Ae^{\sigma t}$, and find A by substitution.

(b) Show that

$$y' + ay = te^{\sigma t}$$

has a particular solution of the form $\phi_p(t) = (A_0 t + A_1)e^{\sigma t}$, and find A_0 and A_1.

(c) Generalize the result in part (b) and write down the form of a particular solution of

$$y' + ay = t^m e^{\sigma t}$$

4. (a) Show that the first order differential equation

$$y' - \sigma y = e^{\sigma t}$$

has a particular solution of the form $\phi_p(t) = Ate^{\sigma t}$, and find A.

(b) Generalize the result in part (a) and write down the form of a particular solution of

$$y' - \sigma y = t^m e^{\sigma t}$$

5. (a) Find the solution $y = \phi(t)$ of the initial-value problem

$$y' + y = u(t)\,, \quad y(0) = y_0$$

where

$$u(t) = \begin{cases} 0\,, & t < 0 \\ \cos t\,, & t > 0 \end{cases}$$

(b) Show that there exists a periodic function $\phi_P(t)$ such that

$$\lim_{t \to \infty} (\phi(t) - \phi_P(t)) = 0$$

independent of y_0.

(c) Find value of y_0 such that $\phi(t) = \phi_P(t), t > 0$.

6. Suppose that the initial-value problem

$$y' + ay = u(t)\,, \quad y(0) = y_0$$

has the solution $y = \phi(t)$. Show that the solution of the initial-value problem

$$y' + ay = u(t - t_0)\,, \quad y(t_0) = y_0\,, \quad y'(t_0) = y_1$$

is $y = \psi(t) = \phi(t - t_0)$. Hint: Substitute $u(t - t_0)$ for $u(t)$ in (2.15).

7. A first order differential equation of the form

$$y' + p(t)y = q(t)y^n$$

where $n \neq 1$, is called a ***Bernoulli equation***.

(a) Show that a change of the dependent variable $x = y^{1-n}$ transforms the Bernoulli equation to a first order linear differential equation in x.

(b) Find the solution of the initial-value problem

$$y' + y + y^2 = 0\,, \quad y(0) = 1$$

and indicate the interval on which the solution is valid.

8. Show, by direct substitution, that if the characteristic equation of the second order linear differential equation in (2.24) has a pair of complex conjugate roots $s_{1,2} = \sigma \mp i\omega$, then each of the functions $\phi_1(t) = e^{\sigma t}\cos\omega t$ and $\phi_2(t) = e^{\sigma t}\sin\omega t$ is a solution of (2.24). Hint: Note that

$$(\sigma + i\omega)^2 + a_1(\sigma + i\omega) + a_2 \;=\; \sigma^2 - \omega^2 + a_1\sigma + a_2 + i(2\sigma\omega + a_1\omega) \;=\; 0$$

9. Show that if $\sigma \in \mathbb{R}$ is not a root of the characteristic equation of the second order differential equation

$$y'' + a_1 y' + a_2 y = e^{\sigma t}$$

then it has a particular solution of the form $\phi_p(t) = Ae^{\sigma t}$, and find A.

10. Show that if $\sigma + i\omega$ is not a root of the characteristic equation of the second order differential equation

$$y'' + a_1 y' + a_2 y = e^{\sigma t}(p\cos\omega t + q\sin\omega t)$$

where $p, q \in \mathbb{R}$, then it has a particular solution of the form

$$\phi_p(t) = e^{\sigma t}(A\cos\omega t + B\sin\omega t)$$

and find A and B.

11. Solve the following initial-value problems

(a) $y'' + 3y' + 2y = e^{-t}\,, \quad y(0) = y'(0) = 0$

(b) $y'' + 2y' + 2y = 10\cos 2t\,, \quad y(0) = -1\,, \quad y'(0) = 4$

(c) $ty'' + 2y' + ty = 0\,, \quad y(\pi) = 0\,, \quad y'(\pi) = -1$. Hint: Let $v = ty$.

12. Find the solution $y = \phi_T(t)$ of the initial-value problem

$$y'' + 3y' + 2y = u_T(t)\,, \quad y(0) = y'(0) = 0$$

for $t > 0$, where $u_T(t)$ is the unit pulse in (2.22) with $t_0 = 0$, and investigate the behavior of the solution as $T \to 0$.

13. Find a general solution of the differential equation

$$t^2y'' - ty' + y = 2t\,, \quad t > 0$$

if it is given that $y = t$ and $y = t \ln t$ are solutions of the associated homogeneous equation.

14. Show the identity

$$\phi(t, t_0)\phi(t_0, \tau) = \phi(t, \tau)$$

for $\phi(t, \tau)$ defined in (2.43).

15. Solve the following exact differential equations

(a) $(3t^2 + y^2)\,dt + 2ty\,dy = 0$

(b) $(ye^{xy} - 4x^3)\,dx + xe^{xy}\,dy = 0$

16. Differentiate $F(t, y)$ in (2.49) with respect to t and with respect to y to show that if M and N satisfy (2.48), then F satisfies (2.47). The expression in (2.49) is obtained by taking the line integral of dF from an arbitrary initial point $(t_0, y_0) \in \boldsymbol{R}$ to $(t, y) \in \boldsymbol{R}$ first along a horizontal line segment from (t_0, y_0) to (t, y_0), and then along a vertical line segment from (t, y_0) to (t, y), and noting that on the horizontal line segment

$$dF(\tau, z) = \frac{\partial F(\tau, y_0)}{\partial \tau}\,d\tau = M(\tau, y_0)\,d\tau$$

and on the vertical line segment

$$dF(\tau, z) = \frac{\partial F(t, z)}{\partial z}\,dz = N(t, z)\,dz$$

Obtain an alternative expression for F by integrating dF first along a vertical line segment from (t_0, y_0) to (t_0, y), and then along a horizontal line segment from (t_0, y) to (t, y).

17. Use formula (2.49) to obtain a function

$$F(t, y) = \int_{t_0}^{t} (y_0 + 1)\,d\tau + \int_{y_0}^{y} (t - z)\,dz$$

for the exact differential equation in Example 2.9, and show that it gives the same implicit solution.

18. For the following differential equation find integrating factors of the given form, and then solve the resulting exact differential equations.

(a) $(t^2 + y^2)\,dt + ty\,dy = 0\,, \quad I(t, y) = f(t)$

(b) $y\,dx + dy = 0\,, \quad I(x, y) = g(x)$

(c) $(u^2v + v^3)\,du - 2uv^2\,dv = 0\,, \quad I(u, v) = u^m v^n$

19. Let M_t denote $\partial M/\partial t$, M_y denote $\partial M/\partial y$, etc.

(a) Show that if

$$p = \frac{M_y - N_t}{N}$$

is a function of t only, then

$$I(t) = e^{\int p(t)\,dt}$$

is an integrating factor for (2.46).

(b) Use the result of part (a) to find an integrating factor for

$$(2t - y^2)\,dt + ty\,dy = 0\,, \quad t > 0$$

and then obtain an implicit solution.

(c) Show that if

$$q = \frac{N_t - M_y}{M}$$

is a function of y only, then

$$I(y) = e^{\int q(y)\,dy}$$

is an integrating factor for (2.46).

(d) Use the result of part (b) to find an integrating factor for

$$y\,dt - (3t + y^4)\,dy = 0\,, \quad y > 0$$

and then obtain an implicit solution.

20. Solve the following separable differential equations

(a) $(2t - 1)y\,dt + dy = 0$

(b) $\sin x \cos y\,dx + \cos x \sin y\,dy = 0$

(c) $4uv\,du + (u^2 + 1)\,dv = 0$

21. Consider a second order differential equation

$$F(t, y', y'') = 0$$

in which the dependent variable y does not appear explicitly.

(a) Show that a change of the dependent variable $x = y'$ transforms the equation to a first order equation in the dependent variable x.

(b) Use the result of part (a) to solve the initial-value problem

$$y'' + y' = 1\,, \quad y(0) = y'(0) = 0$$

22. Consider a second order differential equation

$$F(y, \frac{dy}{dt}, \frac{d^2y}{dt^2}) = 0$$

in which the independent variable t does not appear explicitly.

(a) Show that a change of the variables

$$\frac{dy}{dt} = v\,, \quad \frac{d^2y}{dt^2} = \frac{dv}{dt} = \frac{dv}{dy}\,\frac{dy}{dt} = v\frac{dv}{dy}$$

transforms the equation to a first order equation in the independent variable y and the dependent variable v.

(b) Use the result of part (a) to solve the initial-value problem

$$yy'' - (y')^2 = 0\,, \quad y(0) = 1\,, \quad y'(0) = -1$$

23. Find a general solution of

$$(t-1)y'' - ty' + y = 1$$

Hint: Look for a solution of the form $\phi_c(t) = At + B$ for the associated homogeneous equation.

24. Solve the initial-value problem

$$ty'' + 2y' + ty = 0\,, \quad y(\pi) = 0\,, \quad y'(\pi) = -1$$

Hint: Use a change of the dependent variable as $v = ty$.

25. Solve the initial-value problems in Exercise 2.2 by using the MATLAB function `ode23`. Plot the resulting solutions and the exact solutions obtained in Exercise 2.2 on the same graph. The MATLAB function `ode23` requires a user defined function (call it `myfunction` for future use) that evaluates the vector-valued function $\boldsymbol{f}(t, \boldsymbol{x})$ in (2.56) and returns it as a vector `xdot`.

26. Let $f(t) = \sin t$, whose derivatives are $f'(t) = \cos t$ and $f''(t) = -\sin t$. Let $w_k = f(kh) = \sin kh$ denote the value of f at $t = kh, k = 0, 1, \ldots$, and let

$$f'(kh) \approx \frac{w_{k+1} - w_k}{h} \quad \text{and} \quad f''(kh) \approx \frac{w_{k+2} - 2w_{k+1} + w_k}{h^2}$$

be the Euler approximations of f' and f'' at $t = kh$. Use MATLAB to compute the approximate values of f' and f'' over the interval $0 \le t \le 10$ using a step size of $h = 0.1$, and plot the exact and approximate values of each derivative on the same graph.

27. Write a MATLAB function

`function [tk, wk] = myeuler(ti, t0, tf, h, x0)`

to implement the Euler method, where `h` $= h$ is the step size, `ti` $= t_i = t_0 - Mh$ and `tf` $= t_f = t_0 + Nh$ specify the end points of the interval over which the solution is to be computed, `x0` $= \boldsymbol{x}_0$ is an $n \times 1$ column vector, `tk` is an $N + M + 1$ dimensional array containing the points $t_k, k = -M, \ldots, 0, \ldots, N$, and `wk` is an $n \times (M + N + 1)$ matrix, the columns of which are $\boldsymbol{w}_k$. The function `myeuler` can use the user defined function `myfunction` in Exercise 2.25 that evaluates $\boldsymbol{f}(t, \boldsymbol{x})$. Note that `myeuler` is required to solve a given system of differential equations both forward and backward. It must return only the forward solution if $t_i = t_0$ and only the backward solution if $t_f = t_0$.

28. Solve the initial-value problems in Exercise 2.2 numerically by using the function `myeuler` written in Exercise 2.27 with $h = 0.1$ and $h = 0.5$. Plot the resulting solutions and the exact solutions obtained in Exercise 2.2 on the same graph.

29. Solve the initial-value problems in Exercise 2.11 numerically by using the Euler method. Plot both the exact and numerical solutions on the same graph.

30. Solve the following initial-value problems by using both `myeuler` and `ode23`, and plot the results on the same graph.

(a)

$$y' = t^2 + y^2\,, \quad y(0) = 1$$

(b)

$$y'' - 2ty' + y^2 = 1\,, \quad y(0) = y'(0) = 1$$

31. (a) Transform each of the initial-value problems in Exercise 2.11 into a system of first-order initial-value problems.

(b) Use the MATLAB function `ode23` to solve the systems in part (a). Plot both the exact and numerical solutions on the same graph.

(c) Repeat (b) using the function `myeuler` written in Exercise 2.27.

32. Show that the recursion relations in (2.66) and (2.69) are equivalent, i.e., the sequence w_{1n} produced by (2.69) is the same as the sequence w_n produced by (2.66). Hint: Use (2.69) to obtain an expression for $w_{1,n+2}$ in terms of $w_{1,n+1}$ and w_{1n}.

33. A differential equation together with additional conditions on the solution that are to be satisfied at two or more values of the independent variable is called a ***boundary value problem***.

(a) Show that the boundary value problem

$$y'' + y = 0\,, \quad y(0) = 0\,, \quad y(\pi) = 1$$

has no solution.

(b) Show that the boundary value problem

$$y'' = \lambda y\,, \quad y(0) = 0\,, \quad y(\pi) = 0$$

has a nontrivial solution if and only if $\lambda = -n^2$ for some integer n, in which case the nontrivial solution is

$$y = c_n \sin nt$$

34. (Application) A family of curves defined by the equation

$$F(x, y) = c\,, \quad c \in \mathbb{R}$$

is said to be orthogonal to a second family of curves defined by the equation

$$G(x, y) = d\,, \quad d \in \mathbb{R}$$

if every curve of the first family intersects every curve of the second family at right angles. (Two such families are called orthogonal trajectories of each other.) Assume that every curve of each family has a well-defined gradient at every point on the xy plane. Orthogonality of the curves is equivalent to orthogonality of the gradients at points of intersection, which requires that

$$G_x\,dx + G_y\,dy = -F_y\,dx + F_x\,dy$$

(a) Consider a family of concentric circles defined by the equation

$$x^2 + y^2 = c^2$$

for which $F_x = 2x$ and $F_y = 2y$ at every point (x, y). The orthogonal trajectories must then satisfy

$$-2y\,dx + 2x\,dy = 0$$

Solve the above equation to obtain an expression for the orthogonal trajectories of the family of circles. Plot both families on the same graph, and verify that the curves of the two families indeed intersect at right angles.

(b) Find the orthogonal trajectories of the family of curves defined by

$$2x^2 + y^2 = c^2$$

and plot both the given family and the orthogonal family on the same graph.

(c) Repeat (b) for the family

$$y = ce^x$$

35. (Application) The behavior of a particle of mass m in vertical motion in the air near the surface of the earth is described by the second order linear differential equation

$$my'' = -mg + f(t)$$

where $y(t)$ is the position of the particle at time t measured positive upward from the surface of the earth, mg is the downward gravitational force, and $f(t)$ represents additional external forces acting on the particle.

(a) Assuming $f(t) = 0$, find the solution corresponding to the initial conditions $y(0) = y_0$ and $y'(0) = v_0$.

(b) Find the time $t = t_f$ at which the particle falls on earth and the velocity with which it hits the ground.

(c) Repeat (a) if the air resistance is modeled as $f(t) = -ky'(t)$. Show that if y_0 is large enough, then the velocity of the particle approaches a constant limit v_∞, and find v_∞.

36. (Application) According to Malthusian growth model, a certain population increases at a rate that is proportional to its current value. If $p(t)$ represents the population at time t, then

$$p' = rp$$

where r is a constant birth rate per individual.

(a) Find an expression for $p(t)$ if $p(0) = p_0$.

(b) Find T_d such that $p(T_d) = 2p_0$. T_d is called the doubling time of the population.

(c) Show that $p(t + T_d) = 2p(t)$ for all t.

37. (Application) A more realistic population model, which takes into account the death rate as well as the birth rate is the logistic population model described as

$$p' = r(1 - \frac{p}{C})p$$

where r is the birth rate, and rp/C is the death rate per individual. (The model assumes that the death rate per individual increases in direct proportion to the population due to competition.)

(a) Find an expression for $p(t)$ if $p(0) = p_0$. Plot $p(t)$ for each of the cases $0 < p_0 < C$, $p_0 = C$, and $p_0 > C$.

(b) Show that

$$\lim_{t \to \infty} p(t) = C$$

independent of p_0. C is called the carrying capacity of the environment. (If $p > C$ at any time then $p' < 0$, and $p(t)$ decreases until it eventually reaches C. If $p < C$ then $p' > 0$, and $p(t)$ increases until it reaches C. If $p = C$ then $p' = 0$, and $p(t)$ remains stable at C.)

(c) Show that if p_0 is much smaller than C, then $p(t)$ obtained from the logistic population model can be approximated with that obtained from the Malthusian growth model for small t (that is, as long as $p(t)$ remains small compared with C).

38. (Application) According to Newton's law of cooling, the rate of change of the temperature of an object is proportional to the temperature difference between the object and its surrounding

medium. Denoting the temperature of the object by $T(t)$ and that of the surrounding medium by $T_m(t)$, the law of cooling is expressed as

$$T' = k(T_m(t) - T)$$

where $k > 0$ is a constant. Assume that the $T_m(t) = T_m$ is constant, and the initial temperature of the object is $T(0) = T_0$. Find an expression for $T(t)$ for each of the cases $T_0 < T_m$, $T_0 = T_m$, and $T_0 > T_m$. Are the solutions consistent with our everyday experience?

39. (Application) Consider the electrical circuit shown in Figure 2.11. Let v_s denote the voltage supplied by the source, v_r and v_c denote the voltage drops across the resistor and the capacitor, and i denote the current flowing through the circuit. The behavior of the circuit is determined by the equations

$$C\frac{dv_c}{dt} = i\,, \quad v_r = Ri\,, \quad v_s = v_r + v_c$$

where C is the capacitance and R is the resistance.

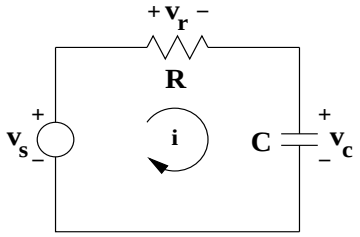

Figure 2.11: An RC circuit.

(a) Eliminate the variables v_r and i from the above equations and obtain a differential equation in the dependent variable v_c.

(b) Let $v_c(0) = 0$. Find $v_c(t)$ for $t \geq 0$ if the supply voltage $v_s(t)$ is a unit step function.

(c) Repeat (b) if $v_s(t) = \cos \omega t, t > 0$. Show that $v_c(t) \to v(t)$ as $t \to \infty$, where v is a periodic function with frequency ω (period $2\pi/\omega$).

40. (Application) Consider the mechanical system shown in Figure 2.12. The force balance on the mass requires that

$$Mx'' + Bx' + Kx = 0$$

where $x(t)$ denotes the displacement of the mass M from the equilibrium position. (Mx'' is the force accelerating the mass, Bx' represents the frictional force, and Kx is the restoring force of the spring.) Dividing the above equation by M we obtain

$$x'' + 2\zeta\omega x' + \omega^2 x = 0$$

where

$$\omega = \sqrt{\frac{K}{M}} \quad \text{and} \quad \zeta = \sqrt{\frac{MB^2}{4K}}$$

are called the natural frequency and the damping ratio of the system, respectively. Suppose that the mass is released at $t = 0$ with an initial displacement $x(0) = x_0$ and initial velocity $x'(0) = 0$. Calculate and plot $x(t)$ for $t \geq 0$ for each of the cases $\zeta = 0$, $0 < \zeta < 1$, $\zeta = 1$, and $1 < \zeta$. Show that when $\zeta = 0$, the system oscillates with natural frequency ω, and that as ζ increases, the oscillations become more and more damped, disappearing completely when $\zeta = 1$. The system is said to be undamped, underdamped, critically damped, and overdamped in the above four cases of ζ.

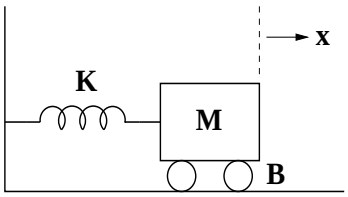

Figure 2.12: A mechanical system.

41. (Application) Rewriting a higher order differential equation in a single dependent variable as an equivalent system is not the only way we come up with a system of first order differential equations. Consider the following problem of formulating the dynamics of two competing populations, say chickens and foxes, that live in a closed environment. Assume that the chickens increase at a rate 0.4 chicken per chicken per unit time in the absence of foxes; but are killed by foxes at a rate 80 chicken per fox per unit time. Foxes, on the other hand, die of hunger at a rate 0.6 fox per fox per unit time if there are no chickens to feed upon; but when there are chickens to eat, they increase at a rate 0.003 fox per chicken per unit time. Let $x_1(t)$ and $x_2(t)$ the number of chickens and foxes at time t. Then the populations of foxes and chickens can be described by a system of two coupled first order differential equations as

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 0.4 & -80 \\ 0.003 & -0.6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix}$$

where x_{10} and x_{20} denote the initial populations of foxes and chicken at $t = 0$.

(a) Use MATLAB command `ode23` to find the solution of the system for $0 \leq t \leq 20$ for each of the following initial populations. In each case, plot $x_1/1000$ and $x_2/10$ on the same graph.

$$\boldsymbol{x}_0 = \begin{bmatrix} 4000 \\ 10 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 8000 \\ 40 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 8000 \\ 50 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 8000 \\ 60 \end{bmatrix}$$

Hint: The reason for plotting $x_1/1000$ and $x_2/10$ rather than x_1 and x_2 is to make the populations comparable so that x_1 does not obscure x_2 when plotted on the same graph. This scaling corresponds to counting chickens in thousands and foxes in tens, and can be done before solving the equations: Let $z_1 = x_1/1000$ and $z_2 = x_2/10$ denote the scaled populations of chickens and foxes. Rewrite the system of equations in z_1 and z_2 and observe that the coefficients in the system become comparable.

(b) Repeat (a) for

$$\boldsymbol{x}_0 = \begin{bmatrix} 9000 \\ 70 \end{bmatrix}$$

How can you modify your model to avoid negative population?

(c) The coefficients of the model are chosen to reflect the delicate balance of nature: If there are enough chickens initially, both populations converge to positive steady state values. Change the coefficient 0.003 to 0.0035 and solve the equation for several initial conditions. Try to interpret the result.

(d) Repeat (c) with the coefficient 0.003 changed to 0.0025.

Chapter 3
Vector Spaces and Linear Transformations

3.1 Vector Spaces

Recall that a vector in the xy plane is a line segment directed from the origin to a point in the plane as shown in Figure 3.1. Recall also that the sum of two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is a vector obtained by the parallelogram rule. Also, for a real number c, $c\boldsymbol{v}$ is a vector whose magnitude is $|c|$ times the magnitude of $\boldsymbol{v}$ and whose direction is the same as the direction of $\boldsymbol{v}$ if $c > 0$, and opposite to the direction of $\boldsymbol{v}$ if $c < 0$.

A convenient way to represent a vector in the xy plane is to consider it as an ordered pair of two real numbers as $\boldsymbol{v} = (\alpha, \beta)$ where α and β are the components of $\boldsymbol{v}$ along the x and y axes. This representation allows us to define the sum of two vectors $\boldsymbol{v}_1 = (\alpha_1, \beta_1)$ and $\boldsymbol{v}_2 = (\alpha_2, \beta_2)$ in terms of their components as

$$\boldsymbol{v}_1 + \boldsymbol{v}_2 = (\alpha_1 + \alpha_2, \beta_1 + \beta_2)$$

and a scalar multiple of a vector $\boldsymbol{v} = (\alpha, \beta)$ as

$$c\boldsymbol{v} = (c\alpha, c\beta)$$

The representation of a vector in the xy plane by a pair also allows us to derive some desirable properties of vector addition and scalar multiplication. For example,

$$\boldsymbol{v}_1 + \boldsymbol{v}_2 = (\alpha_1 + \alpha_2, \beta_1 + \beta_2) = (\alpha_2 + \alpha_1, \beta_2 + \beta_1) = \boldsymbol{v}_2 + \boldsymbol{v}_1$$

and

$$(c + d)\boldsymbol{v} = ((c + d)\alpha, (c + d)\beta) = (c\alpha, c\beta) + (d\alpha, d\beta) = c\boldsymbol{v} + d\boldsymbol{v}$$

Finally, such a representation is useful in expressing a given vector in terms of some special vectors. For example, defining $\boldsymbol{i} = (1, 0)$ and $\boldsymbol{j} = (0, 1)$ to be the unit vectors along the x and y axes, we have

$$\boldsymbol{v} = (\alpha, \beta) = \alpha\,(1, 0) + \beta\,(0, 1) = \alpha\,\boldsymbol{i} + \beta\,\boldsymbol{j}$$

The idea of representing a vector in a plane by an ordered pair can be generalized to vectors in three dimensional xyz space, where we represent a vector $\boldsymbol{v}$ by a triple (α, β, γ), with α, β, and γ corresponding to the components of $\boldsymbol{v}$ along the x, y, and z axes. What about

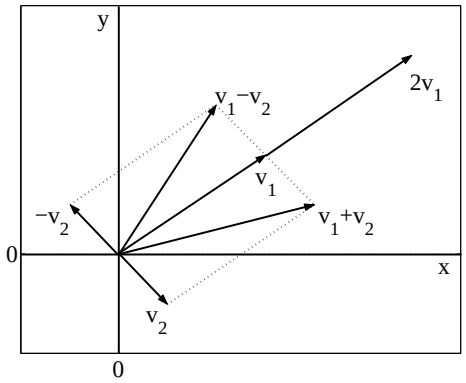

Figure 3.1: Representation of vectors in a plane

a quadruple $(\alpha, \beta, \gamma, \delta)$? Although we cannot visualize it as an arrow in a four dimensional space, we can still define the sum of two such quadruples as well as a scalar multiple of a quadruple in terms of their components. This motivates the need for a more general and abstract definition of a vector.[1]

3.1.1 Definitions

A ***vector space*** $\boldsymbol{\mathcal{X}}$ over a field $\mathbb{F}$ is a non-empty set, elements of which are called vectors, together with two operations called ***addition*** and ***scalar multiplication*** that have the following properties.

Addition operation associates with any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$ a unique vector denoted $\boldsymbol{x} + \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$, and satisfies the following conditions.

A1. $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$.

A2. $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{\mathcal{X}}$.

A3. There exists an element of $\boldsymbol{\mathcal{X}}$, denoted $\mathbf{0}$, such that $\boldsymbol{x} + \mathbf{0} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. $\mathbf{0}$ is called the ***zero vector*** or the ***null vector***.

A4. For any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ there is a vector $-\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ such that $\boldsymbol{x} + (-\boldsymbol{x}) = \mathbf{0}$.

Scalar multiplication operation associates with any vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and any scalar $c \in \mathbb{F}$ a unique vector denoted $c\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and satisfies the following conditions.

S1. $(cd)\boldsymbol{x} = c(d\boldsymbol{x})$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $c, d \in \mathbb{F}$.

S2. $1\boldsymbol{x} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, where 1 is the multiplicative identity of $\mathbb{F}$.

S3. $(c + d)\boldsymbol{x} = c\boldsymbol{x} + d\boldsymbol{x}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $c, d \in \mathbb{F}$.

S4. $c(\boldsymbol{x} + \boldsymbol{y}) = c\boldsymbol{x} + c\boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$ and $c \in \mathbb{F}$.

[1] The reader may be accustomed to defining any directed line segment in the plane, such as an arrow directed from the tip of $\boldsymbol{v}_1$ to the tip of $\boldsymbol{v}_1 + \boldsymbol{v}_2$ in Figure 3.1, as a vector. However, this is just a visual aid, and as far as the addition and scalar multiplication operations just defined are concerned, that arrow is no different from $\boldsymbol{v}_2$. In this sense, $\boldsymbol{v}_2$ represents all arrows that have the same orientation and the same length as $\boldsymbol{v}_2$, which form an equivalence class.

If the field over which a vector space is defined is clear from the context, we omit the phrase "over $\mathbb{F}$" when referring to a vector space.

The following properties of a vector space follow directly from the definition.

a) $\mathbf{0}$ is unique

b) $-\mathbf{0} = \mathbf{0}$

c) $0\,\boldsymbol{x} = \mathbf{0}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

d) $c\mathbf{0} = \mathbf{0}$ for all $c \in \mathbb{F}$

e) $-\boldsymbol{x}$ is unique for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

f) $(-1)\,\boldsymbol{x} = -\boldsymbol{x}$ for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

To prove (a), assume that there are two different vectors $\mathbf{0}_1 \neq \mathbf{0}_2$ satisfying condition A3. Then $\mathbf{0}_2 + \mathbf{0}_1 = \mathbf{0}_2$ (A3 with $\boldsymbol{x} = \mathbf{0}_2$ and $\mathbf{0} = \mathbf{0}_1$), and also $\mathbf{0}_1 + \mathbf{0}_2 = \mathbf{0}_1$ (A3 with $\boldsymbol{x} = \mathbf{0}_1$ and $\mathbf{0} = \mathbf{0}_2$). Then, by A1 we have $\mathbf{0}_2 = \mathbf{0}_1$, contradicting the assumption. Therefore, there can be no two distinct $\mathbf{0}$'s. Other properties can be proved similarly, and are left to the reader as an exercise.

Example 3.1

Consider the set of all ordered n-tuples[2] of the form

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

where $x_1, x_2, \ldots, x_n \in \mathbb{F}$. Defining addition and scalar multiplication operations element-by-element as

$$\begin{aligned}(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) &= (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)\\ c\,(x_1, x_2, \ldots, x_n) &= (cx_1, cx_2, \ldots, cx_n)\end{aligned}$$

and letting

$$\begin{aligned}\mathbf{0} &= (0, 0, \ldots, 0)\\ -(x_1, x_2, \ldots, x_n) &= (-x_1, -x_2, \ldots, -x_n)\end{aligned}$$

all the properties of the vector addition and scalar multiplication are satisfied. Thus the set of all n-tuples of $\mathbb{F}$ is a vector space over $\mathbb{F}$, called the ***n-space*** and denoted $\mathbb{F}^n$. A real n-tuple is $(x_1, x_2, \ldots, x_n)$ is an obvious generalization of the familiar concept of a vector in the plane.

In particular, $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$ can be identified with the real line, the xy plane and the xyz space, respectively.[3]

Example 3.2

The set of $m \times n$ matrices, $\mathbb{F}^{m\times n}$, together with the matrix addition and scalar multiplication operations defined in Section 1.2 is a vector space.[4]

[2] From now on, we will distinguish an ordered set from an unordered set by enclosing its elements with parantheses rather than curly brackets

[3] Note that the set of real numbers is both a field and also a vector space. We distinguish the two by denoting the real field by $\mathbb{R}$ and the vector space of real numbers by $\mathbb{R}^1$.

[4] The reader might ask: "When we multiply two $n \times n$ matrices, are we multiplying two vectors in $\mathbb{F}^{m\times n}$? Can we similarly multiply two vectors in $\mathbb{R}^n$"? The answer is that when we multiply two matrices, we do not view them as vectors, but as something else that we will consider later. Multiplication of vectors is not defined, nor is it needed to construct a vector space.

In particular, $\mathbb{F}^{1\times n}$ and $\mathbb{F}^{n\times 1}$ are vector spaces. This is why we call a row matrix also a row vector, and a column matrix a column vector. In fact, both $\mathbb{F}^{1\times n}$ and $\mathbb{F}^{n\times 1}$ can be identified with $\mathbb{F}^n$ in Example 3.1. In other words, an n-tuple can be viewed as an element of either of the vector spaces $\mathbb{F}^n$, $\mathbb{F}^{1\times n}$ or $\mathbb{F}^{n\times 1}$, in which case it is represented respectively as

$$(x_1, x_2, \ldots, x_n)\,, \quad [\,x_1 \;\; x_2 \;\; \cdots \;\; x_n\,]\,, \quad \text{or} \quad \text{col}\,[\,x_1, x_2, \ldots, x_n\,]$$

* **Example 3.3**

An ordered real n-tuple $(x_1, x_2, \ldots, x_n)$ is a special case of a semi-infinite sequence

$$(x_k)_1^\infty = (x_1, x_2, \ldots)$$

of real numbers. Defining, by analogy to $\mathbb{R}^n$,

$$\begin{aligned} (x_1, x_2, \ldots) + (y_1, y_2, \ldots) &= (x_1 + y_1, x_2 + y_2, \ldots) \\ c\,(x_1, x_2, \ldots) &= (cx_1, cx_2, \ldots) \\ \mathbf{0} &= (0, 0, \ldots) \\ -(x_1, x_2, \ldots) &= (-x_1, -x_2, \ldots) \end{aligned}$$

we observe that the set of all such semi-infinite sequences is a vector space over $\mathbb{R}$.

Similarly, we can extend n-tuples in both directions and consider infinite sequences of the form

$$(x_k)_{-\infty}^{\infty} = (\ldots, x_{-1}, x_0, x_1, \ldots)$$

which form yet another vector space.

* **Example 3.4**

An ordered real n-tuple $(x_1, x_2, \ldots, x_n)$ in Example 3.1 can be viewed as a function $f : \mathbb{N}_n \to \mathbb{R}$, whose domain $\mathbb{N}_n = (\,1, 2, \ldots, n\,)$ is the ordered set of integers from 1 to n, and

$$f[k] = x_k\,, \quad k \in \mathbb{N}_n$$

Similarly, a semi-infinite sequence $(x_k)_1^\infty$ can be viewed as a function whose domain is the set of positive integers $\mathbb{N} = (\,1, 2, \ldots)$, and an infinite sequence $(x_k)_{-\infty}^{\infty}$ as a function whose domain is the set of all integers $\mathbb{Z} = (\,\ldots, -1, 0, 1, \ldots)$.

Consider the set $\mathcal{F}\,(\mathbb{D}, \mathbb{R})$ of all functions $f : \mathbb{D} \to \mathbb{R}$, where $\mathbb{D}$ is any finite or infinite discrete set like $\mathbb{N}_n$, or $\mathbb{N}$, or $\mathbb{Z}$. For $f, g \in \mathcal{F}\,(\mathbb{D}, \mathbb{R})$, we define their sum to be the function $f + g : \mathbb{D} \to \mathbb{R}$ such that

$$(f + g)[k] = f[k] + g[k]\,, \quad k \in \mathbb{D}$$

Likewise, the scalar multiple of f with a scalar c is defined to be the function $cf : \mathbb{D} \to \mathbb{R}$ such that

$$(cf)[k] = cf[k]\,, \quad k \in \mathbb{D}$$

Note that we do nothing new here, but just rephrase the definition of addition and scalar multiplication of n-tuples or sequences using an alternative formulation. We thus reach the conclusion that $\mathcal{F}\,(\mathbb{D}, \mathbb{R})$ is a vector space.

$\mathcal{F}\,(\mathbb{D}, \mathbb{R})$ in Example 3.4 is a typical example of a ***function space***, a vector space whose elements are functions. Other examples of a function space are considered below.

* **Example 3.5**

Consider the set $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ of all real-valued functions $f : \boldsymbol{I} \to \mathbb{R}$ defined on a real interval $\boldsymbol{I}$. For $f, g \in \mathcal{F}(\boldsymbol{I}, \mathbb{R})$ and $c \in \mathbb{R}$, we define the functions $f + g$ and cf pointwise just like we did for $f, g \in \mathcal{F}(\mathbb{D}, \mathbb{R})$:

$$\begin{aligned}(f+g)(t) &= f(t) + g(t)\,, \quad t \in \boldsymbol{I} \\ (cf)(t) &= cf(t)\,, \quad t \in \boldsymbol{I}\end{aligned}$$

The zero function is one with

$$0(t) = 0\,, \quad t \in \boldsymbol{I}$$

and for any $f \in \mathcal{F}(\boldsymbol{I}, \mathbb{R})$, $-f$ is defined pointwise as

$$(-f)(t) = -f(t)\,, \quad t \in \boldsymbol{I}$$

With these definitions, $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ becomes a vector space over $\mathbb{R}$.

The set of all real vector-valued functions $\boldsymbol{f} : \boldsymbol{I} \to \mathbb{R}^{n\times 1}$ is also a vector space over $\mathbb{R}$, denoted $\mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$. A vector-valued function $\boldsymbol{f}$ can also be viewed as a stack of scalar functions as

$$\boldsymbol{f} = \text{col}\,[\, f_1, f_2, \ldots, f_n \,]$$

Note that a function f and its value $f(t)$ at a fixed t are different things. f is a vector, an element of $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$, but $f(t)$ is a scalar, an element of $\mathbb{R}$. This distinction is more apparent in the case of vector-valued functions: If $\boldsymbol{f} \in \mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$ then $\boldsymbol{f}(t) \in \mathbb{R}^{n\times 1}$ for every $t \in \boldsymbol{I}$. Thus, although $\boldsymbol{f}$ and $\boldsymbol{f}(t)$ are both vectors, they are elements of different vector spaces.[5]

Similarly, the set $\mathcal{F}(\boldsymbol{I}, \mathbb{C})$ of all complex-valued functions $f : \boldsymbol{I} \to \mathbb{C}$ and the set $\mathcal{F}(\boldsymbol{I}, \mathbb{C}^{n\times 1})$ of all complex-vector-valued functions $\boldsymbol{f} : \boldsymbol{I} \to \mathbb{C}^{n\times 1}$ defined on a real interval $\boldsymbol{I}$ are vector spaces over $\mathbb{C}$.

3.1.2 Subspaces

A subset $\boldsymbol{\mathcal{U}} \subset \boldsymbol{\mathcal{X}}$ of a vector space is called a ***subspace*** of $\boldsymbol{\mathcal{X}}$ if it is itself a vector space with the same addition and scalar multiplication operations defined on $\boldsymbol{\mathcal{X}}$. To check if a subset is a subspace we need not check all the conditions of a vector space. If $\boldsymbol{\mathcal{U}}$ is a subspace then it must be closed under addition and scalar multiplication. That is, for all $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{\mathcal{U}}$, and $c \in \mathbb{F}$, we must have

$$\boldsymbol{u} + \boldsymbol{v} \in \boldsymbol{\mathcal{U}}\,, \quad c\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$$

Usually these two conditions are combined into a single condition as

$$c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 \in \boldsymbol{\mathcal{U}}$$

for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{\mathcal{U}}$, and $c_1, c_2 \in \mathbb{F}$. Conversely, if $\boldsymbol{\mathcal{U}}$ is closed under vector addition and scalar multiplication, then $-\boldsymbol{u} = (-1)\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$ for all $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$, which in turn implies that $\boldsymbol{u} + (-\boldsymbol{u}) = \boldsymbol{0} \in \boldsymbol{\mathcal{U}}$. Since all other properties of vector addition and scalar multiplication are inherited from $\boldsymbol{\mathcal{X}}$, we conclude that $\boldsymbol{\mathcal{U}} \subset \boldsymbol{\mathcal{X}}$ is a subspace if and only if it is closed under vector addition and scalar multiplication.

[5] Unfortunately, for the lack of an alternative we sometimes use the same notation to denote a function and its value. For example, e^t is used to denote both the exponential function and its value at t.

Example 3.6

Consider the following subset of $\mathbb{R}^3$.

$$\boldsymbol{\mathcal{U}} = \{\, (x, y, x - y) \mid x, y \in \mathbb{R} \,\}$$

For $\boldsymbol{u}_1 = (x_1, y_1, x_1 - y_1), \boldsymbol{u}_2 = (x_2, y_2, x_2 - y_2)$, and $c_1, c_2 \in \mathbb{R}$, we have

$$\begin{aligned} & c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 \\ &= (c_1x_1, c_1y_1, c_1x_1 - c_1y_1) + (c_2x_2, c_2y_2, c_2x_2 - c_2y_2) \\ &= ((c_1x_1 + c_2x_2), (c_1y_1 + c_2y_2), (c_1x_1 + c_2x_2) - (c_1y_1 + c_2y_2)) \in \boldsymbol{\mathcal{U}} \end{aligned}$$

Thus $\boldsymbol{\mathcal{U}}$ is a subspace of $\mathbb{R}^3$. It is the set of all points $(x, y, z) \in \mathbb{R}^3$ that satisfy

$$x - y - z = 0$$

This is the equation of a plane through the origin $\mathbf{0} = (0, 0, 0)$. In $\mathbb{R}^3$, a plane through the origin is represented as the set of all points that satisfy

$$px + qy + rz = 0$$

for some p, q, r, not all zero. It is left to the reader to show that any such plane defines a subspace of $\mathbb{R}^3$. In particular, the equation $x = 0$ defines the yz plane, $y = 0$ the xz plane, and $z = 0$ the xy plane.

Now, consider the set of all points (x, y, z) that satisfy

$$\begin{bmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since each equation above defines a plane through the origin, the points satisfying the above system are on the intersection of these two planes. If (p_1, q_1, r_1) and (p_2, q_2, r_2) are not proportional, then the two equations define distinct planes, and so their intersection is a straight line through the origin. Since the set of solutions of the above system is closed under addition and scalar multiplication, we conclude that any straight line through the origin is also a subspace of $\mathbb{R}^3$.

As an illustration, the first of the equations

$$\begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

describes the subspace $\boldsymbol{\mathcal{U}}$ considered above, and the second describes the subspace

$$\boldsymbol{\mathcal{V}} = \{\, (x, y, -x) \mid x, y \in \mathbb{R} \,\}$$

Their intersection, which is the common solution of these equations, is the straight line described as

$$\boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}} = \{\, (x, 2x, -x) \mid x \in \mathbb{R} \,\}$$

Clearly, a plane or a line not passing through the origin is not a subspace, simply because it does not include the zero vector of $\mathbb{R}^3$.

* **Example 3.7**

The set of polynomials of a complex variable s with complex coefficients is a vector space over $\mathbb{C}$, denoted $\mathbb{P}_C[s]$. The subset $\mathbb{P}_C^n[s]$, consisting of all polynomials with degree less than or equal to n, is a subspace of $\mathbb{P}_C[s]$. However, the set of polynomials with degree equal exactly to n is not a vector space. (Why?)

The set of polynomials in a real variable t with real coefficients is also a vector space, denoted $\mathbb{P}_R[t]$. Clearly, $\mathbb{P}_R[t]$ is a vector space over $\mathbb{R}$.

* **Example 3.8**

Let $\mathcal{C}_m(\boldsymbol{I}, \mathbb{R})$ denote the set of all real-valued functions defined on some real interval $\boldsymbol{I}$ such that $f, f', \ldots, f^{(m)}$ all exist and are continuous on $\boldsymbol{I}$. That is, $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$ is the set of continuous functions, $\mathcal{C}_1(\boldsymbol{I}, \mathbb{R})$ is the set of differentiable functions with a continuous derivative, etc. Also, let $\mathcal{C}_\infty(\boldsymbol{I}, \mathbb{R})$ denote the set of functions that have continuous derivatives of every order. By definition

$$\mathcal{F}(\boldsymbol{I}, \mathbb{R}) \supset \mathcal{C}_0(\boldsymbol{I}, \mathbb{R}) \supset \mathcal{C}_1(\boldsymbol{I}, \mathbb{R}) \supset \cdots \supset \mathcal{C}_\infty(\boldsymbol{I}, \mathbb{R})$$

Each of these sets is closed under the addition and scalar multiplication operations defined for the function space $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ in Example 3.4, and therefore, is a subspace of $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$. The subspaces $\mathcal{C}_m(\boldsymbol{I}, \mathbb{R}^{n\times 1}) \subset \mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$ can be defined similarly as

$$\mathcal{C}_m(\boldsymbol{I}, \mathbb{R}^{n\times 1}) = \{\, \boldsymbol{f} = \text{col}\,[\, f_1, f_2, \ldots, f_n\,] \mid f_i \in \mathcal{C}_m(\boldsymbol{I}, \mathbb{R}),\ i = 1, \ldots, n\,\}$$

3.2 Span and Linear Independence

3.2.1 Span

Let $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k\}$ be a finite subset of a vector space $\boldsymbol{\mathcal{X}}$. An expression of the form

$$c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + \cdots + c_k\boldsymbol{r}_k$$

where $c_1, c_2, \ldots, c_k \in \mathbb{F}$, is called a ***linear combination*** of $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k$. Because of property A2 of vector addition, a linear combination unambiguously defines a vector in $\boldsymbol{\mathcal{X}}$. The set of all linear combinations of $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k$ is called the ***span*** of $\boldsymbol{R}$, denoted span $(\boldsymbol{R})$ or span $(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k)$. Thus

$$\text{span}\,(\boldsymbol{R}) = \{\, c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + \cdots + c_k\boldsymbol{r}_k \mid c_1, c_2, \ldots, c_k \in \mathbb{F}\,\}$$

If span $(\boldsymbol{R}) = \boldsymbol{\mathcal{X}}$, then $\boldsymbol{R}$ is called a ***spanning set***.

The definition of span can be extended to infinite sets. The span of an infinite set of vectors is defined to be the set of all finite linear combinations of vectors of $\boldsymbol{R}$. More precisely,

$$\text{span}\,(\boldsymbol{R}) = \{\, \sum_{i\in\mathbb{I}} c_i\boldsymbol{r}_i \mid \mathbb{I} \text{ is a finite index set}, c_i \in \mathbb{F}, \boldsymbol{r}_i \in \boldsymbol{R}\,\}$$

If $\boldsymbol{u}, \boldsymbol{v} \in$ span $(\boldsymbol{R})$, then $\boldsymbol{u} = \sum a_i\boldsymbol{r}_i$ and $\boldsymbol{v} = \sum b_i\boldsymbol{r}_i$ for some $a_i, b_i \in \mathbb{F}$. Then

$$c\boldsymbol{u} + d\boldsymbol{v} = \sum (ca_i + db_i)\boldsymbol{r}_i \in \text{span}\,(\boldsymbol{R})$$

for any $c, d \in \mathbb{F}$. This shows that span $(\boldsymbol{R})$ is a subspace of $\boldsymbol{\mathcal{X}}$. In fact, it is the smallest subspace that contains all the vectors in $\boldsymbol{R}$.

Example 3.9

Let $\boldsymbol{i} = (1, 0)$ and $\boldsymbol{j} = (0, 1)$ denote the unit vectors along the x and y axes of the xy plane ($\mathbb{R}^2$). Then

$$\text{span}\,(\,\boldsymbol{i}\,) = \{\,(\alpha, 0)\,|\,\alpha \in \mathbb{R}\}$$

and

$$\text{span}\,(\,\boldsymbol{j}\,) = \{\,(0, \beta)\,|\,\beta \in \mathbb{R}\}$$

are the x and y axes, and

$$\text{span}\,(\,\boldsymbol{i}, \boldsymbol{j}\,) = \{\,(\alpha, \beta)\,|\,\alpha, \beta \in \mathbb{R}\}$$

is the whole xy plane.

Example 3.10

In $\mathbb{R}^3$, let

$$\boldsymbol{r}_1 = (0, 0, 1)\,, \quad \boldsymbol{r}_2 = (0, 1, -1)\,, \quad \boldsymbol{r}_3 = (1, 0, 1)\,, \quad \boldsymbol{r}_4 = (1, -1, 2)$$

Then

a) Span of each of the vectors is a straight line through the origin on which that vector lies. For example, span $(\boldsymbol{r}_1) = \{\,(0, 0, c)\mid c \in \mathbb{R}\,\}$, which is the z axis. Since the given vectors are different, each spans a different line.

b) Any two of the given vectors span a plane through the origin that contain those two vectors. For example,

$$\text{span}\,(\boldsymbol{r}_2, \boldsymbol{r}_3) = \{\,(a, b, a - b)\,|\,a, b \in \mathbb{R}\,\}$$

which is the subspace $\boldsymbol{\mathcal{U}}$ in Example 3.6.

c) span $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) = \mathbb{R}^3$, because by definition span $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3) \subset \mathbb{R}^3$, and for any $\boldsymbol{x} = (a, b, c) \in \mathbb{R}^3$

$$\boldsymbol{x} = (b + c - a)\,\boldsymbol{r}_1 + b\,\boldsymbol{r}_2 + a\,\boldsymbol{r}_3 \in \text{span}\,(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$$

so that $\mathbb{R}^3 \subset$ span $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$ also. Similarly, span $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_4) =$ span $(\boldsymbol{r}_1, \boldsymbol{r}_3, \boldsymbol{r}_4) = \mathbb{R}^3$.

d) However, span $(\boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4) =$ span $(\boldsymbol{r}_2, \boldsymbol{r}_3) =$ span $(\boldsymbol{r}_2, \boldsymbol{r}_4) =$ span $(\boldsymbol{r}_3, \boldsymbol{r}_4) = \boldsymbol{\mathcal{U}}$.

e) Finally, span $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4) = \mathbb{R}^3$, simply because

$$\text{span}\,(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4) \supset \text{span}\,(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$$

3.2.2 Linear Independence

A finite set of vectors $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k\}$ is said to be ***linearly independent*** if

$$c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + \cdots + c_k\boldsymbol{r}_k = \boldsymbol{0} \tag{3.1}$$

holds only when $c_1 = c_2 = \cdots = c_k = 0$.

A set is said to be ***linearly dependent*** if it is not linearly independent. Alternatively, a finite set $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k\}$ is linearly dependent if there exist $c_1, \ldots, c_k$, not all 0, that satisfy (3.1).[6]

By definition, a set containing only a single vector $\boldsymbol{r}$ is linearly independent if and only if $\boldsymbol{r} \neq \boldsymbol{0}$.

We have the following results concerning linear independence.

a) If $\boldsymbol{0} \in \boldsymbol{R}$ then $\boldsymbol{R}$ is linearly dependent.

b) If $\boldsymbol{R}$ is linearly independent and $\boldsymbol{S} \subset \boldsymbol{R}$, then $\boldsymbol{S}$ is also linearly independent. Equivalently, if $\boldsymbol{R}$ is linearly dependent and $\boldsymbol{S} \supset \boldsymbol{R}$, then $\boldsymbol{S}$ is also linearly dependent.

c) $\boldsymbol{R}$ is linearly dependent if and only if at least one vector in $\boldsymbol{R}$ can be written as a linear combination of some other vectors in $\boldsymbol{R}$ (assuming, of course, that $\boldsymbol{R}$ contains at least two vectors).

The rest being direct consequences of the definitions, only the necessity part of the last result requires a proof. If $\boldsymbol{R}$ is linearly dependent then there exist $c_1, \ldots, c_k$, not all 0, such that

$$c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + \cdots + c_k\boldsymbol{r}_k = \boldsymbol{0}$$

Suppose $c_p \neq 0$. Then

$$\boldsymbol{r}_p = \sum_{q \neq p} (-c_q/c_p)\boldsymbol{r}_q$$

Property (b) above can be used to define linear dependence and independence of infinite sets. An infinite set is said to be linearly independent if every finite subset of it is linearly independent, and linearly dependent if it has a linearly dependent finite subset.

Example 3.11

The vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ in Example 3.9 are linearly independent, because

$$\boldsymbol{0} = \alpha\boldsymbol{i} + \beta\boldsymbol{j} = (\alpha, \beta) \implies \alpha = \beta = 0$$

Example 3.12

Consider the vectors in Example 3.10. The set $\boldsymbol{R}_1 = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$ is linearly independent, because

$$\boldsymbol{0} = c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + c_3\boldsymbol{r}_3 = (c_3, c_2, c_1 - c_2 + c_3)$$

implies

$$c_1 = c_2 = c_3 = 0$$

Similarly, the sets $\boldsymbol{R}_2 = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_4\}$ and $\boldsymbol{R}_3 = \{\boldsymbol{r}_1, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ are linearly independent. Therefore, all subsets of these three sets, which include all singletons and pairs of $\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3$ and $\boldsymbol{r}_4$, are also linearly independent.

However, the set $\boldsymbol{R}_4 = \{\boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ is linearly dependent, because

$$\boldsymbol{r}_2 - \boldsymbol{r}_3 + \boldsymbol{r}_4 = (0, 0, 0) = \boldsymbol{0}$$

Therefore, $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ is also linearly dependent.

[6] We also say that the vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k$ are linearly independent (dependent) to mean that the set consisting of these vectors is linearly independent (dependent).

Example 3.13

In $\mathbb{C}^3$, let

$$\boldsymbol{x}_1 = (1, i, 0)\,, \quad \boldsymbol{x}_2 = (i, 0, 1)\,, \quad \boldsymbol{x}_3 = (0, 1, 1)$$

Then $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$ is linearly independent, because for $c_1 = a_1 + ib_1$ and $c_2 = a_2 + ib_2$,

$$\begin{aligned} \mathbf{0} &= c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 \\ &= (a_1 + ib_1, -b_1 + ia_1, 0) + (-b_2 + ia_2, 0, a_2 + ib_2) \\ &= ((a_1 - b_2) + i(b_1 + a_2), -b_1 + ia_1, a_2 + ib_2)) \end{aligned}$$

implies $a_1 = b_1 = a_2 = b_2 = 0$, or equivalently, $c_1 = c_2 = 0$.

Similarly, the sets $\{\boldsymbol{x}_1, \boldsymbol{x}_3\}$ and $\{\boldsymbol{x}_2, \boldsymbol{x}_3\}$ are linearly independent. On the other hand, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$ is linearly dependent, because

$$i\,\boldsymbol{x}_1 - \boldsymbol{x}_2 + \boldsymbol{x}_3 = \mathbf{0}$$

* **Example 3.14**

A set of functions $f_1, \ldots, f_k \in \mathcal{F}(\boldsymbol{I}, \mathbb{F})$ is linearly dependent if

$$c_1 f_1 + \cdots + c_k f_k = 0$$

for some scalars $c_1, \ldots, c_k \in \mathbb{F}$, not all 0. This is a functional equality, which is equivalent to

$$c_1 f_1(t) + \cdots + c_k f_k(t) = 0 \quad \text{for all} \;\; t \in \boldsymbol{I} \tag{3.2}$$

Consider the real-valued functions $\phi_1(t) = e^{\sigma_1 t}$ and $\phi_2(t) = e^{\sigma_2 t}$, where $\sigma_1 \neq \sigma_2 \in \mathbb{R}$. Unless $c_1 = c_2 = 0$, the equality

$$c_1 e^{\sigma_1 t} + c_2 e^{\sigma_2 t} = 0$$

can be satisfied for at most a single value of t (the graphs of $c_1 e^{\sigma_1 t}$ and $-c_2 e^{\sigma_2 t}$ either do not intersect, or intersect at a single point). Therefore, ϕ_1 and ϕ_2 are linearly independent on any interval $\boldsymbol{I}$.

Now consider two complex-valued functions $\psi_1(t) = e^{\lambda_1 t}$ and $\psi_2(t) = e^{\lambda_2 t}$, where $\lambda_1 \neq \lambda_2 \in \mathbb{C}$. The graphical argument above is of no use for we cannot plot graphs of complex-valued functions, and we need an algebraic method to test linear independence of $\psi_1(t)$ and $\psi_2(t)$. Such a method is based on the observation that if

$$c_1\psi_1(t) + c_2\psi_2(t) = 0 \quad \text{for all} \;\; t \in \boldsymbol{I}$$

then

$$c_1\psi_1'(t) + c_2\psi_2'(t) = 0 \quad \text{for all} \;\; t \in \boldsymbol{I}$$

provided ψ_1 and ψ_2 are differentiable on $\boldsymbol{I}$. For the given ψ_1 and ψ_2, which are differentiable everywhere, these two equations can be written in matrix form as

$$\begin{bmatrix} e^{\lambda_1 t} & e^{\lambda_2 t} \\ \lambda_1 e^{\lambda_1 t} & \lambda_2 e^{\lambda_2 t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

A simple elementary operation reduces the system to

$$\begin{bmatrix} e^{\lambda_1 t} & e^{\lambda_2 t} \\ 0 & (\lambda_2 - \lambda_1) e^{\lambda_2 t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since $\lambda_2 - \lambda_1 \neq 0$ and $e^{\lambda_2 t} \neq 0$ for all t, the second equation gives $c_2 = 0$. Similarly, since $e^{\lambda_1 t} \neq 0$ for all t, the first equation gives $c_1 = 0$. Hence ψ_1 and ψ_2 too are linearly independent on any interval $\boldsymbol{I}$.

Using the same technique we can show that the real-valued functions $\xi_1(t) = e^{\sigma t}$ and $\xi_2(t) = te^{\sigma t}$ are also linearly independent.

Note that the function pairs in the above three cases are solutions of a second order linear differential equation with constant coefficients whose characteristic polynomial has either the real roots $s_{1,2} = \sigma_{1,2}$ or the complex conjugate roots $s_{1,2} = \lambda_{1,2} = \sigma \mp i\omega$ or a double real root $s = \sigma$. In each case, the corresponding solutions are linearly independent either as elements of $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ or as elements of $\mathcal{F}(\boldsymbol{I}, \mathbb{C})$.

* **Example 3.15**

If $f_j = g_j + ih_j, f_j^* = g_j - ih_j, j = 1, \ldots, k$, are $2k$ linearly independent functions in $\mathcal{F}(\boldsymbol{I}, \mathbb{C})$, then their real and imaginary parts, $g_j, h_j, j = 1, \ldots, k$, are linearly independent in $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$. To show this, suppose that

$$\sum_{j=1}^{k} (a_j g_j + b_j h_j) = 0$$

Noting that

$$g_j = \frac{1}{2}(f_j + f_j^*) \quad \text{and} \quad h_j = \frac{1}{2i}(f_j - f_j^*)$$

the above expression becomes

$$\frac{1}{2}\sum_{j=1}^{k} (c_j f_j + c_j^* f_j^*) = 0$$

where $c_j = a_j - ib_j$. Linear independence of $\{f_j, f_j^* \mid j = 1, \ldots, k\}$ implies $c_j = 0, j = 1, \ldots, k$, and therefore, $a_j = b_j = 0, j = 1, \ldots, k$.

Observe that this example explains why the real and imaginary parts of complex solutions of second order linear differential equation with constant coefficients, whose characteristic polynomial has a pair of complex-conjugate roots, are linearly independent real solutions.

3.2.3 Elementary Operations

Consider an $m \times n$ matrix A partitioned into its rows

$$A = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{bmatrix} \in \mathbb{F}^{m \times n}$$

Since the rows of A are vectors in $\mathbb{F}^{1 \times n}$ (as noted in Example 3.2), the elementary row operations on A discussed in Section 1.4 can be viewed as operations involving the elements of the ordered set $\boldsymbol{R} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m) \subset \mathbb{F}^{1 \times n}$. This observation suggests that similar operations can be defined for any ordered subset of a vector space.

The following operations on a finite ordered set of vectors $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k)$ are called ***elementary operations***.

I: Interchange any two vectors

II: Multiply any vector by a nonzero scalar

III: Add a scalar multiple of a vector to another one

As we discussed in connection with elementary row operations, to every elementary operation there corresponds an inverse operation of the same type such that if $\boldsymbol{R}'$ is obtained from $\boldsymbol{R}$ by a single elementary operation, then $\boldsymbol{R}$ can be recovered from $\boldsymbol{R}'$ by performing the inverse operation.

Let $\boldsymbol{R}'$ be obtained from $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k)$ by a single elementary operation. If it is a Type I or Type II operation, then it is clear that span $(\boldsymbol{R}') =$ span $(\boldsymbol{R})$. Suppose it is a Type III operation that consists of adding α times $\boldsymbol{r}_p$ to $\boldsymbol{r}_q$ for some $p \neq q$. That is,

$$\boldsymbol{r}'_i = \begin{cases} \boldsymbol{r}_i \quad , & i \neq q \\ \boldsymbol{r}_q + \alpha \boldsymbol{r}_p \, , & i = q \end{cases} \tag{3.3}$$

For an arbitrary $\boldsymbol{x} \in$ span $(\boldsymbol{R}')$

$$\begin{aligned} \boldsymbol{x} &= c_1 \boldsymbol{r}'_1 + \cdots + c_p \boldsymbol{r}'_p + \cdots + c_q \boldsymbol{r}'_q + \cdots + c_k \boldsymbol{r}'_k \\ &= c_1 \boldsymbol{r}_1 + \cdots + c_p \boldsymbol{r}_p + \cdots + c_q (\boldsymbol{r}_q + \alpha \boldsymbol{r}_p) + \cdots + c_k \boldsymbol{r}_k \\ &= c_1 \boldsymbol{r}_1 + \cdots + (c_p + \alpha c_q) \boldsymbol{r}_p + \cdots + c_q \boldsymbol{r}_q + \cdots + c_k \boldsymbol{r}_k \end{aligned} \tag{3.4}$$

so that $\boldsymbol{x} \in$ span $(\boldsymbol{R})$. Hence, span $(\boldsymbol{R}') \subset$ span $(\boldsymbol{R})$. Considering the inverse elementary operation, it can similarly be shown that span $(\boldsymbol{R}) \subset$ span $(\boldsymbol{R}')$. Hence span $(\boldsymbol{R}') =$ span $(\boldsymbol{R})$. Obviously, this property also holds if $\boldsymbol{R}'$ is obtained from $\boldsymbol{R}$ by a finite sequence of elementary operations.

Another property of elementary operations is the preservation of linear independence: If $\boldsymbol{R}'$ is obtained from $\boldsymbol{R}$ by a finite sequence of elementary operations, then $\boldsymbol{R}'$ is linearly independent if and only if $\boldsymbol{R}$ is linearly independent. Again, the proof is trivial if $\boldsymbol{R}'$ is obtained from $\boldsymbol{R}$ by a single Type I or Type II elementary operation. Suppose that $\boldsymbol{R}'$ is obtained from $\boldsymbol{R}$ by a single Type III elementary operation as described in (3.3), and consider a linear combination as in (3.4) with $\boldsymbol{x} = \boldsymbol{0}$. If $\boldsymbol{R}$ is linearly independent then all the coefficients in the last linear combination in (3.4) must be zero, which implies that all c_i's are zero, so that $\boldsymbol{R}'$ is also linearly independent. By considering the inverse elementary operation the converse can also be shown to be true.

These properties of elementary operations can be used to characterize the span of a set or to check its linear independence as illustrated by the following example.

Example 3.16

Consider the set $\boldsymbol{R}_1 = (\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$ in Example 3.12. Identifying $\boldsymbol{r}_1, \boldsymbol{r}_2$ and $\boldsymbol{r}_3$ with the rows of the matrix

$$R_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 0 & 1 \end{bmatrix}$$

and performing elementary row operations on R_1, we observe that

$$R_1 \longrightarrow I$$

Thus

$$\boldsymbol{R}_1 \longrightarrow \boldsymbol{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$$

where

$$\boldsymbol{e}_1 = (1, 0, 0)\,, \quad \boldsymbol{e}_2 = (0, 1, 0)\,, \quad \boldsymbol{e}_3 = (0, 0, 1)$$

correspond to rows of I. Since $\boldsymbol{E}$ is linearly independent then so is $\boldsymbol{R}_1$. Also

$$\text{span}\,(\boldsymbol{R}_1) = \text{span}\,(\boldsymbol{E}) = \{\,(x, y, z)\,|\,x, y, z \in \mathbb{R}\,\} = \mathbb{R}^3$$

Now consider the set $\boldsymbol{R}_4 = (\boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4)$. Performing elementary operations on $\boldsymbol{R}_4$ as above, we obtain

$$\boldsymbol{R}_4 \longrightarrow \boldsymbol{S} = (\boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{0})$$

Then

$$\text{span}\,(\boldsymbol{R}_4) = \text{span}\,(\boldsymbol{S}) = \text{span}\,(\boldsymbol{r}_2, \boldsymbol{r}_3) = \boldsymbol{\mathcal{U}}$$

Also, since $\boldsymbol{S}$ is linearly dependent then so is $\boldsymbol{R}_4$.

Note that these results have already been obtained in Examples 3.10 and 3.12.

3.3 Bases and Representations

If $\boldsymbol{R}$ is a spanning set then any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ can be expressed as a linear combination of vectors in $\boldsymbol{R}$. A significant question is whether we need all the vectors in $\boldsymbol{R}$ to be able to do that for every $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. For example, the set $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ in Example 3.10 spans $\mathbb{R}^3$, but so also do $\boldsymbol{R}_1 = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$, $\boldsymbol{R}_2 = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_4\}$, and $\boldsymbol{R}_3 = \{\boldsymbol{r}_1, \boldsymbol{r}_3, \boldsymbol{r}_4\}$. That is, any one of $\boldsymbol{r}_2$, $\boldsymbol{r}_3$, or $\boldsymbol{r}_4$ can be removed from $\boldsymbol{R}$ without losing the spanning property. On the other hand, if $\boldsymbol{r}_1$ is removed from $\boldsymbol{R}$, then the resulting set $\boldsymbol{R}_4 = \{\boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ no longer spans $\mathbb{R}^3$. Apparently, the sets $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ have a property that $\boldsymbol{R}_4$ does not have. Referring to Example 3.12 we find out that the first three sets are linearly independent while the last is not, and this may be a clue.

Lets take another look at one of those linearly independent spanning sets, say $\boldsymbol{R}_1$. If we remove one more vector from $\boldsymbol{R}_1$, then the resulting set of two vectors will only span a plane (a subspace), not the whole space. Thus $\boldsymbol{R}_1$ is a minimal spanning set. On the other hand, if we add any vector $\boldsymbol{r}$ different from $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, and $\boldsymbol{r}_3$ to $\boldsymbol{R}_1$, then the resulting set $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}\}$ will no longer be linearly independent, because $\boldsymbol{r}$ can be expressed as a linear combination of the others. Hence $\boldsymbol{R}_1$ is also a maximal linearly independent set. The same are true also for $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$.

These observations motivate a need to investigate the link between the concepts of span and linear independence.

3.3.1 Basis

The following theorem, which is one of the fundamental results of linear algebra, characterizes a linearly independent spanning set.

Theorem 3.1 *Let $\boldsymbol{R}$ be an ordered subset of a vector space $\boldsymbol{\mathcal{X}}$. Then the following are equivalent.*

a) $\boldsymbol{R}$ is linearly independent and spans $\boldsymbol{\mathcal{X}}$.

b) $\boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$, and no proper subset of $\boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$. (That is, $\boldsymbol{R}$ is a minimal spanning set.)

c) $\boldsymbol{R}$ is linearly independent, and no proper superset of $\boldsymbol{R}$ is linearly independent. (That is, $\boldsymbol{R}$ is a maximal linearly independent set.)

d) Every vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ can be expressed as a linear combination of the vectors of $\boldsymbol{R}$ in a unique way. That is,

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{r}_i$$

for some $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k \in \boldsymbol{R}$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{F}$, all of which are uniquely determined by $\boldsymbol{x}$.

Proof We will show that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a).

(a) $\Rightarrow$ (b):

By the second part of the hypothesis $\boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$. If a proper subset $\boldsymbol{S} \subset \boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$, then there exists a vector $\boldsymbol{r} \in \boldsymbol{R} - \boldsymbol{S}$ that can be written as a linear combination of some vectors in $\boldsymbol{S}$. This implies that $\boldsymbol{R}$ is linearly dependent, contradicting the first part of the hypothesis. Hence no proper subset of $\boldsymbol{R}$ can span $\boldsymbol{\mathcal{X}}$.

(b) $\Rightarrow$ (c):

If $\boldsymbol{R}$ is linearly dependent, then there exists $\boldsymbol{r} \in \boldsymbol{R}$ which can be written as a linear combination of some other vectors in $\boldsymbol{R}$. This implies that $\boldsymbol{R} - \{\boldsymbol{r}\}$ also spans $\boldsymbol{\mathcal{X}}$, contradicting the second part of the hypothesis. Hence $\boldsymbol{R}$ is linearly independent. On the other hand, since every vector $\boldsymbol{x} \notin \boldsymbol{R}$ can be written as a linear combination of vectors of $\boldsymbol{R}$ (because $\boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$), no proper superset of $\boldsymbol{R}$ can be linearly independent.

(c) $\Rightarrow$ (d):

If there exists a nonzero vector $\boldsymbol{x}$ which cannot be expressed as a linear combination of vectors in $\boldsymbol{R}$, then $\boldsymbol{R} \cup \{\boldsymbol{x}\}$ is linearly independent (see Exercise 3.13), contradicting the second part of the hypothesis. Hence every vector can be expressed in terms of the vectors of $\boldsymbol{R}$. Now if a vector $\boldsymbol{x}$ can be expressed as two different linear combinations of the vectors of $\boldsymbol{R}$ as

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{r}_i = \sum_{i=1}^{k} \beta_i \boldsymbol{r}_i$$

then

$$\sum_{i=1}^{k} (\alpha_i - \beta_i) \boldsymbol{r}_i = \boldsymbol{0}$$

where at least one coefficient $\alpha_i - \beta_i$ is nonzero. This means that $\boldsymbol{R}$ is linearly dependent, contradicting the first part of the hypothesis. Hence the expression for $\boldsymbol{x}$ in terms of the vectors of $\boldsymbol{R}$ is unique.

(d) $\Rightarrow$ (a):

By hypothesis $\boldsymbol{R}$ spans $\boldsymbol{\mathcal{X}}$. If $\boldsymbol{R}$ is linearly dependent, then there exists a vector $\boldsymbol{r} \in \boldsymbol{R}$ which can be expressed as

$$\boldsymbol{r} = \sum_{i=1}^{k} c_i \boldsymbol{r}_i$$

for some $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k \in \boldsymbol{R}$, which means that $\boldsymbol{r}$ has two different expressions in terms of $\boldsymbol{r}, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_k \in \boldsymbol{R}$, contradicting the hypothesis. Hence $\boldsymbol{R}$ is also linearly independent.

A set $\boldsymbol{R}$ having the properties in Theorem 3.1 is called a ***basis*** for $\boldsymbol{\mathcal{X}}$. A basis is a generalization of the concept of a coordinate system in a plane to abstract vector spaces. Consider the vectors $\boldsymbol{i} = (1, 0)$ and $\boldsymbol{j} = (0, 1)$ along the x and y axes of the xy plane ($\mathbb{R}^2$). Since any vector $\boldsymbol{x} = (\alpha, \beta)$ has a unique representation as $\boldsymbol{x} = \alpha\boldsymbol{i} + \beta\boldsymbol{j}$, the vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ form a basis for the xy plane. The basis vectors of a vector space play exactly the same role as do the vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ in the xy plane.

Example 3.17

Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ denote columns of I_3. The set $\boldsymbol{E} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ spans $\mathbb{R}^{3\times 1}$, because any $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$ can be expressed as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + x_3\boldsymbol{e}_3 \tag{3.5}$$

$\boldsymbol{E}$ is also linearly independent, because a linear combination of $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ as above is $\boldsymbol{0}$ only if all the coefficients x_1, x_2 and x_3 are zero. Hence $\boldsymbol{E}$ is a basis for $\mathbb{R}^{3\times 1}$, called the ***canonical basis***. Canonical bases for $\mathbb{F}^n$, $\mathbb{F}^{1\times n}$ and $\mathbb{F}^{n\times 1}$ can be defined similarly. For example, the set $\boldsymbol{E}$ in Example 3.16 is the canonical basis for $\mathbb{R}^3$.

We now claim that the set $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$, where

$$\boldsymbol{r}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{r}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{r}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

is also a basis for $\mathbb{R}^{3\times 1}$. To check if an arbitrary vector $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$ can be expressed in terms of the vectors in $\boldsymbol{R}$ we try to solve

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

for α_1, α_2 and α_3. Since the coefficient matrix of the above equation in already in a row echelon form, we obtain a unique solution by back substitution as

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} x_1 - x_2 \\ x_2 - x_3 \\ x_3 \end{bmatrix}$$

Thus

$$\boldsymbol{x} = (x_1 - x_2)\boldsymbol{r}_1 + (x_2 - x_3)\boldsymbol{r}_2 + x_3\boldsymbol{r}_3 \tag{3.6}$$

which shows that $\boldsymbol{R}$ spans $\mathbb{R}^{3\times 1}$. Moreover, since the coefficients of $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$ in the above expression are uniquely determined by $\boldsymbol{x}$, $\boldsymbol{R}$ must also be linearly independent. Indeed,

$$c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + c_3\boldsymbol{r}_3 = \begin{bmatrix} c_1 + c_2 + c_3 \\ c_2 + c_3 \\ c_3 \end{bmatrix} = \boldsymbol{0}$$

implies $c_1 = c_2 = c_3 = 0$. This proves our claim that $\boldsymbol{R}$ is also a basis for $\mathbb{R}^{3\times 1}$.

* **Example 3.18**

In the vector space $\mathbb{P}_C[s]$ of polynomials, let $\boldsymbol{Q} = (q_0, q_1, \ldots)$, where the polynomials q_i are defined as

$$q_i(s) = s^i\,, \quad i = 0, 1, \ldots$$

Since any polynomial $p(s) = c_0 + c_1 s + \cdots + c_n s^n$ can be expressed as

$$p = c_0 q_0 + c_1 q_1 + \cdots + c_n q_n$$

$\boldsymbol{Q}$ spans $\mathbb{P}_C[s]$.

Consider the finite subset $\boldsymbol{Q}_n = \{q_0, q_1, \ldots, q_n\}$ of $\boldsymbol{Q}$, and let p be a linear combination of $q_0, q_1, \ldots, q_n$ expressed as above. If $p = 0$ then $p(s) = p'(s) = p''(s) = \cdots = 0$ for all s. Evaluating at $s = 0$, we get $c_1 = c_2 = \cdots = c_n = 0$, which shows that $\boldsymbol{Q}_n$ is linearly independent. Since any finite subset of $\boldsymbol{Q}$ is a subset of $\boldsymbol{Q}_n$ for some n, it follows that every finite subset of $\boldsymbol{Q}$ is linearly independent. Hence $\boldsymbol{Q}$ is linearly independent, and therefore, it is a basis for $\mathbb{P}_C[s]$.

The reader can show that the set $\boldsymbol{R} = \{r_0, r_1, \ldots\}$, where

$$r_i(s) = 1 + s + \cdots + s^i\,, \quad i = 0, 1, \ldots$$

is also a basis for $\mathbb{P}_C[s]$. In fact, $\boldsymbol{R}$ can be obtained from $\boldsymbol{Q}$ by a sequence of elementary operations.[7]

From the two examples above we observe that a vector space may have a finite or an infinite basis. A vector space with a finite basis is said to be ***finite dimensional***, otherwise, ***infinite dimensional***. Thus $\mathbb{R}^{3\times 1}$ in Example 3.17 is finite dimensional, and $\mathbb{C}[s]$ in Example 3.18 is infinite dimensional. These examples also illustrate that basis for a vector space is not unique.

The following corollary of Theorem 3.1 characterizes bases of a finite dimensional vector space.

Corollary 3.1.1 *Let $\boldsymbol{\mathcal{X}}$ have a finite basis $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n)$. Then*

a) No subset of $\boldsymbol{\mathcal{X}}$ containing more than n vectors is linearly independent.

b) No subset of $\boldsymbol{\mathcal{X}}$ containing less than n vectors spans $\boldsymbol{\mathcal{X}}$.

c) Any basis of $\boldsymbol{\mathcal{X}}$ contains exactly n vectors.

d) Any linearly independent set that contains exactly n vectors is a basis.

e) Any spanning set that contains exactly n vectors is a basis.

[7] Although we defined elementary operations on a finite set only, the definition can easily be extended to infinite sets.

Proof

a) Consider a set of of vector $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m)$, where $m > n$. Since $\boldsymbol{R}$ is a basis, each $\boldsymbol{s}_j$ can be expressed in terms of $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n$ as

$$\boldsymbol{s}_j = \sum_{i=1}^{n} a_{ij}\boldsymbol{r}_i\,, \quad j = 1, 2, \ldots, m$$

Let $A = [\,a_{ij}\,]_{n\times m}$. Since $n < m$, the linear system $A\boldsymbol{c} = \boldsymbol{0}$ has a nontrivial solution, that is, there exist $c_1, c_2, \ldots, c_m$, not all zero, such that

$$\sum_{j=1}^{m} a_{ij}c_j = 0\,, \quad i = 1, 2, \ldots, n$$

Then

$$\sum_{j=1}^{m} c_j\boldsymbol{s}_j = \sum_{j=1}^{m} c_j(\sum_{i=1}^{n} a_{ij}\boldsymbol{r}_i) = \sum_{i=1}^{n}(\sum_{j=1}^{m} a_{ij}c_j)\boldsymbol{r}_i = \boldsymbol{0}$$

which shows that $\boldsymbol{S}$ is linearly dependent.

b) Suppose that a set of vectors $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m)$, where $m < n$, spans $\boldsymbol{\mathcal{X}}$. Then each $\boldsymbol{r}_j$ can be expressed as

$$\boldsymbol{r}_j = \sum_{i=1}^{m} b_{ij}\boldsymbol{s}_i\,, \quad j = 1, 2, \ldots, n$$

Let $B = [\,b_{ij}\,]_{m\times n}$. Since $m < n$, we can show by following the same argument as in part (a) that there exist $c_1, c_2, \ldots, c_n$, not all zero, such that

$$\sum_{j=1}^{n} c_j\boldsymbol{r}_j = \sum_{j=1}^{n} c_j(\sum_{i=1}^{m} b_{ij}\boldsymbol{s}_i) = \sum_{i=1}^{m}(\sum_{j=1}^{n} b_{ij}c_j)\boldsymbol{s}_i = \boldsymbol{0}$$

Since this contradicts the assumption that $\boldsymbol{R}$ is linearly dependent $\boldsymbol{S}$ cannot span $\boldsymbol{\mathcal{X}}$.

c) If $\boldsymbol{S}$ is a basis containing m vectors, then by (a) $m \leq n$, and by (b) $m \geq n$, that is, $m = n$.

d) Let $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$ be a linearly independent set. Then, by (a) no proper superset of $\boldsymbol{S}$ is linearly independent, and by Theorem 3.1, $\boldsymbol{S}$ is a basis.

e) Let $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)$ be a spanning set. Then, by (b) no proper subset of $\boldsymbol{S}$ spans $\boldsymbol{\mathcal{X}}$, and by Theorem 3.1, $\boldsymbol{S}$ is a basis.

By Corollary 3.1(c), all bases of a finite dimensional vector space contain the same number of basis vectors. This fixed number is called the ***dimension*** of $\boldsymbol{\mathcal{X}}$, denoted $\dim(\boldsymbol{\mathcal{X}})$. If $\boldsymbol{\mathcal{X}}$ is a trivial vector space containing only the zero vector, then it has no basis and $\dim(\boldsymbol{\mathcal{X}}) = 0$.

From Example 3.17 we conclude that

$$\dim(\mathbb{F}^n) = \dim(\mathbb{F}^{1\times n}) = \dim(\mathbb{F}^{n\times 1}) = n$$

Example 3.19

From Examples 3.10 and 3.12 we conclude that the sets $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ and $\boldsymbol{R}_3$ are all bases for $\mathbb{R}^3$. Since $\dim(\mathbb{R}^3) = 3$, the set $\boldsymbol{R}$, which contains four vectors, must be linearly dependent. Although the set $\boldsymbol{R}_4$ also contains three vectors, it is not a basis, because it is not linearly independent. Then it cannot span $\mathbb{R}^3$. These results had already been obtained in Example 3.12.

Referring to the same example, we also observe that the set $\{\boldsymbol{r}_2, \boldsymbol{r}_3\}$ is a basis for the subspace $\boldsymbol{\mathcal{U}}$ in Example 3.6. Hence $\dim(\boldsymbol{\mathcal{U}}) = 2$. This is completely expected as $\boldsymbol{\mathcal{U}}$ is essentially the same as the two-dimensional xy plane (or the yz or xz planes) except that it is tilted about the origin. The linearly independent vectors

$$\boldsymbol{s}_1 = (1, 1, 0) \quad \text{and} \quad \boldsymbol{s}_2 = (2, 1, 1)$$

form another basis for $\boldsymbol{\mathcal{U}}$.

Example 3.20

Any two vectors not lying on the same straight line are linearly independent in $\mathbb{R}^2$. Since $\dim(\mathbb{R}^2) = 2$, any two such vectors form a basis for $\mathbb{R}^2$. For example, the vectors $\boldsymbol{u}_1 = (2.0, 1.0)$ and $\boldsymbol{u}_2 = (1.0, 2.0)$ shown in Figure 3.2 are linearly independent and form a basis for $\mathbb{R}^2$. The vectors $\boldsymbol{v}_1 = (1.1, 1.0)$ and $\boldsymbol{v}_2 = (1.0, 1.1)$ shown in the same figure are also linearly independent and form another basis for $\mathbb{R}^2$.

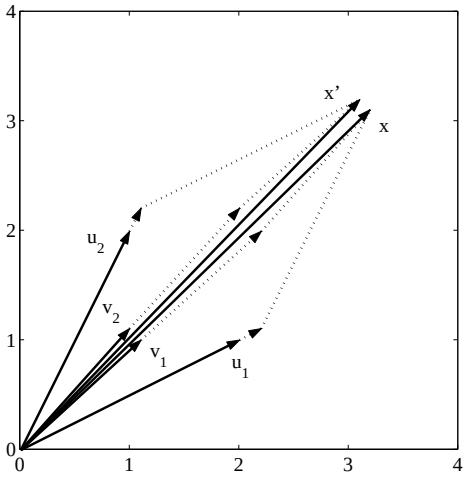

Figure 3.2: Two different bases for $\mathbb{R}^2$

Although $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ and $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ are both bases, they are quite different from a computational point of view. Consider two vectors $\boldsymbol{x} = (3.2, 3.1)$ and $\boldsymbol{x}' = (3.1, 3.2)$ which represent two close points in $\mathbb{R}^2$. We expect that when we express them in terms of a basis, then their corresponding coefficients multiplying the basis vectors should also be close. This is indeed the case for $\boldsymbol{U}$, where

$$\begin{aligned} \boldsymbol{x} &= 1.1\,\boldsymbol{u}_1 + 1.0\,\boldsymbol{u}_2 \\ \boldsymbol{x}' &= 1.0\,\boldsymbol{u}_1 + 1.1\,\boldsymbol{u}_2 \end{aligned}$$

On the other hand, when $\boldsymbol{x}$ and $\boldsymbol{x}'$ are expressed in terms of $\boldsymbol{V}$ as

$$\begin{aligned} \boldsymbol{x} &= 2.0\,\boldsymbol{r}_1' + 1.0\,\boldsymbol{r}_2' \\ \boldsymbol{x}' &= 1.0\,\boldsymbol{r}_1' + 2.0\,\boldsymbol{r}_2' \end{aligned}$$

their corresponding coefficients differ greatly.

To explain the situation we observe that finding the coefficients c_1 and c_2 of a given vector $\boldsymbol{x} = (\alpha, \beta) = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2$ is equivalent to solving the linear system

$$\begin{bmatrix} 1.1 & 1.0 \\ 1.0 & 1.1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Since this system is ill-conditioned, its solution (the coefficients c_1 and c_2) are very sensitive to small changes in α and β. The ill-conditioning of the system results from $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ being very much aligned with each other. We can say that they are closer to being linearly dependent than $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are.[8]

Example 3.21

Any $A \in \mathbb{R}^{2\times 2}$ can be expressed as

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \\ &= a_{11}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + a_{12}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + a_{21}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + a_{22}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= a_{11}E_{11} + a_{12}E_{12} + a_{21}E_{21} + a_{22}E_{22} \end{aligned}$$

Hence the set $\boldsymbol{E} = \{E_{11}, E_{12}, E_{21}, E_{22}\}$ spans $\mathbb{R}^{2\times 2}$. Since it is also linearly independent, it is a basis for $\mathbb{R}^{2\times 2}$. (The same conclusion can also be reached by observing that the coefficients of E_{ij} in the above expression are uniquely determined by A.) Therefore, $\dim(\mathbb{R}^{2\times 2}) = 4$.

In general, $\dim(\mathbb{F}^{m\times n}) = mn$, and the set

$$\{\, E_{ij} \mid 1 \le i \le m, 1 \le j \le n \,\}$$

where $E_{ij} \in \mathbb{F}^{m\times n}$ consists of all 0's except a single 1 in the (i,j)th position, is a basis for $\mathbb{F}^{m\times n}$. Note that if columns of each E_{ij} are stacked to form a $mn \times 1$ column vector $\boldsymbol{e}_{ij}$, then the set $\{\boldsymbol{e}_{ij}\}$ will form the canonical basis for $\mathbb{F}^{mn\times 1}$.

Example 3.22

The sets

$$\mathbb{R}_s^{2\times 2} = \{\, S \in \mathbb{R}^{2\times 2} \mid S \text{ is symmetric} \,\}$$

and

$$\mathbb{R}_q^{2\times 2} = \{\, Q \in \mathbb{R}^{2\times 2} \mid Q \text{ is skew-symmetric} \,\}$$

are subspace of $\mathbb{R}^{2\times 2}$.

The matrices

$$S_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

form a basis for $\mathbb{R}_s^{2\times 2}$, and the matrix

$$Q = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

forms a basis for $\mathbb{R}_q^{2\times 2}$. Hence $\dim(\mathbb{R}_s^{2\times 2}) = 3$ and $\dim(\mathbb{R}_q^{2\times 2}) = 1$.

The following corollary is useful in constructing a basis for a vector space.

[8] We will mention about a measure of linear independence in Chapter 7.

Corollary 3.1.2 *Let dim* $(\boldsymbol{\mathcal{X}}) = n$.

a) Any spanning set containing $m > n$ *vectors can be reduced to a basis by deleting* $m - n$ *vectors from the set.*

b) Any linearly independent set containing $k < n$ *vectors can be completed to a basis by including* $n - k$ *more vectors into the set.*

Proof

a) Let $\boldsymbol{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}, m > n$, be a spanning set. By Corollary 3.1(a), it must be linearly dependent, and therefore, one of its vectors can be expressed in terms of the others. Deleting that vector from $\boldsymbol{R}$ reduces the number of vectors by one without destroying the spanning property. Continuing this process we finally obtain a subset of $\boldsymbol{R}$ which contains exactly n vectors and spans $\boldsymbol{\mathcal{X}}$. By Corollary 3.1(e), it is a basis.
The process of reducing $\boldsymbol{R}$ to a basis can be summarized by an algorithm:

$\boldsymbol{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$
For $i = m : 1$
  If span $(\boldsymbol{R} - \{\boldsymbol{r}_i\}) = \boldsymbol{\mathcal{X}}, \boldsymbol{R} = \boldsymbol{R} - \{\boldsymbol{r}_i\}$
End

b) Let $\boldsymbol{R}_1 = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\}, k < n$, be a linearly independent set, and let $\{\boldsymbol{r}_{k+1}, \ldots, \boldsymbol{r}_{k+n}\}$ be any basis for $\boldsymbol{\mathcal{X}}$. Then $\boldsymbol{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{r}_{k+1}, \ldots, \boldsymbol{r}_{k+n}\}$ is a spanning set with $m = k + n$ elements. Application of the algorithm in part (a) to $\boldsymbol{R}$ reduces it to a basis which includes the first k vectors. Details are worked out in Exercise 3.14.

Example 3.23

Consider Example 3.10 again. The linearly independent set $\{\boldsymbol{r}_1, \boldsymbol{r}_2\}$ can be completed to a basis by adding $\boldsymbol{r}_3$ or $\boldsymbol{r}_4$. The spanning set $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3, \boldsymbol{r}_4\}$ can be reduced to a basis by deleting $\boldsymbol{r}_2$, or $\boldsymbol{r}_3$, or $\boldsymbol{r}_4$.

3.3.2 Representation of Vectors With Respect to A Basis

Let dim $(\boldsymbol{\mathcal{X}}) = n$, and let $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ be an ordered basis for $\boldsymbol{\mathcal{X}}$. Then any vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ can be expressed in terms of the basis vectors as

$$\boldsymbol{x} = \alpha_1 \boldsymbol{r}_1 + \cdots + \alpha_n \boldsymbol{r}_n$$

for some unique scalars $\alpha_i, i = 1, \ldots, n$. The column vector

$$\boldsymbol{\alpha} = \text{col}\,[\,\alpha_1, \alpha_2, \ldots, \alpha_n\,] \in \mathbb{F}^{n \times 1}$$

is called the ***representation*** of $\boldsymbol{x}$ with respect to the basis $\boldsymbol{R}$. This way we establish a one-to-one correspondence between the vectors of $\boldsymbol{\mathcal{X}}$ and the $n \times 1$ vectors of $\mathbb{F}^{n \times 1}$.

Note that although α_i are unique, their locations in $\boldsymbol{\alpha}$ depend on the ordering of the basis vectors. To guarantee that every vector has a unique column representation and that every colum represents a unique vector, it is necessary to associate an order with a basis. For this reson, from now on, whenever we deal with representations of vectors with respect to a basis, we will assume that the basis is ordered.

Example 3.24

Consider the basis $\boldsymbol{E}$ of $\mathbb{R}^{3 \times 1}$ in Example 3.17. From (3.5) we observe that the representation of a

vector $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$ with respect to $\boldsymbol{E}$ is itself. That is why $\boldsymbol{E}$ is called the canonical basis for $\mathbb{R}^{3\times 1}$.

Now consider the basis $\boldsymbol{R}$ in the same example. From (3.6), the representation of $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$ with respect to $\boldsymbol{R}$ is obtained as

$$\boldsymbol{\alpha} = \begin{bmatrix} x_1 - x_2 \\ x_2 - x_3 \\ x_3 \end{bmatrix}$$

Example 3.25

The set $(\boldsymbol{r}_2, \boldsymbol{r}_3)$ in Example 3.10 is a basis for the subspace $\boldsymbol{\mathcal{U}}$ in Example 3.6. The representation of $\boldsymbol{u} = (x, y, x - y) \in \boldsymbol{\mathcal{U}}$ with respect to this basis is obtained by expressing $\boldsymbol{u}$ in terms of $\boldsymbol{r}_2$ and $\boldsymbol{r}_3$ as

$$\boldsymbol{u} = x\,\boldsymbol{r}_2 + y\,\boldsymbol{r}_3$$

which gives

$$\boldsymbol{\alpha} = \begin{bmatrix} x \\ y \end{bmatrix}$$

The set $\boldsymbol{S} = (\boldsymbol{s}_1, \boldsymbol{s}_2)$ in Example 3.19 is also a basis for $\boldsymbol{\mathcal{U}}$. The representation of $\boldsymbol{u} = (x, y, x - y)$ with respect to $\boldsymbol{S}$ is obtained by expressing $\boldsymbol{u}$ in terms of $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ as

$$\boldsymbol{u} = 2y\,\boldsymbol{s}_1 + (x - y)\,\boldsymbol{s}_2$$

to be

$$\boldsymbol{\beta} = \begin{bmatrix} 2y \\ x - y \end{bmatrix}$$

Note that $\dim(\boldsymbol{\mathcal{U}}) = 2$, and therefore, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{2\times 1}$.

* **Example 3.26**

Let $\mathbb{D}_N = (\,0, 1, \ldots, N-1\,)$, and consider the vector space $\mathcal{F}(\mathbb{D}_N, \mathbb{C})$ of complex-valued functions defined on $\mathbb{D}_N$.

Let the functions $e_p \in \mathcal{F}(\mathbb{D}_N, \mathbb{C})$ be defined as

$$e_p[k] = \begin{cases} 1\,, & k = p \\ 0\,, & k \neq p \end{cases} \quad , \quad p = 0, 1, \ldots, N-1$$

Then any $f \in \mathcal{F}(\mathbb{D}_N, \mathbb{C})$ can be expressed in terms of e_p uniquely as

$$f = \sum_{p=0}^{N-1} a_p e_p\,, \quad a_p = f[p]\,, \quad p = 0, 1, \ldots, N-1$$

because

$$(\sum_{p=0}^{N-1} a_p e_p\,)[k] = \sum_{p=0}^{N-1} a_p e_p[k] = \sum_{p=0}^{N-1} f[p] e_p[k] = f[k]\,, \quad k \in \mathbb{D}_N$$

Hence $(\,e_0, e_1, \ldots, e_{N-1}\,)$ is a basis for $\mathcal{F}(\mathbb{D}_N, \mathbb{C})$, and therefore $\dim(\mathcal{F}(\mathbb{D}_N, \mathbb{C})) = N$. From the expression above it also follows that the representation of f with respect to this basis is the column vector

$$\boldsymbol{f} = \text{col}\,[\,a_0, a_1, \ldots, a_{N-1}\,] = \text{col}\,[\,f[0], f[1], \ldots, f[N-1]\,]$$

There is nothing surprising about this result. $\mathcal{F}(\mathbb{D}_N, \mathbb{C})$ is essentially the same as $\mathbb{C}^{N\times 1}$, and a function $f \in \mathcal{F}(\mathbb{D}_N, \mathbb{C})$ is the same as the column vector $\boldsymbol{f}$. The functions $e_0, e_1, \ldots, e_{N-1}$ correspond to the canonical basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_N$ of $\mathbb{C}^{N\times 1}$. That is why the representation of a function with respect to (e_p) is a column vector consisting of the values of f at $k = 0, 1, \ldots, N-1$.

Things become more interesting when we consider another basis for $\mathcal{F}(\mathbb{D}_N, \mathbb{C})$. Let

$$\phi_p[k] = e^{ip\frac{2\pi}{N}k}, \quad p = 0, 1, \ldots, N-1$$

It can be shown (see Exercise 3.20) that any function $f \in \mathcal{F}(\mathbb{D}_N, \mathbb{C})$ can be expressed in terms of ϕ_p uniquely as

$$f = \sum_{p=0}^{N-1} c_p \phi_p, \quad c_p = \frac{1}{N}\sum_{k=0}^{N-1} f[k]\phi_p^*[k], \quad p = 0, 1, \ldots, N-1 \tag{3.7}$$

Hence $(\,\phi_0, \phi_1, \ldots, \phi_{N-1}\,)$ is also a basis for $\mathcal{F}(\mathbb{D}_N, \mathbb{C})$, and the representation of f with respect to (ϕ_p) is

$$\boldsymbol{F} = \text{col}\,[\,c_0, c_1, \ \ldots\ , c_{N-1}\,]$$

The representation of f as a linear combination of ϕ_p is known as the ***discrete Fourier series*** of f, and the coefficients c_p as the discrete Fourier coefficients of f.

As a specific example, suppose $N = 4$. Then the basis functions ϕ_p have the values tabulated below.

	$\phi_0[k]$	$\phi_1[k]$	$\phi_2[k]$	$\phi_3[k]$
$k=0$	1	1	1	1
$k=1$	1	i	-1	$-i$
$k=2$	1	-1	1	-1
$k=3$	1	$-i$	-1	i

Let

$$f[k] = \begin{cases} 2, & k=0 \\ 4, & k=1 \\ -2, & k=2 \\ 0, & k=3 \end{cases}$$

Then the discrete Fourier coefficients of f are computed as

$$\begin{aligned} c_0 &= \tfrac{1}{4}(2+4-2+0) &&= 1 \\ c_1 &= \tfrac{1}{4}(2-4i+2+0) &&= 1-i \\ c_2 &= \tfrac{1}{4}(2-4-2+0) &&= -1 \\ c_3 &= \tfrac{1}{4}(2+4i+2+0) &&= 1+i \end{aligned}$$

Hence the discrete Fourier series of f is

$$f = \phi_0 + (1-i)\phi_1 - \phi_2 + (1+i)\phi_3$$

and the representation of f with respect to (ϕ_p) is

$$\boldsymbol{F} = \begin{bmatrix} 1 \\ 1-i \\ -1 \\ 1+i \end{bmatrix}$$

From the examples above we observe that although the representation of a vector is unique with respect to a given basis, it has a different (but still unique) representation with respect to another basis. We now investigate how different representations of the same vector with respect to different bases are related.

Let $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_n)$ and $\boldsymbol{R}' = (\boldsymbol{r}'_1, \boldsymbol{r}'_2, \ldots, \boldsymbol{r}'_n)$ be two ordered bases for $\boldsymbol{\mathcal{X}}$.[9] Suppose that a vector $\boldsymbol{x}$ has the representations

$$\boldsymbol{\alpha} = \text{col}\,[\,\alpha_1, \ldots, \alpha_n\,] \quad \text{and} \quad \boldsymbol{\alpha}' = \text{col}\,[\,\alpha'_1, \ldots, \alpha'_n\,]$$

with respect to $\boldsymbol{R}$ and $\boldsymbol{R}'$. That is,

$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j \boldsymbol{r}_j = \sum_{i=1}^{n} \alpha'_i \boldsymbol{r}'_i$$

Let the jth basis vector $\boldsymbol{r}_j$ be expressed in terms of the vectors of $\boldsymbol{R}'$ as

$$\boldsymbol{r}_j = \sum_{i=1}^{n} q_{ij} \boldsymbol{r}'_i\,, \quad j = 1, \ldots, n$$

so that it has a representation

$$\boldsymbol{q}_j = \text{col}\,[\,q_{1j}, \ldots, q_{nj}\,]\,, \quad j = 1, \ldots, n$$

with respect to $\boldsymbol{R}'$. Then

$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j \boldsymbol{r}_j = \sum_{j=1}^{n} \alpha_j \,(\sum_{i=1}^{n} q_{ij} \boldsymbol{r}'_i\,) = \sum_{i=1}^{n} (\sum_{j=1}^{n} q_{ij} \alpha_j\,)\, \boldsymbol{r}'_i = \sum_{i=1}^{n} \alpha'_i \boldsymbol{r}'_i$$

By uniqueness of the representation of $\boldsymbol{x}$ with respect to $\boldsymbol{R}'$, we have

$$\alpha'_i = \sum_{j=1}^{n} q_{ij} \alpha_j\,, \quad i = 1, \ldots, n$$

By expressing these equalities in matrix form, we observe that the representations $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}$ are related as

$$\boldsymbol{\alpha}' = Q\,\boldsymbol{\alpha}$$

The matrix

$$Q = [\,\boldsymbol{q}_1 \;\; \boldsymbol{q}_2 \;\; \cdots \;\; \boldsymbol{q}_n\,] = [\,q_{ij}\,]_{n \times n}$$

which is defined uniquely by $\boldsymbol{R}$ and $\boldsymbol{R}'$, is called the ***matrix of change–of–basis*** from $\boldsymbol{R}$ to $\boldsymbol{R}'$.

Now interchange the roles of the bases $\boldsymbol{R}$ and $\boldsymbol{R}'$. Let the jth basis vector $\boldsymbol{r}'_j$ be expressed in terms of the vectors of $\boldsymbol{R}$ as

$$\boldsymbol{r}'_j = \sum_{i=1}^{n} p_{ij} \boldsymbol{r}_i\,, \quad j = 1, \ldots, n$$

[9] Keep in mind that even when $\boldsymbol{R}$ and $\boldsymbol{R}'$ contain exactly the same vectors, if their orderings are different then $\boldsymbol{R}$ and $\boldsymbol{R}'$ are different.

so that it has a representation

$$\boldsymbol{p}_j = \text{col}\,[\,p_{1j}, \ldots, p_{nj}\,], \quad j = 1, \ldots, n$$

with respect to $\boldsymbol{R}$. Defining

$$P = [\,\boldsymbol{p}_1 \;\; \boldsymbol{p}_2 \;\; \cdots \;\; \boldsymbol{p}_n\,] = [\,p_{ij}\,]_{n\times n}$$

to be the matrix of change–of–basis from $\boldsymbol{R}'$ to $\boldsymbol{R}$, we get

$$\boldsymbol{\alpha} = P\,\boldsymbol{\alpha}'$$

The reader might suspect that the matrices Q and P are related. Indeed, since $\boldsymbol{\alpha} = P\boldsymbol{\alpha}' = PQ\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}' = Q\boldsymbol{\alpha} = QP\boldsymbol{\alpha}'$ for all pairs $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathbb{R}^{n\times 1}$, we must have

$$PQ = QP = I_n$$

We will investigate such matrices in Chapter 4.

Example 3.27

Consider Example 3.17. Expressing the vectors of the canonical basis $\boldsymbol{E}$ in terms of $\boldsymbol{R}$ as

$$\begin{aligned} \boldsymbol{e}_1 &= \boldsymbol{r}_1 \\ \boldsymbol{e}_2 &= -\boldsymbol{r}_1 + \boldsymbol{r}_2 \\ \boldsymbol{e}_3 &= -\boldsymbol{r}_2 + \boldsymbol{r}_3 \end{aligned}$$

we obtain the matrix of change–of–basis from $\boldsymbol{E}$ to $\boldsymbol{R}$ as

$$Q = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Hence the representation of a vector $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$ with respect to $\boldsymbol{R}$ is related to its canonical representation $\boldsymbol{x}$ as

$$\boldsymbol{\alpha} = Q\boldsymbol{x} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 - x_2 \\ x_2 - x_3 \\ x_3 \end{bmatrix}$$

which is the same as in Example 3.24.

Since the representations of $\boldsymbol{r}_j$ with respect to $\boldsymbol{E}$ are themselves, the matrix of change–of–basis from $\boldsymbol{R}$ to $\boldsymbol{E}$ is easily obtained as

$$P = [\,\boldsymbol{r}_1 \;\; \boldsymbol{r}_2 \;\; \boldsymbol{r}_3\,] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

If $\boldsymbol{x}$ has a representation

$$\boldsymbol{\alpha} = \text{col}\,[\,a, b, c\,]$$

with respect to $\boldsymbol{R}$, then

$$\boldsymbol{x} = a\boldsymbol{r}_1 + b\boldsymbol{r}_2 + c\boldsymbol{r}_3 = \begin{bmatrix} a+b+c \\ b+c \\ c \end{bmatrix} = P\boldsymbol{\alpha}$$

The reader should verify that $QP = PQ = I$.

3.4 Linear Transformations

Let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be vector spaces over the same field $\mathbb{F}$. A mapping $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ is called a ***linear transformation*** if for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{\mathcal{X}}$ and for all $c_1, c_2 \in \mathbb{F}$

$$\mathcal{A}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2) = c_1\mathcal{A}(\boldsymbol{x}_1) + c_2\mathcal{A}(\boldsymbol{x}_2) \tag{3.8}$$

(3.8) is equivalent to

$$\mathcal{A}(\boldsymbol{x}_1 + \boldsymbol{x}_2) = \mathcal{A}(\boldsymbol{x}_1) + \mathcal{A}(\boldsymbol{x}_2) \tag{3.9}$$

and

$$\mathcal{A}(c\boldsymbol{x}) = c\mathcal{A}(\boldsymbol{x}) \tag{3.10}$$

which are known as ***superposition*** and ***homogeneity***, respectively. (3.9) follows from (3.8) on choosing $c_1 = c_2 = 1$, and (3.10) on choosing $c_1 = 1, c_2 = 0$ and $\boldsymbol{x}_1 = \boldsymbol{x}$. Conversely, (3.9) and (3.10) imply that

$$\mathcal{A}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2) = \mathcal{A}(c_1\boldsymbol{x}_1) + \mathcal{A}(c_2\boldsymbol{x}_2) = c_1\mathcal{A}(\boldsymbol{x}_1) + c_2\mathcal{A}(\boldsymbol{x}_2)$$

$\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ are the ***domain*** and the ***codomain*** of $\mathcal{A}$. A linear transformation from a vector space $\boldsymbol{\mathcal{X}}$ into itself is called a ***linear operator*** on $\boldsymbol{\mathcal{X}}$.

If $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ is a linear transformation then

$$\mathcal{A}(\mathbf{0}_x) = \mathbf{0}_y$$

which follows from (3.8) on taking $c_1 = c_2 = 0$.

Example 3.28

The zero mapping $\mathcal{O} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ defined as

$$\mathcal{O}(\boldsymbol{x}) = \mathbf{0}_y \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

is a linear transformation that satisfies (3.8) trivially.

The identity mapping $\mathcal{I} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ defined as

$$\mathcal{I}(\boldsymbol{x}) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

is also a linear transformation, because

$$\mathcal{I}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2) = c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 = c_1\mathcal{I}(\boldsymbol{x}_1) + c_2\mathcal{I}(\boldsymbol{x}_2)$$

Hence $\mathcal{I}$ is a linear operator on $\mathcal{X}$.

Example 3.29

The mapping $\mathcal{A} : \mathbb{R}^3 \to \mathbb{R}^2$ defined as

$$\mathcal{A}(x_1, x_2, x_3) = (x_1 + x_2, x_2 - x_3)$$

is a linear transformation. For $\boldsymbol{u} = (u_1, u_2, u_3)$, $\boldsymbol{v} = (v_1, v_2, v_3)$ and $c, d \in \mathbb{F}$

$$\begin{aligned}
\mathcal{A}(c\boldsymbol{u} + d\boldsymbol{v}) &= \mathcal{A}(cu_1 + dv_1, cu_2 + dv_2, cu_3 + dv_3) \\
&= (cu_1 + dv_1 + cu_2 + dv_2, cu_2 + dv_2 - cu_3 - dv_3) \\
&= c(u_1 + u_2, u_2 - u_3) + d(v_1 + v_2, v_2 - v_3) \\
&= c\mathcal{A}(\boldsymbol{u}) + d\mathcal{A}(\boldsymbol{v})
\end{aligned}$$

However, none of the mappings

$$\begin{aligned}\mathcal{B}(x_1,x_2,x_3) &= (x_1+x_3, x_2+1)\\ \mathcal{C}(x_1,x_2,x_3) &= (x_1+x_3, x_2^2)\\ \mathcal{D}(x_1,x_2,x_3) &= (x_1x_3, x_2)\end{aligned}$$

is a linear transformation. $\mathcal{B}$ is not linear simply because $\mathcal{B}(\mathbf{0}) \neq \mathbf{0}$. The reader is urged to explain why $\mathcal{C}$ and $\mathcal{D}$ are not linear.

Example 3.30

Let $\mathcal{A} : \mathbb{F}^{n\times 1} \to \mathbb{F}^{m\times 1}$ be defined as

$$\mathcal{A}(\boldsymbol{x}) = A\boldsymbol{x}$$

where A is an $m \times n$ matrix with elements from $\mathbb{F}$. Since

$$A(a\boldsymbol{x} + b\boldsymbol{y}) = aA\boldsymbol{x} + bA\boldsymbol{y}$$

$\mathcal{A}$ is a linear transformation. This example shows that every matrix defines a linear transformation. Thus a linear transformation defined by an $n \times n$ matrix with elements from $\mathbb{F}$ is a linear operator on $\mathbb{F}^{n\times 1}$.

Example 3.31

Let $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$, and $\boldsymbol{\mathcal{Z}}$ be vector spaces over the same field, and let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ and $\mathcal{B} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{Z}}$ be linear transformations. Then the compound mapping $\mathcal{C} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Z}}$ defined as

$$\mathcal{C}(\boldsymbol{x}) = (\mathcal{B} \circ \mathcal{A})(\boldsymbol{x}) = \mathcal{B}(\mathcal{A}(\boldsymbol{x}))$$

is also a linear transformation, because

$$\begin{aligned}\mathcal{C}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2) &= \mathcal{B}(\mathcal{A}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2))\\ &= \mathcal{B}(c_1\mathcal{A}(\boldsymbol{x}_1) + c_2\mathcal{A}(\boldsymbol{x}_2))\\ &= c_1\mathcal{B}(\mathcal{A}(\boldsymbol{x}_1)) + c_2\mathcal{B}(\mathcal{A}(\boldsymbol{x}_2))\\ &= c_1\mathcal{C}(\boldsymbol{x}_1) + c_2\mathcal{C}(\boldsymbol{x}_2)\end{aligned}$$

* **Example 3.32**

In Section 2.5 we defined the differential operator $\mathcal{D}$ as a mapping from a set of functions into itself such that $\mathcal{D}(f) = f'$. We now take a closer look at $\mathcal{D}$.

Recall from Example 3.8 that

$$\mathcal{C}_0(\boldsymbol{I}, \mathbb{R}) \supset \mathcal{C}_1(\boldsymbol{I}, \mathbb{R}) \supset \cdots \supset \mathcal{C}_\infty(\boldsymbol{I}, \mathbb{R})$$

are subspaces of $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$. Also, if $f \in \mathcal{C}_m(\boldsymbol{I}, \mathbb{R})$ then

$$f' \in \mathcal{C}_{m-1}(\boldsymbol{I}, \mathbb{R}),\ f'' \in \mathcal{C}_{m-2}(\boldsymbol{I}, \mathbb{R}),\ \ldots,\ f^{(m)} \in \mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$$

Hence, for any $m > 1$, the differential operator $\mathcal{D}$ is a mapping from $\mathcal{C}_m(\boldsymbol{I}, \mathbb{R})$ into $\mathcal{C}_{m-1}(\boldsymbol{I}, \mathbb{R})$, and therefore, into $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$. The property in (2.34) implies that $\mathcal{D}$ is a linear transformation. Then, as discussed in Example 3.31, the operator $\mathcal{D}^2$ that is defined in terms of $\mathcal{D}$ as

$$\mathcal{D}^2(f) = (\mathcal{D} \circ \mathcal{D})(f) = \mathcal{D}(\mathcal{D}(f)) = \mathcal{D}(f') = f''$$

is also a linear transformation. Consequently, each $\mathcal{D}^k, k = 1, \ldots, n \leq m$, which can be defined recursively as

$$\mathcal{D}^k(f) = (\mathcal{D} \circ \mathcal{D}^{k-1})(f) = \mathcal{D}(\mathcal{D}^{k-1}(f)) = \mathcal{D}(f^{(k-1)}) = f^{(k)}$$

is a linear transformations from $\mathcal{C}_m(\boldsymbol{I}, \mathbb{R})$ into $\mathcal{C}_{m-k}(\boldsymbol{I}, \mathbb{R})$, and therefore, into $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$.

Finally, an nth order linear differential operator $L(\mathcal{D})$ is a linear transformation from $\mathcal{C}_n(\boldsymbol{I}, \mathbb{R})$ into $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$ (see Exercise 3.31). In fact, this is precisely the reason for calling $L(\mathcal{D})$ a linear operator.

* **Example 3.33**

Recall that a vector $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ can also be viewed as a function $f \in \mathcal{F}(\mathbb{N}_n, \mathbb{R})$. The linear operator $\mathcal{A} : \mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times 1}$ defined by a matrix $A \in \mathbb{R}^{n\times n}$ can similarly be interpreted as a a linear operator $\mathcal{A} : \mathcal{F}(\mathbb{N}_n, \mathbb{R}) \to \mathcal{F}(\mathbb{N}_n, \mathbb{R})$ such that the image $g = \mathcal{A}(f)$ of a function f is defined pointwise as

$$g[p] = \sum_{q=1}^{n} a_{pq} f[q]\,, \quad p \in \mathbb{N}_n$$

Now consider the vector space $\mathcal{F}(\mathbb{Z}, \mathbb{R})$ of infinite sequences. We can define a linear operator $\mathcal{H}$ on $\mathcal{F}(\mathbb{Z}, \mathbb{R})$ such that if $g = \mathcal{H}(f)$ then

$$g[p] = \sum_{q=-\infty}^{\infty} h[p, q] f[q]\,, \quad p \in \mathbb{Z} \tag{3.11}$$

where $h[p, q] \in \mathbb{R}$.[10] We can think of $\mathcal{H}$ as defined by an infinitely large matrix H with elements $h[p, q]$ such that

$$\begin{bmatrix} \vdots \\ g[-1] \\ g[\,0\,] \\ g[\,1\,] \\ \vdots \end{bmatrix} = \begin{bmatrix} & \vdots & \vdots & \vdots & \\ \cdots & h[-1,-1] & h[-1,0] & h[-1,1] & \cdots \\ \cdots & h[\,0,-1] & h[\,0,0] & h[\,0,1] & \cdots \\ \cdots & h[\,1,-1] & h[\,1,0] & h[\,1,1] & \cdots \\ & \vdots & \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} \vdots \\ f[-1] \\ f[\,0\,] \\ f[\,1\,] \\ \vdots \end{bmatrix}$$

Let us go one step further, and consider the vector space $\mathcal{F}(\mathbb{R}, \mathbb{R})$ of real-valued functions defined on $\mathbb{R}$. Now the domain of the functions is a continuum, and the infinite summation in (3.11) must be replaced with an integral: We can then define a linear operator $\mathcal{H}$ on $\mathcal{F}(\mathbb{R}, \mathbb{R})$ such that $g = \mathcal{H}(f)$ is characterized by

$$g(t) = \int_{-\infty}^{\infty} h(t, \tau) f(\tau)\, d\tau\,, \quad t \in \mathbb{R} \tag{3.12}$$

In the special cases when $h[p, q] = h[p - q]$ and $h(t, \tau) = h(t - \tau)$, the operators defined by (3.11) and (3.12) are known as ***convolution***, and are widely used in system analysis.

3.4.1 Matrix Representation of Linear Transformations

Let $\dim(\boldsymbol{\mathcal{X}}) = n$, let $\dim(\boldsymbol{\mathcal{Y}}) = m$, let $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ and $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m)$ be ordered bases for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$, and let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ be a linear transformation.

[10] Of course, this definition requires that the infinite series in (3.11) converges for all $p \in \mathbb{Z}$, which puts some restrictions not only on $h[p, q]$ but also on f. These technical difficulties can be worked out by restricting f to a subspace of $\mathcal{F}(\mathbb{Z}, \mathbb{R})$ and by choosing $h[p, q]$ suitably.

Consider $\mathcal{A}(\boldsymbol{r}_j)$. Since it is a vector in $\boldsymbol{\mathcal{Y}}$, it has a unique representation $\boldsymbol{a}_j \in \mathbb{F}^{m\times 1}$ with respect to the basis $\boldsymbol{S}$. That is,

$$\mathcal{A}(\boldsymbol{r}_j) = \sum_{i=1}^{m} a_{ij}\boldsymbol{s}_i\,, \quad j = 1, \ldots, n$$

and

$$\boldsymbol{a}_j = \text{col}\,[\,a_{1j}, a_{2j}, \ldots, a_{mj}\,]\,, \quad j = 1, \ldots, n$$

The $m \times n$ matrix

$$A = [\,\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\cdots\; \boldsymbol{a}_n\,] = [\,a_{ij}\,]$$

which is uniquely defined by $\mathcal{A}$, $\boldsymbol{R}$ and $\boldsymbol{S}$, is called the ***matrix representation of a linear transformation*** of $\mathcal{A}$ with respect to the basis pair $(\boldsymbol{R}, \boldsymbol{S})$.

The significance of the matrix representation of $\mathcal{A}$ is that if $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ has a representation $\boldsymbol{\alpha} \in \mathbb{F}^{n\times 1}$ with respect to $\boldsymbol{R}$ and $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) \in \boldsymbol{\mathcal{Y}}$ has a representation $\boldsymbol{\beta} \in \mathbb{F}^{m\times 1}$ with respect to $\boldsymbol{S}$, then the two representations are related as $\boldsymbol{\beta} = A\boldsymbol{\alpha}$. To show this, suppose

$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j \boldsymbol{r}_j\,, \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{x}) = \sum_{i=1}^{m} \beta_i \boldsymbol{s}_i$$

Then

$$\boldsymbol{y} = \sum_{j=1}^{n} \alpha_j \mathcal{A}(\boldsymbol{r}_j) = \sum_{j=1}^{n} \alpha_j \,(\sum_{i=1}^{m} a_{ij}\boldsymbol{s}_i\,) = \sum_{i=1}^{m} (\sum_{j=1}^{n} a_{ij}\alpha_j\,)\boldsymbol{s}_i = \sum_{i=1}^{m} \beta_i \boldsymbol{s}_i$$

Since the representation of $\boldsymbol{y}$ in terms of $\boldsymbol{s}_i$ is unique, we must have

$$\beta_i = \sum_{j=1}^{n} a_{ij}\alpha_j\,, \quad i = 1, \ldots, m$$

or in matrix form

$$\boldsymbol{\beta} = A\boldsymbol{\alpha}$$

In conclusion, not only does an $m \times n$ matrix define a linear transformation from $\mathbb{F}^{n\times 1}$ into $\mathbb{F}^{m\times 1}$, but also a linear transformation from an n-dimensional vector space $\boldsymbol{\mathcal{X}}$ into an m-dimensional vector space $\boldsymbol{\mathcal{Y}}$ can be represented by an $m \times n$ matrix once a pair of bases for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ are fixed.[11]

Like the column representation of a vector with respect to basis, a linear transformation has different representations with respect to different bases. If $\mathcal{A}$ has a representation A with respect to $(\boldsymbol{R}, \boldsymbol{S})$ and a representation A' with respect to $(\boldsymbol{R}', \boldsymbol{S}')$, then

$$A' = Q_y A P_x$$

[11] The question of whether a similar result can be derived for linear transformations between infinite dimensional vector spaces is beyond the scope of this book.

where Q_y is the matrix of change–of–basis from $\boldsymbol{S}$ to $\boldsymbol{S}'$ in $\boldsymbol{\mathcal{Y}}$, and P_x is the matrix of change–of–basis from $\boldsymbol{R}'$ to $\boldsymbol{R}$ in $\boldsymbol{\mathcal{X}}$. This follows from the fact that if $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are representations of $\boldsymbol{x}$ with respect to $\boldsymbol{R}$ and $\boldsymbol{R}'$, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are representations of $\boldsymbol{y} = \mathcal{A}(\boldsymbol{x})$ with respect to $\boldsymbol{S}$ and $\boldsymbol{S}'$, then

$$A'\boldsymbol{\alpha}' = \boldsymbol{\beta}' = Q_y\boldsymbol{\beta} = Q_yA\boldsymbol{\alpha} = Q_yAP_x\boldsymbol{\alpha}'$$

The relations between the vectors of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ and their representations are summarized by the diagram in Figure 3.3. From the diagram it is clear that A, Q_yA, AP_x and Q_yAP_x all represent the same linear transformation, each with respect to a different pair of bases. Q_yA represents $\mathcal{A}$ with respect to $(\boldsymbol{R}, \boldsymbol{S}')$, and AP_x represents $\mathcal{A}$ with respect to $(\boldsymbol{R}', \boldsymbol{S})$.

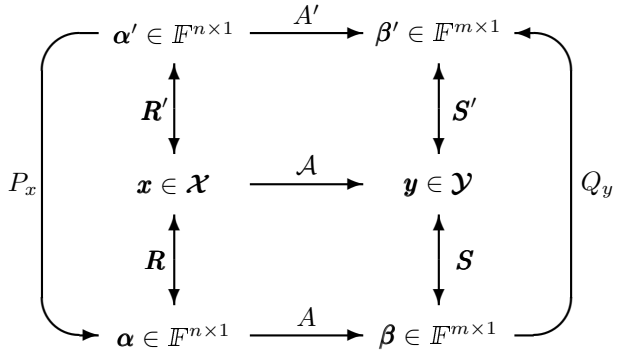

Figure 3.3: Matrix representation of a linear transformation

Example 3.34

Consider the linear transformation $\mathcal{A} : \mathbb{F}^{n\times 1} \to \mathbb{F}^{m\times 1}$ defined by a matrix $A^{m\times n} \in \mathbb{F}^{m\times n}$. Let $\boldsymbol{E}^n = (\,\boldsymbol{e}_1^n, \ldots, \boldsymbol{e}_n^n\,)$ and $\boldsymbol{E}^m = (\,\boldsymbol{e}_1^m, \ldots, \boldsymbol{e}_m^m\,)$ denote the canonical bases for $\mathbb{F}^{n\times 1}$ and $\mathbb{F}^{m\times 1}$, respectively. Then since $A\boldsymbol{e}_j^n = \boldsymbol{a}_j$ (the jth column of A), and since the column representation of $\boldsymbol{a}_j$ with respect to $\boldsymbol{E}^m$ is itself, it follows that the matrix representation of $\mathcal{A}$ with respect to the canonical bases of $\mathbb{F}^{n\times 1}$ and $\mathbb{F}^{m\times 1}$ is the matrix A itself.

Example 3.35

Consider the linear transformation in Example 3.29. If we choose the canonical bases $\boldsymbol{E}^3 = (\,\boldsymbol{e}_1^3, \boldsymbol{e}_2^3, \boldsymbol{e}_3^3\,)$ and $\boldsymbol{E}^2 = (\,\boldsymbol{e}_1^2, \boldsymbol{e}_2^2\,)$ for $\mathbb{F}^3$ and $\mathbb{F}^2$, then

$$\begin{aligned}
\mathcal{A}(\boldsymbol{e}_1^3) &= \mathcal{A}(1,0,0) &&= (1,\ \ 0) &&= \boldsymbol{e}_1^2 \\
\mathcal{A}(\boldsymbol{e}_2^3) &= \mathcal{A}(0,1,0) &&= (1,\ \ 1) &&= \boldsymbol{e}_1^2 + \boldsymbol{e}_2^2 \\
\mathcal{A}(\boldsymbol{e}_3^3) &= \mathcal{A}(0,0,1) &&= (0,-1) &&= -\boldsymbol{e}_2^2
\end{aligned}$$

and the matrix representation of $\mathcal{A}$ is

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

Now suppose we choose

$$\boldsymbol{r}_1 = (1,0,0)\,, \quad \boldsymbol{r}_2 = (0,1,0)\,, \quad \boldsymbol{r}_3 = (-1,1,1)$$

as a basis for $\mathbb{F}^3$, and

$$\boldsymbol{s}_1 = (1,0)\,, \quad \boldsymbol{s}_2 = (1,1)$$

as a basis for $\mathbb{F}^2$. Then, since

$$\begin{aligned} \mathcal{A}(\boldsymbol{r}_1) &= \mathcal{A}(\ 1,0,0) &= (1,0) &= \boldsymbol{s}_1 \\ \mathcal{A}(\boldsymbol{r}_2) &= \mathcal{A}(\ 0,1,0) &= (1,1) &= \boldsymbol{s}_2 \\ \mathcal{A}(\boldsymbol{r}_3) &= \mathcal{A}(-1,1,1) &= (0,0) &= 0 \end{aligned}$$

$\mathcal{A}$ has the matrix representation

$$A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

with respect to $(\boldsymbol{R}, \boldsymbol{S})$.

Let us form the change–of–basis matrices Q_y and P_x. Since

$$\begin{aligned} \boldsymbol{e}_1^2 &= \boldsymbol{s}_1 \\ \boldsymbol{e}_2^2 &= -\boldsymbol{s}_1 + \boldsymbol{s}_2 \end{aligned}$$

the matrix of change–of–basis from $\boldsymbol{E}^2$ to $\boldsymbol{S}$ in $\mathbb{F}^2$ is

$$Q_y = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

The matrix of change–of–basis from $\boldsymbol{R}$ to $\boldsymbol{E}^3$ in $\mathbb{F}^3$ is readily obtained as

$$P_x = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

The reader can easily verify that $A' = Q_y A P_x$.

In Example 3.31 we have seen that if $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ and $\mathcal{B} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{Z}}$ are linear transformations, then the mapping $\mathcal{C} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Z}}$ defined as

$$\mathcal{C}(\boldsymbol{x}) = (\mathcal{B} \circ \mathcal{A})(\boldsymbol{x}) = \mathcal{B}(\mathcal{A}(\boldsymbol{x}))$$

is also a linear transformation. In particular, if $\mathcal{A} : \mathbb{F}^{n\times 1} \to \mathbb{F}^{m\times 1}$ and $\mathcal{B} : \mathbb{F}^{m\times 1} \to \mathbb{F}^{p\times 1}$ are linear transformations defined as

$$\mathcal{A}(\boldsymbol{x}) = A\boldsymbol{x}\,, \quad \mathcal{B}(\boldsymbol{y}) = B\boldsymbol{y}$$

where A and B are $m \times n$ and $p \times m$ matrices then $\mathcal{C}$ is defined as

$$\mathcal{C}(\boldsymbol{x}) = \mathcal{B}(\mathcal{A}(\boldsymbol{x})) = \mathcal{B}(A\boldsymbol{x}) = BA\boldsymbol{x}$$

Conversely, if $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Z}}$ are finite dimensional and $\mathcal{A}$ and $\mathcal{B}$ are represented by matrices A and B with respect to some fixed bases of $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Z}}$, then $\mathcal{C} = \mathcal{B} \circ \mathcal{A}$ is represented by the matrix $C = BA$ with respect to the same bases (see Exercise 3.29). Thus a matrix product can be viewed as the representation of a linear transformation followed by another as illustrated in Figure 3.4.

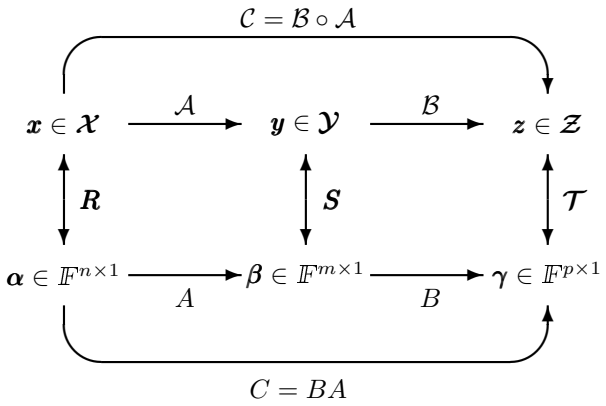

Figure 3.4: An interpretation matrix multiplication

3.4.2 Kernel and Image of a Linear Transformation

Let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ be a linear transformation. The set

$$\ker(\mathcal{A}) = \{\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \mid \mathcal{A}(\boldsymbol{x}) = \mathbf{0} \,\} \subset \boldsymbol{\mathcal{X}}$$

is called the ***kernel*** of $\mathcal{A}$.

Clearly, $\mathbf{0} \in \ker(\mathcal{A})$. Furthermore, if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \ker(\mathcal{A})$ then for any $c_1, c_2 \in \mathbb{F}$

$$\mathcal{A}(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2) = c_1\mathcal{A}(\boldsymbol{x}_1) + c_2\mathcal{A}(\boldsymbol{x}_2) = c_1\mathbf{0} + c_2\mathbf{0} = \mathbf{0}$$

so that $c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 \in \ker(\mathcal{A})$. That is, $\ker(\mathcal{A})$ is closed under vector addition and scalar multiplication. Hence it is a subspace of $\boldsymbol{\mathcal{X}}$, which is also called the ***null space*** of $\mathcal{A}$ and denoted by $\boldsymbol{\mathcal{N}}(\mathcal{A})$. If it is finite dimensional, we define $\nu(\mathcal{A}) = \dim(\ker(\mathcal{A}))$ to be the ***nullity*** of $\mathcal{A}$.

The set

$$\text{im}(\mathcal{A}) = \{\, \boldsymbol{y} \in \boldsymbol{\mathcal{Y}} \mid y = \mathcal{A}(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \,\} \subset \boldsymbol{\mathcal{Y}}$$

is called the ***image*** of $\mathcal{A}$. The reader can easily show that $\text{im}(\mathcal{A})$ is a subspace of $\boldsymbol{\mathcal{Y}}$, which is also called the ***range space*** of $\mathcal{A}$ and denoted as $\boldsymbol{\mathcal{R}}(\mathcal{A})$. If it is finite dimensional, then we define $\rho(\mathcal{A}) = \dim(\text{im}(\mathcal{A}))$ to be the ***rank*** of $\mathcal{A}$.

If $\mathcal{A} : \mathbb{F}^{n\times 1} \to \mathbb{F}^{m\times 1}$ is a linear transformation defined by an $m \times n$ matrix A, then we also use the notation $\ker(A)$ and $\text{im}(A)$ to denote $\ker(\mathcal{A})$ and $\text{im}(\mathcal{A})$.

Example 3.36

Consider the linear transformation $\mathcal{A} : \mathbb{R}^{4\times 1} \to \mathbb{R}^{3\times 1}$ defined by the matrix

$$A = \begin{bmatrix} 1 & 0 & -2 & -2 \\ 1 & -1 & -1 & 1 \\ 0 & -1 & 1 & 3 \end{bmatrix} = [\, \boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \boldsymbol{a}_3 \;\; \boldsymbol{a}_4 \,]$$

Then

$$\ker(\mathcal{A}) = \{\, \boldsymbol{x} \in \mathbb{R}^{4\times 1} \mid A\boldsymbol{x} = \mathbf{0} \,\}$$

that is, $\ker(\mathcal{A})$ is precisely the set of solutions of the homogeneous system $A\boldsymbol{x} = \mathbf{0}$. From the reduced row echelon form of A

$$A \longrightarrow \begin{bmatrix} 1 & 0 & -2 & -2 \\ 0 & 1 & -1 & -3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

we obtain two linearly independent solutions

$$\boldsymbol{\phi}_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{\phi}_2 = \begin{bmatrix} 2 \\ 3 \\ 0 \\ 1 \end{bmatrix}$$

Hence $\ker(\mathcal{A}) = \text{span}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$, and therefore, $\nu(\mathcal{A}) = 2$.

Clearly,

$$\text{im}(\mathcal{A}) = \text{span}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \boldsymbol{a}_4)$$

Performing elementary operations on the columns of A as

$$A \quad \begin{array}{c} 2\mathrm{C}_1 + \mathrm{C}_3 \to \mathrm{C}_3 \\ 2\mathrm{C}_1 + \mathrm{C}_4 \to \mathrm{R}_+ \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 3 \\ 0 & -1 & 1 & 3 \end{bmatrix}$$

$$\begin{array}{c} \mathrm{C}_2 + \mathrm{C}_3 \to \mathrm{C}_3 \\ 3\mathrm{C}_2 + \mathrm{C}_4 \to \mathrm{C}_4 \\ \longrightarrow \end{array} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \quad = \quad [\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \mathbf{0} \;\; \mathbf{0}]$$

we see that $\text{im}(\mathcal{A}) = \text{span}(\boldsymbol{a}_1, \boldsymbol{a}_2)$, and hence $\rho(\mathcal{A}) = 2$.

Example 3.37

Let $\mathcal{T} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ be defined as

$$\mathcal{T}(A) = A + A^t$$

It is easy to see that $\mathcal{T}$ is a linear transformation.

Since $A + A^t = O$ if and only if A is skew-symmetric,

$$\ker(\mathcal{T}) = \mathbb{R}_q^{2\times 2} = \text{span}(Q)$$

where $\mathbb{R}_q^{2\times 2}$ is the subspace in Example 3.22. Hence $\nu(\mathcal{T}) = 1$.

Since $A + A^t$ is symmetric for any A,

$$\text{im}(\mathcal{T}) = \mathbb{R}_s^{2\times 2} = \text{span}(S_1, S_1, S_3)$$

where $\mathbb{R}_s^{2\times 2}$ is the subspace in Example 3.22. Hence $\rho(\mathcal{T}) = 3$.

The kernel and the image of $\mathcal{T}$ can also be characterized using representations. With respect to the basis $\mathcal{E} = \{\, E_{11}, E_{12}, E_{21}, E_{22} \,\}$ of $\mathbb{R}^{2\times 2}$, $\mathcal{T}$ has the representation

$$T = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

It is left to the reader to show that

$$\ker(T) = \text{span}(\boldsymbol{q}) \quad \text{and} \quad \text{im}(T) = \text{span}(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3)$$

where

$$\boldsymbol{q} = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{s}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{s}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{s}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Clearly, $\boldsymbol{q}$ and $\boldsymbol{s}_i$ are the representations of Q and S_i with respect to $\mathcal{E}$.

* 3.4.3 Inverse Transformations

If $\ker(\mathcal{A}) = \{\mathbf{0}\}$ then to every $\boldsymbol{y} \in \text{im}(\mathcal{A})$ there corresponds a unique $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ such that $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}$, that is, $\mathcal{A}$ is one-to-one (see Exercise 3.38). It is then natural to expect that there exists a linear transformation $\hat{\mathcal{A}}_L : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$ that maps the image of every $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ back to $\boldsymbol{x}$ as illustrated in Figure 3.5. Such a linear transformation, if it exists, is called a ***left inverse*** of $\mathcal{A}$.[12]

In general, $\hat{\mathcal{A}}_L$ is not unique because of the arbitrariness in defining $\hat{\mathcal{A}}_L(\boldsymbol{y})$ when $\boldsymbol{y} \notin \text{im}(\mathcal{A})$. However, by the very definition, it has the property that

$$\hat{\mathcal{A}}_L(\mathcal{A}(\boldsymbol{x})) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

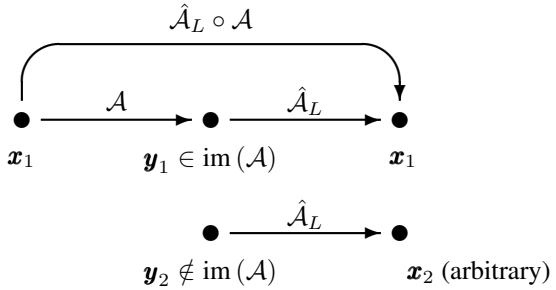

Figure 3.5: Left inverse of a linear transformation

Example 3.38

Let $\mathcal{A} : \mathbb{R}^{2\times 1} \to \mathbb{R}^{3\times 1}$ be defined by the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 0 & 1 \end{bmatrix}$$

[12] The proof of existence of $\hat{\mathcal{A}}_L$ in the general case is beyond the scope of this book. Left inverse of a linear transformation defined by a matrix is studied in Chepter 4.

It can easily be verified that the only solution of $A\boldsymbol{x} = \mathbf{0}$ is the trivial solution $\boldsymbol{x} = \mathbf{0}$, that is, $\ker(A) = \{\mathbf{0}\}$. Let

$$\hat{A}_L = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

Then the mapping $\hat{\mathcal{A}}_L : \mathbb{R}^{3\times 1} \to \mathbb{R}^{2\times 1}$ defined by $\hat{A}_L$ is a left inverse of $\mathcal{A}$, because $\hat{A}_L A = I$ so that

$$\hat{\mathcal{A}}_L(\mathcal{A}(\boldsymbol{x})) = \hat{A}_L A\boldsymbol{x} = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^{2\times 1}$$

The reader can verify that the matrix

$$\hat{A}'_L = \begin{bmatrix} -1 & 1 & -2 \\ 2 & -1 & 2 \end{bmatrix}$$

also defines a left inverse of $\mathcal{A}$.

Now suppose that $\text{im}(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$. Then for every $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ there exists an $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, not necessarily unique, such that $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}$, that is, $\mathcal{A}$ is onto. We can then define a linear transformation $\hat{\mathcal{A}}_R : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$ such that $\hat{\mathcal{A}}_R(\boldsymbol{y}) = \boldsymbol{x}$, where $\boldsymbol{x}$ is any fixed vector that satisfies $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}$. Because of the arbitrariness in choosing $\boldsymbol{x}$ (if there are more than one $\boldsymbol{x}$ that satisfy $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}$), $\hat{\mathcal{A}}_R$ is not unique either. However, it has the property that

$$\mathcal{A}(\hat{\mathcal{A}}_R(\boldsymbol{y})) = \boldsymbol{y}$$

for all $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$. Thus $\hat{\mathcal{A}}_R$ maps every $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ to a vector whose image is $\boldsymbol{y}$ as illustrated in Figure 3.6, and is called a ***right inverse*** of $\mathcal{A}$.[13]

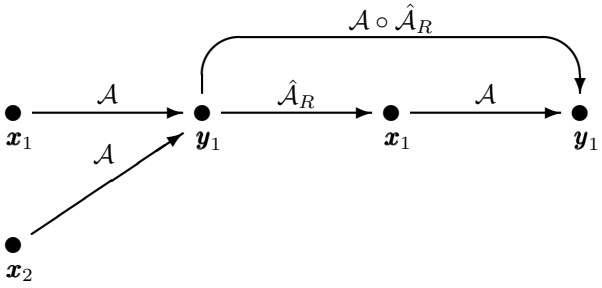

Figure 3.6: Right inverse of a linear transformation

Example 3.39

Let $\mathcal{B} : \mathbb{R}^{3\times 1} \to \mathbb{R}^{2\times 1}$ be defined by the matrix

$$B = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

[13]Right inverse of a linear transformation defined by a matrix is also studied in Chepter 4.

Since $r(B) = 2$, the equation $B\boldsymbol{x} = \boldsymbol{y}$ is consistent for all $\boldsymbol{y}$, that is, $\text{im}\,(B) = \mathbb{R}^{2\times 1}$. Let

$$\hat{B}_R = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}$$

Then the mapping $\hat{\mathcal{B}}_R : \mathbb{R}^{2\times 1} \to \mathbb{R}^{3\times 1}$ defined by $\hat{B}_R$ is a right inverse of $\mathcal{B}$, because $B\hat{B}_R = I$, so that

$$\mathcal{B}(\hat{\mathcal{B}}_R(\boldsymbol{y})) = B\hat{B}_R\boldsymbol{y} = \boldsymbol{y} \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^{2\times 1}$$

As an exercise the reader may try to find a different right inverse of $\mathcal{B}$.

Finally, suppose that $\ker(\mathcal{A}) = \{\mathbf{0}\}$ and also $\text{im}\,(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$. Then $\mathcal{A}$ has both a left inverse $\hat{\mathcal{A}}_L$ and a right inverse $\hat{\mathcal{A}}_R$. Moreover, $\hat{\mathcal{A}}_L$ and $\hat{\mathcal{A}}_R$ are unique.[14] Since $\mathcal{A}(\hat{\mathcal{A}}_R(\boldsymbol{y})) = \boldsymbol{y}$ for all $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, from the definition of left inverse it follows that

$$\hat{\mathcal{A}}_L(\boldsymbol{y}) = \hat{\mathcal{A}}_L(\mathcal{A}(\hat{\mathcal{A}}_R(\boldsymbol{y}))) = \hat{\mathcal{A}}_R(\boldsymbol{y}) \quad \text{for all } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$$

that is, $\hat{\mathcal{A}}_L = \hat{\mathcal{A}}_R$. The unique common left and right inverse of $\mathcal{A}$ is simply called the ***inverse*** of $\mathcal{A}$, denoted $\mathcal{A}^{-1}$.

In summary, if $\ker(\mathcal{A}) = \{\mathbf{0}\}$ and $\text{im}\,(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$ for a linear transformation $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$, then there exists a unique inverse transformation $\mathcal{A}^{-1} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$ such that

$$\mathcal{A}^{-1}(\mathcal{A}(\boldsymbol{x})) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

and

$$\mathcal{A}(\mathcal{A}^{-1}(\boldsymbol{y})) = \boldsymbol{y} \quad \text{for all } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$$

This is somewhat an expected result, because if $\ker(\mathcal{A}) = \{\mathbf{0}\}$ and $\text{im}\,(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$ then $\mathcal{A}$ establishes a one-to-one correspondence between the elements of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$. Such a linear transformation is called an ***isomorphism***, and any two vector spaces related by an isomorphism are called ***isomorphic***. It is left to the reader to show that two finite dimensional vector spaces $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ are isomorphic if and only if $\dim(\boldsymbol{\mathcal{X}}) = \dim(\boldsymbol{\mathcal{Y}})$ (Exercise 3.39).

Example 3.40

Let $\dim(\boldsymbol{\mathcal{X}}) = n$, and let $\boldsymbol{R}$ be a basis for $\boldsymbol{\mathcal{X}}$. Let the unique representation of a vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ with respect to the basis $\boldsymbol{R}$ be $\boldsymbol{\alpha}_x \in \mathbb{F}^{n\times 1}$. Then the mapping $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \mathbb{F}^{n\times 1}$ defined as

$$\mathcal{A}(\boldsymbol{x}) = \boldsymbol{\alpha}_x$$

is a linear transformation as can easily be shown using the definition of the column representation of a vector.

Since $\boldsymbol{\alpha}_x$ is uniquely defined by $\boldsymbol{x}$, $\boldsymbol{x} = \mathbf{0}$ is the only vector whose representation is $\mathbf{0}_{n\times 1}$. Hence $\ker(\mathcal{A}) = \{\mathbf{0}\}$. Also since every $\boldsymbol{\alpha} \in \mathbb{F}^{n\times 1}$ is the representation of some $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\text{im}\,(\mathcal{A}) = \mathbb{F}^{n\times 1}$. Thus $\mathcal{A}$ is a one-to-one mapping from $\boldsymbol{\mathcal{X}}$ onto $\mathbb{F}^{n\times 1}$ (an isomorphism). The inverse of $\mathcal{A}$ is defined as $\mathcal{A}^{-1}(\boldsymbol{\alpha}_x) = \boldsymbol{x}$.

[14] A rigorous proof of this statement is beyond the scope of this book. However, we can argue that since there exists no $y \notin \text{im}\,(\mathcal{A})$, there is no arbitrariness in $\hat{\mathcal{A}}_L$. Also, since for any $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, the vector $\boldsymbol{x}$ that satisfies $\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y}$ is unique, there is no arbitrariness in $\hat{\mathcal{A}}_R$ either.

3.5 Linear Equations

In Chapter 1 we considered linear systems of the form

$$A\boldsymbol{x} = \boldsymbol{b}$$

where A is an $m \times n$ matrix. We have seen that a general solution is of the form

$$\boldsymbol{x} = \boldsymbol{\phi}_p + \boldsymbol{\phi}_c$$

where $\boldsymbol{x} = \boldsymbol{\phi}_p$ is a particular solution, and $\boldsymbol{x} = \boldsymbol{\phi}_c$ is a complementary solution that contains arbitrary constants and satisfies the associated homogeneous equation.

In Chapter 2 we considered first and second order linear differential equations of the form

$$L(\mathcal{D})(y) = u(t)$$

where $L(\mathcal{D})$ is a linear differential operator with constant coefficients. Again, the solution is of the form

$$y = \phi_p(t) + \phi_c(t)$$

where $y = \phi_p$ is a particular solution, and $y = \phi_c$ is a complementary solution.

The similarities between the nature of solutions of linear differential equations and linear systems are striking but not surprising. Both a matrix and a linear differential operator are linear transformations, and both a linear system and a linear differential equation can be viewed as an equation

$$\mathcal{A}(\boldsymbol{x}) = \boldsymbol{y} \tag{3.13}$$

where $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ is a linear transformation and $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ is a given vector. In the case of linear systems $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ are $\mathbb{F}^{n\times 1}$ and $\mathbb{F}^{m\times 1}$, and in the case of linear differential equations, the set of real-valued piece-wise continuous functions defined on some interval. We now recall some definitions of Chapter 1 and Chapter 2.

An equation of the form (3.13), where $\mathcal{A}$ is a linear transformation, is called a ***linear equation***. If $\boldsymbol{y} = \boldsymbol{0}$ then the equation is ***homogeneous***. A vector $\boldsymbol{x} = \boldsymbol{\phi}$ is called a ***solution*** of (3.13) if $\mathcal{A}(\boldsymbol{\phi}) = \boldsymbol{y}$. If (3.13) has no solution, it is said to be ***inconsistent***. Clearly, (3.13) is consistent if and only if $\boldsymbol{y} \in \text{im}\,(\mathcal{A})$.

Consider the homogeneous linear equation

$$\mathcal{A}(\boldsymbol{x}) = \boldsymbol{0} \tag{3.14}$$

which is consistent as $\boldsymbol{x} = \boldsymbol{0}$ is a trivial solution. Clearly, the set of all solutions of (3.14) is $\ker(\mathcal{A})$. Suppose that $\dim(\ker(\mathcal{A})) = \nu(\mathcal{A}) = \nu$, and let $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu\}$ be a basis for $\ker(\mathcal{A})$. Then any solution of (3.14) can be expressed in terms of the basis vectors as

$$\boldsymbol{x} = c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu \tag{3.15}$$

for some choice of the constants $c_1, \ldots, c_\nu$.

Now consider the non-homogeneous linear equation (3.13). Assume that $\boldsymbol{y} \in \text{im}\,(\mathcal{A})$, so that it has at least one particular solution $\boldsymbol{x} = \boldsymbol{\phi}_p$. Then for arbitrary $c_1, \ldots, c_\nu$

$$\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu \tag{3.16}$$

is also a solution, because

$$\mathcal{A}(\boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + \cdots + c_\nu\boldsymbol{\phi}_\nu) = \mathcal{A}(\boldsymbol{\phi}_p) + \sum_{i=1}^{\nu} c_i\mathcal{A}(\boldsymbol{\phi}_i) = \boldsymbol{y} + \sum_{i=1}^{\nu} c_i\mathbf{0} = \boldsymbol{y}$$

Conversely, if $\boldsymbol{x} = \boldsymbol{\phi}$ is any solution of (3.13), then since

$$\mathcal{A}(\boldsymbol{\phi} - \boldsymbol{\phi}_p) = \mathcal{A}(\boldsymbol{\phi}) - \mathcal{A}(\boldsymbol{\phi}_p) = \boldsymbol{y} - \boldsymbol{y} = \mathbf{0}$$

$\boldsymbol{\phi} - \boldsymbol{\phi}_p$ is a solution of the associated homogeneous equation (3.14), and therefore, can be expressed as in (3.15). This, in turn, implies that $\boldsymbol{\phi}$ is of the form (3.16). Thus (3.16) characterizes the solution set of (3.13), and is called a ***general solution***.

Note that neither ϕ_p nor $\phi_i, i = 1, \ldots, \nu$, are unique. If ϕ'_p is another particular solution, and $\phi'_i, i = 1, \ldots, \nu$, form another basis for ker $(\mathcal{A})$, then

$$\boldsymbol{x} = \phi'_p + c'_1\boldsymbol{\phi}'_1 + \cdots + c'_\nu\boldsymbol{\phi}'_\nu \tag{3.17}$$

is also a general solution. Although the expressions in (3.16) and (3.17) are different, they nevertheless define the same family of solutions (see Exercise 3.41).

Example 3.41

Consider the linear system

$$[\,1\;\;2\;\;3\,]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 2$$

whose coefficient matrix is already in reduced row echelon form.

Following the standard procedure of Chapter 1, a general solution is obtained as

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + c_1\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + c_2\begin{bmatrix} -3 \\ 0 \\ 1 \end{bmatrix}$$

where $\boldsymbol{\phi}_p = \text{col}\,[\,2, 0, 0\,]$ is a particular solution, and $\boldsymbol{\phi}_1 = \text{col}\,[\,-2, 1, 0\,]$ and $\boldsymbol{\phi}_2 = \text{col}\,[\,-3, 0, 1\,]$ form a basis for the kernel of the coefficient matrix.

On the other hand, $\boldsymbol{\phi}'_p = \text{col}\,[\,0, 1, 0\,]$ is also a particular solution (obtained from the general solution above by choosing $c_1 = 1$ and $c_2 = 0$), and $\boldsymbol{\phi}'_1 = \text{col}\,[\,-2, 1, 0\,]$ and $\boldsymbol{\phi}'_2 = \text{col}\,[\,0, -3, 2\,]$ form another basis for the kernel of the coefficient matrix. Thus

$$\boldsymbol{x} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c'_1\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + c'_2\begin{bmatrix} 0 \\ -3 \\ 2 \end{bmatrix}$$

is also a general solution.

The reader can verify that any solution obtained from the second expression by choosing arbitrary values for c'_1 and c'_2 can also be obtained from the first expression by choosing $c_1 = 1 + c'_1 - c'_2$ and $c_2 = 2c'_2$, and vice versa.

Linear equations of the form (3.13) are not limited to linear differential equations and linear systems. In the following two examples, we consider different types of linear equations.

Example 3.42

Consider the linear equation

$$\mathcal{T}(A) = A + A^t = S = \begin{bmatrix} 6 & 2 \\ 2 & -4 \end{bmatrix}$$

where $\mathcal{T}$ is the linear transformation considered in Example 3.37. Since the matrix on the right-hand side of the above equation is symmetric, it is in $\text{im}\,(\mathcal{T})$, and hence, the equation is consistent.

A particular solution can be obtained by inspection to be

$$A_p = \frac{1}{2}\,S = \begin{bmatrix} 3 & 1 \\ 1 & -2 \end{bmatrix}$$

(Since S is symmetric, then so is A_p, and hence $A_p + A_p^t = 2A_p = S$.)

The general solution can then be obtained by complementing A_p with $\ker(\mathcal{T})$, which has already been characterized in Example 3.37, as

$$A = \begin{bmatrix} 3 & 1 \\ 1 & -2 \end{bmatrix} + c \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Thus

$$A = \begin{bmatrix} 3 & 0 \\ 2 & -2 \end{bmatrix}$$

is also a solution obtained from the general solution with $c = 1$. Indeed

$$A + A^t = \begin{bmatrix} 3 & 0 \\ 2 & -2 \end{bmatrix} + \begin{bmatrix} 3 & 2 \\ 0 & -2 \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 2 & -4 \end{bmatrix} = S$$

Example 3.43

Suppose that we are interested in finding a sequence $f \in \mathcal{F}(\mathbb{N}, \mathbb{C})$ which satisfies an equation of the form

$$f[k+2] + a_1 f[k+1] + a_2 f[k] = u[k]\,, \quad k \geq 1 \tag{3.18}$$

where a_1, a_2 are fixed (complex) coefficients, and $u \in \mathcal{F}(\mathbb{N}, \mathbb{C})$ is a given sequence. Such an equation is called a (second order) ***difference equation***.[15]

Obtaining a solution to a difference equation is easy: Choose $f[1]$ and $f[2]$ arbitrarily, and calculate $f[3], f[4]$, etc., recursively from (3.18). Thus

$$\begin{aligned} f[1] &= c_1 \\ f[2] &= c_2 \\ f[3] &= -a_2 f[1] - a_1 f[2] + u[1] &&= -a_2 c_1 - a_1 c_2 + u[1] \\ f[4] &= -a_2 f[2] - a_1 f[3] + u[2] &&= (a_1 a_2) c_1 + (a_1^2 - a_2) c_2 + u[2] - a_1 u[1] \end{aligned}$$

and so on. Certainly, any term of a solution sequence can be obtained after sufficient number of substitutions. However, it would be useful to have a formula for the kth term, which could be evaluated without working out all the intermediate terms.

Let us try to formulate the problem as a linear equation. For this purpose we define a shift operator $\Delta : \mathcal{F}(\mathbb{N}, \mathbb{C}) \to \mathcal{F}(\mathbb{N}, \mathbb{C})$ as

$$(\Delta f)[k] = f[k+1]\,, \quad k \in \mathbb{N}$$

[15] Note that the recursion relations we considered in Section 2.8 in connection with numerical solution of differential equations are difference equations with specified initial conditions.

Defining Δ^2, Δ^3, etc., similar to the powers of the differential operator $\mathcal{D}$, the difference equation in (3.18) can be expressed as

$$L(\Delta)(f) = u[k] \tag{3.19}$$

where

$$L(\Delta) = \Delta^2 + a_1\Delta + a_2\mathcal{I}$$

is a polynomial shift operator on $\mathcal{F}(\mathbb{N}, \mathbb{C})$.

It is left to the reader to show that $L(\Delta)$ is a linear operator on $\mathcal{F}(\mathbb{N}, \mathbb{C})$. Hence the linear difference equation (3.18) is a linear equation. Then it must have a general solution of the form

$$f = \phi_p[k] + \phi_c[k]$$

where ϕ_p is any particular solution sequence, and ϕ_c is a complementary solution sequence expressed as a linear combination of the basis vectors of $\ker(L(\Delta))$.

To be more specific, let us consider the difference equation

$$f[k+2] - \frac{5}{6}f[k+1] + \frac{1}{6}f[k] = 1\,, \quad k \in \mathbb{N}$$

or in operator notation

$$(\Delta^2 - \frac{5}{6}\Delta + \frac{1}{6}\mathcal{I})(f) = 1\,, \quad k \in \mathbb{N}$$

Since the right-hand side of the given equation is a constant, we suspect that a constant sequence $f[k] = C$ might be a solution. Substituting the assumed solution into the equation, and noting that $\Delta^2(C) = \Delta(C) = C$, we get

$$C - (5/6)C + (1/6)C = (1/3)C = 1$$

giving $C = 3$. A particular solution is thus obtained as $\phi_p[k] = 3$.

To find the complementary solution, we try a sequence of the form $f[k] = z^k$. Substituting the trial solution into the homogeneous equation, we get

$$L(\Delta)(z^k) = (z^{k+2} - (5/6)z^{k+1} + (1/6)z^k) = 0$$

or equating the corresponding terms

$$z^{k+2} - (5/6)z^{k+1} + (1/6)z^k = z^k(z^2 - (5/6)z + (1/6)) = 0$$

From the last equation, we observe that $\phi[k] = z^k$ is a solution if and only if z is a root of the characteristic equation

$$z^2 - (5/6)z + (1/6) = 0$$

Since the characteristic equation has two real roots, $\mu_1 = 1/2$ and $\mu_2 = 1/3$, each of the sequences

$$\phi_1[k] = (1/2^k) \quad \text{and} \quad \phi_2[k] = (1/3^k)$$

is a solution of the homogeneous equation. It can be shown that ϕ_1 and ϕ_2 are linearly independent sequences, and thus form a basis for $\ker(L(\Delta))$. The general solution of the non-homogeneous difference equation is thus obtained as

$$f = \phi[k] = 3 + c_1/2^k + c_2/3^k\,, \quad c_1, c_2 \in \mathbb{C}$$

* 3.6 Direct Sums and Projections

Consider an arbitrary vector $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$. In terms of the canonical basis vectors $\boldsymbol{e}_1 = (1, 0, 0)$, $\boldsymbol{e}_2 = (0, 1, 0)$ and $\boldsymbol{e}_3 = (0, 0, 1)$, $\boldsymbol{x}$ can be expressed uniquely as

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + x_3\boldsymbol{e}_3 \tag{3.20}$$

Let $\boldsymbol{\mathcal{U}}_i = \text{span}\,(\boldsymbol{e}_i), i = 1, 2, 3$, be the one-dimensional subspaces of $\mathbb{R}^3$ defined by the canonical basis vectors. (They represent the x, y and z axes in the the xyz space). Then we can interpret (3.20) as

$$\boldsymbol{x} = \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3$$

which is a decomposition of $\boldsymbol{x}$ into three components $\boldsymbol{u}_1 = x_1\boldsymbol{e}_1, \boldsymbol{u}_2 = x_2\boldsymbol{e}_2, \boldsymbol{u}_3 = x_3\boldsymbol{e}_3$ in the subspaces $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \boldsymbol{\mathcal{U}}_3$.

Now let $\boldsymbol{\mathcal{V}}_1 = \text{span}\,(\boldsymbol{e}_1, \boldsymbol{e}_2)$ and $\boldsymbol{\mathcal{V}}_2 = \text{span}\,(\boldsymbol{e}_3)$, where $\boldsymbol{\mathcal{V}}_1$ is a two-dimensional subspace (the xy plane) and $\boldsymbol{\mathcal{V}}_2$ is a one-dimensional subspace (the z axis). Then (3.20) can also be written as

$$\boldsymbol{x} = \boldsymbol{v}_1 + \boldsymbol{v}_2$$

which gives a decomposition of $\boldsymbol{x}$ into two components $\boldsymbol{v}_1 = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2$ and $\boldsymbol{v}_2 = x_3\boldsymbol{e}_3$ in the subspaces $\boldsymbol{\mathcal{V}}_1$ and $\boldsymbol{\mathcal{V}}_2$. Obviously, these components are uniquely determined by $\boldsymbol{x}$ and the subspaces $\boldsymbol{\mathcal{V}}_1$ and $\boldsymbol{\mathcal{V}}_2$.

This example suggests that the idea of decomposing a vector into components along the one-dimensional subspaces defined by a given basis can be generalized to a decomposition into components in higher dimensional subspaces.

Let $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \ldots, \boldsymbol{\mathcal{U}}_k$ be subspaces of a vector space $\boldsymbol{\mathcal{X}}$. Their ***algebraic sum*** is defined as

$$\sum_{i=1}^{k} \boldsymbol{\mathcal{U}}_i = \{\, \boldsymbol{x} \,|\, \boldsymbol{x} = \sum_{i=1}^{k} \boldsymbol{u}_i, \boldsymbol{u}_i \in \boldsymbol{\mathcal{U}}_i, i = 1, \ldots, k \,\}$$

It is easy to show that the algebraic sum is also a subspace of $\boldsymbol{\mathcal{X}}$.

The subspaces $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \ldots, \boldsymbol{\mathcal{U}}_k$ are said to be linearly independent if for $\boldsymbol{u}_i \in \boldsymbol{\mathcal{U}}_i, i = 1, \ldots, k$

$$\sum_{i=1}^{k} \boldsymbol{u}_i = \boldsymbol{0}$$

is satisfied only when $\boldsymbol{u}_i = \boldsymbol{0}, i = 1, \ldots, k$.

If $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2, \ldots, \boldsymbol{\mathcal{U}}_k$ are said to be linearly independent their algebraic sum is called a ***direct sum***, denoted $\bigoplus_{i=1}^{k} \boldsymbol{\mathcal{U}}_i$.

Note that algebraic sum of a family of subspaces is a generalization of the concept of span of a set of vectors. Similarly, linear independence of a family of subspaces is a generalization of the concept of linear independence of a set of vectors. The following theorem characterizes two linearly independent subspaces. The extension of the theorem to more than two subspaces is left to the reader as an exercise (see Exercise 3.43).

Theorem 3.2 *Let $\boldsymbol{R} = \{\, \boldsymbol{r}_1, \boldsymbol{r}_2, \ldots \,\}$ and $\boldsymbol{S} = \{\, \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots \,\}$ be bases for the subspaces $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$. Then the following are equivalent.*

a) $\boldsymbol{\mathcal{U}}$ *and* $\boldsymbol{\mathcal{V}}$ *are linearly independent.*

b) $\boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}} = \{\mathbf{0}\}$.

c) $\boldsymbol{T} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots\}$ *is a basis for* $\boldsymbol{\mathcal{U}} + \boldsymbol{\mathcal{V}}$.

Proof We will show that (a) ⇒ (b) ⇒ (c) ⇒ (a).

(a) ⇒ (b):

By contradiction. Suppose that there exists $\mathbf{0} \neq \boldsymbol{x} \in \boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}}$. Let $\boldsymbol{u} = \boldsymbol{x} \in \boldsymbol{\mathcal{U}}$ and $\boldsymbol{v} = -\boldsymbol{x} \in \boldsymbol{\mathcal{V}}$. Then $\boldsymbol{u} \neq \mathbf{0} \neq \boldsymbol{v}$ and $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x} = \mathbf{0}$.

(b) ⇒ (c):

By contradiction. Since span $(\boldsymbol{T}) = \boldsymbol{\mathcal{U}} + \boldsymbol{\mathcal{V}}$, if $\boldsymbol{T}$ is not a basis for $\boldsymbol{\mathcal{U}} + \boldsymbol{\mathcal{V}}$ then it must be linearly dependent, in which case there exist scalars α_i and β_j, not all zero, such that

$$\sum_i \alpha_i \boldsymbol{r}_i + \sum_j \beta_j \boldsymbol{s}_j = \mathbf{0}$$

Let

$$\boldsymbol{x} = \sum_i \alpha_i \boldsymbol{r}_i = -\sum_j \beta_j \boldsymbol{s}_j$$

Then $\boldsymbol{x} \in \boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}}$, and $\boldsymbol{x} \neq \mathbf{0}$ as $\boldsymbol{R}$ and $\boldsymbol{S}$ are linearly independent.

(c) ⇒ (a):

Suppose $\boldsymbol{u} + \boldsymbol{v} = \mathbf{0}$ for some

$$\boldsymbol{u} = \sum_i \alpha_i \boldsymbol{r}_i \quad \text{and} \quad \boldsymbol{v} = \sum_j \beta_j \boldsymbol{s}_j$$

that is,

$$\sum_i \alpha_i \boldsymbol{r}_i + \sum_j \beta_j \boldsymbol{s}_j = \mathbf{0}$$

Since $\boldsymbol{T}$ is a basis, we must have $\alpha_i = 0$ for all i implying $\boldsymbol{u} = \mathbf{0}$, and also $\beta_j = 0$ for all j implying $\boldsymbol{v} = \mathbf{0}$.

Example 3.44

In $\mathbb{R}^3$, let

$$\boldsymbol{\mathcal{U}}_1 = \text{span}\,(\boldsymbol{e}_1, \boldsymbol{e}_2), \quad \boldsymbol{\mathcal{U}}_2 = \text{span}\,(\boldsymbol{e}_2, \boldsymbol{e}_3), \quad \boldsymbol{\mathcal{U}}_3 = \text{span}\,(\boldsymbol{e}_3)$$

where $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ are the canonical basis vectors. Then

$$\boldsymbol{\mathcal{U}}_1 + \boldsymbol{\mathcal{U}}_2 = \boldsymbol{\mathcal{U}}_1 + \boldsymbol{\mathcal{U}}_3 = \text{span}\,(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3) = \mathbb{R}^3$$

$\boldsymbol{\mathcal{U}}_1$ and $\boldsymbol{\mathcal{U}}_2$ are not linearly independent, because $\boldsymbol{e}_2 \in \boldsymbol{\mathcal{U}}_1 \cap \boldsymbol{\mathcal{U}}_2$. However, $\boldsymbol{\mathcal{U}}_1$ and $\boldsymbol{\mathcal{U}}_3$ are linearly independent, because $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ is a basis for $\boldsymbol{\mathcal{U}}_1 + \boldsymbol{\mathcal{U}}_3$. Hence

$$\mathbb{R}^3 = \boldsymbol{\mathcal{U}}_1 \oplus \boldsymbol{\mathcal{U}}_3$$

The reader can interpret these findings by identifying $\mathbb{R}^3$ with the xyz space, and $\boldsymbol{\mathcal{U}}_1, \boldsymbol{\mathcal{U}}_2$ and $\boldsymbol{\mathcal{U}}_3$ with the xy plane, yz plane, and the z axis respectively.

From Theorem 3.2 it is clear that if $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$ is a basis for $\boldsymbol{\mathcal{X}}$, partitioned arbitrarily into two disjoint sets, say

$$\boldsymbol{R} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_k\} \quad \text{and} \quad \boldsymbol{S} = \{\boldsymbol{t}_{k+1}, \boldsymbol{t}_{k+2}, \ldots, \boldsymbol{t}_n\}$$

then

$$\boldsymbol{\mathcal{X}} = \text{span}\,(\boldsymbol{R}) \oplus \text{span}\,(\boldsymbol{S})$$

On the other hand, if $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k\}$ is a basis for a k-dimensional subspace $\boldsymbol{\mathcal{U}}$, then by Corollary 3.2, it can be completed to a basis by including $n - k$ more vectors. Let these additional vectors form a set $\boldsymbol{S}$, and let $\boldsymbol{\mathcal{V}} = \text{span}\,(\boldsymbol{S})$. Then $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}}$. The subspace $\boldsymbol{\mathcal{V}}$ thus constructed is called a ***complement*** of $\boldsymbol{\mathcal{U}}$. Since $\boldsymbol{S}$ can be chosen in many different ways, complement of $\boldsymbol{\mathcal{U}}$ is not unique. For example, for any vector $\boldsymbol{v} = (a, b, c)$ with $c \neq 0$, $\boldsymbol{\mathcal{V}} = \text{span}\,(\boldsymbol{v})$ is a complement of $\boldsymbol{\mathcal{U}}_1$ in Example 3.44, and in particular, so is $\boldsymbol{\mathcal{U}}_3$. However, all complements of $\boldsymbol{\mathcal{U}}$ must have the same dimension.

We also observe that if $\boldsymbol{\mathcal{X}}$ is finite dimensional and is decomposed into a direct sum as $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}}$ then

$$\dim\,(\boldsymbol{\mathcal{X}}) = \dim\,(\boldsymbol{\mathcal{U}}) + \dim\,(\boldsymbol{\mathcal{V}})$$

Let $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}}$, and let $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k)$ and $\boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-k})$ be ordered bases for $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$. Since $\boldsymbol{T} = \boldsymbol{R} \cup \boldsymbol{S}$ is a basis for $\boldsymbol{\mathcal{X}}$, any vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ can be expressed as

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{r}_i + \sum_{j=1}^{n-k} \beta_j \boldsymbol{s}_j = \boldsymbol{u} + \boldsymbol{v} \tag{3.21}$$

The vectors $\boldsymbol{u} \in \boldsymbol{\mathcal{U}}$ and $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}$, which are uniquely defined by $\boldsymbol{x}$, are called the ***components*** of $\boldsymbol{x}$ in $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$.

Let $\mathcal{P} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be a mapping which maps every vector in $\boldsymbol{\mathcal{X}}$ to its component in $\boldsymbol{\mathcal{U}}$, that is,

$$\mathcal{P}(\boldsymbol{x}) = \boldsymbol{u}$$

$\mathcal{P}$ is called a ***projection*** on $\boldsymbol{\mathcal{U}}$ along $\boldsymbol{\mathcal{V}}$. It is left to the reader to show that

a) $\mathcal{P}$ is a linear transformation

b) $\text{im}\,(\mathcal{P}) = \boldsymbol{\mathcal{U}}$

c) $\ker\,(\mathcal{P}) = \boldsymbol{\mathcal{V}}$

The reader should note that the mapping $\mathcal{Q} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ defined as $\mathcal{Q}(\boldsymbol{x}) = \boldsymbol{v}$ is also a projection (on $\boldsymbol{\mathcal{V}}$ along $\boldsymbol{\mathcal{U}}$).

The matrix P that represents $\mathcal{P}$ with respect to the basis $\boldsymbol{T}$ is called a ***projection matrix***. Since

$$\mathcal{P}(\mathcal{P}(\boldsymbol{x})) = \mathcal{P}(\boldsymbol{u}) = \boldsymbol{u} = \mathcal{P}(\boldsymbol{x})$$

for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, it follows that $P^2\boldsymbol{\alpha} = P\boldsymbol{\alpha}$ for any $\boldsymbol{\alpha} \in \mathbb{F}^{n\times 1}$ that stands for the column representation of some vector. Thus if P is a projection matrix then $P^2 = P$. Such a matrix is called ***idempotent***. Conversely, if P is an $n \times n$ idempotent matrix, then

$$\mathbb{F}^{n\times 1} = \text{im}\,(P) \oplus \ker\,(P)$$

and P defines a projection in $\mathbb{F}^{n\times 1}$ on $\text{im}\,(P)$ along $\ker\,(P)$ (see Exercise 3.51).

Example 3.45

In Example 3.44, the projection on $\boldsymbol{\mathcal{U}}_1$ along $\boldsymbol{\mathcal{U}}_3$ is defined by the matrix

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and the projection on $\boldsymbol{\mathcal{U}}_3$ along $\boldsymbol{\mathcal{U}}_1$ is defined by the matrix

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Note that $P^2 = P$ and $Q^2 = Q$.

Now consider the matrix

$$R = \frac{1}{3}\begin{bmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix}$$

Since R is idempotent, it defines a projection in $\mathbb{R}^{3\times 1}$ on

$$\text{im}\,(R) = \text{span}\,(\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}) = \text{span}\,(\,\boldsymbol{u}_1, \boldsymbol{u}_2\,)$$

along

$$\ker\,(R) = \text{span}\,(\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}) = \text{span}\,(\,\boldsymbol{u}_3\,)$$

The components of an arbitrary vector $\boldsymbol{x} = \text{col}\,[\,a, b, c\,]$ in $\text{im}\,(R)$ and $\ker\,(R)$ are

$$\boldsymbol{u} = R\boldsymbol{x} = \frac{1}{3}\begin{bmatrix} 2a+b-c \\ a+2b+c \\ -a+b+2c \end{bmatrix} = \frac{2a+b-c}{3}\,\boldsymbol{u}_1 + \frac{-a+b+2c}{3}\,\boldsymbol{u}_2$$

and

$$\boldsymbol{v} = (I-R)\boldsymbol{x} = \frac{1}{3}\begin{bmatrix} a-b+c \\ -a+b-c \\ a-b+c \end{bmatrix} = \frac{a-b+c}{3}\,\boldsymbol{u}_3$$

Example 3.46

In $\mathbb{P}_C[s]$, let

$$\begin{aligned} \mathbb{P}_C^e[s] &= \{\, p(s) = p_0 + p_1 s^2 + \cdots + p_m s^{2m} \mid m = 0, 1, \ldots \} \\ \mathbb{P}_C^o[s] &= \{\, q(s) = q_0 s + q_1 s^3 + \cdots + q_m s^{2m+1} \mid m = 0, 1, \ldots \} \end{aligned}$$

Then $\mathbb{P}_C[s] = \mathbb{P}_C^e[s] \oplus \mathbb{P}_C^o[s]$.

If

$$r(s) = r_0 + r_1 s + r_2 s^2 + \cdots + r_{2n+1} s^{2n+1}$$

then

$$r_e(s) = r_0 + r_2 s^2 + \cdots + r_{2n} s^{2n}$$

is the projection of r on $\mathbb{P}_C^e[s]$ along $s\mathbb{P}_C^o[s]$.

This example illustrates that direct sum decomposition and projections are not limited to finite dimensional vector spaces.

3.7 Exercises

1. Prove properties (a-f) on p.85 of a vector space.
2. Show that a plane through the origin is a subspace of $\mathbb{R}^3$.
3. Write down the equation of a line passing through two given points $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^n$. Under what conditions on $\boldsymbol{p}, \boldsymbol{q}$ does the line represent a subspace?
4. Discuss how an $m \times n$ real matrix A can be interpreted as a function $f : \mathbb{D} \to \mathbb{R}^{m \times 1}$.
5. Show that if $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ are subspaces of $\boldsymbol{\mathcal{X}}$, then so is $\boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}}$. Is $\boldsymbol{\mathcal{U}} \cup \boldsymbol{\mathcal{V}}$ also a subspace?
6. Show that span (span $(\boldsymbol{R})$) = span $(\boldsymbol{R})$ for any subset $\boldsymbol{R} \subset \boldsymbol{\mathcal{X}}$.
7. Prove facts (a-c) on p.91 concerning linear independence.
8. Let $\boldsymbol{R}$ be a finite set of vectors and let $\boldsymbol{R}'$ be obtained from $\boldsymbol{R}$ by a single Type I or Type II elementary operation.
   (a) Explain why span $(\boldsymbol{R}')$ = span $(\boldsymbol{R})$.
   (b) Explain why $\boldsymbol{R}'$ is linearly independent if and only if $\boldsymbol{R}$ is.
9. Show that $\mathbb{C}$ is a vector space over $\mathbb{R}$, and find a basis for it.
10. Show that the set of all 3×3 real skew-symmetric matrices is a subspace of $\mathbb{R}^{3\times 3}$, and find a basis for it.
11. Show that the set $\boldsymbol{R} = \{\, r_0, r_1, \dots \,\}$ in Example 3.18 is a basis for $\mathbb{R}[s]$.
12. An $n \times n$ matrix N is said to be ***nilpotent of index*** k if $N^k = O$ but $N^{k-1} \neq O$.
    (a) Let $\boldsymbol{v}$ be such that $N^{k-1}\boldsymbol{v} \neq \boldsymbol{0}$. Show that the vectors $\boldsymbol{v}, N\boldsymbol{v}, \dots, N^{k-1}\boldsymbol{v}$ are linearly independent.
    (b) Show that the index of nilpotency cannot exceed n.
13. Suppose $\boldsymbol{R}$ is linearly independent, and $\boldsymbol{x} \notin$ span $(\boldsymbol{R})$. Show that $\boldsymbol{R} \cup \{\, \boldsymbol{x} \,\}$ is also linearly independent. Hint: Consider a finite subset $\boldsymbol{S} \subset \boldsymbol{R} \cup \{\, \boldsymbol{x} \,\}$. If $\boldsymbol{x} \notin \boldsymbol{S}$ then $\boldsymbol{S} \subset \boldsymbol{R}$, and therefore $\boldsymbol{S}$ must be linearly independent. If $\boldsymbol{x} \in \boldsymbol{S}$ then $\boldsymbol{S} = \{\, \boldsymbol{r}_1, \dots, \boldsymbol{r}_k, \boldsymbol{x} \,\}$ for some $\boldsymbol{r}_1, \dots, \boldsymbol{r}_k \in \boldsymbol{R}$.
14. Since the set $\boldsymbol{R} = \{\, \boldsymbol{r}_1, \dots, \boldsymbol{r}_k, \boldsymbol{r}_{k+1}, \dots, \boldsymbol{r}_{k+n} \,\}$ in the proof of Corollary 3.2(b) is linearly dependent, there exist some $c_1, \dots, c_{n+k}$, not all zero, such that

$$\sum_{i=1}^{n+k} c_i \boldsymbol{r}_i = \boldsymbol{0}$$

    Show that at least one of $c_{k+1}, \dots, c_{n+k}$ must be nonzero. Explain why this implies that the set obtained by deleting the corresponding vector from $\boldsymbol{R}$ includes the first k vectors and still spans $\boldsymbol{\mathcal{X}}$. On the basis of this reasoning, explain also why the algorithm in the proof of Corollary 3.2(a) reduces $\boldsymbol{R}$ to a basis that includes $\boldsymbol{R}_1$.
15. Let $\boldsymbol{\mathcal{U}} =$ span $(\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_p)$ and $\boldsymbol{\mathcal{V}} =$ span $(\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_q)$ be subspaces of $\mathbb{R}^{n\times 1}$. Give an algorithm to obtain a basis for $\boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{V}}$.
16. In a two-dimensional vector space $\boldsymbol{\mathcal{X}}$, a vector x has the representation $\boldsymbol{\alpha} =$ col $[\, 1, -1 \,]$ with respect to some ordered basis $\boldsymbol{R} = (\, \boldsymbol{r}_1, \boldsymbol{r}_2 \,)$. Let

$$\boldsymbol{r}'_1 = \boldsymbol{r}_1 + \boldsymbol{r}_2 \quad \text{and} \quad \boldsymbol{r}'_2 = \boldsymbol{r}_1 + 2\boldsymbol{r}_2$$

    (a) Show that $\boldsymbol{R}' = (\, \boldsymbol{r}'_1, \boldsymbol{r}'_2 \,)$ is also a basis for $\boldsymbol{\mathcal{X}}$.
    (b) Find the matrix of change–of–basis Q from $\boldsymbol{R}$ to $\boldsymbol{R}'$, and the matrix of change–of–basis P from $\boldsymbol{R}'$ to $\boldsymbol{R}$. Verify that $QP = PQ = I$.

(c) Obtain the representation $\boldsymbol{\alpha}'$ of x with respect to $\boldsymbol{R}'$.

17. Refer to Example 3.26. Let $N = 4$.

(a) Find the matrix of change–of–basis Q from (e_p) to (ϕ_p).

(b) Verify that $\boldsymbol{F} = Q\,\boldsymbol{f}$ for the sequence f considered in the example.

18. Let $f = (3, -1, -3, 5, 3, 5) \in \mathcal{F}(\mathbb{D}_6, \mathbb{C})$.

(a) Compute the discrete Fourier coefficients of f.

(b) Verify your result by using the MATLAB commands

```
f=[3 -1 -3 5 3 5];
c=fft(f)
```

19. Refer to Example 3.26.

(a) Show that

$$\psi_p[k] = \begin{cases} 1 \;, & p = 0 \\ \cos\frac{\pi}{2}k \,, & p = 1 \\ \sin\frac{\pi}{2}k \,, & p = 2 \\ \cos\pi k \,, & p = 3 \end{cases}$$

is also a basis for $\mathcal{F}(\mathbb{D}_4, \mathbb{C})$.

(b) Find the representation of f with respect to (ψ_p).

20. Show (3.7) for $f \in \mathcal{F}(\mathbb{D}_N, \mathbb{C})$. Hint: Each of the complex numbers

$$s_p = e^{ip\frac{2\pi}{N}} \;, \quad p = 0, 1, \ldots, N-1$$

satisfies

$$s_p^N = 1$$

Use this fact to show that

$$\sum_{p=0}^{N-1} e^{ip\frac{2\pi}{N}q} = \begin{cases} N \,, & q = 0 \\ 0 \;, & q \neq 0 \end{cases}$$

21. Consider $\mathcal{F}(\mathbb{D}, \mathbb{R})$ with $\mathbb{D} = \{\, 1, 2, 3 \,\}$.

(a) Show that (g_1, g_2, g_3), where

$$\begin{aligned} g_1[k] &= 1 \,, \quad k = 1, 2, 3 \\ g_2[k] &= \begin{cases} 1 \,, & k = 1, 2 \\ 0 \,, & k = 3 \end{cases} \\ g_3[k] &= \begin{cases} 0 \,, & k = 1 \\ 1 \,, & k = 2, 3 \end{cases} \end{aligned}$$

is a basis for $\mathcal{F}(\mathbb{D}, \mathbb{R})$.

(b) Find the column representation of

$$f[k] = k \,, \quad k = 1, 2, 3$$

with respect to (g_1, g_2, g_3).

22. Prove that $\mathcal{F}(\mathbb{N}, \mathbb{R})$ is infinite dimensional. Hint: Assume that it has a finite dimension, say, M, and try to find a subspace of $\mathcal{F}(\mathbb{N}, \mathbb{R})$ with dimension larger than M.

23. Let $\mathcal{A} : \mathbb{R}^{2\times 1} \to \mathbb{R}^{2\times 1}$ be a linear transformation defined by the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Can you find a basis for $\mathbb{R}^{2\times 1}$ with respect to which $\mathcal{A}$ is represented by a diagonal matrix?

24. Let $\mathcal{A} : \mathbb{R}^3 \to \mathbb{R}^2$ be defined as $\mathcal{A}(x, y, z) = (x + y, y - 2z)$. Choose arbitrary bases for $\mathbb{R}^3$ and $\mathbb{R}^2$ (other than the canonical bases) and obtain the matrix representation of $\mathcal{A}$ with respect to these bases.

25. Let

$$\mathbb{P}^3_R[t] = \{\, p(t) = p_0 + p_1 t + p_2 t^2 + p_3 t^3 \,|\, p_0, p_1, p_2, p_3 \in \mathbb{R} \,\}$$

(a) Find a basis $\boldsymbol{R}$ for $\mathbb{P}^3_R[t]$.

(b) Find the column representation of $q(s) = 1 + t^2 - 2t^3$ with respect to $\boldsymbol{R}$.

(c) Let $\mathcal{A} : \mathbb{P}^3_R[t] \to \mathbb{P}^3_R[t]$ be defined as $\mathcal{A}(p) = p'$, where p' denotes the derivative of p with respect to t. Show that $\mathcal{A}$ is a linear transformation, and find its matrix representation with respect to $\boldsymbol{R}$.

26. Let $\mathcal{A} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ be defined as $\mathcal{A}(X) = CX$, where

$$C = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}$$

Find the matrix representation of $\mathcal{A}$ with respect to the basis in Example 3.21.

27. (a) Show that the matrices

$$M_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, M_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, M_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, M_4 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

form an ordered basis $\boldsymbol{\mathcal{M}} = (M_1, M_2, M_3, M_4)$ for $\mathbb{R}^{2\times 2}$.

(b) Let $\mathcal{A} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ be defined as in Example 3.37. Find $Y = \mathcal{A}(X)$ for

$$X = \begin{bmatrix} 1 & 0 \\ 2 & -3 \end{bmatrix}$$

(c) Find the column representations $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of X and Y with respect to $\boldsymbol{\mathcal{M}}$.

(d) Find the matrix representation A of $\mathcal{A}$ with respect to $\boldsymbol{\mathcal{M}}$, and show that $\boldsymbol{\beta} = A\boldsymbol{\alpha}$.

28. In special theory of relativity, the space and time coordinates of an object measured in two coordinate systems moving in the x direction at a constant relative speed are related by the ***Lorentz transformation***

$$\begin{aligned} \mathcal{L}(v): \quad x' &= k_v(x - vt) \\ t' &= k_v(-\frac{v}{c^2}x + t) \end{aligned}$$

where c is the speed of light, $v < c$ is the relative speed of the coordinate systems, and $k_v = 1/\sqrt{1 - (v/c)^2}$.

(a) Show that the Lorentz transformation is a linear transformation from $\mathbb{R}^2$ into itself, mapping (x, t) to (x', t').

(b) Find the matrix representation $L(v)$ of the Lorentz transformation with respect to the canonical basis in $\mathbb{R}^2$.

(c) Show that $L(u)L(v) = L(w)$ for some w, and find w in terms of u and v.

29. Let $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n), \boldsymbol{S} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m)$ and $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_p)$ be bases for the vector spaces $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{Z}}$, respectively, and let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ and $\mathcal{B} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{Z}}$ be linear transformations. Show that if $\mathcal{A}$ and $\mathcal{B}$ are represented by the matrices A and B with respect to the given bases, then $\mathcal{C} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Z}}$ defined as $\mathcal{C}(\boldsymbol{x}) = \mathcal{B}(\mathcal{A}(\boldsymbol{x}))$ is represented by the matrix $C = BA$. Hint: By definition

$$\mathcal{A}(\boldsymbol{r}_j) = \sum_{k=1}^{m} a_{kj}\boldsymbol{s}_k\,, \quad \mathcal{B}(\boldsymbol{s}_k) = \sum_{i=1}^{p} b_{ik}\boldsymbol{t}_i\,, \quad \mathcal{C}(\boldsymbol{r}_j) = \sum_{i=1}^{p} c_{ij}\boldsymbol{t}_i$$

Find an expression for c_{ij} in terms of b_{ik} and a_{kj}.

30. Let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be vector spaces over the same field $\mathbb{F}$.

(a) Show that the set of all linear transformations from $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{Y}}$ is also vector space over $\mathbb{F}$. This vector space is denoted by $\mathcal{L}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$. Define clearly the addition and scalar multiplication operations on $\mathcal{L}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$, as well as the null vector and the additive inverse.

(b) Find $\dim(\mathcal{L}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}))$ if $\dim(\boldsymbol{\mathcal{X}}) = n$ and $\dim(\boldsymbol{\mathcal{Y}}) = m$.

31. In the light of the previous exercise, explain why $L(\mathcal{D})$ is a linear transformation from $\mathcal{C}_n(\boldsymbol{I}, \mathbb{R})$ into $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$.

32. Let

$$f[k] = \begin{cases} 1 + (-1)^k\,, & k \geq 0 \\ 0\quad\;\; , & k < 0 \end{cases}$$

(a) Find the sequence g defined by (3.11) for

$$h[p,q] = \begin{cases} 1/2\,, & q = p \;\text{ or }\; q = p-1 \\ 0\;\;\, , & q \neq p, p-1 \end{cases}$$

Plot f and g pointwise for $-2 \leq k \leq 5$.

(b) Repeat part(a) for

$$h[p,q] = \begin{cases} 1/2\;\;\, , & q = p \\ -1/2\;\;\, , & q = p-1 \\ 0\;\;\, , & q \neq p, p-1 \end{cases}$$

33. Let

$$h(t,\tau) = \begin{cases} 1\,, & t-1 < \tau < t \\ 0\,, & \tau < t-1 \;\text{ or }\; \tau > t \end{cases}$$

and

$$f(t) = \begin{cases} e^{-t}\,, & t > 0 \\ 0\;\;\, , & t < 0 \end{cases}$$

Find the function g defined by (3.12). Plot f and g for $-1 \leq t \leq 5$.

34. Show that if $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ is a linear transformation, then $\text{im}(\mathcal{A})$ is a subspace of $\boldsymbol{\mathcal{Y}}$.

35. Let $\mathcal{A} : \mathbb{R}^3 \to \mathbb{R}^3$ be defined as $\mathcal{A}(x,y,z) = (x+y, x+2z, 2x+y+2z)$. Find bases for $\text{im}(\mathcal{A})$ and $\ker(\mathcal{A})$.

36. Let $\mathcal{A}$ be the linear transformation defined in Exercise 3.25. Find bases for for $\text{im}\,(\mathcal{A})$ and $\ker\,(\mathcal{A})$.

37. Find a 3×3 real matrix A such that $\mathbf{0} \neq \ker\,(A) \subset \text{im}\,(A) \neq \mathbb{R}^{3\times 1}$.

38. Let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ be a linear transformation.

    (a) Show that if $\{\, \mathcal{A}(\boldsymbol{x}_1), \ldots, \mathcal{A}(\boldsymbol{x}_k) \,\}$ is linearly independent, then so is $\{\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \,\}$.

    (b) Show that $\mathcal{A}$ is one-to-one if and only if $\ker\,(\mathcal{A}) = \{\mathbf{0}\}$. Hint: Suppose that corresponding to some $\boldsymbol{y} \in \text{im}\,(\mathcal{A})$ there exist $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ such that $\mathcal{A}(\boldsymbol{x}_1) = \mathcal{A}(\boldsymbol{x}_2) = \boldsymbol{y}$. Let $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2 \neq \mathbf{0}$, and consider $\mathcal{A}(\boldsymbol{x})$.

    (c) Show that if $\mathcal{A}$ is a one-to-one linear transformation, and $\{\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \,\}$ is linearly independent, then $\{\, \mathcal{A}(\boldsymbol{x}_1), \ldots, \mathcal{A}(\boldsymbol{x}_k) \,\}$ is also linearly independent.

39. (a) Let $\dim\,(\boldsymbol{\mathcal{X}}) = n$ and let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ be an isomorphism. Prove that $\dim\,(\boldsymbol{\mathcal{Y}}) = n$. Hint: Let $(\, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n \,)$ be a basis for $\boldsymbol{\mathcal{X}}$. Show that $(\, \mathcal{A}(\boldsymbol{r}_1), \ldots, \mathcal{A}(\boldsymbol{r}_n) \,)$ is a basis for $\boldsymbol{\mathcal{Y}}$.

    (b) Let $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ be vector spaces over $\mathbb{F}$, and let $(\, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n \,)$ and $(\, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \,)$ be bases for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$. Define $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ such that

$$\boldsymbol{x} = \sum_{i=1}^{n} c_i \boldsymbol{r}_i \implies \mathcal{A}(\boldsymbol{x}) = \sum_{i=1}^{n} c_i \boldsymbol{s}_i$$

    Show that $\mathcal{A}$ is an isomorphism.

40. Let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$, where $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ are finite dimensional. Show that there exist bases for $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ with respect to which $\mathcal{A}$ has a matrix representation

$$A = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

41. Let ϕ_p and ϕ_p' be any two particular solutions of (3.13), and let $(\, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\nu \,)$ and $(\, \boldsymbol{\phi}_1', \ldots, \boldsymbol{\phi}_\nu' \,)$ be any two bases for $\ker\,(A)$. Show that

$$\boldsymbol{x} = \boldsymbol{\phi}_p + c_1\boldsymbol{\phi}_1 + c_2\boldsymbol{\phi}_2 + \cdots + c_\nu\boldsymbol{\phi}_\nu$$

    and

$$\boldsymbol{x} = \boldsymbol{\phi}_p' + c_1'\boldsymbol{\phi}_1' + c_2'\boldsymbol{\phi}_2' + \cdots + c_\nu'\boldsymbol{\phi}_\nu'$$

    define the same family of solutions, and therefore, are both general solutions. Hint: Since $\boldsymbol{\phi}_p'$ is a solution, it is a member of the first family. Also, each $\boldsymbol{\phi}_j'$ can be expressed in terms of $\boldsymbol{\phi}_i\,, \quad i = 1, \ldots, \nu$.

42. Show that the polynomial shift operator $L(\Delta)$ in Example 3.43 is a linear operator on the vector space of $\mathcal{F}(\mathbb{N}, \mathbb{C})$.

43. State and prove Theorem 3.2 for more than two subspaces $\boldsymbol{\mathcal{U}}_i, i = 1, \ldots, k$. Hint: Part (b) will be

    (b) $\boldsymbol{\mathcal{U}}_i \cap \sum_{j\neq i} \boldsymbol{\mathcal{U}}_j = \{\, \mathbf{0} \,\}\,, \quad i = 1, \ldots, k$

44. (a) Show that the set $\boldsymbol{\mathcal{U}} = \{\, \text{col}\,[\, x, x, y \,] \mid x, y \in \mathbb{R} \,\}$ is a subspace of $\boldsymbol{R}^{3\times 1}$.

    (b) Find a basis for $\boldsymbol{\mathcal{U}}$. What is the dimension of $\boldsymbol{\mathcal{U}}$?

    (c) Obtain the representation of the vector $\boldsymbol{u} = \text{col}\,[\, 1, 1, 2 \,] \in \boldsymbol{\mathcal{U}}$ with respect to the basis chosen in (b).

    (d) Characterize another subspace $\boldsymbol{\mathcal{V}}$ such that $\boldsymbol{R}^{3\times 1} = \boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}}$

45. Let $\boldsymbol{\mathcal{U}} = \text{span}\,(\boldsymbol{u}_1, \boldsymbol{u}_2)$, where

$$\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

(a) Characterize a subspace $\boldsymbol{\mathcal{V}}$ such that $\boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}} = \mathbb{R}^{3\times 1}$.

(b) Find the projection of $\boldsymbol{x} = \text{col}\,[\,0, 2, 1\,]$ on $\boldsymbol{\mathcal{U}}$ along $\boldsymbol{\mathcal{V}}$.

(c) Find a matrix P such that for any $\boldsymbol{x} \in \mathcal{R}^{3\times 1}$, $P\boldsymbol{x}$ is the projection of $\boldsymbol{x}$ on $\boldsymbol{\mathcal{U}}$.

46. Let $\boldsymbol{\mathcal{U}} = \text{span}\,(\boldsymbol{u}_1, \boldsymbol{u}_2)$, $\boldsymbol{\mathcal{V}} = \text{span}\,(\boldsymbol{v}_1, \boldsymbol{v}_2)$, where

$$\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

(a) Show that $\boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}} = \mathbb{R}^{4\times 1}$.

(b) Find the projection of $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3, x_4\,]$ on $\boldsymbol{\mathcal{U}}$ along $\boldsymbol{\mathcal{V}}$.

(c) Construct a matrix P such that for any $\boldsymbol{x} \in \mathbb{R}^{4\times 1}$, the projection of $\boldsymbol{x}$ on $\boldsymbol{\mathcal{U}}$ along $\boldsymbol{\mathcal{V}}$ is $P\boldsymbol{x}$.

47. Let $\boldsymbol{\mathcal{X}}$ be the set of all semi-infinite sequences $f \in \mathcal{F}(\mathbb{N}, \mathbb{R})$ such that

$$f[k+2] = f[k] + f[k+1]\,, \quad k \in \mathbb{N}$$

Such a sequence is known as a ***Fibonacci sequence***.

(a) Show that $\boldsymbol{\mathcal{X}}$ is a vector space over $\mathbb{R}$.

(b) Show that the sequences

$$s_1 = (1, 1, 2, 3, 5, 8, 13, ...) \quad \text{and} \quad s_2 = (-1, 1, 0, 1, 1, 2, 3, ...)$$

form a basis for $\boldsymbol{\mathcal{X}}$.

(c) Find the projection of the sequence $f = (1, 2, 3, 5, 8, 13, ...)$ on span (s_2) along span (s_1).

(d) Let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \mathbb{R}^2$ be defined as $\mathcal{A}(f) = (f[3], f[4])$. Find the matrix representation of $\mathcal{A}$ with respect to the basis of $\boldsymbol{\mathcal{X}}$ in part (b) and the canonical basis of $\mathbb{R}^2$.

48. Let $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{U}} \oplus \boldsymbol{\mathcal{V}}$, and let $\boldsymbol{R} = (\,\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k\,)$ and $\boldsymbol{S} = (\,\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-k}\,)$ be bases for $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ so that $\boldsymbol{T} = (\,\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n-k}\,)$ is a basis for $\boldsymbol{\mathcal{X}}$. Let $\mathcal{P}$ be a projection on $\boldsymbol{\mathcal{U}}$ along $\boldsymbol{\mathcal{V}}$.

(a) Find the matrix representation of $\mathcal{P}$ with respect to $\boldsymbol{T}$ if $\mathcal{P}$ is interpreted as a linear transformation from $\boldsymbol{\mathcal{X}}$ into itself.

(b) Find the matrix representation of $\mathcal{P}$ with respect to $(\boldsymbol{T}, \boldsymbol{R})$ if $\mathcal{P}$ is interpreted as a linear transformation from $\boldsymbol{\mathcal{X}}$ into $\boldsymbol{\mathcal{U}}$.

49. Show that $\mathcal{P} : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$\mathcal{P}(\alpha, \beta) = (\frac{\alpha - \beta}{2}\,,\ \frac{\beta - \alpha}{2})$$

is a projection. Characterize im $(\mathcal{P})$ and ker $(\mathcal{P})$. Illustrate the decomposition of a vector into components in im $(\mathcal{P})$ and ker $(\mathcal{P})$ with the help of a picture.

50. Prove facts (a-c) on p.124 concerning a projection.

51. Show that if $P \in \mathbb{F}^{n\times n}$ is idempotent, then it defines a projection on $\text{im}\,(P)$ along $\ker(P)$. Hint: For any $\boldsymbol{x} \in \mathbb{F}^{n\times 1}$, let $\boldsymbol{u} = P\boldsymbol{x} \in \text{im}\,(P)$ and $\boldsymbol{v} = (I - P)\boldsymbol{x}$, and show that $\boldsymbol{v} \in \ker(P)$.

52. A projection in the xyz space projects every point onto the plane described by

$$x + y + z = 0$$

along its normal $\boldsymbol{n} = \text{col}\,[\,1, 1, 1\,]$.

(a) Find the projection $\boldsymbol{u}$ of the point $\boldsymbol{x} = \text{col}\,[\,a, b, c\,]$ on the plane.

(b) Find a matrix P such that $\boldsymbol{u} = P\boldsymbol{x}$.

(c) Verify that $P^2 = P$.

Chapter 4
Rank, Inverse and Determinants

4.1 Row and Column Spaces and The Rank

Let A be an $m \times n$ matrix partitioned into its rows:

$$A = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{bmatrix}$$

where $\boldsymbol{\alpha}_i \in \mathbb{F}^{1\times n}, i = 1, \ldots, m$. The span of the rows of A is a subspace of $\mathbb{F}^{1\times n}$, and is called the ***row space*** of A, denoted rs (A):

$$\text{rs}\,(A) = \text{span}\,(\,\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m\,) \subset \mathbb{F}^{1\times n}$$

If $\{\,\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m\,\}$ is linearly independent then it is a basis for rs (A). Otherwise, it can be reduced to a basis by means of elementary operations. From the discussion in Section 1.4 and Section 3.2.1 it is clear that if

$$R = \begin{bmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_r \\ O \end{bmatrix}$$

is the reduced row echelon form of A then

a) rs (A) = rs (R)

b) $\{\,\boldsymbol{r}_1, \ldots, \boldsymbol{r}_r\,\}$ is a basis for rs (A)

c) dim (rs (A)) $= r$

Thus the row rank of a matrix defined in Section 1.4 is the dimension of its row space.

Now let us partition A into its columns:

$$A = [\,\boldsymbol{a}_1 \;\; \boldsymbol{a}_2 \;\; \cdots \;\; \boldsymbol{a}_n\,]$$

where $\boldsymbol{a}_j \in \mathbb{F}^{m\times 1}, j = 1, \ldots, n$. The span of the columns of A, which is a subspace of $\mathbb{F}^{m\times 1}$, is called the ***column space*** of A, denoted cs (A):

$$\text{cs}\,(A) = \text{span}\,(\,\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\,) \subset \mathbb{F}^{m\times 1}$$

Again, if $\{ \boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n \}$ is linearly independent then it is a basis for cs (A). Otherwise, it can be reduced to a basis by means of elementary column operations on A (see Exercise 1.33). If

$$C = [\, \boldsymbol{c}_1 \ \cdots \ \boldsymbol{c}_\rho \ O \,]$$

is the reduced column echelon form of A, where ρ (the number of nonzero columns of C) is the column rank of A, then

a) cs $(A) =$ cs (C)

b) $\{ \boldsymbol{c}_1, \ldots, \boldsymbol{c}_\rho \}$ a basis for cs (A)

c) dim (cs $(A)) = \rho$

It is interesting to examine the relation between the row rank and the column rank of a matrix. Consider the reduced row echelon form of A, and rename the basic columns of A as $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r$ and the non-basic columns as $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_\nu$, where $\nu = n - r$. Then with the notation of Section 1.5

$$[\, B \ \ G \,] \xrightarrow{\text{e.r.o.}} \begin{bmatrix} I_r & H \\ O & O \end{bmatrix} \tag{4.1}$$

or equivalently

$$\left[\, B \ \ \boldsymbol{g}_j \,\right] \xrightarrow{\text{e.r.o.}} \begin{bmatrix} I_r & \boldsymbol{h}_j \\ O & \boldsymbol{0} \end{bmatrix}, \quad j = 1, \ldots, \nu \tag{4.2}$$

where B and G are $m \times r$ and $m \times \nu$ submatrices of A consisting of its basic and non-basic columns respectively. (4.2) implies that each of the $m \times r$ systems

$$B\boldsymbol{u} = \boldsymbol{g}_j\,, \quad j = 1, \ldots, \nu$$

is consistent and has a solution $\boldsymbol{u} = \boldsymbol{h}_j$, that is,

$$\boldsymbol{g}_j = B\boldsymbol{h}_j\,, \quad j = 1, \ldots, \nu$$

This shows that every non-basic column can be written as a linear combination of the basic columns. In other words, $\boldsymbol{g}_j \in$ cs $(B)\,, j = 1, \ldots, \nu$. Thus

$$\text{cs}\,(A) = \text{cs}\,[\, B \ \ G \,] = \text{cs}\,(B)$$

Moreover, since $r(B) = r$, the only solution of the homogeneous system $B\boldsymbol{u} = \boldsymbol{0}$ is the trivial solution $\boldsymbol{u} = \boldsymbol{0}$. This shows that columns of B are also linearly independent, and therefore, form a basis for cs (A). Since B has r columns, we have

R1. $\rho(A) = r(A)$

The common value of the row and column ranks of A is simply called the ***rank*** of A. Thus the row and column spaces of a given matrix, which are subspaces of different vector spaces, have the same dimension r, which is the maximum number of linearly independent rows and also the maximum number of linearly independent columns of that matrix.

Recall that the image of an $m \times n$ matrix A is defined as

$$\text{im}\,(A) = \{\, \boldsymbol{y} \,|\, \boldsymbol{y} = A\boldsymbol{x},\ \boldsymbol{x} \in \mathbb{F}^{n\times 1}\}$$

Since for any $\boldsymbol{x}$, $A\boldsymbol{x}$ is a linear combination of the columns of A (coefficients being the elements of $\boldsymbol{x}$), it follows that

$$\text{im}\,(A) = \text{cs}\,(A)$$

That explains why we use the same term "rank" for both the dimension of the image of a linear transformation and the dimension of the column space of a matrix that defines a linear transformation.

Example 4.1

Let us find bases for the row and column spaces of the matrix

$$A = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 3 & 3 & -2 & 5 \\ 2 & 2 & -1 & 3 \end{bmatrix}$$

The reduced row echelon form of A is

$$A \xrightarrow{\text{e.r.o.}} \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = R$$

from which we conclude that $r = 2$, that rows 1 and 2 of R form a basis for the row space of A, and that the basic columns 1 and 3 of A form a basis for the column space of A. Let us verify these conclusions.

Any $\boldsymbol{\alpha} \in \text{rs}\,(A)$ is of the form

$$\begin{aligned} \boldsymbol{\alpha} &= c_1 [\, 1 \quad 1 \ -1 \quad\ 2\,] + c_2 [\, 3 \quad 3 \ -2 \quad 5\,] + c_3 [\, 2 \quad 2 \ -1 \quad 3\,] \\ &= (c_1 + 3c_2 + 2c_3)[\, 1 \quad 1 \quad 0 \quad 1\,] + (-c_1 - 2c_2 - c_3)[\, 0 \quad 0 \quad 1 \ -1\,] \end{aligned}$$

Thus rows 1 and 2 of R span the row space of A. Since rows 1 and 2 of R are also linearly independent, they form a basis for rs (A).

Any $\boldsymbol{y} \in \text{cs}\,(A)$ is of the form

$$\begin{aligned} \boldsymbol{y} &= c_1 \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + c_3 \begin{bmatrix} -1 \\ -2 \\ -1 \end{bmatrix} + c_4 \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix} \\ &= (c_1 + c_2 + c_4) \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + (c_3 - c_4) \begin{bmatrix} -1 \\ -2 \\ -1 \end{bmatrix} \end{aligned} \tag{4.3}$$

so that columns 1 and 3 of A (the basic columns) span the column space of A. It is easy to verify that they are also linearly independent, and therefore, form a basis for cs (A).

We can also find a basis for cs (A) by considering its reduced column echelon form, which is obtained by the sequence elementary column operations described below.

$$\begin{bmatrix} 1 & 1 & -1 & 2 \\ 3 & 3 & -2 & 5 \\ 2 & 2 & -1 & 3 \end{bmatrix} \quad \begin{matrix} -\text{C}_1 + \text{C}_2 \to \text{C}_2 \\ \text{C}_1 + \text{C}_3 \to \text{C}_3 \\ -2\text{C}_1 + \text{C}_4 \to \text{C}_4 \\ \longrightarrow \end{matrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 3 & 0 & 1 & -1 \\ 2 & 0 & 1 & -1 \end{bmatrix}$$

$$\begin{matrix} \text{C}_2 \longleftrightarrow \text{C}_3 \\ \text{C}_1 - 3\text{C}_2 \to \text{C}_1 \\ \text{C}_2 + \text{C}_4 \to \text{C}_4 \\ \longrightarrow \end{matrix} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{bmatrix}$$

Thus the nonzero columns 1 and 2 of the reduced column echelon form of A form a basis for its column space. This can also be verified by observing that a typical vector in the column space given in (4.3) can be expressed as

$$\begin{aligned}\boldsymbol{y} &= c_1\begin{bmatrix}1\\3\\2\end{bmatrix}+c_2\begin{bmatrix}1\\3\\2\end{bmatrix}+c_3\begin{bmatrix}-1\\-2\\-1\end{bmatrix}+c_4\begin{bmatrix}2\\5\\3\end{bmatrix}\\ &= (c_1+c_2-c_3+2c_4)\begin{bmatrix}1\\0\\-1\end{bmatrix}+(3c_1+3c_2-2c_3+5c_4)\begin{bmatrix}0\\1\\1\end{bmatrix}\end{aligned}$$

A square matrix of order n is called ***nonsingular*** if $r(A) = n$, and ***singular*** if $r(A) < n$. By definition, rows (columns) of a nonsingular matrix are linearly independent, and its reduced row (column) echelon form is I.

Let A be an $m \times n$ matrix with $r(A) = r$. Let B be an $m \times r$ submatrix consisting of any r linearly independent columns of A (for example, the submatrix B in (4.1) that consists of the basic columns of A). Since $r(B) = r$, B has r linearly independent rows. Let C be an $r \times r$ submatrix of B consisting of such linearly independent rows. Then $r(C) = r$, so that C is a nonsingular submatrix of A. This shows that if $r(A) = r$, then A contains an $r \times r$ nonsingular submatrix. Now suppose $r < \min\{m, n\}$, and consider any $k \times k$ submatrix of A with $k > r$. Since any k columns of A are linearly dependent, so are the columns of this submatrix, and therefore, it must be singular. We thus conclude that

R2. the rank of a matrix is the order of its largest nonsingular submatrix.

Example 4.2

Since the matrix A in Example 4.1 has rank 2, it must have a nonsingular submatrix of order 2, and all square submatrices of order 3 must be singular.

Indeed, the 2×2 submatrix

$$\begin{bmatrix}1 & -1\\3 & -2\end{bmatrix}$$

consisting of first and second rows and first and third columns is nonsingular as can easily be shown by observing that its reduced row echelon form is I_2.

On the other hand, the rows of A are linearly dependent (the second row is the sum of the other two). Then the rows of every 3×3 submatrix of A are also linearly dependent; that is, every 3×3 submatrix of A is singular. The reader should verify this observation by computing ranks of all possible 3×3 submatrices.

Let $C = AB$. If $\boldsymbol{y} \in \text{cs}\,(C)$ then $\boldsymbol{y} = C\boldsymbol{x}$ for some $\boldsymbol{x}$ so that

$$\boldsymbol{y} = C\boldsymbol{x} = AB\boldsymbol{x} = A(B\boldsymbol{x}) = A\boldsymbol{z} \in \text{cs}\,(A)$$

Hence $\text{cs}\,(C) \subset \text{cs}\,(A)$, and therefore, $r(C) \leq r(A)$. Similarly, $\text{rs}\,(C) \subset \text{rs}\,(B)$, and $r(C) \leq r(B)$. As a result, we have

R3. $r(AB) \leq \min\{\, r(A), r(B)\,\}$

Example 4.3

Let $C = AB$, where

$$A = \begin{bmatrix} 1 & 1 & -1 & 2 \\ 3 & 3 & -2 & 5 \\ 2 & 2 & -1 & 3 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

We established in Example 4.1 that $r(A) = 2$. Also, from

$$B \xrightarrow{\text{e.c.o.}} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

we get $r(B) = 3$. Hence we must have $r(C) \leq \min\{\, 2, 3 \,\}$, that is, $r(C) = 0, 1$ or 2. Indeed, computing C as

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

we find $r(C) = 1$.

Example 4.4

Computing the rank of a matrix may pose numerical difficulties, similar to those encountered when dealing with ill-conditioned systems, when some rows or columns are nearly linearly dependent.

Consider the matrix

$$A = \begin{bmatrix} 0.9502 & 0.2312 & 0.7189 \\ 0.6067 & 0.4859 & 0.1208 \\ 0.8913 & 0.7621 & 0.1292 \end{bmatrix}$$

A calculator that operates with 4-digit floating point arithmetic computes the reduced row echelon form of the matrix using Gaussian elimination as

$$R = \begin{bmatrix} 1.0000 & 0.0000 & 0.9999 \\ 0.0000 & 1.0000 & -0.9998 \\ 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

and therefore, its row rank as $r = 2$. On the other hand, the same calculator computes the reduced column echelon form of A as $C = I$, and therefore, its column rank as $\rho = 3$. Apparently, Gaussian elimination is not reliable in computation of the rank.

The fact is that A is nonsingular, and therefore, has rank $r = 3$. (The reader can verify this by using MATLAB's built-in function `rank`). However, the matrix

$$\tilde{A} = \begin{bmatrix} 0.9502 & 0.2312 & 0.7190 \\ 0.6067 & 0.4859 & 0.1208 \\ 0.8913 & 0.7621 & 0.1292 \end{bmatrix}$$

which differs from A only in the fourth decimal digit of the element in the $(1, 3)$ position has rank $\tilde{r} = 2$. (The third column of $\tilde{A}$ is the difference of the first two; hence $\tilde{A}$ has only two linearly independent columns.) Thus although A has rank $r = 3$, it is very close to a matrix with rank $\tilde{r} = 2$. Whether A should be viewed as having rank two or rank three depends on the numerical accuracy desired in the particular application it appears.

4.2 Inverse

In Section 3.3.2 we stated the following facts concerning a linear transformation $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ without proof:

a) If $\mathcal{A}$ is one-to-one ($\ker(\mathcal{A}) = \{\mathbf{0}\}$) then it has a left inverse $\hat{\mathcal{A}}_L : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$, not necessarily unique, such that

$$\hat{\mathcal{A}}_L(\mathcal{A}(\boldsymbol{x})) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

b) If $\mathcal{A}$ is onto ($\text{im}(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$) then it has a right inverse $\hat{\mathcal{A}}_R : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$, not necessarily unique, such that

$$\mathcal{A}(\hat{\mathcal{A}}_R(\boldsymbol{y})) = \boldsymbol{y} \quad \text{for all } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$$

c) If $\mathcal{A}$ is both one-to-one and onto then it has a unique inverse $\hat{\mathcal{A}} : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{X}}$ such that

$$\hat{\mathcal{A}}(\mathcal{A}(\boldsymbol{x})) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}} \quad \text{and} \quad \mathcal{A}(\hat{\mathcal{A}}(\boldsymbol{y})) = \boldsymbol{y} \quad \text{for all } \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$$

In this section, we will prove these statements for a linear transformation defined by a matrix. To be precise, we first define left inverse, right inverse, and (two-sided) inverse of a matrix:

a) A matrix $\hat{A}_L$ that satisfies $\hat{A}_L A = I$ is called a ***left inverse*** of A.

b) A matrix $\hat{A}_R$ that satisfies $A\hat{A}_R = I$ is called a ***right inverse*** of A.

c) A matrix $\hat{A}$ that satisfies $\hat{A}A = A\hat{A} = I$ is called an ***inverse*** of A.

It is a simple exercise to show that for $A \in \mathbb{F}^{m\times n}$, $\ker(A) = \{\mathbf{0}\}$ if and only if $r(A) = n$ and $\text{im}(\mathcal{A}) = \boldsymbol{\mathcal{Y}}$ if and only if $r(A) = m$. With this observation, we state facts (a)-(c) above and few additional facts as a theorem, whose proof will be given in the following subsections.

Theorem 4.1 *Let $A \in \mathbb{F}^{m\times n}$. Then*

a) A has a left inverse $\hat{A}_L$ if and only if $r(A) = n$.

b) A has a right inverse $\hat{A}_R$ if and only if $r(A) = m$.

c) A has an inverse $\hat{A}$ if and only if $r(A) = m = n$, that is, A is square and nonsingular.

d) If $r(A) = m = n$ then $\hat{A}_L$, $\hat{A}_R$ and $\hat{A}$ are unique and $\hat{A}_L = \hat{A}_R = \hat{A}$.

Note that since $r(A) \le \min\{m, n\}$, A can have a left inverse only when $n \le m$ and a right inverse only when $m \le n$.

Assuming that parts (a)-(c) of Theorem 4.1 are true, part (d) can be proved by a simple argument: If $r(A) = m = n$ then A has an inverse $\hat{A}$, which is certainly also a left inverse. Suppose that A has another left inverse $\hat{A}_L$ that satisfies $\hat{A}_L A = I$. Then postmultiplying both sides with $\hat{A}$ we obtain

$$\hat{A} = (\hat{A}_L A)\hat{A} = \hat{A}_L(A\hat{A}) = \hat{A}_L I = \hat{A}_L$$

contradicting the assumption. It can similarly be shown that $\hat{A}$ is the only right inverse of A.

Because of the fact stated in Theorem 4.1(c), a nonsingular matrix is also called ***invertible***. It is customary to denote the unique inverse of a square, nonsingular matrix A by A^{-1}.

4.2.1 Elementary Matrices

A matrix obtained from the identity matrix by a single elementary row or column operation is called an ***elementary matrix***. Corresponding to the three types of elementary operations there are three types of elementary matrices. For example,

$$E_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}$$

are 4×4 elementary matrices of Type I, Type II, and Type III, respectively. E_1 is obtained from I_4 by interchanging the first and the third rows (or columns), E_2 by multiplying the second row (or column) by the scalar $c \neq 0$, and E_3 by adding 2 times the second row to the forth row (or 2 times the fourth column to the second column).

It is left to the reader to show that an elementary row operation on an $m \times n$ matrix A can be represented by premultiplying A with the corresponding $m \times m$ elementary matrix. For example, if

$$A = \begin{bmatrix} 1 & 0 & 3 & -1 \\ -2 & 1 & -4 & 3 \\ 3 & -2 & -1 & 0 \end{bmatrix} \xrightarrow{2R_1 + R_2 \to R_2} \begin{bmatrix} 1 & 0 & 3 & -1 \\ 0 & 1 & 2 & 1 \\ 3 & -2 & -1 & 0 \end{bmatrix} = B$$

then $B = EA$, where

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.4}$$

is an elementary matrix obtained from I by the same elementary row operation. Similarly, an elementary column operation on an $m \times n$ matrix A can be represented by postmultiplying A with the corresponding $n \times n$ elementary matrix. Moreover, if E is an $n \times n$ elementary matrix that represents an elementary operation on the rows of a square matrix of order n, then E^t is also an elementary matrix that represents the same operation on the corresponding columns of A.

Let E be an elementary matrix that represents an elementary row operation on I, and let $\hat{E}$ represent the inverse operation. Then clearly,

$$\hat{E}E = I$$

On the other hand, the same E can also be considered as representing an elementary column operation on I, and $\hat{E}$ the inverse operation. Then

$$E\hat{E} = I$$

As a result $\hat{E}$ is the unique inverse of E, that is,

$$E^{-1} = \hat{E}$$

Moreover, E^{-1} is also an elementary matrix of the same type as E. For example, the inverse of the elementary matrix E in (4.4) is

$$E^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which represents the inverse operation operation of adding -2 times the first row to the second row.

If $E_1, E_2, \ldots, E_k$ are elementary matrices of order n, the product

$$E_s = E_k \cdots E_2 E_1$$

represents a sequence of elementary row operations on I_n. Then the product

$$\hat{E}_s = E_1^{-1} E_2^{-1} \cdots E_k^{-1}$$

represents a sequence of elementary row operations that undo the operations represented by E_s, so that $\hat{E}_s E_s = I$. Similarly, $E_s \hat{E}_s = I$. We thus conclude that $\hat{E}_s = E_s^{-1}$, that is,

$$(E_k \cdots E_2 E_1)^{-1} = E_1^{-1} E_2^{-1} \cdots E_k^{-1} \tag{4.5}$$

An elementary matrix of Type I is also called an ***elementary permutation matrix*** for the obvious reason that it permutes (reorders) the rows or columns of the matrix that it multiplies. The reader can show that if P is an elementary permutation matrix then

$$P^{-1} = P^t$$

Let $P_s = P_k \cdots P_2 P_1$, where $P_1, P_2, \ldots, P_k$ are elementary permutation matrices. Since a permutation followed by another permutation is also a permutation, we can conveniently call P_s a ***permutation matrix***.[1] Note that a permutation matrix contains a single 1 in every row and column. The inverse of such a permutation matrix can be found by means of (4.5) to be

$$P_s^{-1} = P_1^{-1} P_2^{-2} \cdots P_k^{-1} = P_1^t P_2^t \cdots P_k^t = (P_k \cdots P_2 P_1)^t = P_s^t$$

Example 4.5

If

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \; E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}, \; A = PE = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

then

$$A^{-1} = E^{-1} P^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 \end{bmatrix}$$

The reader should verify that $A^{-1}A = AA^{-1} = I$.

4.2.2 Left, Right and Two-Sided Inverses

If $A \in \mathbb{F}^{m \times n}$ has a left inverse $\hat{A}_L \in \mathbb{F}^{n \times m}$ so that $\hat{A}_L A = I_n$ then

$$r(\hat{A}_L A) = n \leq r(A) \leq \min\{\, m, n \,\}$$

[1] Permutations are discussed in Section 4.4.

and we must have $r(A) = n$. This proves the necessity part of Theorem 4.1(a). (As a byproduct, we also find that $r(\hat{A}_L) = n$.) Conversely, if $r(A) = n$ then

$$E_q \cdots E_2 E_1 A = QA = R = \begin{bmatrix} I_n \\ O \end{bmatrix} \tag{4.6}$$

for some elementary matrices $E_1, \ldots, E_q$, where R is the reduced row echelon form of A, and Q represents the sequence of elementary row operations used to transform A into R. Partitioning rows of Q in accordance with the partitioning of R, (4.6) can be written as

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} A = \begin{bmatrix} Q_1 A \\ Q_2 A \end{bmatrix} = \begin{bmatrix} I_n \\ O \end{bmatrix} \tag{4.7}$$

from which we observe that

$$\hat{A}_L = Q_1 = R^t Q$$

is a left inverse of A. Thus we not only prove the sufficiency part of Theorem 4.1(a), but also give a method to construct a left inverse when the sufficiency condition is satisfied.

Example 4.6

Consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 0 & 1 \end{bmatrix}$$

in Example 3.38. The elementary row operations that transform A into its reduced row echelon form can be summarized as

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & -1 & 0 \\ -2 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

A left inverse of A is thus obtained as

$$\hat{A}_L = \begin{bmatrix} 3 & -1 & 0 \\ -2 & 1 & 0 \end{bmatrix}$$

Note that $\hat{A}_L$ above is different from both of the left inverses in Example 3.38.

The proof of part (b) of Theorem 4.1 follows similar lines: If $A \in \mathbb{F}^{m \times n}$ has a right inverse $\hat{A}_R \in \mathbb{F}^{n \times m}$ so that $A\hat{A}_R = I_m$ then from

$$r(A\hat{A}_R) = m \leq r(A) \leq \min\{m, n\}$$

we get $r(A) = m$. (We also have $r(\hat{A}_R) = m$.) On the other hand, if $r(A) = m$ then

$$AE_1 E_2 \cdots E_p = AP = C = [\, I_m \;\; O \,] \tag{4.8}$$

where C is the reduced column echelon form of A, and P represents the sequence of elementary column operations used to transform A into C. Partitioning columns of P in accordance with the partitioning of C, (4.8) can be written as

$$A\,[\,P_1 \quad P_2\,] = [\,AP_1 \quad AP_2\,] = [\,I_m \quad O\,] \tag{4.9}$$

from which a right inverse of A is obtained as

$$\hat{A}_R = P_1 = PC^t$$

Example 4.7

The matrix

$$B = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

considered in Example 3.39 can be transformed into its reduced column echelon form as

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

A right inverse of B is

$$\hat{B}_R = \begin{bmatrix} 1 & -2 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Finally, part (c) of Theorem 4.1 follows from parts (a) and (b): If A has an inverse $\hat{A}$ then it is also a left inverse and a right inverse so that $r(A) = m = n$, that is, A is square and nonsingular. Conversely, if $r(A) = m = n$ then the row echelon form of A is I_n so that (4.6) reduces to

$$E_q \cdots E_2 E_1 A = QA = I_n \tag{4.10}$$

Thus Q is a left inverse of A. Premultiplying both sides of (4.10) with the product $E_1^{-1}E_2^{-1}\cdots E_q^{-1}$, we obtain

$$A = E_1^{-1}E_2^{-2}\cdots E_q^{-1} \tag{4.11}$$

which implies that

$$AQ = E_1^{-1}E_2^{-2}\cdots E_q^{-1}E_q\cdots E_2E_1 = I$$

that is, Q is also a right inverse of A. Hence if $r(A) = m = n$ then A has an inverse

$$A^{-1} = Q$$

This completes the proof of Theorem 4.1.

MATLAB provides a built-in command `inv` to compute the unique inverse of a square, nonsingular matrix.

From (4.10) it follows that

$$E_k \cdots E_2 E_1 [\, A \;\; I \,] = [\, I \;\; Q \,] = [\, I \;\; A^{-1} \,]$$

The expression above provides a convenient method to find the inverse of a nonsingular matrix by means of elementary operations as illustrated by the following example.

Example 4.8

Show that the matrix

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & -1 & 2 \\ 3 & 0 & 5 \end{bmatrix}$$

is nonsingular, and then find its inverse.

We form the augmented matrix $[\, A \;\; I \,]$, and perform elementary row operations to reduce A into its reduced row echelon form.

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & 1 & 0 & 0 \\ 2 & -1 & 2 & 0 & 1 & 0 \\ 3 & 0 & 5 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ 0 & 3 & 5 & -3 & 0 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 2 & -1 & 1 & 0 \\ 0 & 1 & 2 & -2 & 1 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 5 & -5 & 2 \\ 0 & 1 & 0 & 4 & -5 & 2 \\ 0 & 0 & 1 & -3 & 3 & -1 \end{array}\right]$$

Since the reduced row echelon form of A is I, it is nonsingular, and

$$A^{-1} = \begin{bmatrix} 5 & -5 & 2 \\ 4 & -5 & 2 \\ -3 & 3 & -1 \end{bmatrix}$$

MATLAB gives the same answer. The reader can also verify that $A^{-1}A = AA^{-1} = I$.

(4.11) shows that every nonsingular matrix can be expressed as a product of elementary matrices. Since elementary operations do not change the rank of a matrix, we reach the following result.

R4. If A is nonsingular then $r(AB) = r(B)$ and $r(CA) = r(C)$.

Some special matrices have special inverses. For example, a diagonal matrix

$$D = \text{diag}\,[\, d_1, d_2, \ldots, d_n \,]$$

is nonsingular if and only if $d_i \neq 0$ for all i, in which case

$$D^{-1} = \text{diag}\,[\, 1/d_1, 1/d_2, \ldots, 1/d_n \,]$$

In particular, $I^{-1} = I$.

The following properties of inverse are easy to show, and are left to the reader.

I1. If A is nonsingular then so is A^h, and $(A^h)^{-1} = (A^{-1})^h$.

I2. If $A_1, A_2, \ldots, A_k$ are nonsingular matrices of order n then their product is also nonsingular, and

$$(A_k \cdots A_2 A_1)^{-1} = A_1^{-1} A_2^{-2} \cdots A_k^{-1}$$

Note that the third property above is a generalization of (4.5) stated for a product of elementary matrices to a product of arbitrary nonsingular matrices.

4.2.3 Generalized Inverse

If $\hat{A}_G$ is a left inverse or a right inverse or a two-sided inverse of A, then it certainly satisfies both of the relations

$$A\hat{A}_G A = A\,, \quad \hat{A}_G A \hat{A}_G = \hat{A}_G \tag{4.12}$$

If none of the rank conditions of Theorem 4.1 holds, then A does not have a left or a right inverse, nor a two-sided inverse. However, it may still be possible to construct a matrix $\hat{A}_G$ that satisfies the above relations. Such a matrix, if it exists, is called a ***generalized inverse*** of A.

Let $A \in \mathbb{F}^{m\times n}$ with $r(A) = r$, and let

$$QA = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} A = \begin{bmatrix} R_1 \\ O \end{bmatrix} = R$$

where R is the reduced row echelon form of A and Q represents the sequence of elementary row operations that transform A into R. Since $r(R) = r(R_1) = r$

$$RP = \begin{bmatrix} R_1 \\ O \end{bmatrix} [\, P_1 \quad P_2 \,] = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix} = N$$

where N is the reduced column echelon form of R, and P represents the sequence of elementary column operations that transform R into N. The matrix

$$N = QAP = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} A \, [\, P_1 \quad P_2 \,] = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix} \tag{4.13}$$

is called the ***normal form*** of A.[2]

Let

$$\hat{A}_G = PN^tQ = P_1 Q_1$$

Noting that

$$A = Q^{-1} N P^{-1} \tag{4.14}$$

[2]The normal form can also be obtained by first obtaining the reduced column echelon form C of A and then finding the reduced row echelon form of C.

$NN^tN = N$, and $N^tNN^t = N^t$, straightforward multiplications give

$$\begin{aligned} A\hat{A}_GA &= (Q^{-1}NP^{-1})(PN^tQ)(Q^{-1}NP^{-1}) \\ &= Q^{-1}NN^tNP^{-1} = Q^{-1}NP^{-1} = A \\ \hat{A}_GA\hat{A}_G &= (PN^tQ)A(PN^tQ) \\ &= PN^tNN^tQ = PN^tQ = \hat{A}_G \end{aligned}$$

Hence $\hat{A}_G$ is a generalized inverse of A.

The MATLAB command `pinv(A)` computes a special generalized inverse of A, which reduces to a left inverse when $r = n$ and to a right inverse when $r = m$.

Example 4.9

The matrix

$$A = \begin{bmatrix} 1.0 & -0.8 & 0.6 \\ -0.5 & 0.4 & -0.3 \end{bmatrix}$$

can be reduced to its normal form by means of elementary operations that are summarized as

$$\begin{bmatrix} 1.0 & 0.0 \\ 0.5 & 1.0 \end{bmatrix} \begin{bmatrix} 1.0 & -0.8 & 0.6 \\ -0.5 & 0.4 & -0.3 \end{bmatrix} \begin{bmatrix} 1.0 & 0.8 & -0.6 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

A generalized inverse of A is then obtained as

$$\hat{A}_G = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The normal form of A can also be obtained by a different sequence of elementary operations:

$$\begin{bmatrix} 0.0 & -2.0 \\ 1.0 & 2.0 \end{bmatrix} \begin{bmatrix} 1.0 & -0.8 & 0.6 \\ -0.5 & 0.4 & -0.3 \end{bmatrix} \begin{bmatrix} 1.0 & -0.8 & 0.6 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

which result in a different generalized inverse

$$\hat{A}_G = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & -2 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

The MATLAB command `pinv` computes yet another generalized inverse

$$\hat{A}_G = \begin{bmatrix} 0.4000 & -0.2000 \\ -0.3200 & 0.1600 \\ 0.2400 & -0.1200 \end{bmatrix}$$

* 4.3 Equivalence and Similarity

Recall from Section 3.2.1 that if Q is the matrix of transition from a basis $\boldsymbol{R}$ of an n-dimensional vector space $\boldsymbol{\mathcal{X}}$ to another basis $\boldsymbol{R}'$, and if P is the matrix of transition from

$\boldsymbol{R}'$ to $\boldsymbol{R}$, then $QP = PQ = I_n$. This shows that Q and P are both nonsingular, and are inverses of each other. Conversely, any nonsingular $n \times n$ matrix Q can be viewed as a matrix of transition from a basis $\boldsymbol{R}$ to another basis $\boldsymbol{R}'$ in some n-dimensional vector space $\boldsymbol{\mathcal{X}}$, say $\mathbb{F}^{n\times 1}$. In other words, if $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ is a basis for $\boldsymbol{\mathcal{X}}$, and $\boldsymbol{r}'_j$ are defined as

$$\boldsymbol{r}'_j = \sum_{i=1}^{n} p_{ij}\boldsymbol{r}_i\,, \quad j = 1, \ldots, n \tag{4.15}$$

where $P = [\,p_{ij}\,]$ is nonsingular, then $\boldsymbol{R}' = (\boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_n)$ is also a basis for $\boldsymbol{\mathcal{X}}$ (see Exercise 4.22). Moreover, the matrix of transition from $\boldsymbol{R}$ to $\boldsymbol{R}'$ is precisely $Q = P^{-1}$.

Also recall from Section 3.3.1 that if a linear transformation $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ has a matrix representation A with respect to a pair of bases $(\boldsymbol{R}, \boldsymbol{S})$ and a representation A' with respect to another pair $(\boldsymbol{R}', \boldsymbol{S}')$, then

$$A' = Q_m A P_n \tag{4.16}$$

where Q_m is the nonsingular matrix of transition from $\boldsymbol{S}$ to $\boldsymbol{S}'$ in $\boldsymbol{\mathcal{Y}}$, and P_n is the nonsingular matrix of transition from $\boldsymbol{R}'$ to $\boldsymbol{R}$ in $\boldsymbol{\mathcal{X}}$. (Subscripts n and m refer to the dimensions of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$.) In particular, if $\mathcal{A} : \mathbb{F}^{n\times 1} \to \mathbb{F}^{m\times 1}$ is a linear transformation defined by an $m \times n$ matrix A, then its representation with respect to the canonical bases $(\boldsymbol{E}^n, \boldsymbol{E}^m)$ of $\mathbb{F}^{n\times 1}$ and $\mathbb{F}^{m\times 1}$ is the A matrix itself (see Example 3.34). Thus if A' is an $m \times n$ matrix that is related to A as in (4.16) then it represents the same linear transformation with respect to a different pair of bases, which are uniquely defined by the matrices P_n and $P_m = Q_m^{-1}$.

From the discussion in Section 4.1 we observe that two $m \times n$ matrices A' and A are row equivalent if they are related as $A' = Q_m A$, where Q_m is an $m \times m$ nonsingular matrix that stands for the elementary row operations performed on A to obtain A'. Thus all row equivalent $m \times n$ matrices represent the same linear transformation with respect to a fixed basis in $\mathbb{F}^{n\times 1}$ and different bases in $\mathbb{F}^{m\times 1}$. Their common (unique) reduced row echelon form can be considered as a canonical form that represents the equivalence class formed by these row equivalent matrices. Similarly, two $m \times n$ matrices A' and A are column equivalent if they are related as $A' = AP_n$, where P_n is an $n \times n$ nonsingular matrix that stands for the elementary column operations performed on A to obtain A'. All column equivalent $m \times n$ matrices represent the same linear transformation with respect to a fixed basis in $\mathbb{F}^{m\times 1}$ and different bases in $\mathbb{F}^{n\times 1}$. Their common (unique) reduced column echelon form is a canonical form that represents the equivalence class formed by column equivalent matrices. Combining the two types of equivalence, we call A and A' ***equivalent*** if they are related as in (4.16) for some nonsingular matrices Q_m and P_n. Thus equivalent matrices represent the same linear transformation with respect to different bases, and their common (unique) normal form is a canonical form that represents the equivalence class formed by equivalent matrices.

Example 4.10

The reduced row and column echelon forms of

$$A = \begin{bmatrix} 1 & 2 & 1 & 4 \\ 2 & 4 & 1 & 5 \\ 3 & 6 & 2 & 9 \end{bmatrix}$$

are

$$R = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

R represents the class of matrices that are row equivalent to A, and C those that are column equivalent to A.

The reduced column echelon form of R

$$N = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

which is also the reduced row echelon form of C, represents all matrices that are equivalent to A, that is, all 3×4 matrices with rank $r = 2$.

When a square matrix A of order n is viewed as the representation of a linear transformation from an n-dimensional vector space $\boldsymbol{\mathcal{X}}$ into another n-dimensional vector space $\boldsymbol{\mathcal{Y}}$, then by choosing suitable bases for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$, A can be transformed into its normal form N as in (4.13). However, if it is viewed as a linear operator on $\mathbb{F}^{n\times 1}$, that is, as a linear transformation from $\mathbb{F}^{n\times 1}$ into itself, then it is natural to use the same basis in both its domain and codomain. In this case, the equivalence relation in (4.13) becomes

$$A' = P^{-1}AP \tag{4.17}$$

Two square matrices related as in (4.17) are called ***similar***. Thus similarity is a special case of equivalence. We will discuss similarity transformations in detail in the next chapter.

Example 4.11

Let a linear transformation $\mathcal{A} : \mathbb{R}^{2\times 1} \to \mathbb{R}^{2\times 1}$ be represented by the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}$$

with respect to the canonical basis $\boldsymbol{E}$. The vectors

$$\boldsymbol{r}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{r}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

are linearly independent, and form a basis for $\mathbb{R}^{2\times 1}$. Since representations of $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ with respect to $\boldsymbol{E}$ are themselves, the matrix of change of basis from $\boldsymbol{E}$ to $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2)$ is

$$P = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

Hence the representation of A with respect to $\boldsymbol{R}$ is found as

$$A' = P^{-1}AP = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

4.4 LU Decomposition

Some applications require solving an $n \times n$ system of equations

$$A\boldsymbol{x} = \boldsymbol{b} \tag{4.18}$$

for several values of $\boldsymbol{b}$. Since the elementary row operations involved in reducing A into a row echelon form are independent of $\boldsymbol{b}$, it would be a waste of time to repeat the same operations for each new value of $\boldsymbol{b}$.

LU decomposition is an algorithm, based on Gaussian Elimination, for factoring a non-singular matrix A into a product

$$A = L\,U \tag{4.19}$$

where L is a lower triangular matrix with unity diagonal elements and U is an upper triangular matrix.

With A factored as in (4.19), (4.18) is written as

$$LU\boldsymbol{x} = \boldsymbol{b} \tag{4.20}$$

Defining $\boldsymbol{z} = U\boldsymbol{x}$, the last equation is decomposed into two $n \times n$ systems

$$\begin{aligned} L\,\boldsymbol{z} &= \boldsymbol{b} \\ U\boldsymbol{x} &= \boldsymbol{z} \end{aligned} \tag{4.21}$$

Since L is lower triangular, for any given $\boldsymbol{b}$ the first system in (4.21) can easily be solved for $\boldsymbol{z}$ by means of forward substitutions. Once $\boldsymbol{z}$ is obtained, the second system in (4.21), whose coefficient matrix is upper triangular, can be solved by means of backward substitutions to obtain a solution for $\boldsymbol{x}$. If (4.18) is to be solved for a different $\boldsymbol{b}$, all we have to do is to solve the two systems in (4.21) using simple forward and backward substitutions.

Example 4.12

We illustrate the LU decomposition algorithm on

$$A = \begin{bmatrix} 2 & -2 & 1 \\ -4 & 3 & -3 \\ 6 & -8 & 4 \end{bmatrix}$$

Let $A_1 = A$. The Gaussian Elimination algorithm applied to the first column of A_1 yields

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & 1 \\ -4 & 3 & -3 \\ 6 & -8 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & -2 & 1 \end{bmatrix}$$

which we write in compact form as $L_1 A_1 = A_2$.

Now the algorithm applied to the second column of A_2 yields

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -2 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & 3 \end{bmatrix}$$

or $L_2 A_2 = U$.

Thus $L_2 L_1 A = U$, and therefore, $A = L_1^{-1} L_2^{-1} U = LU$ where

$$L = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

From the example above we observe that the first column elements of the matrix L, which are the first column elements of L_1^{-1}, can be obtained directly from the first column elements of A_1 as

$$l_{i1} = a_{i1}^{(1)} / a_{11}^{(1)}\,, \quad i = 2, \ldots, n$$

Similarly, the second column elements of L are those that appear in L_2^{-1}, and can be obtained from the second column elements of A_2 as

$$l_{i2} = a_{i2}^{(2)}/a_{22}^{(2)}, \quad i = 3, \ldots, n$$

and so on. We also observe that L and U can be stored on the original matrix A, L on the lower left half of A and U on the upper right half and on the diagonal (since the diagonal elements of L are all 1, they need not be stored). These observations lead to the basic LU decomposition algorithm given in Table 4.1, which overwrites A with L and U.

Table 4.1: Basic LU Decomposition Algorithm

```
1.  For j = 1 : n − 1
2.      For i = j + 1 : n
3.          μ_ij = a_ij / a_jj
4.          a_ij ← μ_ij
5.          For q = j + 1 : n
6.              a_iq ← a_iq − μ_ij a_jq
7.          End
8.      End
9.  End
```

Clearly, the algorithm requires that the pivot element must be nonzero at every step. If $a_{jj} = 0$ at the jth step, then to continue the reduction the jth row of A must be interchanged with a row below it to bring a nonzero element to the pivot position. Even if $a_{jj} \neq 0$, for reasons of numerical accuracy, the pivot element is chosen to be the largest element in magnitude among $\{ a_{pj} : p \geq j \}$.[3] Since row interchanges can conveniently be represented by premultiplying A with a permutation matrix P, LU decomposition of A with row interchanges is equivalent to basic LU decomposition of the permuted matrix PA. Rather than using a permutation matrix P to keep track of the row interchanges, a permutation list $\mathbb{P}$ serves the same purpose. With row interchanges, the LU decomposition algorithm is modified as in Table 4.2.

MATLAB provides the build in function for obtaining the LU decomposition. The command `[L,U,P]=lu(A)` returns the matrices involved.

Example 4.13

Obtain the LU decomposition of

$$A = \begin{bmatrix} 1 & 0 & 2 & 2 \\ -2 & -4 & 2 & 0 \\ 4 & 8 & 0 & 4 \\ 2 & 8 & -2 & 6 \end{bmatrix}$$

with partial pivoting.

[3] This is known as partial pivoting.

Table 4.2: LU Decomposition with Partial Pivoting

```
1.  ℙ = 𝕀_n
2.  For j = 1 : n − 1
3.      Find p ≥ j such that |a_pj| = max{ |a_ij| : i ≥ j }
4.      Interchange the jth row of A with the pth row
5.      Interchange the jth element of ℙ with the pth element
6.      For i = j + 1 : n
7.          μ_ij = a_ij / a_jj
8.          a_ij ← μ_ij
9.          For q = j + 1 : n
10.             a_iq ← a_iq − μ_ij a_jq
11.         End
12.     End
13. End
```

The steps of the algorithm are summarized below.

$j = 1: \quad p = 3\,, \quad \mathbb{P} \;\rightarrow\; \{\,3, 2, 1, 4\,\}$

$$A \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ -2 & -4 & 2 & 0 \\ 1 & 0 & 2 & 2 \\ 2 & 8 & -2 & 6 \end{bmatrix} \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ -1/2 & 0 & 2 & 2 \\ 1/4 & -2 & 2 & 1 \\ 1/2 & 4 & -2 & 4 \end{bmatrix}$$

$j = 2: \quad p = 4\,, \quad \mathbb{P} \rightarrow \{\,3, 4, 1, 2\,\}$

$$A \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ 1/2 & 4 & -2 & 4 \\ 1/4 & -2 & 2 & 1 \\ -1/2 & 0 & 2 & 2 \end{bmatrix} \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ 1/2 & 4 & -2 & 4 \\ 1/4 & -1/2 & 1 & 3 \\ -1/2 & 0 & 2 & 2 \end{bmatrix}$$

$j = 3: \quad p = 4\,, \quad \mathbb{P} \rightarrow \{\,3, 4, 2, 1\,\}$

$$A \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ 1/2 & 4 & -2 & 4 \\ -1/2 & 0 & 2 & 2 \\ 1/4 & -1/2 & 1 & 3 \end{bmatrix} \;\longrightarrow\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ 1/2 & 4 & -2 & 4 \\ -1/2 & 0 & 2 & 2 \\ 1/4 & -1/2 & 1/2 & 2 \end{bmatrix}$$

Thus

$$L \;=\; \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ -1/2 & 0 & 1 & 0 \\ 1/4 & -1/2 & 1/2 & 1 \end{bmatrix}, \quad U \;=\; \begin{bmatrix} 4 & 8 & 0 & 4 \\ 0 & 4 & -2 & 4 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

and

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

The reader can verify that

$$LU = PA$$

4.5 Determinant of a Square Matrix

4.5.1 Permutations

A sequence of integers $\mathbb{J}_n = (j_1, j_2, \ldots, j_n)$ in which each integer from 1 to n appears only once is called a ***permutation*** of $\mathbb{N}_n = (1, 2, \ldots, n)$. The sequence $\mathbb{N}_n$ is called the ***natural order*** of the integers form 1 to n. There are $n!$ permutations of n integers including the natural order.

In every permutation other than the natural order there is at least one integer which is followed by one or more smaller integers. The total number of integers that follow a larger integer is called the number of ***inversions*** in a permutation. For example, the permutation $(4, 6, 1, 5, 2, 3)$ contains nine inversions since 4 is followed by $1, 2$, and 3; 6 is followed by $1, 5, 2$ and 3; and 5 is followed by 2 and 3. The ***sign*** of a permutation $\mathbb{J}_n$ is defined as $s(\mathbb{J}_n) = (-1)^k$, where k is the total number of inversions. That is, a permutation has a positive sign if it contains an even number of inversions, and a negative sign if it contains an odd number of inversions.

The interchange of any two integers in a permutation is called a ***transposition***. A transposition involving adjacent integers is an adjacent transposition. If the adjacent integers j_p and j_{p+1} of a permutation $\mathbb{J}_n$ are interchanged, then the total number of inversions is either increased or decreased by exactly one depending on whether $j_p < j_{p+1}$ or $j_p > j_{p+1}$. Thus an adjacent transposition changes the sign of a permutation. Now consider the transposition of any two integers j_p and j_q with $p < q$. This can be achieved by first placing j_p between j_{q-1} and j_q by means of $q - p - 1$ forward adjacent transpositions, and then placing j_q between j_{p-1} and j_{p+1} by means of $q - p$ backward adjacent transpositions. Thus transposition of j_p and j_q requires a total of $2q - 2p - 1$ adjacent transpositions. Since $2q - 2p - 1$ is an odd number, we conclude that any transposition changes the sign of a permutation.

Finally, we note that if a permutation has a total of k inversions, then it can be reduced to the natural order by k adjacent transpositions. To show this, suppose that in the permutation, n is followed by i_n smaller integers, $n-1$ by i_{n-1} smaller integers, etc. Then the total number of inversions is $k = i_n + i_{n-1} + \cdots + i_2$. The integer n can be brought into its natural position by i_n adjacent transpositions, at each step interchanging n with the next integer. Then, $n - 1$ can be put into its natural position by i_{n-1} adjacent transpositions, etc. Thus the permutation can be put into natural order by $i_n + i_{n-1} + \cdots + i_2 = k$ adjacent transpositions.

4.5.2 Determinants

Let A be a square matrix of order n. The scalar associated with A

$$\det A = \sum_{\mathbb{J}_n=(j_1,\ldots,j_n)} s(\mathbb{J}_n) a_{1j_1} a_{2j_2} \cdots a_{nj_n} \tag{4.22}$$

where the sum is taken over all $n!$ permutations of $(1, \ldots, n)$, is called the ***determinant*** of A.

Thus the determinant of a 2×2 matrix is

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

and the determinant of a 3×3 matrix is

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{array}{r} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{array} \tag{4.23}$$

Consider a typical product term $a_{1j_1}a_{2j_2}\cdots a_{nj_n}$ in (4.22). Reordering the elements so that the column indices appear in natural order, we obtain a product term $a_{i_1 1}a_{i_2 2}\cdots a_{i_n n}$, where $\mathbb{I}_n = (i_1, \ldots, i_n)$ is another permutation of the integers $(1, \ldots, n)$. Clearly, $s(\mathbb{J}_n) = s(\mathbb{I}_n)$ as the same transpositions are involved in putting $\mathbb{J}_n$ into natural order and the natural order into $\mathbb{I}_n$. Also, to each product term $a_{1j_1}a_{2j_2}\cdots a_{nj_n}$ in (4.22) there corresponds a unique product term $a_{i_1 1}a_{i_2 2}\cdots a_{i_n n}$. This shows that the determinant of A can also be expressed as

$$\det A = \sum_{\mathbb{I}_n=(i_1,\ldots,i_n)} s(\mathbb{I}_n) a_{i_1 1} a_{i_2 2} \cdots a_{i_n n} \tag{4.24}$$

where the sum is again over all $n!$ permutations. The expressions in (4.22) and (4.24) are called the ***row expansion*** and the ***column expansion*** of $\det A$.

MATLAB function `det` computes the determinant of a square matrix.

The following properties of determinants follow from the definition.

D1. $\det A^t = \det A$.

D2. If B is obtained from A by a Type I elementary row (column) operation, then $\det B = -\det A$.

D3. If any two rows (columns) of A are identical, then $\det A = 0$.

D4. If B is obtained from A by multiplying a row (column) by a scalar c (Type II elementary operation), then $\det B = c \cdot \det A$. As a consequence, if A contains a zero row (column), then $\det A = 0$.

D5. If a row of A is expressed as the sum of two rows as

$$A = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}'_p + \boldsymbol{\alpha}''_p \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix}$$

then $\det A = \det A' + \det A''$, where

$$A' = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}'_p \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix}, \quad A'' = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}''_p \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix}$$

A corresponding property holds for columns of A.

D6. If B is obtained from A by a Type III elementary operation, then $\det B = \det A$.

To prove property D1, we note that if $A^t = B = [\, b_{ij} \,]$ so that $b_{ij} = a_{ji}$ for all (i, j) then

$$\begin{aligned} \det A^t &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) b_{1j_1} b_{2j_2} \cdots b_{nj_n} \\ &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{j_1 1} a_{j_2 2} \cdots a_{j_n n} \;=\; \det A \end{aligned}$$

To prove property D2, suppose B is obtained by interchanging the pth and qth rows of A. Then

$$\begin{aligned} \det B &= \sum_{\mathbb{I}_n} s(i_1, \ldots, i_p, \ldots, i_q, \ldots, i_n) b_{i_1 1} \cdots b_{i_p p} \cdots b_{i_q q} \cdots b_{i_n n} \\ &= \sum_{\mathbb{I}_n} -s(i_1, \ldots, i_q, \ldots, i_p, \ldots, i_n) a_{i_1 1} \cdots a_{i_q p} \cdots a_{i_p q} \cdots a_{i_n n} \\ &= -\det A \end{aligned}$$

The same property also holds if B is obtained from A by interchanging any two columns, because then B^t will be obtained from A^t by interchanging the corresponding rows so that $\det B = \det B^t = -\det A^t = -\det A$. In fact, because of property D1, any result about the determinant of a matrix involving its rows is also valid for its columns, and need not be proved separately.

To prove D3, let B be obtained from A by interchanging the identical rows. Then $B = A$, so that $\det B = \det A$. However, by property D2, we also have $\det B = -\det A$. So, $\det A = 0$.

To prove D4, let B be obtained from A by multiplying pth row by a scalar c. If $A = [\, a_{ij} \,]$ and $B = [\, b_{ij} \,]$, then

$$b_{ij} = \begin{cases} a_{ij}\,, & i \neq p \\ c a_{pj}\,, & i = p \end{cases}$$

and

$$\begin{aligned} \det B &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) b_{1j_1} \cdots b_{pj_p} \cdots b_{nj_n} \\ &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots c a_{pj_p} \cdots a_{nj_n} \\ &= c \cdot \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots a_{pj_p} \cdots a_{nj_n} \;=\; c \cdot \det A \end{aligned}$$

Property D5 follows from

$$\begin{aligned}\det A &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots (a'_{pj_p} + a''_{pj_p}) \cdots a_{nj_n} \\ &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots a'_{pj_p} \cdots a_{nj_n} + \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots a''_{pj_p} \cdots a_{nj_n} \\ &= \det A' + \det A''\end{aligned}$$

Finally, to prove D6, let B be obtained from A by adding c times row p to row q, that is,

$$B = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \\ \vdots \\ \boldsymbol{\alpha}_q + c\boldsymbol{\alpha}_p \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \\ \vdots \\ \boldsymbol{\alpha}_q \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \\ \vdots \\ c\boldsymbol{\alpha}_p \\ \vdots \\ \boldsymbol{\alpha}_n \end{bmatrix} = A + C$$

Then, by D5 and D3, $\det B = \det A + \det C = \det A$.

Using the definition and the properties above, we can find explicit expressions for the determinants of some special matrices. For example, if A is a block upper triangular matrix of the form

$$A = \begin{bmatrix} B & \boldsymbol{c} \\ \mathbf{0} & a \end{bmatrix}$$

then since $a_{nj_n} = 0$ when $j_n \neq n$, (4.22) reduces to

$$\begin{aligned}\det A &= \sum_{\mathbb{J}_n} s(\mathbb{J}_n) a_{1j_1} \cdots a_{n-1,j_{n-1}} a_{nj_n} = \sum_{\mathbb{J}_{n-1}} s(\mathbb{J}_{n-1}) a_{1j_1} \cdots a_{n-1,j_{n-1}} a_{nn} \\ &= a \cdot \sum_{\mathbb{J}_{n-1}} s(\mathbb{J}_{n-1}) b_{1j_1} \cdots b_{n-1,j_{n-1}} = a \cdot \det B\end{aligned}$$

Similarly, if

$$A = \begin{bmatrix} a & \mathbf{0} \\ \boldsymbol{\gamma} & B \end{bmatrix}$$

then

$$\det A = a \cdot \det B$$

Using this result repeatedly we observe that if A is a lower triangular matrix then

$$\begin{aligned}\det \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} &= a_{11} . \det \begin{bmatrix} a_{22} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ a_{n2} & \cdots & a_{nn} \end{bmatrix} \\ &\vdots \\ &= a_{11} a_{22} \cdots a_{nn}\end{aligned}$$

Obviously, the same is true for an upper triangular matrix. An immediate consequence of this result is that

$$\det(\,\text{diag}\,[\,d_1, d_2, \ldots, d_n\,]\,) = d_1 d_2 \cdots d_n$$

As a special case, we have

$$\det I = 1$$

4.5.3 Laplace Expansion of Determinants

Consider the determinant of the 3×3 matrix in (4.23). Grouping the product terms on the right-hand side of the expression, we can write the determinant as

$$\begin{aligned}
&\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \\
&= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \\
&= a_{11} \det \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12} \det \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + a_{13} \det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}
\end{aligned}$$

In the above expression, each determinant multiplying a first row element a_{1j} is precisely the determinant of a 2×2 submatrix of A which is obtained by deleting the first row and the jth column of A. A different grouping of the product terms in (4.23) would result in a similar expression. We now generalize this observation to matrices of arbitrary order.

Let $A = [\,a_{ij}\,]$ be a square matrix of order n. The determinant of the $k \times k$ submatrix of A obtained by deleting any $n-k$ rows and any $n-k$ columns of A is called a ***minor*** of A. Let A_{ij} denote the $(n-1) \times (n-1)$ submatrix of A obtained by deleting the ith row and the jth column of A, and let the corresponding minor be denoted by $m_{ij}^A = \det A_{ij}$. The signed minor $(-1)^{i+j} m_{ij}^A$ is called the ***cofactor*** of the element a_{ij}. We then have

D7. For any fixed $1 \le i \le n$,

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} m_{ij}^A \tag{4.25}$$

Alternatively, for any fixed $1 \le j \le n$,

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} m_{ij}^A \tag{4.26}$$

The expressions in (4.25) and (4.26) are called the ***Laplace expansion*** of $\det A$ with respect to the ith row and the jth column.

To prove (4.25) let us first consider the special case of $i = n$, and express the last row of A as the sum of n rows, the jth one of which contains all 0's except a_{nj} at its jth position. Then, by repeated use of property D5, we can express $\det A$ as

$$\det A = \det A_1 + \cdots + \det A_j + \cdots + \det A_n$$

where

$$A_j = \begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,j} & \cdots & a_{n-1,n} \\ 0 & \cdots & a_{nj} & \cdots & 0 \end{bmatrix}$$

Let B_j be obtained from A_j by moving the jth column to the last position by means of $n-j$ adjacent transpositions, so that

$$\det A_j = (-1)^{n-j} \cdot \det B_j = (-1)^{n+j} \cdot \det B_j\,, \quad j = 1, \ldots, n$$

We also observe that

$$B_j = \begin{bmatrix} A_{nj} & \boldsymbol{b}_j \\ \mathbf{0} & a_{nj} \end{bmatrix}$$

where $\boldsymbol{b}_j = \text{col}\,[\,a_{1j}, \ldots, a_{n-1,j}\,]$, so that

$$\det B_j = a_{nj} \cdot \det A_{nj} = a_{nj} m_{nj}^A$$

Hence

$$\det A = \sum_{j=1}^{n} \det A_j = \sum_{j=1}^{n} (-1)^{n+j} a_{nj} m_{nj}^A$$

This establishes (4.25) for $i = n$.

Now, for any fixed i, let B be the matrix obtained from A by moving the ith row to the nth position by means of $n-i$ adjacent transpositions, so that $\det B = (-1)^{n-i} \det A$. Also, for any j, $b_{nj} = a_{ij}$, and the submatrix B_{nj} is the same as the submatrix A_{ij}. Thus

$$\det A = (-1)^{n-i} \det B = (-1)^{i-n} \sum_{j=1}^{n} (-1)^{n+j} b_{nj} m_{nj}^B = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} m_{ij}^A$$

which proves (4.25) for an arbitrary i. Finally, (4.26) follows from (4.25) on using property D1.

Example 4.14

Find the determinant of

$$A = \begin{bmatrix} 3 & 0 & -1 & 2 \\ -1 & 1 & 3 & 0 \\ 2 & 2 & 0 & 4 \\ -4 & 0 & 1 & 1 \end{bmatrix}$$

Since the second column contains most zeros, we prefer to expand $\det A$ with respect to the second

column, because we do not need to calculate the cofactors of zero elements. Thus

$$\begin{aligned}
&\det A \\
&= (-1)^{2+2}.1.\det\begin{bmatrix} 3 & -1 & 2 \\ 2 & 0 & 4 \\ -4 & 1 & 1 \end{bmatrix} + (-1)^{3+2}.2.\det\begin{bmatrix} 3 & -1 & 2 \\ -1 & 3 & 0 \\ -4 & 1 & 1 \end{bmatrix} \\
&= \{(-1)^{1+2}.(-1).\det\begin{bmatrix} 2 & 4 \\ -4 & 1 \end{bmatrix} + (-1)^{3+2}.1.\det\begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}\} \\
&\quad -2.\{(-1)^{1+3}.2.\det\begin{bmatrix} -1 & 3 \\ -4 & 1 \end{bmatrix} + (-1)^{3+3}.1.\det\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}\} \\
&= \{(2+16)-(12-4)\} - 2.\{2.(-1+12)+(9-1)\} = -50
\end{aligned}$$

Example 4.15

Laplace expansion becomes even more convenient when coupled with elementary operations of Type III as we illustrate below.

$$\begin{aligned}
\det\begin{bmatrix} i & 0 & 1+i & i \\ 0 & i & 1-i & -1 \\ i & -i & -1 & 1 \\ -1 & 0 & 2+i & i \end{bmatrix} &= \det\begin{bmatrix} i & 0 & 1+i & i \\ 0 & i & 1-i & -1 \\ i & 0 & -i & 0 \\ -1 & 0 & 2+i & i \end{bmatrix} \\
&= i\cdot\det\begin{bmatrix} i & 1+i & i \\ i & -i & 0 \\ -1 & 2+i & i \end{bmatrix} \\
&= i\cdot\det\begin{bmatrix} i & 1+i & i \\ i & -i & 0 \\ -1-i & 1 & 0 \end{bmatrix} \\
&= i^2\cdot\det\begin{bmatrix} i & -i \\ -1-i & 1 \end{bmatrix} \\
&= -1
\end{aligned}$$

From properties D2, D4 and D6 we observe that for an elementary matrix

$$\det E = \begin{cases} -1, & \text{if } E \text{ is a Type I elementary matrix} \\ c, & \text{if } E \text{ is a Type II elementary matrix} \\ 1, & \text{if } E \text{ is a Type III elementary matrix} \end{cases}$$

This observation allows us to conclude that if E is an elementary matrix then

a) $\det E \neq 0$

b) $\det(EA) = \det(AE) = (\det E)(\det A)$

Let $R = E_k \cdots E_1 A$ be the reduced row echelon form of A, so that $\det R = (\det E_k)\cdots(\det E_1)(\det A)$. If A is singular then $\det R = 0$ (as it contains one or more zero rows), and we must have $\det A = 0$. If, on the other hand, A is nonsingular, then $R = I$ so that $\det R = 1$, and therefore, $\det A \neq 0$. We thus obtain the following result.

D8. A is nonsingular if and only if $\det A \neq 0$.

Now consider a product AB. If A is singular then so is AB, and we have $\det AB = 0$. If A is nonsingular then we can represent it as a product of elementary matrices as $A = E_k \cdots E_1$. Then $AB = E_k \cdots E_1 B$ so that $\det AB = (\det E_k) \cdots (\det E_1)(\det B) = (\det A)(\det B)$. In either case, we have

D9. $\det(AB) = (\det A)(\det B)$

An immediate consequence of Property D9 is that for a nonsingular matrix A

$$\det A^{-1} = (\det A)^{-1}$$

4.5.4 Cramer's Rule and a Formula for A^{-1}

Let A be a nonsingular matrix of order n. The Cramer's rule states that the unique solution of the system

$$A\boldsymbol{x} = \boldsymbol{b} \tag{4.27}$$

is given by

$$x_j = \frac{\det B_j}{\det A}\ , \quad j = 1, 2, \ldots, n$$

where B_j is obtained by replacing the jth column of A with $\boldsymbol{b}$, that is,

$$B_j = \begin{bmatrix} a_{11} & \ldots & a_{1,j-1} & b_1 & a_{1,j+1} & \ldots & a_{1n} \\ a_{21} & \ldots & a_{2,j-1} & b_2 & a_{2,j+1} & \ldots & a_{2n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \ldots & a_{n,j-1} & b_n & a_{n,j+1} & \ldots & a_{nn} \end{bmatrix}$$

Expanding $\det B_j$ with respect to the jth column we get

$$x_j = \frac{1}{\det A} \sum_{p=1}^{n} (-1)^{p+j} b_p m_{pj}^A\ , \quad j = 1, 2, \ldots, n \tag{4.28}$$

That x_j's given by (4.28) satisfy the system can be verified by substitution and using the properties of determinants as follow.

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} x_j &= \frac{1}{\det A} \sum_{j=1}^{n} a_{ij} \left[\sum_{p=1}^{n} (-1)^{p+j} b_p m_{pj}^A \right] \\ &= \frac{1}{\det A} \sum_{p=1}^{n} b_p \left[\sum_{j=1}^{n} (-1)^{p+j} a_{ij} m_{pj}^A \right] \\ &= \frac{1}{\det A} \sum_{p=1}^{n} \delta_{ip} (\det A) b_p \\ &= b_i\ , \quad i = 1, 2, \ldots, n \end{aligned}$$

where we used the symbol

$$\delta_{ip} = \begin{cases} 1\,, & p = i \\ 0\,, & p \neq i \end{cases}$$

to express the fact that

$$\sum_{i=1}^{n} (-1)^{p+j} a_{ij} m_{pj}^{A} = \begin{cases} \det A\,, & p = i \\ 0 \quad , & p \neq i \end{cases}$$

Example 4.16

Cramer's rule applied to a 2×2 system

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

gives

$$x_1 = \frac{1}{\det A} \det \begin{bmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{bmatrix} = \frac{a_{22}b_1 - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}$$

$$x_2 = \frac{1}{\det A} \det \begin{bmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{bmatrix} = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{12}a_{21}}$$

or in matrix form

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\det A} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

Since the unique solution of the system $A\boldsymbol{x} = \boldsymbol{b}$ is $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, the matrix on the right-hand side of the last expression in Example 4.16 must be the inverse of the coefficient matrix. This example shows that the Cramer's rule can be used to obtain a formula for the inverse of a 2×2 nonsingular matrix in terms of determinants. Let us generalize this result to higher order matrices.

Let the inverse of an $n \times n$ nonsingular matrix $A = [\, a_{ij} \,]$ be expressed in terms of its columns as

$$A^{-1} = [\, \hat{\boldsymbol{a}}_1 \;\; \hat{\boldsymbol{a}}_2 \;\; \cdots \;\; \hat{\boldsymbol{a}}_n \,]$$

Since

$$I = [\, \boldsymbol{e}_1 \;\; \boldsymbol{e}_2 \;\; \cdots \;\; \boldsymbol{e}_n \,] = AA^{-1} = [\, A\hat{\boldsymbol{a}}_1 \;\; A\hat{\boldsymbol{a}}_2 \;\; \cdots \;\; A\hat{\boldsymbol{a}}_n \,]$$

each $\hat{\boldsymbol{a}}_j$ is the unique solution of the system $A\boldsymbol{x} = \boldsymbol{e}_j$. Thus if

$$\hat{\boldsymbol{a}}_j = \text{col}\,[\, \hat{a}_{1j}, \ldots, \hat{a}_{nj} \,]$$

then by Cramer's rule we have

$$\begin{aligned} \hat{a}_{ij} &= \frac{1}{\det A} \sum_{p=1}^{n} (-1)^{p+i} \delta_{pj} m_{pi}^{A} \\ &= \frac{1}{\det A} (-1)^{i+j} m_{ji}^{A}\,, \quad i, j = 1, \ldots, n \end{aligned}$$

We thus obtain the formula

$$A^{-1} = \frac{1}{\det A} \text{ adj } A \tag{4.29}$$

for the inverse of A, where

$$\text{adj } A = [\,(-1)^{i+j} m_{ij}^{A}\,]^t$$

is called the ***adjugate*** of A. Note that adj A is the transpose of a matrix consisting of the cofactors of A.

Example 4.17

The determinant and the minors of the matrix

$$A = \begin{bmatrix} 1 & 2 & -1 \\ -1 & 0 & 3 \\ 1 & 2 & 0 \end{bmatrix}$$

are found as

$$\det A = \det \begin{bmatrix} 1 & 2 & -1 \\ -1 & 0 & 3 \\ 0 & 0 & 1 \end{bmatrix} = \det \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} = 2$$

and

$$m_{11} = \det \begin{bmatrix} 0 & 3 \\ 2 & 0 \end{bmatrix} = -6 \qquad m_{12} = \det \begin{bmatrix} -1 & 3 \\ 1 & 0 \end{bmatrix} = -3$$

$$m_{13} = \det \begin{bmatrix} -1 & 0 \\ 1 & 2 \end{bmatrix} = -2 \qquad m_{21} = \det \begin{bmatrix} 2 & -1 \\ 2 & 0 \end{bmatrix} = 2$$

$$m_{22} = \det \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} = 1 \qquad m_{23} = \det \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} = 0$$

$$m_{31} = \det \begin{bmatrix} 2 & -1 \\ 0 & 3 \end{bmatrix} = 6 \qquad m_{32} = \det \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix} = 2$$

$$m_{33} = \det \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} = 2$$

Then

$$A^{-1} = \frac{1}{2} \begin{bmatrix} -6 & 3 & -2 \\ -2 & 1 & 0 \\ 6 & -2 & 2 \end{bmatrix}^t = \frac{1}{2} \begin{bmatrix} -6 & -2 & 6 \\ 3 & 1 & -2 \\ -2 & 0 & 2 \end{bmatrix}$$

In practice, formula (4.29) is seldom used to calculate the inverse of a matrix. Gaussian Elimination (as in Example 4.8) is preferred for reasons of efficiency and numerical accuracy.

4.6 Exercises

1. Find bases for the row and the column spaces of the coefficient matrices in Exercise 1.22.
2. Prove that if two $m \times n$ matrices R_1 and R_2 in reduced row echelon form are row equivalent, then $R_1 = R_2$. Explain how this result implies that the reduced row echelon form of a matrix is unique. Hint: Since $\text{rs}\,(R_1) = \text{rs}\,(R_2)$, $r_1 = r_2$. Also, column indices of the leading entries of R_1 and R_2 must be the same.
3. Use MATLAB command `rank(A)` to find the rank of matrices in Exercise 1.22. Do the results agree with the results you obtained in Exercise 4.1?
4. A famous example of ill-conditioned matrices are ***Hilbert*** matrices.[4] A Hilbert matrix of order n is defined as

   $$H_n = [\,\frac{1}{i+j-1}\,]_{n\times n}$$

   Thus

   $$H_2 = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix} \quad \text{and} \quad H_3 = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}$$

   It is known that Hilbert matrices are nonsingular, that is, $r(H_n) = n$.

   (a) Use MATLAB command `rref` to find reduced row echelon forms and ranks of H_n for $n = 10, 11, 12, 13$.

   (b) Use MATLAB command `rank` to find ranks of H_n for $n = 10, 11, 12, 13$. Apparently, MATLAB does not use the reduced row echelon form to compute the rank of a matrix.[5]
5. Show that an elementary row (column) operation on an $m \times n$ matrix A is equivalent to premultiplying (postmultiplying) A with the corresponding elementary matrix.
6. Show that if E is an elementary matrix which represents an elementary operation on the rows of a square matrix A, then E^t represents the same operation on the corresponding columns of A.
7. Write down the inverses of the following elementary matrices.

   $$E_1 = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

8. Show that if P is a permutation matrix, then $P^{-1} = P^t$.
9. (a) Referring to (4.7), show that $\hat{A}_L = Q_1 + XQ_2$ is a left inverse of A for arbitrary choice of X.

   (b) Express each of the left inverses of A considered in Example 3.38 as above.
10. (a) Referring to (4.9), show that $\hat{A}_R = P_1 + P_2Y$ is a right inverse of A for arbitrary choice of Y.

    (b) Express the right inverse of B considered in Example 3.39 as above.
11. Let $A \in \mathbb{F}^{m\times n}$. Show that

    (a) $\ker\,(A) = \{\,\mathbf{0}\,\}$ if and only if $r(A) = n$.

    (b) $\text{im}\,(A) = \mathbb{F}^{m\times 1}$ if and only if $r(A) = m$.

[4] An application involving Hilbert matrices is considered in Exercise 7.33.

[5] We will consider the algorithm used by MATLAB in Chapter 8.

12. (a) Find inverses of the following matrices by using Gaussian Elimination.

    (b) Use MATLAB command `inv` to compute the inverses of the same matrices.

$$A = \begin{bmatrix} 3 & 2 & -1 \\ 1 & 1 & 1 \\ 2 & 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1+i & 2i & -1 & -2i \\ 1-2i & -1-i & 1+i & 1 \\ 1-2i & -1 & 1+i & 1-i \\ 2i & -1 & -1-i & 1+i \end{bmatrix}$$

13. Write the matrices in Exercise 4.12 as products of elementary matrices.

14. Find inverses of

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix}_{n \times n}$$

    by inspection.

15. Execute the following MATLAB commands and comment on the result.

```
A=hilb(10); B=inv(A);
C=A*B
```

16. Use MATLAB command `pinv` to find a left inverse of the A matrix in Example 4.6 and a right inverse of the B matrix in Example 4.7. Verify that $\hat{A}_L A = I$ and $B\hat{B}_R = I$.

17. Find two different generalized inverses of the matrix

$$A = \begin{bmatrix} 1+i & 1 & i \\ 1 & 1-i & i \\ i & -1 & 1+i \end{bmatrix}$$

18. (a) Referring to (4.13), show that $\hat{A}_G = (P_1 + P_2 Y)(Q_1 + XQ_2)$ is a generalized inverse of A for arbitrary choices of X and Y.

    (b) Express the generalized inverse computed by MATLAB in Example 4.9 as above.

19. Show that if $r(A) = r$ and $\hat{A}_G$ is a generalized inverse of A then $r(\hat{A}_G) = r$.

20. Let

$$A = \begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} \end{bmatrix}$$

    where A_{11} and A_{22} are square submatrices.

    (a) Show that A is nonsingular if and only if A_{11} and A_{22} are both nonsingular.

    (b) Assuming A is nonsingular, find A^{-1} in terms of A_{11}^{-1}, A_{22}^{-1} and A_{21}.

21. Let $A \in \mathbb{F}^{m \times n}$ and $B \in \mathbb{F}^{n \times m}$ be such that $I_m + AB$ is nonsingular.

    (a) Show that $I_n + BA$ is also nonsingular. Hint: If $I_n + BA$ is singular, then $(I_n + BA)\boldsymbol{c} = \boldsymbol{0}$ for some $\boldsymbol{c} \neq \boldsymbol{0}$. Premultiply both sides with A.

    (b) Show that $(I_m + AB)^{-1} A = A(I_n + BA)^{-1}$.

(c) Verify (b) for

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

22. Show that if $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$ is an ordered basis for $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{r}'_j$ are as defined in (4.15), where $P = [\,p_{ij}\,]$ is nonsingular, then $\boldsymbol{R}' = (\boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_n)$ is also a basis for $\boldsymbol{\mathcal{X}}$. Hint: Show that $\boldsymbol{R}'$ is linearly independent.

23. (a) Obtain normal forms of the coefficient matrices in Exercise 1.22.

(b) Verify your results by using the MATLAB commands

```
R=rref(A);      % Reduced row echelon form of A
N=rref(R')';    % Reduced column echelon form of R
```

24. Find bases for $\mathbb{R}^{4\times 1}$ and $\mathbb{R}^{3\times 1}$ with respect to which the matrix A in Example 4.10 has the representation R, C or N.

25. Let A be an $m \times n$ matrix with $r(A) = r$. Show that it can be expressed as $A = BC$, where B is an $m \times r$ matrix with full column rank (that is, $r(B) = r$) and C is an $r \times n$ matrix with full row rank (that is, $r(C) = r$). Hint: Partition Q^{-1} and P^{-1} in (4.14) suitably.

26. Obtain LU decompositions of the matrices in Exercises 4.12 and 4.14

(a) without row interchanges

(b) with arbitrary row interchanges.

27. Use MATLAB command `lu` to find the LU decompositions of the matrices in Exercises 4.12 and 4.14.

28. Explain why permuting rows of A does not cause any difficulty in the use of LU decomposition in solving linear systems. Hint: (4.20) becomes

$$LU\boldsymbol{x} = P\boldsymbol{b}$$

and accordingly, the first equation in (4.21) has to be slightly modified.

29. Find the solution of $A\boldsymbol{x} = \boldsymbol{b}$ where A is the matrix in Example 4.12 and

$$\boldsymbol{b} = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

by forward and backward substitutions.

30. (a) Obtain an LU decomposition of

$$A = \begin{bmatrix} 2 & -1 & 1 & 3 \\ -2 & 1 & -1 & -1 \\ 0 & 1 & 1 & -2 \\ 4 & -3 & 3 & 8 \end{bmatrix}$$

(b) Find the solution of $A\boldsymbol{x} = \boldsymbol{b}$ by forward and backward substitutions for

(i) $\boldsymbol{b} = \begin{bmatrix} 1 \\ -3 \\ 4 \\ 0 \end{bmatrix}$ (ii) $\boldsymbol{b} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$

31. Execute the following MATLAB commands and comment on the result.

```
A=hilb(5);
[L,U,P]=lu(A);
C=PA-L*U
```

32. Find determinants of the following matrices.

$$A = \begin{bmatrix} 1 & 0 & 2 & 3 \\ -2 & 0 & -4 & -5 \\ 2 & 1 & 4 & 0 \\ 0 & 1 & 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{bmatrix}_{n \times n}$$

33. Find the determinant of the matrix in Exercise 4.30.

34. Show that if P is a permutation matrix then $\det P = \mp 1$.

35. (a) Use MATLAB command `det` to find the determinant of the A matrix in Exercise 4.32.

    (b) Use MATLAB command `rand` to generate several random matrices (of various orders), and compute their determinants using the `det` command. Observe that all the matrices you generated randomly are nonsingular.

    (c) Generate Hilbert matrices of order $n = 2, \ldots, 20$, using the MATLAB command `hilb`, and compute their determinants using the MATLAB command `det`. Comment on the result.

36. Show that for the block lower triangular matrix A in Exercise 4.20

$$\det A = (\det A_{11})(\det A_{22})$$

Hint: The result is obvious if A_{11} is singular. If A_{11} is nonsingular, let $A_{11} = E_1 \cdots E_k$, where E_j's are elementary matrices.

37. Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^{3\times 1}$ be fixed linearly independent vectors. Show that the set of all $\boldsymbol{x} \in \mathbb{R}^{3\times 1}$ for which

$$\det [\, \boldsymbol{x} \;\; \boldsymbol{p} \;\; \boldsymbol{q} \,] = 0$$

is a subspace of $\mathbb{R}^{3\times 1}$, and find a basis for it.

38. Show that the equation

$$\det \begin{bmatrix} 1 & x & y \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{bmatrix} = 0$$

describes a straight line through the points (x_1, y_1) and (x_2, y_2) in the xy plane.

39. (a) Let $\boldsymbol{a} \in \mathbb{R}^{n\times 1}$ and $q \in \mathbb{R}$ be given. Obtain a linear system in $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ whose solutions are exactly the same as the solutions of

$$\det \begin{bmatrix} 1 & \boldsymbol{a}^T \\ \boldsymbol{x} & I_n \end{bmatrix} = q$$

    (b) Find all solutions of the above equation for $\boldsymbol{a} = \text{col}\,[\,1, 2, 3\,]$ and $q = 0$.

40. Let

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ r_1 & r_2 & \cdots & r_n \\ \vdots & \vdots & & \vdots \\ r_1^{n-1} & r_2^{n-1} & \cdots & r_n^{n-1} \end{bmatrix}$$

Use induction on n to show that

$$\det V = \prod_{i=2}^{n} \prod_{j=1}^{i-1} (r_i - r_j)$$

V is called a ***Vandermonde's matrix***.

41. Solve the following linear system of equations by using the Cramer's rule.

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & 6 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$$

42. Use formula (4.29) to calculate inverses of A in Exercise 4.12 and C in Exercise 4.14.

Chapter 5
Structure of Square Matrices

5.1 Eigenvalues and Eigenvectors

Recall from Section 2.4 that a second order linear homogeneous differential equation with constant coefficients

$$y'' + a_1 y' + a_2 y = 0 \tag{5.1}$$

has a complex solution of the form $y = \phi(t) = e^{st}$. Also recall from from Section 2.7 that (5.1) is equivalent to a system of two first order differential equations

$$\boldsymbol{x}' = A\boldsymbol{x} \tag{5.2}$$

where

$$A = \begin{bmatrix} 0 & 1 \\ -a_2 & -a_1 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} y \\ y' \end{bmatrix}$$

that has a solution of the form

$$\boldsymbol{x} = \boldsymbol{\phi}(t) = \begin{bmatrix} \phi(t) \\ \phi'(t) \end{bmatrix} = \begin{bmatrix} e^{st} \\ se^{st} \end{bmatrix} = e^{st} \begin{bmatrix} 1 \\ s \end{bmatrix} = e^{st}\boldsymbol{v} \tag{5.3}$$

where

$$\boldsymbol{v} = \text{col}\,[\,1, s\,]$$

Substituting $\boldsymbol{x} = e^{st}\boldsymbol{v}$ and $\boldsymbol{x}' = se^{st}\boldsymbol{v}$ into (5.2) and cancelling out the nonzero terms e^{st}, we obtain

$$A\boldsymbol{v} = s\boldsymbol{v} \tag{5.4}$$

(5.4) provides a necessary and sufficient condition to be satisfied by s in order for $\boldsymbol{x} = e^{st}\boldsymbol{v}$ to be solution of (5.2), and therefore, must be related to the characteristic equation of (5.1). Indeed, writing (5.4) in open form we get

$$\begin{bmatrix} 0 & 1 \\ -a_2 & -a_1 \end{bmatrix} \begin{bmatrix} 1 \\ s \end{bmatrix} = \begin{bmatrix} s \\ -a_1 s - a_2 \end{bmatrix} = s \begin{bmatrix} 1 \\ s \end{bmatrix} = \begin{bmatrix} s \\ s^2 \end{bmatrix}$$

The first equation above is an identity, and the second is the characteristic equation of (5.1):

$$s^2 + a_1 s + a_2 = 0$$

In the study above, (5.2) is derived from the second order differential equation in (5.1). This not only results in a coefficient matrix A having a special structure, but also constrains the vector $\boldsymbol{v}$ to the form in (5.3). Now suppose that we started directly with (5.2) where no special structure was imposed on the coefficient matrix A. Then we could still assume a solution of the form $\boldsymbol{x} = e^{st}\boldsymbol{v}$, where $\boldsymbol{v}$ is not restricted to a special form, and end up with (5.4).

Systems of differential equations is just one example where we come across the matrix equation (5.4). There are many other significant problems that lead to a similar equation. In fact, (5.4) is just a special case of a more general equation

$$\mathcal{A}(\boldsymbol{x}) = s\boldsymbol{x} \tag{5.5}$$

involving a linear operator $\mathcal{A}$ defined on a vector space $\boldsymbol{\mathcal{X}}$. This equation simply asks for such $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ that are not changed (except for a scaling) by $\mathcal{A}$, which is a significant question in many scientific and engineering problems.

Eigenvalues and Eigenvectors of A Square Matrix

Let $A \in \mathbb{C}^{n\times n}$. A complex scalar $s = \lambda$ that satisfies (5.4) for some nonzero vector $\boldsymbol{v} \in \mathbb{C}^{n\times 1}$ is called an ***eigenvalue*** of A, and $\boldsymbol{v}$ is called an ***eigenvector*** of A associated with the eigenvalue λ.

Rewriting (5.4) as

$$(sI - A)\boldsymbol{v} = \boldsymbol{0}$$

we get an $n \times n$ homogeneous linear system. By Theorem 1.1 it has a nonzero solution if and only if the coefficient matrix is singular, or equivalently, if and only if

$$\det(sI - A) = 0$$

Treating s as a parameter, we observe that $sI - A$ is a matrix whose elements are simple polynomials in s. The diagonal elements of $sI - A$ are $s - a_{ii}$, and the off-diagonal elements are $-a_{ij}$. Therefore, $\det(sI - A)$ is also a polynomial in s, called the ***characteristic polynomial*** of A, denoted $d(s)$. It is not too difficult to show that $d(s)$ is an nth degree polynomial with unity leading coefficient (see Exercise 5.2). That is,

$$d(s) = \det(sI - A) = s^n + d_1 s^{n-1} + \cdots + d_n$$

We thus reach the conclusion that equation (5.4) has a nonzero solution if and only if s is a root of the ***characteristic equation***

$$d(s) = 0 \tag{5.6}$$

Since $d(s)$ is an nth degree polynomial, by the fundamental theorem of algebra, the characteristic equation (5.6) has exactly n complex roots counting the multiplicities of the repeated roots (if any). Suppose that it has k distinct roots $\lambda_1, \ldots, \lambda_k$ with multiplicities $n_1, \ldots, n_k$, so that

$$d(s) = \prod_{i=1}^{k} (s - \lambda_i)^{n_i} \tag{5.7}$$

where

$$\sum_{i=1}^{k} n_i = n$$

Then each λ_i is an eigenvalue of A with ***algebraic multiplicity*** n_i. Any nonzero solution $\boldsymbol{v} = \boldsymbol{v}_i$ of

$$A\boldsymbol{v} = \lambda_i \boldsymbol{v}$$

is an eigenvector associated with λ_i. There are infinitely many eigenvectors associated with an eigenvalue. In fact, together with the zero vector, the set of all eigenvectors associated with λ_i form a subspace

$$\boldsymbol{\mathcal{K}}_i = \ker (A - \lambda_i I)$$

which is called the ***eigenspace*** of A associated with λ_i. The dimension of $\boldsymbol{\mathcal{K}}_i$

$$\nu_i = \dim (\boldsymbol{\mathcal{K}}_i)$$

is called the ***geometric multiplicity*** of the eigenvalue λ_i. Clearly, ν_i is the maximum number of linearly independent eigenvectors associated with the corresponding eigenvalue.

Some immediate results concerning eigenvalues can be derived from the definition and from properties of determinants, and are listed below.

a) A is singular if and only if it has a zero eigenvalue.

b) A and A^t have the same characteristic polynomial, and therefore, the same eigenvalues. (However, eigenvectors of A^t are, in general, different from those of A.)

c) Eigenvalues of a lower (upper) triangular matrix are its diagonal elements.

Property (a) follows directly from the definition of an eigenvalue. Property (b) is a consequence of the equality

$$\det (sI - A^t) = \det (sI - A)^t = \det (sI - A)$$

Finally, property (c) follows from

$$\begin{aligned} \det (sI - A) &= \det \begin{bmatrix} s - a_{11} & 0 & \cdots & 0 \\ -a_{21} & s - a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ -a_{n1} & -a_{n2} & \cdots & s - a_{nn} \end{bmatrix} \\ &= (s - a_{11})(s - a_{22}) \cdots (s - a_{nn}) \end{aligned}$$

Note that property (c) is also true for a diagonal matrix. In particular, I_n has the only eigenvalue $\lambda_1 = 1$ with algebraic multiplicity $n_1 = n$. Also, since $I_n - \lambda_1 I_n = O$, $\nu_1 = n$, and any nonzero vector is an eigenvector of I.

Example 5.1

Let us find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 2 & -1 \\ -2 & 3 \end{bmatrix}$$

The characteristic polynomial of A is

$$d(s) = \det \begin{bmatrix} s-2 & 1 \\ 2 & s-3 \end{bmatrix} = (s-2)(s-3) - 2 = s^2 - 5s + 4 = (s-1)(s-4)$$

Hence A has two simple eigenvalues:

$$\lambda_1 = 1\,, \quad \lambda_2 = 4$$

An eigenvector associated with $\lambda_1 = 1$ is obtained by solving the equation

$$(A - \lambda_1 I)\boldsymbol{v} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathbf{0}$$

to be

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Similarly, solving

$$(A - \lambda_2 I)\boldsymbol{v} = \begin{bmatrix} -2 & -1 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \mathbf{0}$$

we obtain an eigenvector associated with $\lambda_2 = 4$ as

$$\boldsymbol{v}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Note that any scalar multiple of $\boldsymbol{v}_1$ is also an eigenvector associated with $\lambda_1 = 1$, and the same is true for any scalar multiple of $\boldsymbol{v}_2$. Thus

$$\mathcal{K}_1 = \text{span}\,(\boldsymbol{v}_1) \quad \text{and} \quad \mathcal{K}_2 = \text{span}\,(\boldsymbol{v}_2)$$

Example 5.2

The characteristic equation of the matrix

$$A = \begin{bmatrix} 0 & 1 \\ -5 & 2 \end{bmatrix}$$

is

$$d(s) = \det \begin{bmatrix} s & -1 \\ 5 & s-2 \end{bmatrix} = s^2 - 2s + 5 = 0$$

The eigenvalues of A are the complex conjugate roots

$$\lambda_1 = 1 + 2i\,, \quad \lambda_2 = \lambda_1^* = 1 - 2i$$

of the characteristic equation. The reader can verify that

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1+2i \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 1-2i \end{bmatrix}$$

are associated eigenvectors. This example shows that even when A is real its eigenvalues and eigenvectors may be complex. However, they appear in conjugate pairs. ∎

Eigenvalues and Eigenvectors of A Linear Operator

Since any linear operator $\mathcal{A}$ on a finite dimensional vector space $\boldsymbol{\mathcal{X}}$ over a field $\mathbb{F}$ is represented by a square matrix $A \in \mathbb{F}^{n\times n}$, the concept of eigenvalues and eigenvectors can be generalized to such operators as illustrated by the following example.

Example 5.3

Let $\mathcal{A}$ denote the reflection of the xy plane in the line $y = x$. That is, $\mathcal{A}(\boldsymbol{v})$ is the mirror image of the vector $\boldsymbol{v}$ with respect to the reflecting line. It is not difficult to show that $\mathcal{A}$ is a linear operator on the xy plane.

If a vector $\boldsymbol{r}$ lies in this line then it is mapped into itself, that is, $\mathcal{A}(\boldsymbol{r}) = \boldsymbol{r}$. It follows that $\lambda = 1$ is an eigenvalue of $\mathcal{A}$, with $\boldsymbol{r}$ being an associated eigenvector. On the other hand, if a vector $\boldsymbol{s}$ lies in the line $y = -x$, which is orthogonal to the line $y = x$, then $\mathcal{A}(\boldsymbol{s}) = -\boldsymbol{s}$ so that $\lambda = -1$ is also an eigenvalue of $\mathcal{A}$ with $\boldsymbol{s}$ being an associated eigenvector.

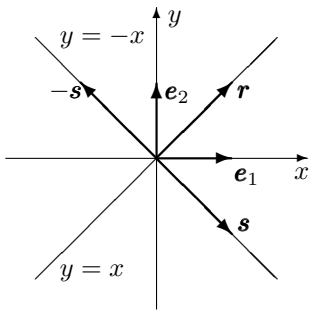

Figure 5.1: Reflection of the xy plane in the line $y = x$.

The statements above, which are illustrated in Figure 5.1, can formally be shown by representing $\mathcal{A}$ with a matrix. If we identify the xy plane with $\mathbb{R}^{2\times 1}$, then since the unit vectors along the x and y axes are mapped into each other, we have

$$\mathcal{A}(\boldsymbol{e}_1) = \boldsymbol{e}_2\,, \quad \mathcal{A}(\boldsymbol{e}_2) = \boldsymbol{e}_1$$

and $\mathcal{A}$ is represented by the matrix

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

with respect to the canonical basis. The characteristic polynomial of A is

$$\det \begin{bmatrix} s & -1 \\ -1 & s \end{bmatrix} = s^2 - 1 = (s-1)(s+1)$$

Hence A has the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$. It is easy to verify that

$$\boldsymbol{v}_1 = \boldsymbol{r} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\,, \quad \boldsymbol{v}_2 = \boldsymbol{s} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

are associated eigenvectors.

Now let $\mathcal{B}$ denote a rotation in the xy plane through an angle $0 < \theta < \pi$ counter-clock-wise. Then there exists no vector in the plane whose image is a (real) multiple of itself. Hence $\mathcal{B}$ has no eigenvectors in the plane. The reader can show that $\mathcal{B}$ is represented by the matrix

$$B = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

with respect to the canonical basis. Calculating the characteristic polynomial of B as

$$\det \begin{bmatrix} s - \cos\theta & \sin\theta \\ -\sin\theta & s - \cos\theta \end{bmatrix} = s^2 - 2\,\cos\theta\, s + 1$$

we observe that B has a pair of complex conjugate eigenvalues

$$\lambda_{1,2} = \cos\theta \mp i \sin\theta$$

Since B has no real eigenvalues, it has no real eigenvectors, verifying our observation. It does, however, have a pair of complex conjugate eigenvectors, which are

$$\boldsymbol{v}_{1,2} = \begin{bmatrix} 1 \\ \mp i \end{bmatrix}$$

The concept of eigenvalues and eigenvectors can also be generalized to linear operators on infinite dimensional vector spaces:

* **Example 5.4**

Let

$$\boldsymbol{\mathcal{X}} = \{\, \phi \,|\, \phi \in \mathcal{C}_\infty(\mathbb{R}, \mathbb{R})\,,\ \phi(0) = \phi(\pi) = 0 \,\}$$

Clearly, $\boldsymbol{\mathcal{X}}$ is a subspace of $\mathcal{C}_\infty(\mathbb{R}, \mathbb{R})$. Let $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be defined as

$$\mathcal{A}(\phi) = \phi''$$

Then the eigenvalue problem

$$\mathcal{A}(\phi) = s\phi$$

is equivalent to determining those s for which the boundary value problem

$$y'' = sy\,, \quad y(0) = y(\pi) = 0$$

has a nonzero solution.

This boundary value problem, which has already been considered in Exercise 2.33, has a nonzero solution if and only if

$$s = \lambda_n = -n^2\,, \quad n = 1, 2, \ldots$$

in which case a solution is

$$y = \phi_n(t) = \sin nt\,, \quad n = 1, 2, \ldots$$

We thus conclude that the operator $\mathcal{A}$ has infinitely many eigenvalues $\lambda_n = -n^2$ with associated eigenvectors $\phi_n(t) = \sin nt$.

When $\mathcal{A}$ is a linear operator defined on a function space as in this example, its eigenvectors are also called ***eigenfunctions***.

In the rest of this chapter, we will deal with linear operators defined by square matrices only.

5.2 The Cayley-Hamilton Theorem

Let $A \in \mathbb{F}^{n\times n}$, and let

$$p(s) = p_0 s^m + p_1 s^{m-1} + \cdots + p_{m-1}s + p_m$$

be a polynomial in s with $p_i \in \mathbb{F}, i = 1, \ldots, m$. We define the matrix $p(A) \in \mathbb{F}^{n\times n}$ as

$$p(A) = p_0 A^m + p_1 A^{m-1} + \cdots + p_{m-1}A + p_m I$$

From the definition it follows that for arbitrary polynomials p and q

$$p(A) + q(A) = (p+q)(A) \quad \text{and} \quad p(A)q(A) = q(A)p(A) = (pq)(A)$$

where $p + q$ and pq are polynomials obtained from p and q by the familiar rules.

We now state and prove one of the key theorems of matrix algebra.

Theorem 5.1 (Cayley-Hamilton) *Let the characteristic polynomial of A be $d(s)$. Then $d(A) = O$.*

Proof Consider

$$(sI - A)^{-1} = \frac{1}{\det(sI - A)} \operatorname{adj}(sI - A) = \frac{1}{d(s)} B(s) \tag{5.8}$$

Since the elements of $B(s) = \operatorname{adj}(sI - A)$ are cofactors of the elements of $sI - A$, they are polynomials of degree not exceeding $n - 1$ (see Exercise 5.2). Thus we can write

$$B(s) = s^{n-1}B_1 + \cdots + sB_{n-1} + B_n$$

for some $B_i \in \mathbb{F}^{n\times n}$. Premultiplying both sides of (5.8) with $d(s)(sI - A)$ we obtain

$$d(s)I = (sI - A)B(s)$$

or in open form as

$$s^n I + \sum_{i=1}^{n-1} s^{n-i} d_i I + d_n I = s^n B_1 + \sum_{i=1}^{n-1} s^{n-i}(B_{i+1} - AB_i) - AB_n$$

Equating the coefficient matrices of the like powers of s, we get

$$\begin{aligned} B_1 &= I \\ B_2 &= AB_1 + d_1 I \\ &\vdots \\ B_n &= AB_{n-1} + d_{n-1} I \\ O &= AB_n + d_n I \end{aligned}$$

Substituting B_1 into the equation for B_2, the resulting expression for B_2 into the equation for B_3, and so on, we obtain

$$\begin{aligned} B_2 &= A + d_1 I \\ B_3 &= A^2 + d_1 A + d_2 I \\ &\vdots \\ O &= A^n + d_1 A^{n-1} + \cdots + d_n I \;=\; d(A) \end{aligned}$$

completing the proof.

Example 5.5

The characteristic polynomial of

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

is

$$d(s) = s^3 - 2s^2 - s + 2 = (s-1)^2(s-2)$$

Forming

$$\begin{aligned} A - I &= \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \end{bmatrix} \\ (A-I)^2 &= A - I \\ A - 2I &= \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 1 & -1 & 0 \end{bmatrix} \end{aligned}$$

it is easily verified that

$$d(A) = (A-I)^2(A-2I) = O$$

Now consider the polynomial $\alpha(s) = (s-1)(s-2) = s^2 - 3s + 2$, which has a smaller degree than the characteristic polynomial. Since $(A-I)^2 = A - I$, it follows that we also have

$$\alpha(A) = (A-I)(A-2I) = O$$

The least degree polynomial $\alpha(s)$ with a unity leading coefficient for which $\alpha(A) = O$ is called the ***minimum polynomial*** of A. Properties of the minimum polynomial are discussed in Exercise 5.9.

The Cayley-Hamilton theorem implies that the nth power of an $n \times n$ matrix can be expressed as a linear combination of smaller powers as

$$A^n = -d_1 A^{n-1} - \cdots - d_n I$$

It then follows by induction that any power of A, and therefore, any polynomial $p(A)$ can be written as a linear combination of powers up to $n-1$. If p is any polynomial of degree higher than $n-1$, then dividing p with the characteristic polynomial d, we obtain

$$p(s) = d(s)q(s) + r(s)$$

where q is the quotient polynomial and r is the remainder polynomial with degree not exceeding $n-1$. Then

$$p(A) = d(A)q(A) + r(A) = r(A)$$

where the last equality follows from the fact that $d(A) = O$. Thus $p(A)$ can be evaluated by calculating $r(A)$, which involves powers of A up to at most $n-1$.

Example 5.6

The reflection matrix

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

in Example 5.3 has the property that $A^2 = I$. Hence $A^{2k} = I$, $k = 1, 2, \ldots$, that is, reflecting a vector an even number of times in the line $y = x$ we end up with the original vector. Let us verify this observation for A^{100} using the Cayley-Hamilton theorem.

The characteristic polynomial of A has already been found as $d(s) = s^2 - 1$. Dividing $p(s) = s^{100}$ with $d(s)$ we get

$$p(s) = s^{100} = (s^2 - 1)(s^{98} + s^{96} + \cdots + s^2 + 1) + 1 = d(s)q(s) + r(s)$$

Hence $r(s) = 1$, and $p(A) = A^{100} = r(A) = I$.

Example 5.7

Let us evaluate A^q and find $\lim_{q \to \infty} A^q$, if it exists, for

$$A = \begin{bmatrix} 2/3 & 1/2 \\ 1/3 & 1/2 \end{bmatrix}$$

The characteristic polynomial of A is obtained as

$$d(s) = s^2 - \frac{7}{6}\, s + \frac{1}{6} = (s - \frac{1}{6})(s - 1)$$

Let

$$s^q = (s - \frac{1}{6})(s - 1)q(s) + r_0 s + r_1\,, \quad q \geq 2$$

Evaluating both sides at the eigenvalues $s = 1/6$ and $s = 1$, we get

$$\begin{aligned} \frac{1}{6^q} &= \frac{1}{6}\, r_0 + r_1 \\ 1 &= r_0 + r_1 \end{aligned}$$

from which we obtain

$$\begin{aligned} r_0 &= \frac{6}{5}\,(1 - \frac{1}{6^q}) \\ r_1 &= \frac{6}{5}\,(\frac{1}{6^q} - \frac{1}{6}) \end{aligned}$$

Thus

$$A^q = r_0 A + r_1 I$$

and

$$\lim_{q \to \infty} A^q = (\lim_{q \to \infty} r_0)\, A + (\lim_{q \to \infty} r_1)\, I = \frac{6}{5}\, A - \frac{1}{5}\, I = \begin{bmatrix} 3/5 & 3/5 \\ 2/5 & 2/5 \end{bmatrix}$$

5.3 The Diagonal Form

Recall from Section 4.3 that two matrices $A, A' \in \mathbb{F}^{n\times n}$ are said to be similar if they are related as

$$A' = P^{-1}AP$$

for some nonsingular matrix $P \in \mathbb{F}^{n\times n}$ which represents a change of basis in $\mathbb{F}^{n\times 1}$. Since similar matrices represent the same linear operator with respect to different bases, they are expected to share some common characteristics. From

$$\begin{aligned}\det(sI - P^{-1}AP) &= \det[P^{-1}(sI - A)P] \\ &= (\det\ P^{-1}) \cdot \det(sI - A) \cdot (\det\ P) \\ &= \det(sI - A)\end{aligned}$$

it follows that similar matrices have the same characteristic polynomial, and therefore, the same eigenvalues. It is then natural to look for a canonical form that displays the eigenstructure of similar matrices. Such a canonical form can conveniently represent the whole equivalence class of similar matrices.

Consider a diagonal matrix

$$D = \text{diag}[\,d_1, \ldots, d_n\,] \in \mathbb{F}^{n\times n}$$

where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Clearly, the eigenvalues of D are $\lambda_i = d_i, i = 1, \ldots, n$, and D looks like a good candidate to characterize the class of matrices that have the same eigenvalues. As we showed above, all matrices that are similar to D have the same eigenvalues $\lambda_i = d_i$. The difficult problem is to determine whether a matrix having the eigenvalues $\lambda_i = d_i$ is similar to D.

Example 5.8

Let $D_1 = \text{diag}[\,1, 2\,]$. We can construct infinitely many matrices that are similar to D. For example,

$$A_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}$$

and

$$A_2 = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 5 & -2 \\ 6 & -2 \end{bmatrix}$$

are both similar to D. In fact, all 2×2 matrices with eigenvalues $\lambda_1 = d_1 = 1, \lambda_2 = d_2 = 2$ are similar to D as we will prove shortly.

On the other hand, there is no other matrix that is similar to $D_2 = I_2$, simply because $P^{-1}IP = I$ for any nonsingular P. Although the matrix

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

has the same eigenvalues $\lambda_1 = \lambda_2 = 1$ as D_2 does, it is not similar to D_2.

Example 5.8 shows that eigenvalues alone do not provide enough information to determine if a matrix is similar to a diagonal matrix. That information is in the eigenvectors.

First consider a diagonal matrix $D = \text{diag}\,[\,d_1, \ldots, d_n\,]$. Since

$$D\boldsymbol{e}_i = d_i\boldsymbol{e}_i$$

it follows that $\boldsymbol{v}_i = \boldsymbol{e}_i$ is an eigenvector of D associated with the eigenvalue $\lambda_i = d_i$. Thus eigenvectors of a diagonal matrix are linearly independent in $\mathbb{F}^{n\times 1}$.

Now, suppose A is similar to D. Then $P^{-1}AP = D$, or equivalently, $AP = PD$ for some nonsingular matrix P. Partitioning P into its columns as

$$P = [\,\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n\,] \tag{5.9}$$

we have

$$AP = [\,A\boldsymbol{v}_1 \ \cdots \ A\boldsymbol{v}_n\,] = PD = [\,d_1\boldsymbol{v}_1 \ \cdots \ d_n\boldsymbol{v}_n\,]$$

or columnwise

$$A\boldsymbol{v}_i = d_i\boldsymbol{v}_i\,, \quad i = 1, \ldots, n$$

Thus each diagonal element d_i of D is an eigenvalue of A and the corresponding column $\boldsymbol{v}_i$ of P is an eigenvector of A associated with it. Since P is nonsingular, $\boldsymbol{v}_i$ are linearly independent.

Conversely, if A has the linearly independent eigenvectors $\boldsymbol{v}_i$ associated with the eigenvalues $\lambda_i, i = 1, \ldots, n$, then the matrix P in (5.9) is nonsingular, and

$$AP = [\,A\boldsymbol{v}_1 \ \cdots \ A\boldsymbol{v}_n\,] = [\,\lambda_1\boldsymbol{v}_1 \ \cdots \ \lambda_n\boldsymbol{v}_n\,] = PD$$

Thus $P^{-1}AP = D$, where

$$D = \text{diag}\,[\,\lambda_1, \ldots, \lambda_n\,]$$

We summarize this important result as a theorem:

Theorem 5.2 *A matrix $A \in \mathbb{F}^{n\times n}$ is similar to a diagonal matrix D if and only if it has n linearly independent eigenvectors in $\mathbb{F}^{n\times 1}$, in which case the diagonal elements of D are the eigenvalues of A.*

The diagonal matrix D in Theorem 5.2 is called the ***diagonal form***, and the matrix P consisting of the eigenvectors is called a ***modal matrix*** of A. D represents the equivalence class of matrices that are similar to A in the sense that every matrix that is similar to A has the same diagonal form D up to a reordering of the diagonal elements.

When an $n\times n$ matrix A has n linearly independent eigenvectors, the MATLAB command

```
[ P,D ]=eig(A)
```

returns a modal matrix and the diagonal form of A.

Example 5.9

The eigenvectors of the real matrix A in Example 5.1 can easily be shown to be linearly independent in $\mathbb{R}^{2\times 1}$. The modal matrix formed from the eigenvectors

$$P_A = \begin{bmatrix} 1 & 1 \\ 1 & -2 \end{bmatrix}$$

is nonsingular. Computing P_A^{-1} it is easily verified that

$$P_A^{-1} A P_A = D = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

MATLAB returns a different modal matrix

$$P = \begin{bmatrix} -0.7071 & 0.4472 \\ -0.7071 & -0.8944 \end{bmatrix}$$

but the same diagonal form D.

Now, consider a nonsingular matrix

$$Q = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} \quad \text{with} \quad Q^{-1} = \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}$$

and let

$$B = Q^{-1} A Q = \begin{bmatrix} -8 & -27 \\ 4 & 13 \end{bmatrix}$$

Since B is similar to A, it must have the same diagonal form. Indeed, obtaining the characteristic polynomial of B as

$$\det(sI - B) = \det \begin{bmatrix} s+8 & 27 \\ -4 & s-13 \end{bmatrix} = (s+8)(s-13) + 108 = s^2 - 5s + 4$$

we observe that the eigenvalues of B are the same as those of A: $\lambda_1 = 1$ and $\lambda_2 = 4$. The reader can verify that

$$P_B = \begin{bmatrix} 3 & -9 \\ -1 & 4 \end{bmatrix}$$

is a modal matrix of B and that

$$P_B^{-1} B P_B = D = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

That is, the diagonal form of B is the same as the diagonal form of A.

Finally, let

$$C = \begin{bmatrix} 0 & -1 \\ 4 & 5 \end{bmatrix}$$

which also has the same eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 4$. It is easy to show that

$$P_C = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

is a modal matrix of C, and that

$$P_C^{-1} C P_C = D = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

Since C has the same diagonal form, it must be similar to A. Indeed,

$$C = P_C D P_C^{-1} = P_C P_A^{-1} A P_A P_C^{-1} = P^{-1} A P$$

where

$$P = P_A P_C^{-1} = \begin{bmatrix} 5/3 & 2/3 \\ 1/3 & -1/3 \end{bmatrix}$$

Example 5.10

The characteristic polynomial of the complex matrix

$$A = \begin{bmatrix} 2+i & -1-i \\ 2 & -1 \end{bmatrix}$$

is

$$d(s) = s^2 - (1+i)s + i = (s-1)(s-i)$$

The eigenvectors associated with $\lambda_1 = 1$ and $\lambda_2 = i$ can be found as

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} 1+i \\ 2 \end{bmatrix}$$

$\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent in $\mathbb{C}^{2\times 1}$. Constructing a modal matrix

$$P = \begin{bmatrix} 1 & 1+i \\ 1 & 2 \end{bmatrix}$$

from the eigenvectors and computing P^{-1}, we obtain the diagonal form of A as

$$P^{-1} A P = D = \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}$$

Example 5.11

The real matrix A in Example 5.2 has no real eigenvectors, and therefore, it can not be diagonalized by a change of basis in $\mathbb{R}^{2\times 1}$. However, it has two linearly independent complex eigenvectors as given in Example 5.2. Treating A as a complex matrix, we can construct a modal matrix

$$P = \begin{bmatrix} 1 & 1 \\ 1+2i & 1-2i \end{bmatrix}$$

from the eigenvectors. The transformation

$$P^{-1} A P = \begin{bmatrix} 1+2i & 0 \\ 0 & 1-2i \end{bmatrix} = D$$

results in a diagonal matrix that is similar to A in $\mathbb{C}^{2\times 1}$. Note that although A is real, its diagonal form D is complex.

From Example 5.11 we observe that some real matrices that can not be diagonalized when considered as a linear operator on $\mathbb{R}^{n\times 1}$ may be diagonalized when considered as a linear operator on $\mathbb{C}^{n\times 1}$. For this reason, we will first concentrate on complex matrices, and treat all matrices as linear operators on $\mathbb{C}^{n\times 1}$ even if they are real.

5.3.1 Complex Diagonal Form

Let $A \in \mathbb{C}^{n\times n}$ have k distinct eigenvalues $\lambda_i \in \mathbb{C}$ with algebraic multiplicities n_i. Consider the eigenspaces $\boldsymbol{\mathcal{K}}_i \subset \mathbb{C}^{n\times 1}$ associated with λ_i. We claim that these subspaces are linearly independent. To prove the claim, pick arbitrary vectors $\boldsymbol{v}_i \in \boldsymbol{\mathcal{K}}_i$ and set

$$\boldsymbol{v}_1 + \cdots + \boldsymbol{v}_k = \mathbf{0} \tag{5.10}$$

Let us define the polynomials

$$p_q(s) = \prod_{\substack{j=1\\ j\neq q}}^{k} (s - \lambda_j)\,, \quad q = 1, \ldots, k$$

Since all factors $(s - \lambda_i)$ except $(s - \lambda_q)$ are included in $p_q(s)$, $p_q(\lambda_i) = 0$ for $i \neq q$ but $p_q(\lambda_q) \neq 0$. Consider the products

$$p_q(A)\boldsymbol{v}_i = (A - \lambda_1 I)\cdots(A - \lambda_{q-1}I)(A - \lambda_{q+1}I)\cdots(A - \lambda_k I)\boldsymbol{v}_i$$

Since the matrices $(A - \lambda_i I)$ and $(A - \lambda_j I)$ commute for all (i, j) and

$$(A - \lambda_j I)\boldsymbol{v}_i = (\lambda_i - \lambda_j)\boldsymbol{v}_i$$

we have

$$p_q(A)\boldsymbol{v}_i = p_q(\lambda_i)\boldsymbol{v}_i = \begin{cases} p_q(\lambda_q)\boldsymbol{v}_q\,, & i = q \\ \mathbf{0} \quad\;, & i \neq q \end{cases}$$

Hence premultiplying both sides of (5.10) with $p_q(A)$ we obtain

$$p_q(\lambda_q)\boldsymbol{v}_q = \mathbf{0}\,, \quad q = 1, \ldots, k$$

Since $p_q(\lambda_q) \neq 0$ the last equality implies

$$\boldsymbol{v}_q = \mathbf{0}\,, \quad q = 1, \ldots, k$$

proving the claim.

An immediate consequence of this result is that if A has n distinct eigenvalues (in which case they are simple zeros of the characteristic polynomial, that is, $n_i = 1, i = 1, \ldots, n$) then it has n linearly independent eigenvectors, and hence it can be diagonalized by a similarity transformation. This has already been illustrated in Examples 5.9-5.11. Since almost all square matrices with independent elements have simple eigenvalues, in practice almost all matrices can be diagonalized.[1] However, some problems may lead to matrices whose elements are so interdependent that they have repeated eigenvalues. For completeness of the theory, in the rest of this section and in the following section we will study such matrices.

We first answer the question of under what conditions matrices with repeated eigenvalues can be diagonalized:

[1] The reader may try to verify this statement by generating several random matrices by MATLAB and computing their eigenvalues.

Corollary 5.2.1 *Let $A \in \mathbb{C}^{n\times n}$ have the characteristic polynomial in (5.7), where $\lambda_i \neq \lambda_j$ for $i \neq j$, and let dim $(\boldsymbol{\mathcal{K}}_i) = \nu_i$, $i = 1, \ldots, k$. Then the following are equivalent.*

a) A *is similar (in $\mathbb{C}^{n\times 1}$) to a diagonal matrix D.*

b) $\nu_i = n_i\,, \quad i = 1, \ldots, k.$

c) $\mathbb{C}^{n\times 1} = \boldsymbol{\mathcal{K}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{K}}_k.$

Proof We will show that (a) ⇒ (b) ⇒ (c) ⇒ (a).

(a) ⇒ (b):

Let $P^{-1}AP = D$ for some diagonal matrix D. Since D has the same characteristic polynomial as A, we can assume without loss of generality that

$$D = \begin{bmatrix} \lambda_1 I_{n_1} & & \\ & \ddots & \\ & & \lambda_k I_{n_k} \end{bmatrix} \tag{5.11}$$

Let P be partitioned into its columns as

$$P = [\, P_1 \ \cdots \ P_k \,] \tag{5.12}$$

where P_i is an $n \times n_i$ block of P corresponding to the ith diagonal block of D. From

$$AP = [\, AP_1 \ \cdots \ AP_k \,] = PD = [\, \lambda_1 P_1 \ \cdots \ \lambda_k P_k \,]$$

we observe that each column of P_i is an eigenvector of A associated with the eigenvalue $\lambda_i, i = 1, \ldots, k$. Hence cs $(P_i) \subset \boldsymbol{\mathcal{K}}_i$, which implies that

$$n_i \leq \nu_i\,, \quad i = 1, \ldots, k$$

However, since $\boldsymbol{\mathcal{K}}_i$ are linearly independent and $\bigoplus \boldsymbol{\mathcal{K}}_i \subset \mathbb{C}^{n\times 1}$, we have

$$\sum n_i \leq \sum \nu_i = \text{dim}\,(\bigoplus \boldsymbol{\mathcal{K}}_i\,) \leq n$$

Then the only possibility is that $\nu_i = n_i, i = 1, \ldots, k$.

(b) ⇒ (c):

This follows directly from the facts that $\boldsymbol{\mathcal{K}}_i$ are linearly independent and that

$$\text{dim}\,(\bigoplus \boldsymbol{\mathcal{K}}_i\,) = \sum \nu_i = \sum n_i = n$$

(c) ⇒ (a):

Let the columns of $n \times n_i$ matrices

$$P_i = [\, \boldsymbol{v}_{i1} \ \cdots \ \boldsymbol{v}_{in_i} \,]$$

form bases for $\boldsymbol{\mathcal{K}}_i, i = 1, \ldots, k$, and let P be constructed from P_i as in (5.12). Then columns of P form a basis for $\mathbb{C}^{n\times 1}$, and therefore, P is nonsingular. Also,

$$A\boldsymbol{v}_{ij} = \lambda_i \boldsymbol{v}_{ij} \implies AP_i = \lambda_i P_i \implies AP = PD$$

where D is as given by (5.11), and the result follows.

Corollary 5.2.1 simply states that a matrix can be diagonalized if and only if the geometric multiplicity of each eigenvalue equals its algebraic multiplicity, in which case there exist sufficiently many linearly independent eigenvectors associated with each multiple eigenvalue. We illustrate this with the following example.

Example 5.12

The matrix

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

has the characteristic polynomial

$$d(s) = s^4 - 2s^3 + 2s^2 - 2s + 1 = (s-1)^2(s^2+1)$$

Hence $\lambda_1 = 1$ with $n_1 = 2$, $\lambda_2 = i$ with $n_2 = 1$, and $\lambda_3 = -i$ with $n_3 = 1$.

From

$$(A - \lambda_1 I) = \begin{bmatrix} 0 & 2 & -2 & -2 \\ 0 & -2 & 2 & 2 \\ 0 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix} \xrightarrow{\text{e.r.o.}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

we find that $\dim(\boldsymbol{\mathcal{K}}_1) = 2 = n_1$. Also, since $\boldsymbol{\mathcal{K}}_2$ and $\boldsymbol{\mathcal{K}}_3$ each contain at least one eigenvector, $\dim(\boldsymbol{\mathcal{K}}_2) = 1 = n_2$ and $\dim(\boldsymbol{\mathcal{K}}_3) = 1 = n_3$. Then by Corollary 5.2.1 A is similar to a diagonal matrix.

To construct a modal matrix for A, we need to find four linearly independent eigenvectors; two associated with λ_1, and one associated with each of λ_2 and λ_3. Eigenvectors associated with λ_1 can be found from the reduced row echelon form of $A - \lambda_1 I$ above as

$$\boldsymbol{v}_{11} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{v}_{12} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

An eigenvector associated with λ_2 can be found from

$$(A - \lambda_2 I) = \begin{bmatrix} 1-i & 2 & -2 & -2 \\ 0 & -1-i & 2 & 2 \\ 0 & -1 & 2-i & 1 \\ 1 & 1 & -1 & -i \end{bmatrix} \xrightarrow{\text{e.r.o.}} \begin{bmatrix} 1 & 0 & 0 & -2i \\ 0 & 1 & 0 & 2i \\ 0 & 0 & 1 & i \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

as

$$\boldsymbol{v}_2 = \begin{bmatrix} -2 \\ 2 \\ 1 \\ i \end{bmatrix}$$

Finally, since $\lambda_3 = \lambda_2^*$, we choose $\boldsymbol{v}_3 = \boldsymbol{v}_2^*$.

Thus

$$P = \begin{bmatrix} 0 & 0 & -2 & -2 \\ 1 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & i & -i \end{bmatrix}, \quad P^{-1} = \frac{1}{4} \begin{bmatrix} 2 & 0 & 4 & 0 \\ 2 & 4 & -4 & 0 \\ -1+i & 2i & -2i & -2i \\ -1-i & -2i & 2i & 2i \end{bmatrix}$$

and

$$P^{-1}AP = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & i & \\ & & & -i \end{bmatrix}$$

* 5.3.2 Invariant Subspaces

Let us take a closer look at the role the eigenspaces $\boldsymbol{\mathcal{K}}_i$ play in diagonalization of a complex matrix: If $\boldsymbol{v} \in \boldsymbol{\mathcal{K}}_i$ then $\boldsymbol{v}$ is an eigenvector of A associated with λ_i, and

$$A\boldsymbol{v} = \lambda_i \boldsymbol{v} \in \boldsymbol{\mathcal{K}}_i$$

In other words, the eigenspaces $\boldsymbol{\mathcal{K}}_i$ have the property that

$$\boldsymbol{v} \in \boldsymbol{\mathcal{K}}_i \implies A\boldsymbol{v} \in \boldsymbol{\mathcal{K}}_i$$

A subspace $\boldsymbol{\mathcal{V}} \subset \mathbb{C}^{n\times 1}$ is said to be ***invariant*** under A, or A-invariant, if

$$A\boldsymbol{v} \in \boldsymbol{\mathcal{V}} \quad \text{for all } \boldsymbol{v} \in \boldsymbol{\mathcal{V}}$$

Thus each eigenspace $\boldsymbol{\mathcal{K}}_i$ is an A-invariant subspace.

Suppose that we have a direct sum decomposition

$$\mathbb{C}^{n\times 1} = \boldsymbol{\mathcal{V}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_\kappa \tag{5.13}$$

where $\boldsymbol{\mathcal{V}}_i$ are A-invariant subspaces with $\dim(\boldsymbol{\mathcal{V}}_i) = \eta_i$, so that $\sum \eta_i = n$. Let columns of the $n \times \eta_i$ matrices

$$T_i = [\, \boldsymbol{t}_{i1} \ \cdots \ \boldsymbol{t}_{i\eta_i} \,]$$

form bases for $\boldsymbol{\mathcal{V}}_i$. Then their union, that is, the columns of the $n \times n$ matrix

$$T = [\, T_1 \ \cdots \ T_\kappa \,]$$

form a basis for $\mathbb{C}^{n\times 1}$, and hence T is nonsingular.

Since $\boldsymbol{t}_{ij} \in \boldsymbol{V}_i$ and $\boldsymbol{V}_i$ is A-invariant, $A\boldsymbol{t}_{ij} \in \boldsymbol{V}_i$ and hence $A\boldsymbol{t}_{ij}$ can be written as a linear combination of the columns of T_i. That is,

$$A\boldsymbol{t}_{ij} = T_i \boldsymbol{f}_{ij}\,, \quad i = 1, \ldots, \kappa\,;\ j = 1, \ldots, \eta_i$$

for some $\eta_i \times 1$ column vector $\boldsymbol{f}_{ij}$. Then

$$AT_i = [\, A\boldsymbol{t}_{i1} \ \cdots \ A\boldsymbol{t}_{i\eta_i} \,] = \left[\, T_i\boldsymbol{f}_{i1} \ \cdots \ T_i\boldsymbol{f}_{i\eta_i} \,\right] = T_i \left[\, \boldsymbol{f}_{i1} \ \cdots \ \boldsymbol{f}_{i\eta_i} \,\right] = T_i F_i$$

where F_i are $\eta_i \times \eta_i$ matrices, and

$$AT = [\, AT_1 \ \cdots \ AT_\kappa \,] = [\, T_1F_1 \ \cdots \ T_\kappa F_\kappa \,] = TF$$

where

$$F = \begin{bmatrix} F_1 & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & F_\kappa \end{bmatrix}$$

Since T is nonsingular we have

$$T^{-1}AT = F = \text{diag}\,[\,F_1\ \cdots\ F_\kappa\,] \tag{5.14}$$

Thus a direct sum decomposition of $\mathbb{C}^{n\times 1}$ into A-invariant subspaces as in (5.13) results in a block diagonal form of A as in (5.14).

The diagonal form in Corollary 5.2.1 is a special case of this result, where $\kappa = k$, $\boldsymbol{\mathcal{V}}_i = \boldsymbol{\mathcal{K}}_i$, $\eta_i = n_i$ and $T_i = P_i$, $i = 1, \ldots, k$. However, there is more in Corollary 5.2.1: Since the columns of

$$P_i = [\,\boldsymbol{v}_{i1}\ \cdots\ \boldsymbol{v}_{in_i}\,]$$

are eigenvectors that satisfy $A\boldsymbol{v}_{ij} = \lambda_i \boldsymbol{v}_{ij}, j = 1, \ldots, n_i$, they induce a further decomposition of $\boldsymbol{\mathcal{K}}_i$ as

$$\boldsymbol{\mathcal{K}}_i = \boldsymbol{\mathcal{V}}_{i1} \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_{in_i}$$

where each one-dimensional subspace $\boldsymbol{\mathcal{V}}_{ij} = \text{span}\,(\boldsymbol{v}_{ij})$ is also A-invariant. That is why we have $F_i = \lambda_i I_{n_i}, \quad i = 1, \ldots, k$ and $F = D$. In other words, condition (c) of Corollary 5.2.1 provides the finest possible direct sum decomposition

$$\mathbb{C}^{n\times 1} = \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{n_i} \boldsymbol{\mathcal{V}}_{ij}$$

into one-dimensional A-invariant subspaces, which leads to the simplest possible form (the diagonal form) of A.

* 5.3.3 Real Semi-Diagonal Form

We now consider the problem of transforming a real matrix A into a simple form by a suitable choice of a basis for $\mathbb{R}^{n\times 1}$.

We have no difficulty if A has only real eigenvalues: We can choose its eigenvectors to be real and restrict the eigenspaces $\boldsymbol{\mathcal{K}}_i$ to be subspaces of $\mathbb{R}^{n\times 1}$ rather than $\mathbb{C}^{n\times 1}$. Then Corollary 5.2.1 remains valid even with $\mathbb{C}^{n\times 1}$ replaced with $\mathbb{R}^{n\times 1}$, and we can diagonalize A provided the conditions on the dimension of the eigenspaces are satisfied. The difficulty arises if A has one or more complex eigenvalues, in which case the corresponding eigenspaces are no longer subspaces of $\mathbb{R}^{n\times 1}$, and Corollary 5.2.1 becomes void. However, the argument in the preceeding subsection suggests that it may still be possible to decompose $\mathbb{R}^{n\times 1}$ into a direct sum of A-invariant subspaces and choose suitable bases for these subspaces to come up with a simple representation of A.

Example 5.13

Consider the real matrix

$$A = \begin{bmatrix} 0 & 1 \\ -5 & 2 \end{bmatrix}$$

in Example 5.2, which has a pair of complex conjugate eigenvalues $\lambda_{1,2} = 1 \mp 2i$.

Clearly, $\mathbb{R}^{2\times 1}$ cannot be decomposed into smaller A-invariant subspaces, because then A would be similar to a real diagonal matrix whose diagonal elements would have to be the eigenvalues of A. However, we may try to construct a basis for $\mathbb{R}^{2\times 1}$ and transform A into a special

form. For this purpose, let us separate the complex conjugate eigenvectors of A into their real and imaginary parts as

$$\boldsymbol{v}_{1,2} = \begin{bmatrix} 1 \\ 1 \mp 2i \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mp i \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \boldsymbol{u} \mp i\boldsymbol{w}$$

Since $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are linearly independent in $\mathbb{C}^{2\times 1}$, $\boldsymbol{u}$ and $\boldsymbol{w}$ are linearly independent in $\mathbb{R}^{2\times 1}$. Thus

$$P_R = [\,\boldsymbol{u} \;\; \boldsymbol{w}\,] = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

is nonsingular. The representation of A with respect to the basis $(\boldsymbol{u}, \boldsymbol{w})$ is obtained as

$$P_R^{-1} A P_R = D_R = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$$

Although D_R is no simpler than A, it has a special structure: Its diagonal elements are the real parts and the off-diagonal elements are the imaginary parts of the complex eigenvalues. Of course, this is not just a coincidence as we explain next.

Suppose that A has m distinct pairs of complex conjugate eigenvalues

$$\begin{array}{rclrcl} \lambda_1 & = & \sigma_1 + i\omega_1\,, & \lambda_2 & = & \sigma_1 - i\omega_1 \\ & \vdots & & & \vdots & \\ \lambda_{2m-1} & = & \sigma_m + i\omega_m\,, & \lambda_{2m} & = & \sigma_m - i\omega_m \end{array} \tag{5.15}$$

with associated complex conjugate pairs of eigenvectors

$$\begin{array}{rclrcl} \boldsymbol{v}_1 & = & \boldsymbol{u}_1 + i\boldsymbol{w}_1\,, & \boldsymbol{v}_2 & = & \boldsymbol{u}_1 - i\boldsymbol{w}_1 \\ & \vdots & & & \vdots & \\ \boldsymbol{v}_{2m-1} & = & \boldsymbol{u}_m + i\boldsymbol{w}_m\,, & \boldsymbol{v}_{2m} & = & \boldsymbol{u}_m - i\boldsymbol{w}_m \end{array} \tag{5.16}$$

and $n - 2m$ distinct real eigenvalues

$$\lambda_{2m+1}, \ldots, \lambda_n$$

with associated real eigenvectors

$$\boldsymbol{v}_{2m+1}, \ldots, \boldsymbol{v}_m$$

Equating the real and imaginary parts of

$$A(\boldsymbol{u}_i + i\boldsymbol{w}_i) = (\sigma_i + i\omega_i)(\boldsymbol{u}_i + i\boldsymbol{w}_i)\,, \quad i = 1, \ldots, m$$

we obtain

$$A\,[\,\boldsymbol{u}_i \;\; \boldsymbol{w}_i\,] = [\,\boldsymbol{u}_i \;\; \boldsymbol{w}_i\,] \begin{bmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{bmatrix}, \quad i = 1, \ldots, m$$

or in compact form

$$AP_{Ri} = P_{Ri} D_{Ri}\,, \quad i = 1, \ldots, m \tag{5.17}$$

where

$$P_{Ri} = [\, \boldsymbol{u}_i \;\; \boldsymbol{w}_i \,] \;, \quad D_{Ri} = \begin{bmatrix} \sigma_i & \omega_i \\ -\omega_i & \sigma_i \end{bmatrix}$$

Note that the diagonal elements of D_{Ri} are the real parts and the off-diagonal elements are the imaginary parts of the corresponding conjugate pair of complex eigenvalues (as we already observed in Example 5.13). Thus with

$$P_R = [\, P_{R1} \;\cdots\; P_{Rm} \;\; \boldsymbol{v}_{2m+1} \;\cdots\; \boldsymbol{v}_n \,] \tag{5.18}$$

we have

$$AP_R = P_R D_R$$

where

$$D_R = \begin{bmatrix} D_{R1} & & & & & & \\ & \ddots & & & & & \\ & & D_{Rm} & & & & \\ & & & \lambda_{2m+1} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_n \end{bmatrix} \tag{5.19}$$

It can be shown that P_R is also nonsingular (see Exercise 5.29), so that

$$P_R^{-1} A P_R = D_R$$

The matrix P_R is called a real modal matrix of A, and D_R the real semi-diagonal form of A.

The structure of P_R in (5.18) implies a decomposition of $\mathbb{R}^{n\times 1}$ as

$$\mathbb{R}^{n\times 1} = \boldsymbol{\mathcal{V}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_m \oplus \boldsymbol{\mathcal{V}}_{2m+1} \oplus \cdots \boldsymbol{\mathcal{V}}_n \tag{5.20}$$

where

$$\boldsymbol{\mathcal{V}}_i = \begin{cases} \text{span}\,(\boldsymbol{u}_i, \boldsymbol{w}_i)\,, & i = 1, \ldots, m \\ \text{span}\,(\boldsymbol{v}_i) \quad , & i = 2m+1, \ldots, n \end{cases}$$

(5.17) guarantees that each $\boldsymbol{\mathcal{V}}_i, i = 1, \ldots, m$, is a two-dimensional A-invariant subspace of $\mathbb{R}^{n\times 1}$ associated with a pair of complex conjugate eigenvalues. Since the one-dimensional subspaces $\boldsymbol{\mathcal{V}}_i, i = 2m+1, \ldots, n$, associated with the real eigenvalues are also A-invariant, this decomposition yields a block diagonal representation of A as in (5.19).

Example 5.14

The matrix

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

has the characteristic polynomial

$$d(s) = s^3 - s^2 + 2 = (s+1)(s^2 - 2s + 2)$$

The eigenvalues are $\lambda_1 = -1$, $\lambda_{2,3} = 1 \mp i$, with the associated eigenvectors

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_{2,3} = \boldsymbol{u}_2 \mp i\boldsymbol{w}_2 = \begin{bmatrix} 1 \\ 1 \mp i \\ \mp 2i \end{bmatrix}$$

A real modal matrix is

$$P_R = [\,\boldsymbol{v}_1 \;\; \boldsymbol{u}_2 \;\; \boldsymbol{w}_2\,] = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 1 \\ 1 & 0 & 2 \end{bmatrix}$$

which results in the real semidiagonal form of A:

$$P_R^{-1} A P_R = D_R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}$$

If a real matrix $A \in \mathbb{R}^{n\times n}$ has multiple complex eigenvalues, then as long as $\nu_i = n_i$ for all the distinct eigenvalues, a real modal matrix can be constructed from the real and imaginary parts of the complex eigenvectors, and A can be semi-diagonalized by a similarity transformation in $\mathbb{R}^{n\times 1}$.

Example 5.15

The matrix

$$A = \begin{bmatrix} -5 & 3 & 0 & 3 \\ -3 & 1 & 3 & 3 \\ 3 & -3 & -2 & 0 \\ -3 & 0 & -3 & -2 \end{bmatrix}$$

has the characteristic polynomial

$$d(s) = s^4 + 8s^3 + 42s^2 + 104s + 169 = (s^2 + 4s + 13)^2$$

and hence the complex conjugate eigenvalues

$$\lambda_{1,2} = -2 \mp 3i\,, \quad n_{1,2} = 2$$

Obtaining the reduced row echelon form of $A - \lambda_1 I$ as

$$A - \lambda_1 I = \begin{bmatrix} -3-3i & 3 & 0 & 3 \\ -3 & 3-3i & 3 & 3 \\ 3 & -3 & -3i & 0 \\ -3 & 0 & -3 & -3i \end{bmatrix} \xrightarrow{\text{e.r.o.}} \begin{bmatrix} 1 & 0 & 1 & i \\ 0 & 1 & 1+i & i \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

we find two linearly independent eigenvectors associated with λ_1:

$$\boldsymbol{v}_{11} = \boldsymbol{u}_1 + i\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 1+i \\ -1 \\ 0 \end{bmatrix}, \quad \boldsymbol{v}_{12} = \boldsymbol{u}_2 + i\boldsymbol{w}_2 = \begin{bmatrix} i \\ i \\ 0 \\ -1 \end{bmatrix}$$

Since $\lambda_2 = \lambda_1^*$ we choose eigenvectors associated with λ_2 to be the complex conjugates of $\boldsymbol{v}_{11}$ and $\boldsymbol{v}_{12}$: $\boldsymbol{v}_{21} = \boldsymbol{v}_{11}^* = \boldsymbol{u}_1 - i\boldsymbol{w}_1\,, \quad \boldsymbol{v}_{22} = \boldsymbol{v}_{12}^* = \boldsymbol{u}_2 - i\boldsymbol{w}_2$. Constructing a real modal matrix

$$P_R = [\, P_{R1} \;\; P_{R2} \,] = [\, \boldsymbol{u}_1 \;\; \boldsymbol{w}_1 \mid \boldsymbol{u}_2 \;\; \boldsymbol{w}_2 \,] = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

from the real and imaginary parts of $\boldsymbol{v}_{11}$ and $\boldsymbol{v}_{12}$, we obtain the real semi-diagonal form of A as

$$D_R = P_R^{-1} A P_R = \begin{bmatrix} -2 & 3 & 0 & 0 \\ -3 & -2 & 0 & 0 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & -3 & -2 \end{bmatrix}$$

* 5.4 The Jordan Form

5.4.1 The Complex Jordan Form

Corollary 5.2.1 implies that if A has an eigenvalue λ_i with $n_i > 1$ for which $\nu_i < n_i$, then it can not be diagonalized. However, by a careful choice of the basis vectors of $\mathbb{C}^{n\times 1}$, it can still be transformed into a simple form as stated by the following theorem. The proof of the theorem is beyond the scope of this book, and is omitted.

Theorem 5.3 (The Jordan Form) *Let A have the characteristic polynomial in (5.7), where $\lambda_i \neq \lambda_j$ for $i \neq j$, and let dim $(\mathcal{K}_i) = \nu_i\,,\ i = 1, \ldots, k$. Then there exists a nonsingular matrix P such that*

$$P^{-1}AP = J = \textit{diag}\,[\, J_1, \ldots, J_k \,] \tag{5.21}$$

where

a) each block J_i is an $n_i \times n_i$ matrix that consists of ν_i subblocks

$$J_i = \textit{diag}\,[\, J_{i1}, \ldots, J_{i\nu_i}] \;, \quad i = 1, \ldots, k$$

b) each subblock J_{ij} is an $n_{ij} \times n_{ij}$ matrix, and

$$J_{ij} \;=\; \lambda_i\,, \qquad \textit{if } n_{ij} = 1$$

$$J_{ij} \;=\; \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix}, \qquad \textit{if } n_{ij} > 1$$

c) $n_{i1} \geq \cdots \geq n_{i\nu_i}$ and $\sum_{j=1}^{\nu_i} n_{ij} = n_i\,, \quad i = 1, \ldots, k$

*J is called the **Jordan form** of A, and is unique up to a reordering of the **Jordan blocks** J_i and the **Jordan subblocks** J_{ij}. P is called a **modal matrix** of A.*

The general structure of the Jordan form is illustrated below for a typical case.

$$
\begin{aligned}
&n_1 = \nu_1 = n_{11} = 1 \\
&n_2 = \nu_2 = 2\,, \quad n_{21} = n_{22} = 1 \\
&n_3 = 2\,, \quad \nu_3 = 1\,, \quad n_{31} = 2 \\
&n_4 = 3\,, \quad \nu_4 = 2\,, \quad n_{41} = 2, n_{42} = 1
\end{aligned}
\qquad
\left[\begin{array}{c|c|cc|cc|cc}
\lambda_1 & & & & & & & \\
\hline
 & \lambda_2 & & & & & & \\
 & & \lambda_2 & & & & & \\
\hline
 & & & \lambda_3 & 1 & & & \\
 & & & & \lambda_3 & & & \\
\hline
 & & & & & \lambda_4 & 1 & \\
 & & & & & & \lambda_4 & \\
\hline
 & & & & & & & \lambda_4
\end{array}\right]
$$

Note that the algebraic multiplicity n_i of an eigenvalue determines the size of the associated Jordan block J_i, and the geometric multiplicity ν_i determines the number of the subblocks J_{ij} in J_i. If $\nu_i = n_i$ for an eigenvalue λ_i (as for λ_1 and λ_2 above), then the corresponding Jordan block J_i consists of n_i subblocks each of which is a scalar; that is, $J_{ij} = \lambda_i, j = 1, \ldots, n_i$, and $J_i = \lambda_i I_{n_i}$. Thus if $\nu_i = n_i$ for all eigenvalues, then the Jordan form reduces to a diagonal matrix. In other words, the diagonal form of A (if it exists) is a special Jordan form.

We now provide an interpretation of the Jordan form in terms of a direct sum decomposition of $\mathbb{C}^{n\times 1}$. For this purpose let us partition P as

$$P = [\, P_1 \ \cdots \ P_k \,]$$

where P_i are $n \times n_i$ matrices that are associated with the Jordan blocks P_i. Similarly, if $\nu_i > 1$ then P_i can further be partitioned as

$$P_i = [\, P_{i1} \ \cdots \ P_{i\nu_i} \,]$$

where P_{ij} are $n \times n_{ij}$ matrices that are associated with the Jordan subblocks J_{ij}. Partitioning of P corresponds to a direct sum decomposition of $\mathbb{C}^{n\times 1}$ as

$$\mathbb{C}^{n\times 1} = \boldsymbol{\mathcal{V}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_k \tag{5.22}$$

where

$$\boldsymbol{\mathcal{V}}_i = \mathrm{cs}\,(P_i)\,, \quad i = 1, \ldots, k$$

Similarly, partitioning of P_i corresponds to a direct sum decomposition of $\boldsymbol{\mathcal{V}}_i$ as

$$\boldsymbol{\mathcal{V}}_i = \boldsymbol{\mathcal{V}}_{i1} \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_{i\nu_i}\,, \quad i = 1, \ldots, k \tag{5.23}$$

where

$$\boldsymbol{\mathcal{V}}_{ij} = \mathrm{cs}\,(P_{ij})\,, \quad i = 1, \ldots, k\,;\ j = 1, \ldots, \nu_i$$

Combining (5.22) and (5.23) we obtain

$$\mathbb{C}^{n\times 1} = \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{\nu_i} \boldsymbol{\mathcal{V}}_{ij} \tag{5.24}$$

Thus the Jordan form is related to a direct sum decomposition of $\mathbb{C}^{n\times 1}$.

Rewriting (5.21) as $AP = PJ$ and performing block multiplication, we get

$$AP_i = P_i J_i\,, \quad i = 1, \ldots, k \tag{5.25}$$

If $\boldsymbol{v} \in \boldsymbol{\mathcal{V}}_i$ then $\boldsymbol{v} = P_i\boldsymbol{\alpha}_i$ for some $\boldsymbol{\alpha}_i \in \mathbb{C}^{n_i \times 1}$, and by (5.25) we have

$$A\boldsymbol{v} = AP_i\boldsymbol{\alpha}_i = P_i(J_i\boldsymbol{\alpha}_i) \in \boldsymbol{\mathcal{V}}_i$$

This shows that each $\boldsymbol{\mathcal{V}}_i$ in (5.22) is an A-invariant subspace. Similarly, from (5.25) we get

$$AP_{ij} = P_{ij}J_{ij}\,, \quad i = 1, \ldots, k\,, \quad j = 1, \ldots, \nu_i \tag{5.26}$$

which shows that each $\boldsymbol{\mathcal{V}}_{ij}$ in (5.23) is also an A-invariant subspace. Consequently, (5.24) gives a direct sum decomposition of $\mathbb{C}^{n\times 1}$ into A-invariant subspaces. In fact, it is the finest decomposition of $\mathbb{C}^{n\times 1}$ in the sense that no $\boldsymbol{\mathcal{V}}_{ij}$ can further be decomposed into smaller A-invariant subspaces. In this sense, the Jordan form of A is the simplest matrix that is similar to A. Like the diagonal form of a diagonalizable matrix, it represents the equivalence class of matrices that are similar to A.

Next we investigate the relation of the subspaces $\boldsymbol{\mathcal{V}}_{ij}$ in (5.24) to the eigenvalues of A.

We first consider the special case where $\nu_i = n_i$ for an eigenvalue λ_i. In this case $J_i = \lambda_i I_{n_i}$ as mentioned before, and therefore, $AP_i = P_iJ_i = \lambda_iP_i$. That is,

$$P_i = [\, P_{i1} \;\cdots\; P_{in_i}\,] = [\, \boldsymbol{v}_{i1} \;\cdots\; \boldsymbol{v}_{in_i}\,]$$

where each $\boldsymbol{v}_{ij}$ is an eigenvector associated with λ_i. In terms of subspaces, we have

$$\boldsymbol{\mathcal{V}}_i = \boldsymbol{\mathcal{K}}_i = \bigoplus_{j=1}^{n_i} \boldsymbol{\mathcal{V}}_{ij}$$

where

$$\boldsymbol{\mathcal{V}}_{ij} = \text{span}\,(\boldsymbol{v}_{ij})$$

are one-dimensional A-invariant subspaces. In particular, if $\nu_i = n_i = 1$ then the corresponding Jordan block J_i contains only a single subblock $J_i = J_{i1} = \lambda_i$. In this case $\boldsymbol{\mathcal{V}}_i = \boldsymbol{\mathcal{K}}_i = \text{span}\,(\boldsymbol{v}_i)$ is a one-dimensional A-invariant eigenspace that cannot be decomposed any further.

Another special case is when $\nu_i = 1 < n_i$ for an eigenvalue λ_i. In this case, J_i consists of a single Jordan subblock J_{i1} of size $n_{i1} = n_i$, and is of the form specified in Theorem 5.3, that is,

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_i \end{bmatrix}_{n_i \times n_i}$$

Subtracting λ_iP_i from both sides of (5.25), we obtain

$$(A - \lambda_i I)P_i = P_i(J_i - \lambda_i I) = P_iQ_i \tag{5.27}$$

where

$$Q_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n_i \times n_i}$$

Let

$$P_i = [\, \boldsymbol{v}_{i1} \ \cdots \ \boldsymbol{v}_{in_i} \,]$$

Equating the columns of the products on both sides of (5.27), we get

$$\begin{aligned} (A - \lambda_i I)\boldsymbol{v}_{i1} &= \mathbf{0} \\ (A - \lambda_i I)\boldsymbol{v}_{iq} &= \boldsymbol{v}_{i,q-1}\,, \quad q = 2, \ldots, n_i \end{aligned} \tag{5.28}$$

A vector $\boldsymbol{w} \in \mathbb{C}^{n\times 1}$ for which

$$\begin{aligned} (A - \lambda_i I)^r \boldsymbol{w} &= \mathbf{0} \\ (A - \lambda_i I)^{r-1} \boldsymbol{w} &\neq \mathbf{0} \end{aligned}$$

for some integer $r \geq 1$ is called a ***generalized eigenvector*** of rank r associated with λ_i.[2] A generalized eigenvector of rank one is an ordinary eigenvector. If $\boldsymbol{w}$ is a generalized eigenvector of rank $r > 1$, then $(A - \lambda_i I)\boldsymbol{w}$ is a generalized eigenvector of rank $r - 1$, because

$$\begin{aligned} (A - \lambda_i I)^{r-1}(A - \lambda_i I)\boldsymbol{w} &= (A - \lambda_i I)^r \boldsymbol{w} &&= \mathbf{0} \\ (A - \lambda_i I)^{r-2}(A - \lambda_i I)\boldsymbol{w} &= (A - \lambda_i I)^{r-1} \boldsymbol{w} &&\neq \mathbf{0} \end{aligned}$$

Thus a generalized eigenvector $\boldsymbol{w}$ of rank r generates a sequence of generalized eigenvectors

$$\boldsymbol{v}_1 = (A - \lambda_i I)^{r-1}\boldsymbol{w}\,,\ \boldsymbol{v}_2 = (A - \lambda_i I)^{r-2}\boldsymbol{w}\,,\ \ldots\,,\ \boldsymbol{v}_r = \boldsymbol{w}$$

of ranks $1, 2, \ldots, r$, such that

$$\begin{aligned} (A - \lambda_i I)\boldsymbol{v}_1 &= \mathbf{0} \\ (A - \lambda_i I)\boldsymbol{v}_q &= \boldsymbol{v}_{q-1}\,, \quad q = 2, \ldots, r \end{aligned}$$

Together with the zero vector, the set of all generalized eigenvectors associated with a multiple eigenvalue λ_i is a subspace, called the ***generalized eigenspace*** of A associated with λ_i, and denoted by $\boldsymbol{\mathcal{L}}_i$. Like an eigenspace $\boldsymbol{\mathcal{K}}_i$, a generalized eigenspace $\boldsymbol{\mathcal{L}}_i$ is also A-invariant (see Exercise 5.12).

(5.28) implies that when $\nu_i = 1 < n_i$, the columns $\boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{in_i}$ of P_i form a sequence of generalized eigenvectors associated with λ_i, generated by $\boldsymbol{w} = \boldsymbol{v}_{in_i}$. It can be shown that they form a basis for the generalized eigenspace $\boldsymbol{\mathcal{L}}_i$. That is, $\boldsymbol{\mathcal{V}}_i = \boldsymbol{\mathcal{L}}_i$, and it can not be further decomposed into smaller A-invariant subspaces.

[2] The rank of a generalized eigenvector must not be confused with the rank of a matrix.

Example 5.16

The matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ 2 & -2 & 3 \end{bmatrix}$$

has the characteristic polynomial

$$d(s) = (s-1)^3$$

Thus $\lambda_1 = 1$ is the only eigenvalue of A with $n_1 = 3$, and the Jordan form of A consists of a single Jordan block J_1.

Since

$$A - I = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & 2 \\ 2 & -2 & 2 \end{bmatrix} \overset{\text{e.r.o.}}{\longrightarrow} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

$\nu_1 = 1$, which means that J_1 consists of a single Jordan subblock J_{11}. Thus the Jordan form of A is

$$J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

To find a modal matrix of A all we need to do is to find a generalized eigenvector $\boldsymbol{w}$ of rank $n_1 = 3$ associated with $\lambda_1 = 1$. For this pupose we compute $(A-I)^2$ and $(A-I)^3$:

$$(A-I)^2 = \begin{bmatrix} 2 & -2 & 2 \\ 2 & -2 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$(A-I)^3 = O$$

A vector $\boldsymbol{w}$ for which $(A-I)^3\boldsymbol{w} = \mathbf{0}$ but $(A-I)^2\boldsymbol{w} \neq \mathbf{0}$ can be found as

$$\boldsymbol{w}_1 = \text{col}\,[\,0, 0, 1\,]$$

which yields a modal matrix

$$P_1 = [\,(A-I)^2\boldsymbol{w}_1 \quad (A-I)\boldsymbol{w}_1 \quad \boldsymbol{w}_1\,] = \begin{bmatrix} 2 & 0 & 0 \\ 2 & 2 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

Computing

$$P_1^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & 2 \end{bmatrix}$$

we can verify that $P_1^{-1}AP_1 = J$.

Note that since A has only one eigenvalue, there is no decomposition of $\mathbb{C}^{n\times 1}$ into A-invariant subspaces, and therefore, the Jordan form of A consists of a single block. However, it has a much simpler form than A because of the appropriate choice of the basis vectors.

A different choice of $\boldsymbol{w}$ gives a different modal matrix. For example, the choice

$$\boldsymbol{w}_2 = \text{col}\,[\,1, 0, 0\,]$$

yields

$$P_2 = [\,(A-I)^2\boldsymbol{w}_2 \quad (A-I)\boldsymbol{w}_2 \quad \boldsymbol{w}_2\,] = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 1 & 0 \\ 0 & 2 & 0 \end{bmatrix}$$

that also satisfies $P_2^{-1}AP_2 = J$.

This example also shows that although the Jordan form is unique, the modal matrix is not. The reader may try to find different generalized eigenvectors, and hence different modal matrices, and verify that they all result in the same Jordan form.

A class of matrices for which $\nu_i = 1$ for all eigenvalues (simple or multiple) are those in companion form. A matrix of the form

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix} \tag{5.29}$$

is said to be in ***companion form***. It is left to the reader as an exercise (see Exercise 5.24) to show that if A is of the form in (5.29), then

a) the characteristic polynomial of A is

$$d(s) = s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n$$

that is, the coefficients of the characteristic polynomial are the negatives of the last row elements of A in reverse order

b) $\nu_i = 1\,, \quad i = 1, \ldots, k$

c) the vectors

$$\boldsymbol{v}_{i1} = \boldsymbol{v}(\lambda_i)\,,\ \boldsymbol{v}_{i2} = \frac{1}{1!}\,\boldsymbol{v}'(\lambda_i)\,,\ \ldots\,,\ \boldsymbol{v}_{in_i} = \frac{1}{(n_i-1)!}\,\boldsymbol{v}^{(n_i-1)}(\lambda_i) \tag{5.30}$$

form a sequence of generalized eigenvectors associated with λ_i, where

$$\boldsymbol{v}(s) = \text{col}\,[\,1, s, \ldots, s^{n-1}\,]$$

Example 5.17

The matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & -5 & 4 \end{bmatrix}$$

is in companion form. Its characteristic polynomial can be written from the last row elements as

$$d(s) = s^3 - 4s^2 + 5s - 2 = (s-1)^2(s-2)$$

From

$$\boldsymbol{v}(s) = \begin{bmatrix} 1 \\ s \\ s^2 \end{bmatrix}, \quad \boldsymbol{v}'(s) = \begin{bmatrix} 0 \\ 1 \\ 2s \end{bmatrix}$$

we obtain a sequence of generalized eigenvectors associated with $\lambda_1 = 1$ as

$$\boldsymbol{v}_{11} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_{12} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

and a simple eigenvector associated with $\lambda_2 = 2$ as

$$\boldsymbol{v}_2 = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

Thus

$$P = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 4 \end{bmatrix}, \quad P^{-1} = \begin{bmatrix} 0 & 2 & -1 \\ -2 & 3 & -1 \\ 1 & -2 & 1 \end{bmatrix}$$

and

$$P^{-1}AP = J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Analysis of the case $1 < \nu_i < n_i$ is more complicated than the cases $\nu_i = n_i$ and $\nu_i = 1$ considered above, and is omitted. A worked example is given Exercise 5.28.

5.4.2 The Real Jordan Form

When A is a real matrix with real eigenvalues, we can choose its eigenvectors and generalized eigenvectors also real. We can then repeat the decomposition of the previous section by replacing $\mathbb{C}^{n\times 1}$ with $\mathbb{R}^{n\times 1}$. This leads to a real modal matrix and a real Jordan form. However, if A has complex eigenvalues then its eigenvectors and generalized eigenvectors, and therefore, its Jordan form will also be complex. In such a case, using the real and imaginary parts of the eigenvectors and generalized eigenvectors, we can construct a real modal matrix resulting in a real Jordan form. As in real semi-diagonal form, this corresponds to decomposing $\mathbb{R}^{n\times 1}$ into A-invariant subspaces.

Example 5.18

The matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4 & 8 & -8 & 4 \end{bmatrix}$$

which is in companion form, has the characteristic polynomial

$$d(s) = s^4 - 4s^3 + 8s^2 - 8s + 4 = (s^2 - 2s + 2)^2$$

Thus A has a pair of complex conjugate eigenvalues $\lambda_{1,2} = 1 \mp i$ with multiplicities $n_{1,2} = 2$. A complex modal matrix can be formed as

$$\begin{aligned} P &= \begin{bmatrix} \boldsymbol{v}(\lambda_1) & \boldsymbol{v}'(\lambda_1) & \boldsymbol{v}(\lambda_2) & \boldsymbol{v}'(\lambda_2) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1+i & 1 & 1-i & 1 \\ 2i & 2+2i & -2i & 2-2i \\ -2+2i & 6i & -2-2i & -6i \end{bmatrix} \end{aligned}$$

that yields a diagonal Jordan form

$$J = P^{-1}AP = \begin{bmatrix} 1+i & 1 & 0 & 0 \\ 0 & 1+i & 0 & 0 \\ 0 & 0 & 1-i & 1 \\ 0 & 0 & 0 & 1-i \end{bmatrix}$$

On the other hand, by constructing a real modal matrix as

$$\begin{aligned} P_R &= \begin{bmatrix} \text{Re}\, \boldsymbol{v}(\lambda_1) & \text{Im}\, \boldsymbol{v}(\lambda_1) & \text{Re}\, \boldsymbol{v}'(\lambda_1) & \text{Im}\, \boldsymbol{v}'(\lambda_1) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 2 & 2 & 2 \\ -2 & 2 & 0 & 6 \end{bmatrix} \end{aligned}$$

and computing its inverse, we obtain the real Jordan form of A as

$$J_R = P_R^{-1}AP_R = \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} J_{R1} & I \\ O & J_{R1} \end{bmatrix}$$

5.5 Function of a Matrix

Let $A \in \mathbb{C}^{n\times n}$ have the characteristic polynomial $d(s)$ in (5.7), and let the function $f : \mathbb{C} \to \mathbb{C}$ be analytic[3] in an open disc $\boldsymbol{D} = \{ s \in \mathbb{C} : |s| < r \}$ containing the eigenvalues of A. Let p be any polynomial such that

$$p^{(j)}(\lambda_i) = f^{(j)}(\lambda_i), \quad i = 1, \ldots, k, \quad j = 0, \ldots, n_i - 1 \tag{5.31}$$

We then define the matrix $f(A)$ to be

$$f(A) = p(A)$$

A polynomial p that satisfies (5.31) is called an ***interpolating polynomial*** of A for the function f.

There are three issues that must be considered in connection with this definition. The first is the existence of an interpolating polynomial. Let

$$p(s) = p_0 s^{n-1} + \cdots + p_{n-2}s + p_{n-1} = \boldsymbol{v}^t(s)\boldsymbol{p}$$

[3]The reader is referred to a book on complex calculus for a definition and properties of analytic functions. For our purpose, it suffices to know that many functions of interest, such as polynomials, exponential, trigonometric and hyperbolic functions, are analytic.

where

$$\boldsymbol{v}(s) = \text{col}\,[\,1, s, \ldots, s^{n-1}\,], \quad \boldsymbol{p} = \text{col}\,[\,p_{n-1}, p_{n-2}, \ldots, p_0\,]$$

Then (5.31) can be written in matrix form as

$$P^t \boldsymbol{p} = \boldsymbol{f} \tag{5.32}$$

where

$$P = [\,\boldsymbol{v}(\lambda_1)\ \cdots\ \boldsymbol{v}^{(n_1-1)}(\lambda_1)\ \cdots\ \boldsymbol{v}(\lambda_k)\ \cdots\ \boldsymbol{v}^{(n_k-1)}(\lambda_k)\,]$$

and

$$\boldsymbol{f} = \text{col}\,[\,f(\lambda_1)\ \cdots\ f^{(n_1-1)}(\lambda_1)\ \cdots\ f(\lambda_k)\ \cdots\ f^{(n_k-1)}(\lambda_k)\,]$$

Thus P is the modal matrix (with columns scaled) of a matrix in companion form and having $d(s)$ as its characteristic polynomial, and therefore, it is nonsingular. Hence (5.32) has a unique solution $\boldsymbol{p}$ that gives the coefficients of $p(s)$. This proves the existence of an interpolating polynomial.

The second issue concerned with the definition of function of a matrix is the uniqueness of $f(A)$. Obviously, there are infinitely many polynomials that satisfy (5.31). If p_1 and p_2 are any two such polynomials, then we must have $p_1(A) = p_2(A)$ for $f(A)$ to be defined uniquely. This is indeed the case, for if $p = p_1 - p_2$ then

$$p^{(j)}(\lambda_i) = 0\,, \quad i = 1, \ldots, k\,, \quad j = 0, \ldots, n_i - 1$$

which means that λ_i is a zero of p of multiplicity at least n_i. Then p must be of the form

$$p(s) = d(s)q(s)$$

for some polynomial q, and therefore,

$$p(A) = p_1(A) - p_2(A) = d(A)q(A) = O$$

that is, $p_1(A) = p_2(A)$.

The third concern with the definition of $f(A)$ is an issue of consistency. From complex calculus it is known that a function that is analytic in some open disc $\boldsymbol{D}$ in the complex plane has a power series representation

$$f(s) = \sum_{m=0}^{\infty} c_m s^m$$

which converges for all $s \in \boldsymbol{D}$. Then we expect that the infinite matrix series

$$\sum_{m=0}^{\infty} c_m A^m$$

should also converge to $f(A)$. The proof of the fact that

$$f(A) = \sum_{m=0}^{\infty} c_m A^m \tag{5.33}$$

is worked out in Exercise 5.31.

Example 5.19

Let us find $\sin \frac{\pi}{2} A$ for

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

Since A is in companion form, we immediately write

$$d(s) = s^2 - 2s + 1 = (s-1)^2$$

Thus the only eigenvalue is $\lambda_1 = 1$ with $n_1 = 2$. Let $p(s) = p_0 s + p_1$ be an interpolating polynomial for $f(s) = \sin \frac{\pi}{2} s$. Then

$$\begin{array}{rclcrclcl} p(\lambda_1) & = & p_0 + p_1 & = & f(\lambda_1) & = & \sin \dfrac{\pi}{2} & = & 1 \\ p'(\lambda_1) & = & p_0 & = & f'(\lambda_1) & = & \dfrac{\pi}{2} \cos \dfrac{\pi}{2} & = & 0 \end{array}$$

Thus $p_0 = 0, p_1 = 1$, and

$$\sin \frac{\pi}{2} A = p(A) = p_0 A + p_1 I = I$$

Example 5.20

Let us calculate e^{At} for

$$A = \begin{bmatrix} -4 & -3 \\ 2 & 1 \end{bmatrix}$$

The characteristic polynomial is $d(s) = (s+1)(s+2)$. Let $p(s) = p_0 s + p_1$ be an interpolating polynomial for $f(s) = e^{st}$. Then from

$$\begin{array}{rclcrcl} p(-1) & = & -p_0 + p_1 & = & f(-1) & = & e^{-t} \\ p(-2) & = & -2p_0 + p_1 & = & f(-2) & = & e^{-2t} \end{array}$$

we get

$$p_0 = e^{-t} - e^{-2t}, \quad p_1 = 2e^{-t} - e^{-2t}$$

Thus

$$e^{At} = p_0 A + p_1 I = \begin{bmatrix} -2e^{-t} + 3e^{-2t} & -3e^{-t} + 3e^{-2t} \\ 2e^{-t} - 2e^{-2t} & 3e^{-t} - 3e^{-2t} \end{bmatrix}$$

The reader can verify that

$$\frac{d}{dt} e^{At} = A e^{At}$$

where derivative of e^{At} is obtained by differentiating its elements individually. This is an expected result, which can be derived by differentiating the power series

$$e^{At} = \sum_{m=0}^{\infty} \frac{t^m}{m!} A^m$$

term-by-term as

$$\frac{d}{dt} e^{At} = \sum_{m=1}^{\infty} \frac{t^{m-1}}{(m-1)!} A^m = A \sum_{m=0}^{\infty} \frac{t^m}{m!} A^m = A e^{At}$$

Example 5.21

Let P be a projection matrix so that $P^2 = P$. Since

$$e^{st} = 1 + st + \frac{1}{2!}s^2t^2 + \frac{1}{3!}s^3t^3 + \cdots$$

we have

$$\begin{aligned} e^{Pt} &= I + Pt + \frac{1}{2!}P^2t^2 + \frac{1}{3!}P^3t^3 + \cdots \\ &= I + (t + \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \cdots)P \\ &= I + (e^t - 1)P \end{aligned}$$

Let A and B be similar matrices related as

$$A = PBP^{-1}$$

for some nonsingular matrix P. Then

$$A^m = (PBP^{-1})^m = PB^mP^{-1}, \quad m = 0, 1, \ldots$$

as can easily be shown by induction on m, so that

$$p(A) = p(PBP^{-1}) = Pp(B)P^{-1}$$

for any polynomial p. Let f be a given function, and let p be an interpolating polynomial of B for f. Since A and B have the same characteristic polynomial, it follows from (5.31) that p is also an interpolating polynomial of A for f. Hence we have

$$f(A) = p(A) = Pp(B)P^{-1} = Pf(B)P^{-1} \tag{5.34}$$

This property allows us to evaluate $f(A)$ using the Jordan (or diagonal) form of A as we explain below.

Let

$$B = \begin{bmatrix} B_1 & & \\ & \ddots & \\ & & B_k \end{bmatrix}$$

Then

$$B^m = \begin{bmatrix} B_1^m & & \\ & \ddots & \\ & & B_k^m \end{bmatrix}, \quad m = 0, 1, \ldots$$

Since any polynomial of B is a linear combination of a finite number of powers of B, it follows that

$$p(B) = \begin{bmatrix} p(B_1) & & \\ & \ddots & \\ & & p(B_k) \end{bmatrix}$$

for any polynomial p. Now let f be a given function and let p be an interpolating polynomial of B for f so that $f(B) = p(B)$. Since any eigenvalue λ of B_1 must be an eigenvalue of B and since the multiplicity of λ in the characteristic polynomial of B_1 can not exceed its multiplicity in the characteristic polynomial of B, it follows from (5.31) that p is also an interpolating polynomial of B_1 for f. The same is also true for all diagonal blocks of B. Hence $p(B_i) = f(B_i), i = 1, \ldots, k$, which implies that

$$f(B) = p(B) = \begin{bmatrix} p(B_1) & & \\ & \ddots & \\ & & p(B_k) \end{bmatrix} = \begin{bmatrix} f(B_1) & & \\ & \ddots & \\ & & f(B_k) \end{bmatrix}$$

Combining the above expression with (5.34) we observe that if A has the Jordan form $J = P^{-1}AP = \text{diag}\,[\,J_{ij}\,]$, then $A = PJP^{-1}$, and therefore

$$f(A) = Pf(J)P^{-1} = P \cdot \text{diag}\,[\,f(J_{ij})\,] \cdot P^{-1} \tag{5.35}$$

In particular, if A is diagonalizable then $J = D = \text{diag}\,[\,d_i\,]$ and (5.35) reduces to

$$f(A) = P \cdot \text{diag}\,[\,f(d_i)\,] \cdot P^{-1}$$

If A is not diagonalizable, for (5.35) to be useful in evaluating $f(A)$ we need an expression for the function of a Jordan subblock

$$J = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \lambda \end{bmatrix}_{n \times n} \tag{5.36}$$

where we omitted the subscripts for simplicity. For a given function f, let

$$p(s) = \sum_{q=0}^{n-1} \frac{f^{(q)}(\lambda)}{q!}(s - \lambda)^q$$

Since

$$p^{(j)}(s) = \sum_{q=j}^{n-1} \frac{f^{(q)}(\lambda)}{(q-j)!}(s - \lambda)^{q-j}\,, \quad j = 0, 1, \ldots, n-1$$

it follows that

$$p^{(j)}(\lambda) = f^{(j)}(\lambda)\,, \quad j = 0, 1, \ldots, n-1$$

Thus p is an interpolating polynomial of J for f, and therefore

$$f(J) = p(J) = \sum_{q=0}^{n-1} \frac{f^{(q)}(\lambda)}{q!}(J - \lambda I)^q = \sum_{q=0}^{n-1} \frac{f^{(q)}(\lambda)}{q!} Q^q \tag{5.37}$$

where

$$Q = J - \lambda I = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n\times n}$$

Recalling from Exercise 1.19 that the line of 1's above the diagonal of Q shifts one position upwards with each power of Q, we obtain from (5.37)

$$f(J) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2!}f''(\lambda) & \cdots & \frac{1}{(n-1)!}f^{(n-1)}(\lambda) \\ 0 & f(\lambda) & f'(\lambda) & \cdots & \frac{1}{(n-2)!}f^{(n-2)}(\lambda) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & f'(\lambda) \\ 0 & 0 & 0 & \cdots & f(\lambda) \end{bmatrix} \tag{5.38}$$

Example 5.22

Let us calculate e^{At} for the matrix A in Example 5.16.

Since we have already obtained the Jordan form of A, we can immediately write

$$\begin{aligned} e^{At} &= P_1 e^{Jt} P_1^{-1} \\ &= \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 \\ 2 & 2 & 0 \\ 0 & 2 & 1 \end{bmatrix}\begin{bmatrix} e^t & te^t & t^2e^t/2 \\ 0 & e^t & te^t \\ 0 & 0 & e^t \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & -2 & 2 \end{bmatrix} \\ &= e^t\begin{bmatrix} t^2-t+1 & -t^2+t & t^2 \\ t^2+t & -t^2-t+1 & t^2+2t \\ 2t & -2t & 2t+1 \end{bmatrix} \end{aligned}$$

To check the result, let $p(s) = \alpha s^2 + \beta s + \gamma$ be an interpolating polynomial for $f(s) = e^{st}$. Since A has the only eigenvalue $\lambda_1 = 1$ with $n_1 = 3$, (5.31) becomes

$$\begin{array}{lllllll} p(1) &=& \alpha+\beta+\gamma &=& f(1) &=& e^t \\ p'(1) &=& 2\alpha+\beta &=& f'(1) &=& te^t \\ p''(1) &=& 2\alpha &=& f''(1) &=& t^2e^t \end{array}$$

Note that f' denotes derivative with respect to s so that

$$f'(1) = [\,\frac{d}{ds}\,e^{st}\,]_{\,s=1} = [\,te^{st}\,]_{\,s=1} = te^t$$

Similarly,

$$f''(1) = t^2e^t$$

Solving for α, β, γ, we get

$$\alpha = \frac{t^2}{2}\,e^t\,, \quad \beta = (t - t^2)\,e^t\,, \quad \gamma = (1 - t + \frac{t^2}{2})\,e^t$$

and

$$e^{At} = \alpha A^2 + \beta A + \gamma I$$

gives the same result.

We finally note a couple of properties of function of a matrix. The first is that since any function of a matrix is defined through an interpolating polynomial, we have

$$f(A)g(A) = g(A)f(A)$$

for any f and g. A second property, which can be shown using the Jordan form of A, is that if A has the characteristic polynomial in (5.7), then $f(A)$ has the characteristic polynomial

$$\prod_{i=1}^{k}(s - f(\lambda_i))^{n_i}$$

In other words, the eigenvalues of $f(A)$ are $f(\lambda_i)$ with the same multiplicities in the characteristic polynomial. For example, the eigenvalues of $A + I$ are $\lambda_i + 1$, those of A^m are λ_i^m, etc.

5.6 Exercises

1. Find eigenvalues and eigenvectors of the following matrices

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \qquad B = \begin{bmatrix} 4 & 3 \\ -6 & -2 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix} \qquad D = \begin{bmatrix} 4i & 3 \\ -3 & -4i \end{bmatrix}$$

$$E = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 1 & 0 \\ 2 & 0 & 5 \end{bmatrix} \qquad F = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & 2 \end{bmatrix}$$

$$G = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \qquad H = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 3 \end{bmatrix}$$

$$K = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & -2 & 2 \\ 0 & 1 & -2 & 2 \end{bmatrix} \qquad M = \begin{bmatrix} -1 & 1 & 0 & 1 \\ 0 & -3 & 0 & -2 \\ -2 & 0 & -1 & 1 \\ 0 & 2 & 0 & 1 \end{bmatrix}$$

2. (a) Show that $\det(sI - A)$ is an nth degree polynomial in s with a unity leading coefficient. Hint: Let $A(s) = sI - A = [\, a_{ij}(s) \,]$ and consider the product terms in

$$\det A(s) = \sum_{\boldsymbol{J}_n} s(\boldsymbol{J}_n) a_{1j_1}(s) \cdots a_{nj_n}(s)$$

   (b) Show that elements of $B(s) = [\, b_{ij}(s) \,] = \text{adj}\,(sI - A)$ are polynomials in s of degree not exceeding $n - 1$. Hint: Recall that

$$b_{ij}(s) = (-1)^{i+j} m_{ji}^{B}(s)$$

3. Verify the Cayley-Hamilton theorem for the matrices in Exercise 5.1.

4. (a) Use the Cayley-Hamilton theorem to calculate A^{100} for

$$A = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$$

   (b) Check your result by MATLAB.

5. An $n \times n$ real matrix $A = [\, a_{ij} \,]$ whose elements satisfy

$$a_{ij} \geq 0 \quad \text{for all} \;\; (i, j)\,, \quad \sum_{i=1}^{n} a_{ij} = 1 \quad \text{for all} \;\; j$$

   is called a ***Markov matrix***. Show that $s = 1$ is an eigenvalue of a Markov matrix.

6. Prove the ***Gersgorin's theorem***:*The eigenvalues of an $n \times n$ complex matrix $A = [\, a_{ij} \,]$ are located in the union of the n discs*

$$\boldsymbol{D}_k : |\, s - a_{kk} \,| \leq \rho_k = \sum_{j \neq k}^{n} |\, a_{kj} \,|\,, \quad k = 1, 2, \ldots, n$$

   Hint: If λ is an eigenvalue of A with an associated eigenvector $\boldsymbol{x} = \text{col}\,[\, x_1, x_2, \ldots, x_n \,]$ then

$$(\lambda - a_{ii}) x_i = \sum_{j \neq i}^{n} a_{ij} x_j\,, \quad i = 1, 2, \ldots, n$$

   Suppose

$$\max_i |\, x_i \,| = |\, x_k \,|$$

   and consider the absolute values of both sides of the above equality with $i = k$.

7. Use Gershgorin's theorem to find upper and lower bounds on the magnitudes of the eigenvalues of the matrix

$$A = \begin{bmatrix} 4 & -1 & 0 & 0 \\ 1 & -4 & i & 0 \\ 0 & -i & 4i & -1 \\ 0 & 0 & 1 & -4i \end{bmatrix}$$

8. An $n \times n$ matrix $A = [\, a_{ij} \,]$ whose elements satisfy

$$|\, a_{kk} \,| < \sum_{j \neq k}^{n} |\, a_{kj} \,|\,, \quad k = 1, 2, \ldots, n$$

   is said to be ***diagonally dominant***.

   (a) Show that a diagonally dominant matrix is nonsingular.

   (b) Show that if the diagonal elements of a diagonally dominant matrix A are real and positive, then all the eigenvalues of A have positive real parts.

9. Prove the following properties of the minimum polynomial $\alpha(s)$ of a matrix A having a characteristic polynomial as given in (5.7).

(a) Show that $\alpha(s)$ is unique. Hint: Suppose that

$$\alpha(s) = s^m + \alpha_1 s^{m-1} + \cdots + \alpha_m$$

and

$$\alpha'(s) = s^m + \alpha'_1 s^{m-1} + \cdots + \alpha'_m$$

are two distinct minimum polynomials of A satisfying $\alpha(A) = \alpha'(A) = O$. (Note that α and α' must have the same degree, otherwise one of them can not be a minimum polynomial.) Define $r(s) = \alpha(s) - \alpha'(s)$. Then $r \neq 0$ and $r(A) = O$. Argue how this fact leads to a contradiction.

(b) Show that every eigenvalue λ_i of A is a zero of $\alpha(s)$. Hint: Divide $\alpha(s)$ with $(s - \lambda_i)$ and write

$$\alpha(s) = q_i(s)(s - \lambda_i) + r_i$$

where r_i is a constant remainder. Then

$$q_i(A)(A - \lambda_i I) + r_i I = \alpha(A) = O$$

Postmultiply both sides by an eigenvector $\boldsymbol{v}_i$ associated with λ_i to show that $r_i = 0$.

(c) Show that $\alpha(s)$ divides $d(s)$. Hint: Divide $d(s)$ with $\alpha(s)$ and write

$$d(s) = q(s)\alpha(s) + r(s)$$

and show that $r(s) = 0$.

Properties in (b) and (c) imply that the minimum polynomial must be of the form

$$\alpha(s) = \prod_{i=1}^{k} (s - \lambda_i)^{m_i}, \quad 1 \leq m_i \leq n_i$$

10. Given that the polynomials

$$\begin{aligned} p_1(s) &= s^7 + 2s^6 - 2s^5 - 4s^4 + 2s^3 + 5s^2 + 2s \\ p_2(s) &= s^2(s+1)^2(s+2) \end{aligned}$$

satisfy $p_1(A) = p_2(A) = O$ for a 10×10 matrix A.

(a) What can you say about the degree of the minimum polynomial of A?

(b) What are the remainder polynomials when p_1 and p_2 are divided by the minimum polynomial?

(c) How many distinct eigenvalues can A have?

(d) If A has 3 distinct eigenvalues, what are they? What is the minimum polynomial in this case?

11. Show that if

$$\mathbb{C}^{n\times 1} = \boldsymbol{\mathcal{U}}_1 \oplus \boldsymbol{\mathcal{U}}_2$$

where $\boldsymbol{\mathcal{U}}_1$ is A-invariant, then A is similar to a matrix of the form

$$F = \begin{bmatrix} F_{11} & F_{12} \\ O & F_{22} \end{bmatrix}$$

Hint: Let

$$P = [\, P_1 \;\; P_2 \,]$$

where columns of P_1 and P_2 form bases for $\boldsymbol{\mathcal{U}}_1$ and $\boldsymbol{\mathcal{U}}_2$.

12. Prove that if λ_i is an eigenvalue with multiplicity n_i in $d(s)$ then $\dim(\boldsymbol{\mathcal{K}}_i) \leq n_i$. Hint: Let $\dim(\boldsymbol{\mathcal{K}}_i) = \eta$, and let $\boldsymbol{\mathcal{V}}_i$ be a complement of $\boldsymbol{\mathcal{K}}_i$ so that

$$\mathbb{C}^{n\times 1} = \boldsymbol{\mathcal{K}}_i \oplus \boldsymbol{\mathcal{V}}_i$$

and use the result of Exercise 5.11 with $\boldsymbol{\mathcal{U}}_1 = \boldsymbol{\mathcal{K}}_i$ and $\boldsymbol{\mathcal{U}}_2 = \boldsymbol{\mathcal{V}}_i$.

13. Prove, without using the Jordan form, that every square matrix $A \in \mathbb{C}^{n\times n}$ is similar to an upper triangular matrix U whose diagonal elements are the eigenvalues of A. Hint: Let $A_1 = A$, and let $\boldsymbol{v}_1$ be an eigenvector of A_1 associated with an eigenvalue λ_1. Choose $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ such that

$$P_1 = [\,\boldsymbol{v}_1 \;\; \boldsymbol{v}_2 \;\cdots\; \boldsymbol{v}_n\,]$$

is nonsingular. Show that

$$P_1^{-1} A_1 P_1 = \begin{bmatrix} \lambda_1 & \boldsymbol{\alpha}_1 \\ \mathbf{0} & A_2 \end{bmatrix}$$

where $\boldsymbol{\alpha}_1$ is a row $(n-1)$-vector and A_2 is of order $n-1$. Then use induction on n.

14. The matrices

$$Q_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

that occur in quantum mechanics are called ***Pauli spin matrices***. Find eigenvalues and eigenvectors of Pauli spin matrices, and show that they are similar.

15. Find the Jordan forms of the matrices in Exercise 5.1.

16. Use MATLAB command `[P,J]=eig(A)` to find a modal matrix and the Jordan form of the matrices in Exercise 5.1 and comment on the results.

17. Find eigenvalues, eigenvectors, a modal matrix, and the Jordan form of

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

18. Use MATLAB command `A=rand(n,n)` to generate random matrices of different order, and use the command `[P,J]=eig(A)` to find their Jordan form. This exercise shows that almost all square matrices have distinct eigenvalues, and are therefore diagonalizable.

19. Find the Jordan forms of the following matrices.

$$A = \begin{bmatrix} \sigma & & & & & \\ & \sigma & 1 & & & \\ & & \sigma & & & \\ & & & \mu & & \\ & & & & \sigma & \end{bmatrix} \qquad B = \begin{bmatrix} \sigma & & 1 & & & \\ & \mu & & 1 & & \\ & & \sigma & & & \\ & & & \mu & & \\ & & & & \sigma & \end{bmatrix}$$

20. A linear transformation $\mathcal{A} : \mathbb{R}^2 \to \mathbb{R}^2$ is defined as

$$\mathcal{A}(x, y) = (x + y, -x + 3y)$$

(a) Find the matrix representation of $\mathcal{A}$ with respect to the basis $\boldsymbol{v}_1 = (1, 1)$ and $\boldsymbol{v}_2 = (1, 2)$ in $\mathbb{R}^2$.

(b) Can you find a basis with respect to which $\mathcal{A}$ has a diagonal representation? If yes, find it; otherwise, explain why.

21. Let $(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_4)$ be a basis for a four-dimensional complex vector space $\boldsymbol{\mathcal{X}}$, and let a linear operator $\mathcal{A}$ on $\boldsymbol{\mathcal{X}}$ be defined as

$$\mathcal{A}(\boldsymbol{u}_1) = \boldsymbol{u}_4\,, \quad \mathcal{A}(\boldsymbol{u}_2) = \boldsymbol{u}_1\,, \quad \mathcal{A}(\boldsymbol{u}_3) = \boldsymbol{u}_2\,, \quad \mathcal{A}(\boldsymbol{u}_4) = \boldsymbol{u}_3$$

Find eigenvalues and eigenvectors of $\mathcal{A}$. Hint: First obtain a matrix representation of $\mathcal{A}$ with respect to the given basis.

22. Let $\boldsymbol{\mathcal{X}} = \mathbb{R}_1[s] = \{\, p(s) = as + b \,|\, a, b \in \mathbb{R} \,\}$, and let a linear transformation $\mathcal{A} : \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{X}}$ be defined as

$$\mathcal{A}(as + b) = bs + a$$

Find a basis for $\boldsymbol{\mathcal{X}}$ with respect to which the matrix representation of $\mathcal{A}$ is in Jordan form.

23. Show that the generalized eigenspace $\boldsymbol{\mathcal{L}}_i$ of an eigenvalue λ_i of a matrix A is A-invariant. Hint: First show that $\boldsymbol{\mathcal{L}}_i$ is a subspace, and then show that if $\boldsymbol{w} \in \boldsymbol{\mathcal{L}}_i$ is a generalized eigenvector of rank r, then so is $A\boldsymbol{w}$.

24. Let A be an $n \times n$ matrix in companion form as given in (5.29). Define

$$C(s) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ s & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ s^{n-1} & s^{n-2} & \cdots & 1 \end{bmatrix}$$

(a) Show that

$$\det\,(sI - A) = \det\,[(sI - A)C(s)] = s^n + a_1 s^{n-1} + \cdots + a_n$$

(b) Show that

$$r(\lambda_i I - A) = r[(\lambda_i I - A)C(\lambda_i)] = n - 1$$

for every eigenvalue of A, so that $\nu_i = 1, i = 1, \ldots, k$.

(c) Let $\boldsymbol{y}(s) = (sI - A)\boldsymbol{v}(s)$. Calculate $\boldsymbol{y}(s)$ and show that

$$\boldsymbol{y}(\lambda_i) = \boldsymbol{y}'(\lambda_i) = \cdots = \boldsymbol{y}^{(n_i - 1)}(\lambda_i) = \boldsymbol{0}$$

Use this result together with

$$\begin{aligned} \boldsymbol{y}'(s) &= \boldsymbol{v}(s) + (sI - A)\boldsymbol{v}'(s) \\ \boldsymbol{y}''(s) &= 2\boldsymbol{v}'(s) + (sI - A)\boldsymbol{v}''(s) \end{aligned}$$

and so on, to show that the vectors in (5.30) form a sequence of generalized eigenvectors associated with λ_i.

25. Find the Jordan form of the following matrix in companion form.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & -2 & -3 & -2 \end{bmatrix}$$

26. Let $A \in \mathbb{C}^{n \times n}$ have n distinct eigenvalues λ_i with the associated eigenvectors $\boldsymbol{v}_i, i = 1, 2, ..., n$, and let $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2 + \cdots + \boldsymbol{v}_n$. Show that $\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{n-1}\boldsymbol{v}$ are linearly independent.

27. Let $A = \boldsymbol{b}\boldsymbol{c}^T$, where $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{c}^T\boldsymbol{b} \neq 0$. Obtain the Jordan form of A. Hint: Show that $\lambda_1 = \boldsymbol{c}^T\boldsymbol{b}$ is an eigenvalue of A and $\boldsymbol{v}_1 = \boldsymbol{b}$ is an associated eigenvector. What are the other eigenvalues?

28. The matrix

$$A = \begin{bmatrix} 0 & 4 & -4 \\ 1 & 0 & 2 \\ 2 & -4 & 6 \end{bmatrix}$$

has the characteristic polynomial

$$d(s) = (s-2)^3$$

Thus it has a single eigenvalue $\lambda_1 = 2$ with multiplicity $n_1 = 3$. From

$$A - 2I = \begin{bmatrix} -2 & 4 & -4 \\ 1 & -2 & 2 \\ 2 & -4 & 4 \end{bmatrix} \xrightarrow{R} \begin{bmatrix} 1 & -2 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we find that $1 < \nu_1 = 2 < 3$. Then the Jordan form of A must have a single Jordan block consisting of $\nu_1 = 2$ subblocks of orders two and one respectively (since the sum of the orders must be $n_1 = 3$, this is the only possibility). Find suitable generalized eigenvectors to form a modal matrix P, and show that $P^{-1}AP = J$. Hint: You need a generalized eigenvector of rank two and an ordinary eigenvector that is linearly independent of the vectors of the sequence generated by the generalized eigenvector.

29. Show that the real modal matrix P_R in (5.18) is nonsingular. Hint: Suppose that

$$a_1\boldsymbol{u}_1 + b_1\boldsymbol{w}_1 + \cdots + a_m\boldsymbol{u}_m + b_m\boldsymbol{w}_m + c_{2m+1}\boldsymbol{v}_{2m+1} + \cdots + c_n\boldsymbol{v}_n = \boldsymbol{0}$$

Then

$$c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_{2m-1}\boldsymbol{v}_{2m-1} + c_{2m}\boldsymbol{v}_{2m} + c_{2m+1}\boldsymbol{v}_{2m+1} + \cdots + c_n\boldsymbol{v}_n = \boldsymbol{0}$$

for some $c_i, i = 1, \ldots, 2m$.

30. Find the real Jordan forms of

(a) the matrix in Exercise 5.25

(b)

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

31. (a) Prove (5.33) for a diagonal matrix D.

(b) Prove (5.33) for a single Jordan subblock J given in (5.36).

(c) Use the results of parts (a) and (b), together with (5.35), to prove (5.33) for an arbitrary matrix A.

32. Find a general expression for e^{At} if

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$$

Note that A is in real Jordan form with eigenvalues $\lambda_{1,2} = \sigma \mp i\omega$.

33. Calculate e^{At} for

    (a) $A = \begin{bmatrix} -2 & 2 \\ 1 & -3 \end{bmatrix}$ by diagonalizing A,

    (b) $A = \begin{bmatrix} -1 & 1 \\ -2 & -3 \end{bmatrix}$ by using an interpolating polynomial,

    (c) $A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$ by any method.

34. Calculate $\sin(\pi A)$ and $\cos(\pi A)$ for

$$A = \begin{bmatrix} 0 & 1 \\ -1/6 & 5/6 \end{bmatrix}$$

    and verify the equality $\sin^2(\pi A) + \cos^2(\pi A) = I$. Check your results by MATLAB.

35. Show that the matrices $f(A)$ and $g(A)$ commute for arbitrary functions f and g.

36. Let A be an $n \times n$ real matrix with distinct eigenvalues $\lambda_i, i = 1, 2, ..., n$.

    (a) Show that $A = \sum_{i=1}^{n} \lambda_i Q_i$ for some matrices Q_i which satisfy

$$Q_i^2 = Q_i\,, \quad Q_i Q_j = 0\,, i \neq j\,, \quad \text{and} \quad \sum_{i=1}^{n} Q_i = I$$

    Hint: Start with $A = PDP^{-1}$, partition P into columns and P^{-1} into rows.

    (b) Show that with Q_i as above,

$$f(A) = \sum_{i=1}^{n} f(\lambda_i) Q_i$$

    for any function f for which $f(A)$ is defined.

37. Verify the result of Exercise 5.36 for the matrix A in Exercise 5.33(a) and for the function $f(s) = e^{st}$.

38. Show that if A satisfies $A^2 = -A$, then

$$e^{At} = (1 - e^{-t})A + I$$

Chapter 6
Linear Differential Equations

6.1 Systems of Linear Differential Equations

A system of n linear first order differential equations (SLDE) in n functions $x_1, \ldots, x_n$ of a real variable t has the general form

$$\begin{aligned} x_1' &= a_{11}(t)x_1 + \cdots + a_{1n}(t)x_n + u_1(t) \\ &\vdots \\ x_n' &= a_{n1}(t)x_1 + \cdots + a_{nn}(t)x_n + u_n(t) \end{aligned} \tag{6.1}$$

where $a_{ij}(t)$ and $u_i(t)$ are given real-valued functions. The SLDE in (6.1) can be written in compact form as

$$\boldsymbol{x}' = A(t)\boldsymbol{x} + \boldsymbol{u}(t) \tag{6.2}$$

with the obvious definitions of $\boldsymbol{x}$, $A(t)$ and $\boldsymbol{u}(t)$.

Suppose that $A(t)$ and $\boldsymbol{u}(t)$ are piecewise continuous on some interval $\boldsymbol{I} = (t_i, t_f)$. A vector-valued function $\boldsymbol{\phi} : \boldsymbol{I} \to \mathbb{R}^{n\times 1}$ is called a solution of (6.2) on $\boldsymbol{I}$ if

$$\boldsymbol{\phi}'(t) = A(t)\boldsymbol{\phi}(t) + \boldsymbol{u}(t)$$

for all $t \in \boldsymbol{I}$ except the discontinuity points of the right-hand side of (6.2),[1] where the derivative of $\boldsymbol{\phi} = \text{col}\,[\,\phi_1, \ldots, \phi_n\,]$ is defined element-by-element as

$$\boldsymbol{\phi}' = \text{col}\,[\,\phi_1', \ldots, \phi_n'\,]$$

Together with n initial conditions, $x_1(t_0) = x_{10}, \ldots, x_n(t_0) = x_{n0}$, (6.2) becomes an initial-value problem

$$\boldsymbol{x}' = A(t)\boldsymbol{x} + \boldsymbol{u}(t)\,, \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{6.3}$$

where

$$\boldsymbol{x}_0 = \text{col}\,[\,x_{10}, \ldots, x_{n0}\,]$$

Then a vector-valued function $\boldsymbol{\phi}(t)$ defined on some interval $\boldsymbol{I}$ that includes t_0 is called a solution of (6.3) if it satisfies both the differential equation and the initial condition. The following theorem, the proof of which is given in Appendix B, presents an existence and uniqueness result about the solution of (6.3).

[1] Strictly speaking, a solution $\boldsymbol{\phi}$ must have a continuous derivative on the interval of interest. In other words, we require that $\boldsymbol{\phi} \in \mathcal{C}_1(\boldsymbol{I}, \mathbb{R}^{n\times 1})$. However, as we discussed in connection with Theorem 2.1, we can extend the definition of a solution to include piecewise differentiable continuous functions provided that they are continuously differentiable on every subinterval of $\boldsymbol{I}$ that does not contain a discontinuity point of any element of A or $\boldsymbol{u}$.

Theorem 6.1 *Suppose that the elements of $A(t)$ and $\boldsymbol{u}(t)$ are piecewise continuous on some interval $\boldsymbol{I}$ and $t_0 \in \boldsymbol{I}$. Then the initial value problem (6.3) has a unique, continuous solution $\boldsymbol{x} = \boldsymbol{\phi}(t)$ on $\boldsymbol{I}$.*

6.1.1 Homogeneous SLDE

Consider the homogeneous SLDE

$$\boldsymbol{x}' = A(t)\boldsymbol{x} \tag{6.4}$$

where elements of $A(t)$ are piecewise continuous on an interval $\boldsymbol{I}$. The following theorem characterizes solutions of (6.4).

Theorem 6.2 *The set of solutions of (6.4) is an n-dimensional vector space over $\mathbb{R}$.*

Proof If $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are any two solutions of (6.4), then so is $\boldsymbol{\psi} = c_1\boldsymbol{\phi}_1 + c_2\boldsymbol{\phi}_2$ for any $c_1, c_2 \in \mathbb{R}$, because

$$\boldsymbol{\psi}'(t) = c_1\boldsymbol{\phi}_1'(t) + c_2\boldsymbol{\phi}_2'(t) = c_1A(t)\boldsymbol{\phi}_1(t) + c_2A(t)\boldsymbol{\phi}_2(t) = A(t)\boldsymbol{\psi}(t)$$

Hence the set of solutions of (6.4) is a subspace of $\mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$, and therefore, it is a vector space over $\mathbb{R}$.

Let $\boldsymbol{R} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be an ordered basis for $\mathbb{R}^{n\times 1}$, and let $\boldsymbol{\phi}_i(t)$ denote the unique solution of (6.4) that satisfy the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_i$ for some arbitrary $t_0 \in \boldsymbol{I}$. We claim that the set of solutions $(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)$ is linearly independent in $\mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$.[2] To prove the claim, suppose that

$$\boldsymbol{\psi} = c_1\boldsymbol{\phi}_1 + \cdots + c_n\boldsymbol{\phi}_n = \boldsymbol{0}$$

Then $\boldsymbol{\psi}(t) = \boldsymbol{0}$ for all $t \in \boldsymbol{I}$, and in particular,

$$\boldsymbol{\psi}(t_0) = c_1\boldsymbol{x}_1 + \cdots + c_n\boldsymbol{x}_n = \boldsymbol{0}$$

Since $\boldsymbol{R}$ is linearly independent, the last equality implies $c_1 = \cdots = c_n = 0$ proving the claim.

Let $\boldsymbol{\phi}$ be any solution of (6.4), and suppose that $\boldsymbol{\phi}(t_0) = \boldsymbol{x}_0$. Since $\boldsymbol{R}$ is a basis for $\mathbb{R}^{n\times 1}$,

$$\boldsymbol{x}_0 = \alpha_1\boldsymbol{x}_1 + \cdots + \alpha_n\boldsymbol{x}_n$$

for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Let

$$\boldsymbol{\psi} = \alpha_1\boldsymbol{\phi}_1 + \cdots + \alpha_n\boldsymbol{\phi}_n$$

Then $\boldsymbol{\psi}$ is a solution of (6.4) that satisfies the initial condition

$$\boldsymbol{x}(t_0) = \boldsymbol{\psi}(t_0) = \alpha_1\boldsymbol{\phi}_1(t_0) + \cdots + \alpha_n\boldsymbol{\phi}_n(t_0) = \alpha_1\boldsymbol{x}_1 + \cdots + \alpha_n\boldsymbol{x}_n = \boldsymbol{x}_0$$

By uniqueness of the solution of the initial-value problem consisting of (6.4) and the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$, we must have

$$\boldsymbol{\phi} = \boldsymbol{\psi} = \alpha_1\boldsymbol{\phi}_1 + \cdots + \alpha_n\boldsymbol{\phi}_n$$

This shows that the solutions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ also span the solution space, and therefore, they form a basis for it. We thus conclude that the solution space is n-dimensional.

[2] From now on, when talking about linear independence of a set of solutions, we will not mention the underlying vector space $\mathcal{F}(\boldsymbol{I}, \mathbb{R}^{n\times 1})$.

Let $\boldsymbol{\phi}_i(t), i = 1, \ldots, n$, be any n linearly independent solutions of (6.4) corresponding to a linearly independent set of initial conditions specified at some arbitrary t_0. Then the family of solutions

$$\boldsymbol{x} = c_1\boldsymbol{\phi}_1(t) + \cdots + c_n\boldsymbol{\phi}_n(t) \tag{6.5}$$

includes all solutions of (6.4), and therefore, characterizes a general solution. The general solution can conveniently be expressed in matrix form as

$$\boldsymbol{x} = X(t)\boldsymbol{c} \tag{6.6}$$

where

$$X(t) = [\,\boldsymbol{\phi}_1(t)\ \ \boldsymbol{\phi}_2(t)\ \cdots\ \boldsymbol{\phi}_n(t)\,]$$

and $\boldsymbol{c} \in \mathbb{R}^{n\times 1}$ is arbitrary. The matrix $X(t)$ is called a ***fundamental matrix*** of (6.4) and also of (6.2). Note that a fundamental matrix satisfies the matrix differential equation

$$X' = A(t)X \tag{6.7}$$

For convenience, we sometimes use the notation $\boldsymbol{x} = \boldsymbol{\phi}(t, t_0, \boldsymbol{x}_0)$ to denote the unique solution of (6.4) corresponding to the initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$. Similarly, we use the notation

$$X(t, t_0, X_0) = [\,\boldsymbol{\phi}(t, t_0, \boldsymbol{x}_1)\ \ \boldsymbol{\phi}(t, t_0, \boldsymbol{x}_2)\ \cdots\ \boldsymbol{\phi}(t, t_0, \boldsymbol{x}_n)\,]$$

to denote a fundamental matrix consisting of the solutions that correspond to an ordered linearly independent set of initial conditions $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ which form a nonsingular matrix

$$X_0 = [\,\boldsymbol{x}_1\ \ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n\,]$$

The special fundamental matrix corresponding to $X_0 = I$ is called the ***state transition matrix***, denoted $\boldsymbol{\Phi}(t, t_0)$. That is,

$$\boldsymbol{\Phi}(t, t_0) = [\,\boldsymbol{\phi}(t, t_0, \boldsymbol{e}_1)\ \ \boldsymbol{\phi}(t, t_0, \boldsymbol{e}_2)\ \cdots\ \boldsymbol{\phi}(t, t_0, \boldsymbol{e}_n)\,]$$

where $\boldsymbol{e}_j$ are columns of I_n. By definition, $X(t, t_0, X_0)$ and $\boldsymbol{\Phi}(t, t_0)$ are the unique solutions of the matrix differential equation (6.7) satisfying the initial conditions $X(t_0) = X_0$ and $X(t_0) = I$, respectively.

Given a fundamental matrix $X(t, t_0, X_0)$, by expressing an arbitrary initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ in terms of the columns of X_0 as $\boldsymbol{x}_0 = X_0\boldsymbol{\alpha}$, we observe that the solution corresponding to $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ is given by

$$\boldsymbol{\phi}(t, t_0, \boldsymbol{x}_0) = X(t, t_0, X_0)\boldsymbol{\alpha} = X(t, t_0, X_0)X_0^{-1}\boldsymbol{x}_0 \tag{6.8}$$

In particular,

$$\boldsymbol{\phi}(t, t_0, \boldsymbol{x}_0) = \boldsymbol{\Phi}(t, t_0)\boldsymbol{x}_0 \tag{6.9}$$

From (6.9) we observe that the unique solution of the homogeneous SLDE in (6.4) corresponding to an initial value $\boldsymbol{x}(t_0) = \boldsymbol{0}$ is $\boldsymbol{x} = \boldsymbol{0}$.

The expressions in (6.8) and (6.9) imply that that any fundamental matrix can be obtained from the state transition matrix as

$$X(t, t_0, X_0) = \mathbf{\Phi}(t, t_0)X_0$$

and conversely, given any fundamental matrix $X(t, t_0, X_0)$, we have

$$\mathbf{\Phi}(t, t_0) = X(t, t_0, X_0)X_0^{-1}$$

These relations also imply that if $X_1(t) = X(t, t_0, X_{10})$ and $X_2(t) = X(t, t_0, X_{20})$ are any two fundamental matrices, then

$$X_2(t) = \mathbf{\Phi}(t, t_0)X_{20} = X_1(t)X_{10}^{-1}X_{20} = X_1(t)Q$$

where $Q = X_{10}^{-1}X_{20}$ is nonsingular. Conversely, if $X_1(t) = X(t, t_0, X_{10})$ is a fundamental matrix and Q is any nonsingular matrix then

$$X_2(t) = X_1(t)Q = X(t, t_0, X_{10})Q = X(t, t_0, X_{10}Q)$$

is also a fundamental matrix.

An important property of a fundamental matrix $X(t, t_0, X_0)$ is that it is nonsingular for all $t, t_0 \in \boldsymbol{I}$. To show this suppose that $X(t_1, t_0, X_0)$ is singular for some $t_0, t_1 \in \boldsymbol{I}$, so that

$$X(t_1, t_0, X_0)\boldsymbol{c} = \mathbf{0}$$

for some $\boldsymbol{c} \neq \mathbf{0}$. Let $\boldsymbol{\psi}(t) = X(t; t_0, X_0)\boldsymbol{c}$. Then $\boldsymbol{\psi}(t)$ is the solution of the initial-value problem

$$\boldsymbol{x}' = A(t)\boldsymbol{x}\,, \quad \boldsymbol{x}(t_1) = \mathbf{0}$$

and therefore, $\boldsymbol{\psi}(t) = \mathbf{0}$ for all $t \in \boldsymbol{I}$. This implies that $\boldsymbol{\psi}(t_0) = X_0\boldsymbol{c} = \mathbf{0}$, contradicting the fact that X_0 is nonsingular. Thus $X(t, t_0, X_0)$ is nonsingular for all $t, t_0 \in \boldsymbol{I}$. In particular, $\mathbf{\Phi}(t, t_0)$ is nonsingular for all $t, t_0 \in \boldsymbol{I}$.

Another property of the state transition matrix is that

$$\mathbf{\Phi}(t, t_1)\mathbf{\Phi}(t_1, t_0) = \mathbf{\Phi}(t, t_0) \tag{6.10}$$

for all $t, t_0, t_1 \in \boldsymbol{I}$. To show this, let $\mathbf{\Psi}(t) = \mathbf{\Phi}(t, t_1)\mathbf{\Phi}(t_1, t_0)$. Then

$$\frac{d}{dt}\mathbf{\Psi}(t) = [\frac{d}{dt}\mathbf{\Phi}(t, t_1)]\mathbf{\Phi}(t_1, t_0) = A(t)\mathbf{\Phi}(t, t_1)\mathbf{\Phi}(t_1, t_0) = A(t)\mathbf{\Psi}(t)$$

and

$$\mathbf{\Psi}(t_1) = \mathbf{\Phi}(t_1, t_1)\mathbf{\Phi}(t_1, t_0) = \mathbf{\Phi}(t_1, t_0)$$

Hence $\mathbf{\Psi}(t)$ is the solution of the initial-value problem

$$X' = A(t)X\,, \quad X(t_1) = \mathbf{\Phi}(t_1, t_0)$$

Obviously, $\mathbf{\Phi}(t, t_0)$ is also a solution of this problem, and by uniqueness of the solution we must have $\mathbf{\Psi}(t) = \mathbf{\Phi}(t, t_1)\mathbf{\Phi}(t_1, t_0) = \mathbf{\Phi}(t, t_0)$.

Letting $t = t_0$ in (6.10), we obtain

$$\mathbf{\Phi}(t_0, t_1)\mathbf{\Phi}(t_1, t_0) = \mathbf{\Phi}(t_0, t_0) = I$$

This gives an explicit expression for the inverse of the state state transition matrix as

$$\mathbf{\Phi}^{-1}(t, t_0) = \mathbf{\Phi}(t_0, t) \tag{6.11}$$

Example 6.1

Consider the system of two first order linear differential equations

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

It can be verified by substitution that

$$\boldsymbol{\phi}_1(t) = \begin{bmatrix} \cos t \\ \cos t - \sin t \end{bmatrix}, \quad \boldsymbol{\phi}_2(t) = \begin{bmatrix} \sin t \\ \cos t + \sin t \end{bmatrix}$$

are two linearly independent solutions for $-\infty < t < \infty$.

A fundamental matrix is constructed from $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ as

$$X(t) = \begin{bmatrix} \cos t & \sin t \\ \cos t - \sin t & \cos t + \sin t \end{bmatrix}$$

from which the state transition matrix can be obtained as

$$\begin{aligned} \boldsymbol{\Phi}(t, t_0) &= X(t)X^{-1}(t_0) \\ &= \begin{bmatrix} \cos t & \sin t \\ \cos t - \sin t & \cos t + \sin t \end{bmatrix} \begin{bmatrix} \cos t_0 + \sin t_0 & -\sin t_0 \\ -\cos t_0 + \sin t_0 & \cos t_0 \end{bmatrix} \\ &= \begin{bmatrix} \cos(t - t_0) - \sin(t - t_0) & \sin(t - t_0) \\ -2\sin(t - t_0) & \cos(t - t_0) + \sin(t - t_0) \end{bmatrix} \end{aligned}$$

Note that $\boldsymbol{\Phi}(t_0, t_0) = I$.

The solution corresponding to $\boldsymbol{x}(0) = \boldsymbol{x}_0 = \text{col}\,[\,1, 0\,]$ is

$$\begin{aligned} \boldsymbol{x} = \boldsymbol{\Phi}(t, 0)\boldsymbol{x}_0 &= \begin{bmatrix} \cos t - \sin t & \sin t \\ -2\sin t & \cos t + \sin t \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \cos t - \sin t \\ -2\sin t \end{bmatrix} \end{aligned}$$

and the solution corresponding to $\boldsymbol{x}(\pi/2) = \boldsymbol{x}_1 = \text{col}\,[\,0, 1\,]$ is

$$\begin{aligned} \boldsymbol{x} = \boldsymbol{\Phi}(t, \pi/2)\boldsymbol{x}_1 &= \begin{bmatrix} \sin(t - \pi/2) \\ \cos(t - \pi/2) + \sin(t - \pi/2) \end{bmatrix} \\ &= \begin{bmatrix} -\cos t \\ -\cos t + \sin t \end{bmatrix} \end{aligned}$$

6.1.2 Non-Homogeneous SLDE

We now turn our attention to the non-homogeneous SLDE in (6.2). Following the method of variation of parameters, we assume a solution of the form

$$\boldsymbol{x} = X(t)\boldsymbol{v}(t)$$

where $X(t)$ is any fundamental matrix, and

$$\boldsymbol{v}(t) = \text{col}\,[\,v_1(t), \ldots, v_n(t)\,]$$

Substituting $\boldsymbol{x}$ and $\boldsymbol{x}'$ into (6.2), we obtain after simplification

$$X(t)\boldsymbol{v}'(t) = \boldsymbol{u}(t) \tag{6.12}$$

so that

$$\boldsymbol{v}'(t) = X^{-1}(t)\boldsymbol{u}(t)$$

Hence

$$\boldsymbol{v}(t) = \int X^{-1}(t)\boldsymbol{u}(t)\,dt = \boldsymbol{V}(t) + \boldsymbol{c}$$

where $\boldsymbol{V}(t)$ is any antiderivative of $X^{-1}(t)\boldsymbol{u}(t)$, and $\boldsymbol{c} \in \mathbb{R}^{n\times 1}$ is arbitrary. Thus a general solution of (6.2) is obtained as

$$\boldsymbol{x} = X(t)(\boldsymbol{V}(t) + \boldsymbol{c}) = \boldsymbol{\phi}_p(t) + \boldsymbol{\phi}_c(t) \tag{6.13}$$

where $\boldsymbol{\phi}_p(t)$ is a particular solution, and $\boldsymbol{\phi}_c(t)$ is the complementary solution.

When an initial condition $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ is specified along with (6.2), then it is convenient to choose $X(t) = \boldsymbol{\Phi}(t, t_0)$, and

$$\boldsymbol{V}(t) = \int_{t_0}^{t} \boldsymbol{\Phi}(t_0, \tau)\boldsymbol{u}(\tau)\,d\tau$$

Then $\boldsymbol{V}(t_0) = \boldsymbol{0}$, and (6.13) evaluated at $t = t_0$ gives $\boldsymbol{c} = \boldsymbol{x}_0$. Thus the required solution is obtained as

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{\Phi}(t, t_0)(\boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t_0, \tau)\boldsymbol{u}(\tau)\,d\tau) \\ &= \boldsymbol{\Phi}(t, t_0)\boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\boldsymbol{u}(\tau)\,d\tau \;=\; \boldsymbol{\Phi}_o(t) + \boldsymbol{\Phi}_u(t) \end{aligned} \tag{6.14}$$

Note that the expressions in (6.14) are generalization of the expressions in (2.44) and (2.45) to vector case: $\boldsymbol{\Phi}_o(t)$ is the part of the solution due to $\boldsymbol{x}_0$ and $\boldsymbol{\Phi}_u(t)$ is the part due to $\boldsymbol{u}(t)$.

Example 6.2

Let us find the solution of

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} u(t)\,, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

for a unit step input

$$u(t) = \begin{cases} 0\,, & t < 0 \\ 1\,, & t > 0 \end{cases}$$

When $u_1, \ldots, u_n$ in (6.1) are proportional as in this example, so that $\boldsymbol{u}(t) = \boldsymbol{b}u(t)$ for some $\boldsymbol{b} = \text{col}\,[\,b_1, \ldots, b_n\,]$, then it is more convenient to express (6.2) as

$$\boldsymbol{x}' = A\boldsymbol{x} + \boldsymbol{b}u(t)$$

A fundamental matrix for the associated homogeneous SLDE has already been obtained in Example 6.1.

For $t < 0$, $u(t) = 0$ and the given SLDE reduces to a homogeneous one whose solution is $\boldsymbol{x} = X(t)\boldsymbol{c}$, where $\boldsymbol{c} = \text{col}\,[\,c_1, c_2\,]$ is arbitrary. Using the continuity of the solution, $\boldsymbol{c}$ can be evaluated from the initial condition as

$$\boldsymbol{x}(0) = X(0)\boldsymbol{c} = \boldsymbol{x}_0 = \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{c} = \boldsymbol{0}$$

Hence

$$\boldsymbol{x} = \boldsymbol{0}\,, \quad t \leq 0$$

For $t > 0$, $u(t) = 1$ and the solution is of the form $\boldsymbol{x} = X(t)\boldsymbol{v}(t)$, where

$$\begin{aligned} \boldsymbol{v}'(t) &= X^{-1}(t)\boldsymbol{b}u(t) = \begin{bmatrix} \cos t + \sin t & -\sin t \\ -\cos t + \sin t & \cos t \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} \cos t - \sin t \\ \cos t + \sin t \end{bmatrix} \end{aligned}$$

Integrating $\boldsymbol{v}'(t)$, we get

$$\boldsymbol{v}(t) = \begin{bmatrix} \sin t + \cos t + c_1 \\ \sin t - \cos t + c_2 \end{bmatrix}$$

Thus a general solution is obtained as

$$\begin{aligned} \boldsymbol{x} &= X(t)\boldsymbol{v}(t) = \begin{bmatrix} \cos t & \sin t \\ \cos t - \sin t & \cos t + \sin t \end{bmatrix} \begin{bmatrix} \sin t + \cos t + c_1 \\ \sin t - \cos t + c_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 + c_1 \cos t + c_2 \sin t \\ (c_1 + c_2)\cos t + (c_2 - c_1)\sin t \end{bmatrix} \end{aligned}$$

Initial conditions give

$$\boldsymbol{x}(0) = \begin{bmatrix} 1 + c_1 \\ c_1 + c_2 \end{bmatrix} = \boldsymbol{0} \quad \Longrightarrow \quad \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Hence

$$\boldsymbol{x} = \begin{bmatrix} 1 - \cos t + \sin t \\ 2\sin t \end{bmatrix}, \quad t \geq 0$$

Alternatively, using the state transition matrix $\boldsymbol{\Phi}$ obtained in Example 6.1 in the expression in (6.14) with $t_0 = 0$ and $\boldsymbol{x}_0 = \boldsymbol{0}$, we get

$$\boldsymbol{x} = \int_0^t \boldsymbol{\Phi}(t,\tau)\boldsymbol{b}u(\tau)\,d\tau = \int_0^t \begin{bmatrix} \cos(t-\tau) + \sin(t-\tau) \\ 2\cos(t-\tau) \end{bmatrix} u(\tau)\,d\tau$$

For $t \leq 0$, $u(\tau) = 0$ on the interval of integration, so that $\boldsymbol{x} = \boldsymbol{0}$. For $t \geq 0$, $u(\tau) = 1$ on the interval of integration and we get

$$\boldsymbol{x} = \begin{bmatrix} -\sin(t-\tau) + \cos(t-\tau) \\ -2\sin(t-\tau) \end{bmatrix}_{\tau=0}^{\tau=t} = \begin{bmatrix} 1 - \cos t + \sin t \\ 2\sin t \end{bmatrix}$$

6.1.3 SLDE With Constant Coefficients

As in the case of higher order linear differential equations with non-constant coefficients, it may not be possible to obtain a fundamental matrix of (6.2) when the coefficient matrix is non-constant. However, when $A(t) = A$, a constant matrix, then the state transition matrix of the SLDE in (6.4) is given explicitly by the matrix exponential function

$$\boldsymbol{\Phi}(t, t_0) = e^{A(t-t_0)} \tag{6.15}$$

This follows from the facts that

$$\frac{d}{dt} e^{A(t-t_0)} = A e^{A(t-t_0)}$$

as shown in Example 5.20 and that

$$e^{A(t-t_0)} \,|_{\, t=t_0} = I$$

Thus the solution of the constant-coefficient-SLDE

$$\boldsymbol{x}' = A\boldsymbol{x} + \boldsymbol{u}(t)\,, \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0 \tag{6.16}$$

is obtained by substituting $\boldsymbol{\Phi}(t, \tau) = e^{A(t-\tau)}$ in (6.14) as

$$\boldsymbol{x} = e^{A(t-t_0)}\boldsymbol{x}_0 + \int_{t_0}^{t} e^{A(t-\tau)}\boldsymbol{u}(\tau)\, d\tau = \boldsymbol{\phi}_o(t) + \boldsymbol{\phi}_u(t) \tag{6.17}$$

Example 6.3

The SLDE in Example 6.1 has a constant coefficient matrix

$$A = \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix}$$

The eigenvalues of A are $\lambda_{1,2} = \mp i$. Let $p(s) = \alpha s + \beta$ be an interpolating polynomial for $f(s) = e^{st}$. Then from

$$\begin{array}{rcccl} i\,\alpha + \beta & = & e^{it} & = & \cos t + i\, \sin t \\ -i\,\alpha + \beta & = & e^{-it} & = & \cos t - i\, \sin t \end{array}$$

we obtain $\alpha = \sin t$ and $\beta = \cos t$. Thus

$$e^{At} = \alpha A + \beta I = \begin{bmatrix} \cos t - \sin t & \sin t \\ -2\sin t & \cos t + \sin t \end{bmatrix}$$

Hence

$$\begin{array}{rcl} \boldsymbol{\Phi}(t, t_0) & = & e^{A(t-t_0)} \\ & = & \begin{bmatrix} \cos(t-t_0) - \sin(t-t_0) & \sin(t-t_0) \\ -2\sin(t-t_0) & \cos(t-t_0) + \sin(t-t_0) \end{bmatrix} \end{array}$$

which is the same as the one constructed in Example 6.1 from a given pair of linearly independent solutions.

6.2 Modal Decomposition of Solutions

Consider a homogeneous SLDE with constant coefficients

$$\boldsymbol{x}' = A\boldsymbol{x}\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{6.18}$$

where we assume, without loss of generality, that $t_0 = 0$.[3] For simplicity in notation, let us denote the solution of (6.18) by $\boldsymbol{x} = \boldsymbol{\phi}(t, \boldsymbol{x}_0)$ by omitting the unnecessary argument $t_0 = 0$. Thus

$$\boldsymbol{\phi}(t, \boldsymbol{x}_0) = e^{At}\boldsymbol{x}_0 \tag{6.19}$$

6.2.1 Complex Modes

Suppose λ is an eigenvalue of A with an associated eigenvector $\boldsymbol{v}$, and suppose $\boldsymbol{x}_0 = \boldsymbol{v}$. Then since

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

we have

$$e^{At}\boldsymbol{v} = e^{\lambda t}\boldsymbol{v}$$

so that the solution of (6.18) corresponding to $\boldsymbol{x}_0 = \boldsymbol{v}$ is

$$\boldsymbol{x} = \boldsymbol{\phi}(t, \boldsymbol{v}) = e^{At}\boldsymbol{v} = e^{\lambda t}\boldsymbol{v} \tag{6.20}$$

(6.20) implies that

$$\boldsymbol{\phi}(t, \boldsymbol{v}) \in \text{span}\,(\boldsymbol{v}) \quad \text{for all} \;\; t$$

that is, a (complex) solution that starts from an eigenvector $\boldsymbol{v}$ remains in the direction of that eigenvector for all t.

The observation above allows us to decompose the solution starting from an arbitrary initial condition into components with specific properties. We first consider the simpler case where A has simple distinct eigenvalues λ_i with associated linearly independent eigenvectors $\boldsymbol{v}_i$, $i = 1, \ldots, n$. Then any arbitrary initial condition can be decomposed uniquely as

$$\boldsymbol{x}_0 = \alpha_1\boldsymbol{v}_1 + \cdots + \alpha_n\boldsymbol{v}_n = \boldsymbol{x}_{01} + \cdots + \boldsymbol{x}_{0n} \tag{6.21}$$

where

$$\boldsymbol{x}_{0i} = \alpha_i\boldsymbol{v}_i$$

are components of $\boldsymbol{x}_0$ in one-dimensional eigenspaces $\boldsymbol{\mathcal{K}}_i = \text{span}\,(\boldsymbol{v}_i)$. The corresponding solution is then obtained as

$$\begin{aligned} \boldsymbol{x} = \boldsymbol{\phi}(t, \boldsymbol{x}_0) &= \alpha_1 e^{At}\boldsymbol{v}_1 + \cdots + \alpha_n e^{At}\boldsymbol{v}_n \\ &= \alpha_1 e^{\lambda_1 t}\boldsymbol{v}_1 + \cdots + \alpha_n e^{\lambda_n t}\boldsymbol{v}_n \\ &= \boldsymbol{\phi}(t, \boldsymbol{x}_{01}) + \cdots + \boldsymbol{\phi}(t, \boldsymbol{x}_{0n}) \end{aligned} \tag{6.22}$$

[3] To study the case $t_0 \neq 0$ all we have to do is to replace t with $t - t_0$ in the expression for the state transition matrix.

where each

$$\boldsymbol{\phi}(t, \boldsymbol{x}_{0i}) = e^{\lambda_i t}\boldsymbol{x}_{0i} = \alpha_i e^{\lambda_i t}\boldsymbol{v}_i\,, \quad i = 1, \ldots, n \tag{6.23}$$

is itself a solution corresponding to an initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_{0i} = \alpha_i \boldsymbol{v}_i$. Thus a decomposition of the initial condition $\boldsymbol{x}_0$ into components $\boldsymbol{x}_{0i}$ as in (6.20) results in a corresponding decomposition of the solution into components $\boldsymbol{\phi}(t, \boldsymbol{x}_{0i})$ as in (6.22). Both decompositions are a result of the fact that

$$\mathcal{C}^{n\times 1} = \boldsymbol{\mathcal{K}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{K}}_n$$

Each solution component $\boldsymbol{\phi}(t; \boldsymbol{x}_{0i})$ that starts at $t = 0$ from a point $\boldsymbol{x}_{0i}$ in the A-invariant subspace $\boldsymbol{\mathcal{K}}_i$ and remains there for all t, is called a ***mode*** of the SLDE in (6.18), and the decomposition of a solution into its modes as in (6.22) is called the ***modal decomposition*** of solution.

Example 6.4

Let us find the solution of (6.18) for

$$A = \begin{bmatrix} -3 & -2 \\ -1 & -2 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and decompose it into its modes.

The eigenvalues and eigenvectors of A can be computed as $\lambda_1 = -4$, $\lambda_2 = -1$, and

$$\boldsymbol{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

The initial condition can easily be decomposed into its components along the eigenvectors as

$$\boldsymbol{x}_0 = \boldsymbol{v}_1 + \boldsymbol{v}_2$$

Hence the solution, decomposed into its modes, is

$$\boldsymbol{x} = \boldsymbol{\phi}(t, \boldsymbol{x}_0) = e^{-4t}\boldsymbol{v}_1 + e^{-t}\boldsymbol{v}_2 = \begin{bmatrix} 2e^{-4t} - e^{-t} \\ e^{-4t} + e^{-t} \end{bmatrix}$$

The modal decomposition of the solution is illustrated in Figure 6.1. The reader should try to obtain the same solution from (6.19) by calculating the matrix function e^{At}.

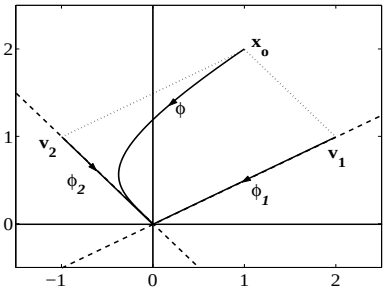

Figure 6.1: Modal decomposition of solution

We now extend the modal decomposition of solution to the general case. Let A have the characteristic polynomial

$$d(s) = \prod_{i=1}^{k} (s - \lambda_i)^{n_i}$$

where λ_i are the distinct eigenvalues with multiplicities n_i. Recall from Section 5.4 that the eigenstructure of A induces a direct sum decomposition of $\mathbb{C}^{n\times 1}$ as

$$\mathbb{C}^{n\times 1} = \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{n_i} \boldsymbol{\mathcal{V}}_{ij}$$

where each $\boldsymbol{\mathcal{V}}_{ij}$ is an A-invariant subspace spanned by eigenvectors and/or generalized eigenvectors associated with the eigenvalue λ_i. Moreover, if the columns of the $n \times n_{ij}$ matrix P_{ij} form a basis for $\boldsymbol{\mathcal{V}}_{ij}$ then

$$AP_{ij} = P_{ij} J_{ij} \tag{6.24}$$

where J_{ij} is the corresponding Jordan subblock. Since the columns of the modal matrix P constructed from P_{ij} form a basis for $\mathbb{C}^{n\times 1}$, any initial condition $\boldsymbol{x}_0$ can be expressed as

$$\boldsymbol{x}_0 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} P_{ij} \boldsymbol{\alpha}_{ij} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \boldsymbol{x}_{0ij} \tag{6.25}$$

for some uniquely determined $\boldsymbol{\alpha}_{ij} \in \mathbb{C}^{n_{ij}\times 1}$. Then the solution $\boldsymbol{\phi}(t; \boldsymbol{x}_0)$ has a corresponding decomposition

$$\boldsymbol{\phi}(t, \boldsymbol{x}_0) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \boldsymbol{\phi}(t, \boldsymbol{x}_{0ij}) \tag{6.26}$$

where

$$\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij}) = e^{At} \boldsymbol{x}_{0ij} = e^{At} P_{ij} \boldsymbol{\alpha}_{ij}$$

(6.24) implies that $A^m P_{ij} = P_{ij} J_{ij}^m$ for all m, which in turn implies that $p(A)P_{ij} = P_{ij} p(J_{ij})$ for any polynomial p, and therefore, $f(A)P_{ij} = P_{ij} f(J_{ij})$ for any function f, as we have already discussed in Section 5.6. In particular,

$$\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij}) = e^{At} P_{ij} \boldsymbol{\alpha}_{ij} = P_{ij} e^{J_{ij} t} \boldsymbol{\alpha}_{ij} \tag{6.27}$$

and (6.26) becomes

$$\boldsymbol{\phi}(t; \boldsymbol{x}_0) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} P_{ij} e^{J_{ij} t} \boldsymbol{\alpha}_{ij} \tag{6.28}$$

Now modes of the solution are $\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij})$ in (6.27). As in the simple eigenvalues case, the mode $\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij})$ starts from $\boldsymbol{x}_{0ij} \in \boldsymbol{\mathcal{V}}_{ij}$ at $t = 0$ and remains in $\boldsymbol{\mathcal{V}}_{ij}$ for all t.

Recall from Section 5.6 that

$$e^{J_{ij}t} = e^{\lambda_i t} \begin{bmatrix} 1 & t & \cdots & t^{n_{ij}-1}/(n_{ij}-1)! \\ 0 & 1 & \cdots & t^{n_{ij}-2}/(n_{ij}-2)! \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{6.29}$$

which implies that the mode $\boldsymbol{\phi}(t; \boldsymbol{x}_{0ij})$ consists of $e^{\lambda_i t}, te^{\lambda_i t}, \ldots, t^{n_{ij}-1}e^{\lambda_i t}$ terms. That is, it is of the form

$$\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij}) = e^{\lambda_i t}\boldsymbol{\beta}_{ij1} + te^{\lambda_i t}\boldsymbol{\beta}_{ij2} + \cdots + t^{n_{ij}-1}e^{\lambda_i t}\boldsymbol{\beta}_{ijn_{ij}}$$

for some $\boldsymbol{\beta}_{ijm} \in \mathbb{C}^{n\times 1}$. Of course, not all the terms have to be present in the expression above. However, it can be shown that if the $t^q e^{\lambda_i t}$ term appears in $\boldsymbol{\phi}(t, \boldsymbol{x}_{0ij})$ then all the $t^m e^{\lambda_i t}, m < q$, terms must also appear (see Exercise 6.16).

When the eigenvalues of A are simple, that is, when $\nu_i = n_i = 1$ for all i, we have $P_i = P_{i1} = \boldsymbol{v}_i$ and $J_i = J_{i1} = \lambda_i$, where $\boldsymbol{v}_i$ are the eigenvectors of A associated with the eigenvalues λ_i. Then (6.25), (6.27) and (6.28) reduce to (6.21), (6.23) and (6.22) respectively.

Another case of special interest is when $\nu_i = 1$ for all eigenvalues even when they are multiple. In this case, each Jordan block consists of a single subblock, that is

$$P_i = P_{i1} = [\, \boldsymbol{v}_i \cdots \boldsymbol{v}_{in_i} \,], \quad J_i = J_{i1} = \begin{bmatrix} \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & \lambda_i & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}_{n_i \times n_i}$$

where $\boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{in_i}$ form a sequence of generalized eigenvectors associated with λ_i. Then (6.25) becomes

$$\boldsymbol{x}_0 = \sum_{i=1}^{k} \boldsymbol{x}_{0i} = \sum_{i=1}^{k} P_i \boldsymbol{\alpha}_i$$

and consequently, the modal decomposition in (6.28) takes the form

$$\boldsymbol{\phi}(t, \boldsymbol{x}_0) = \sum_{i=1}^{k} P_i e^{J_i t} \boldsymbol{\alpha}_i$$

Note that the ith mode consists of $e^{\lambda_i t}, te^{\lambda_i t}, \ldots, t^{n_i-1}e^{\lambda_i t}$ terms, in general.

Example 6.5

Let us find the solution of (6.18) for the A matrix in Example 5.17, and for the initial condition $\boldsymbol{x}_0 = \text{col}\,[\,1, 0, 0\,]$.

A modal matrix for and the Jordan form of A are already found in Example 5.17. Expressing $\boldsymbol{x}_0$ in terms the blocks of the modal matrix as

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} [\,1\,] = P_1\boldsymbol{\alpha}_1 + \boldsymbol{v}_2\alpha_2$$

we obtain the solution in terms of its modes as

$$\begin{aligned}\boldsymbol{x} &= P_1 e^{J_1 t}\boldsymbol{\alpha}_1 + \boldsymbol{v}_2 e^{\lambda_2 t}\alpha_2 \\ &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} e^t & te^t \\ 0 & e^t \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} + e^{2t} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \\ &= e^t \begin{bmatrix} -2t \\ -2t-2 \\ -2t-4 \end{bmatrix} + e^{2t} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}\end{aligned}$$

6.2.2 Real Modes

If all eigenvalues of A are real, then choosing the modal matrix P also real we obtain a modal decomposition of any solution in which all the modes are real. However, if some eigenvalues of A are complex, then the eigenvectors associated with complex eigenvalues will also be complex. Although there is nothing to prevent us to treat $\boldsymbol{x}_0$ as an element of $\mathbb{C}^{n\times 1}$, and obtain a modal decomposition of the solution as in the previous subsection, the modes will turn out to be complex-valued in general. To decompose the solution into real components, we need to consider the complex eigenvalues and eigenvectors in conjugate pairs.

Let us first consider the case of simple eigenvalues. Suppose that A has m distinct pairs of complex conjugate eigenvalues $\lambda_{2i-1,2i} = \sigma_i \mp i\omega_i$ with associated complex conjugate pairs of eigenvectors $\boldsymbol{v}_{2i-1,2i} = \boldsymbol{u}_i \mp i\boldsymbol{w}_i, i = 1,\ldots,m$, and $n-2m$ real eigenvalues λ_i, with associated real eigenvectors $\boldsymbol{v}_i, i = 2m+1,\ldots,n$. Then, when a real initial condition vector is expressed in terms of the eigenvectors as in (6.21), it turns out that the coefficients of complex conjugate eigenvectors also appear in conjugate pairs, that is, if

$$\boldsymbol{x}_0 = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i$$

then

$$\alpha_{2i-1,2i} = \beta_i \mp i\gamma_i\,, \quad i = 1,\ldots,m$$

Now consider a complex conjugate pair of solution components

$$\begin{aligned}&\alpha_{2i-1}e^{\lambda_{2i-1}t}\boldsymbol{v}_{2i-1} + \alpha_{2i}e^{\lambda_{2i}t}\boldsymbol{v}_{2i} \\ &\quad= 2\,\mathrm{Re}\,\{\,\alpha_{2i-1}e^{\lambda_{2i-1}t}\boldsymbol{v}_{2i-1}\,\} \\ &\quad= 2\,\mathrm{Re}\,\{\,(\beta_i + i\gamma_i)e^{\sigma_i t}(\cos\omega_i t + i\sin\omega_i t)(\boldsymbol{u}_i + i\boldsymbol{w}_i)\,\} \\ &\quad= 2\,e^{\sigma_i t}(\beta_i\cos\omega_i t - \gamma_i\sin\omega_i t)\boldsymbol{u}_i - 2\,e^{\sigma_i t}(\gamma_i\cos\omega_i t + \beta_i\sin\omega_i t)\boldsymbol{w}_i \qquad (6.30)\end{aligned}$$

We observe that these two solution components, which are themselves complex solutions, add up to a real solution that lies in the two-dimensional subspace defined by the real and imaginary parts $\boldsymbol{u}_i$ and $\boldsymbol{w}_i$ of the associated eigenvectors $\boldsymbol{v}_{2i-1}$ and $\boldsymbol{v}_{2i}$, and thus define a two-dimensional mode. Repeating this for all conjugate pairs of complex modes, the solution can be decomposed as

$$\boldsymbol{\phi}(t,\boldsymbol{x}_0) = \sum_{i=1}^{m}\boldsymbol{\phi}_i(t) + \sum_{i=2m+1}^{n}\boldsymbol{\phi}_i(t) \qquad (6.31)$$

where $\boldsymbol{\phi}_i(t)$ is of the form (6.30) for $i = 1, \ldots, m$, and of the form (6.23) for $i = 2m + 1, \ldots, n$. Observe that the solution consists of $e^{\sigma_i t} \cos \omega_i t$ and $e^{\sigma_i t} \sin \omega_i t$ terms corresponding to complex eigenvalues, and $e^{\lambda_i t}$ terms corresponding to real eigenvalues.

The decomposition of the solution into real modes corresponds to direct sum decomposition of $\mathbb{R}^{n\times 1}$ as

$$\mathbb{R}^{n\times 1} = \boldsymbol{\mathcal{V}}_1 \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_m \oplus \boldsymbol{\mathcal{V}}_{2m+1} \oplus \cdots \oplus \boldsymbol{\mathcal{V}}_n$$

where each $\boldsymbol{\mathcal{V}}_i = \text{span}\,(\boldsymbol{u}_i, \boldsymbol{w}_i)$, $i = 1, \ldots, m$, is a two-dimensional A-invariant subspace associated with a pair of complex conjugate eigenvalues, and each $\boldsymbol{\mathcal{V}}_i = \text{span}\,(\boldsymbol{v}_i)$, $i = 2m + 1, \ldots, n$, is a one-dimensional simple eigenspace associated with a real eigenvalue.

Example 6.6

Let us find the solution of (6.18) for

$$A = \begin{bmatrix} -1 & -8 & 8 \\ 8 & -1 & -8 \\ 0 & 0 & -1 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

and decompose it into its modes.

The eigenvalues of A are $\lambda_{1,2} = -1 \mp 8i$, $\lambda_3 = -1$, and the associated eigenvectors are

$$\boldsymbol{v}_{1,2} = \begin{bmatrix} \mp i \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Since

$$\boldsymbol{x}_0 = \frac{1}{2i}\boldsymbol{v}_1 - \frac{1}{2i}\boldsymbol{v}_2 + \boldsymbol{v}_3$$

the corresponding solution is

$$\boldsymbol{x} = \frac{1}{2i}\, e^{\lambda_1 t}\boldsymbol{v}_1 - \frac{1}{2i}\, e^{\lambda_2 t}\boldsymbol{v}_2 + e^{\lambda_3 t}\boldsymbol{v}_3$$

We observe that although A and $\boldsymbol{x}_0$ are both real, the modes of the solution associated with the complex eigenvalues are also complex. However, since λ_1 and λ_2 as well as $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are conjugate pairs, then so are the corresponding modes. Hence they add up to a real solution. Indeed, the solution can be expressed as

$$\begin{aligned}
\boldsymbol{x} &= \text{Im}\,\{\, e^{\lambda_1 t}\boldsymbol{v}_1 \,\} + e^{\lambda_3 t}\boldsymbol{v}_3 \\
&= \text{Im}\,\{\, e^{-t}(\cos 8t + i \sin 8t) \begin{bmatrix} i \\ 1 \\ 0 \end{bmatrix} \,\} + e^{-t} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
&= e^{-t}(\, \cos 8t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \sin 8t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}) + e^{-t} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \\
&= \boldsymbol{\phi}_1(t) + \boldsymbol{\phi}_3(t)
\end{aligned}$$

Note that $\boldsymbol{\phi}_1(t)$ lies in the two-dimensional subspace spanned by $\boldsymbol{u}_1 = \text{Re}\,\{\, \boldsymbol{v}_1 \,\}$ and $\boldsymbol{w}_1 = \text{Im}\,\{\, \boldsymbol{v}_1 \,\}$, and $\boldsymbol{\phi}_3(t)$ lies in the one-dimensional subspace spanned by $\boldsymbol{v}_3$. The decomposition of the solution into its real modes is illustrated in Figure 6.2.

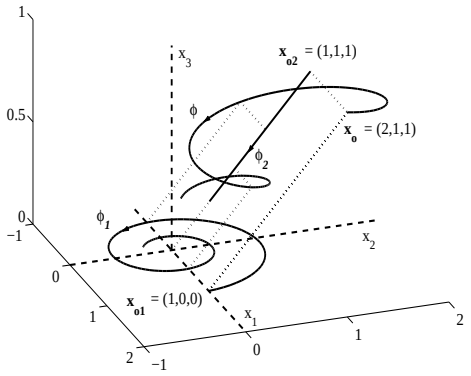

Figure 6.2: Modal decomposition of solution

The analysis of the general case of repeated complex eigenvalues is similar to that of the simple eigenvalue case, and is left as an exercise to the reader. It is worth to mention that when $\nu_i = 1$ for a repeated complex eigenvalue $\lambda_i = \sigma_i + i\omega_i$ with multiplicity n_i, the corresponding real mode consists of

$$e^{\sigma_i t}\cos\omega_i t, e^{\sigma_i t}\sin\omega_i t, \ldots, t^{n_i-1}e^{\sigma_i t}\cos\omega_i t, t^{n_i-1}e^{\sigma_i t}\sin\omega_i t$$

terms, as we illustrate by the following example.

Example 6.7

Let us find the modes of the solution of (6.18) for the A matrix considered in Example 5.18 and for the initial condition $\boldsymbol{x}_0 = \text{col}\,[\,0, 1, 2, 0\,]$.

A complex modal matrix for and the Jordan form of A are already obtained in Example 5.18 as

$$P = [\,\boldsymbol{v}_{11} \quad \boldsymbol{v}_{12} \quad \boldsymbol{v}_{11}^* \quad \boldsymbol{v}_{12}^*\,] = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1+i & 1 & 1-i & 1 \\ 2i & 2+2i & -2i & 2-2i \\ -2+2i & 6i & -2-2i & -6i \end{bmatrix}$$

and

$$J = P^{-1}AP = \begin{bmatrix} J_1 & O \\ O & J_1^* \end{bmatrix} = \begin{bmatrix} 1+i & 1 & 0 & 0 \\ 0 & 1+i & 0 & 0 \\ 0 & 0 & 1-i & 1 \\ 0 & 0 & 0 & 1-i \end{bmatrix}$$

Decomposition of $\boldsymbol{x}_0$ into its components along the columns of P gives

$$\boldsymbol{x}_0 = \text{Re}\,\{\boldsymbol{v}_{12}\} = \frac{1}{2}\boldsymbol{v}_{12} + \frac{1}{2}\boldsymbol{v}_{12}^* = \frac{1}{2}P_1\boldsymbol{e}_2 + \frac{1}{2}P_1^*\boldsymbol{e}_2$$

Then the corresponding solution is

$$\begin{aligned}
\boldsymbol{x} &= \boldsymbol{\phi}(t, \boldsymbol{x}_0) = \frac{1}{2} P_1 e^{J_1 t} \boldsymbol{e}_2 + \frac{1}{2} P_1^* e^{J_1^* t} \boldsymbol{e}_2 = \mathrm{Re}\,\{P_1 e^{J_1 t} \boldsymbol{e}_2\} \\
&= \mathrm{Re}\,\{ \begin{bmatrix} 1 & 0 \\ 1+i & 1 \\ 2i & 2+2i \\ -2+2i & 6i \end{bmatrix} \begin{bmatrix} e^{(1+i)t} & te^{(1+i)t} \\ 0 & e^{(1+i)t} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \} \\
&= \mathrm{Re}\,\{e^t(\cos t + i \sin t)(\begin{bmatrix} 1 \\ 1+i \\ 2i \\ -2+2i \end{bmatrix} t + \begin{bmatrix} 0 \\ 1 \\ 2+2i \\ 6i \end{bmatrix})\} \\
&= e^t \cos t \begin{bmatrix} t \\ t+1 \\ 2 \\ -2t \end{bmatrix} + e^t \sin t \begin{bmatrix} 0 \\ -t \\ -2t-2 \\ -2t-6 \end{bmatrix}
\end{aligned}$$

Note that the solution consists of $e^t \cos t$, $te^t \cos t$, $e^t \sin t$, and $te^t \sin t$ terms whose coefficient vectors span $\mathbb{R}^{4\times 1}$. No matter what the initial condition is, these four terms always appear in the solution, because $\mathbb{R}^{4\times 1}$ has no decomposition into smaller A-invariant subspaces.

6.3 nth Order Linear Differential Equations

Recall that an nth order linear differential equation (LDE) in an unknown function y of an independent variable t is of the form

$$y^{(n)} + a_1(t) y^{(n-1)} + \cdots + a_{n-1}(t) y' + a_n(t) y = u(t) \tag{6.32}$$

where the coefficients $a_i(t)$ and $u(t)$ are given real-valued function defined on some interval $\boldsymbol{I}$.

As we have already considered in Section 2.7, defining new variables

$$x_1 = y\,,\; x_2 = y'\,,\; \ldots\,,\; x_n = y^{(n-1)} \tag{6.33}$$

we obtain an equivalent system of n first order linear differential equations

$$\begin{aligned}
x_1' &= x_2 \\
x_2' &= x_3 \\
&\vdots \\
x_{n-1}' &= x_n \\
x_n' &= -a_n(t)x_1 - a_{n-1}(t)x_2 - \cdots - a_1(t)x_n + u(t)
\end{aligned}$$

which can be written in matrix form as

$$\boldsymbol{x}' = A(t)\boldsymbol{x} + \boldsymbol{b}u(t) \tag{6.34}$$

where

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix},\; A(t) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ -a_n(t) & -a_{n-1}(t) & \cdots & -a_1(t) \end{bmatrix},\; \boldsymbol{b} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

If $y = \phi(t)$ is a solution of (6.32), then

$$\boldsymbol{x} = \boldsymbol{\phi}(t) = \text{col}\,[\,\phi(t), \phi'(t), \ldots, \phi^{(n-1)}(t)\,] \tag{6.35}$$

is a solution of (6.34). Conversely, if

$$\boldsymbol{x} = \boldsymbol{\phi}(t) = \text{col}\,[\,\phi_1(t), \phi_2(t), \ldots, \phi_n(t)\,]$$

is a solution of (6.34), then $y = \phi_1(t)$ is a solution of (6.32), and furthermore, $\phi_1'(t) = \phi_2(t), \ldots, \phi_{n-1}'(t) = \phi_n(t)$. That is, any solution of (6.34) must be of the form in (6.35). This observation allows us to apply what we already know about the solution of SLDE's to obtain solutions and their properties of LDE's.

6.3.1 Homogeneous Linear Differential Equations

Consider an nth order homogeneous LDE

$$y^{(n)} + a_1(t)y^{(n-1)} + \cdots + a_{n-1}(t)y' + a_n(t)y = 0 \tag{6.36}$$

which is transformed into a homogeneous SLDE

$$\boldsymbol{x}' = A(t)\boldsymbol{x} \tag{6.37}$$

by means of the change of variables in (6.33).

Let $X(t)$ be a fundamental matrix of (6.37) consisting of the solutions $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$. Then it must be of the form

$$\begin{aligned} X(t) &= [\,\boldsymbol{\phi}_1(t)\ \boldsymbol{\phi}_2(t)\ \cdots\ \boldsymbol{\phi}_n(t)\,] \\ &= \begin{bmatrix} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \\ \phi_1'(t) & \phi_2'(t) & \cdots & \phi_n'(t) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(n-1)}(t) & \phi_2^{(n-1)}(t) & \cdots & \phi_n^{(n-1)}(t) \end{bmatrix} \end{aligned} \tag{6.38}$$

for some functions $\phi_1, \ldots, \phi_n$, each of which is a solution of (6.36). We claim that $\phi_1, \ldots, \phi_n$ are linearly independent. To prove the claim, let

$$c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n = 0$$

which means

$$c_1\phi_1(t) + c_2\phi_2(t) + \cdots + c_n\phi_n(t) = 0 \quad \text{for all } \ t \in \boldsymbol{I}$$

Differentiating both sides $n-1$ times at any $t \in \boldsymbol{I}$, we get

$$X(t)\boldsymbol{c} = \boldsymbol{0}$$

where

$$\boldsymbol{c} = \text{col}\,[\,c_1, c_2, \ldots, c_n\,]$$

Since $X(t)$ is nonsingular for all $t \in \boldsymbol{I}$, the last equality implies $c_1 = c_2 = \cdots = c_n = 0$, proving linear independence of $\phi_1, \ldots, \phi_n$.

Suppose that ϕ is any solution of (6.36). Then

$$\boldsymbol{\phi} = \text{col}\,[\,\phi, \phi', \ldots, \phi^{(n-1)}\,]$$

is a solution of (6.37) and can be expressed as

$$\boldsymbol{\phi}(t) = X(t)\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}_i(t)$$

for some $\alpha_i \in \mathbb{R}$, where $X(t)$ is the fundamental matrix in (6.38). Considering the first element of $\boldsymbol{\phi}$, we have

$$\phi(t) = \sum_{i=1}^{n} \alpha_i \phi_i(t)$$

This shows that the set of solutions of (6.36) is an n-dimensional subspace of $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ having $\{\,\phi_1, \phi_2, \ldots, \phi_n\,\}$ as a basis, and that a general solution of (6.36) is expressed as

$$y = c_1\phi_1(t) + c_2\phi_2(t) + \cdots + c_n\phi_n(t) \tag{6.39}$$

Example 6.8

Consider the third order differential equation

$$y''' - 2y'' - y' + 2y = 0 \tag{6.40}$$

which is equivalent to

$$\boldsymbol{x}' = A\boldsymbol{x}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & 1 & 2 \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} y \\ y' \\ y'' \end{bmatrix}$$

The coefficient matrix A is in companion form with the characteristic polynomial

$$d(s) = s^3 - 2s^2 - s + 2 = (s-1)(s+1)(s-2)$$

Constructing a modal matrix as

$$P = [\,\boldsymbol{v}(\lambda_1)\ \ \boldsymbol{v}(\lambda_2)\ \ \boldsymbol{v}(\lambda_3)\,] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 1 & 4 \end{bmatrix}$$

we obtain

$$P^{-1}AP = D = \text{diag}\,[\,1, -1, 2\,], \quad A = PDP^{-1}$$

Since e^{At} is a fundamental matrix, then so is

$$\begin{aligned} X(t) &= e^{At}P = Pe^{Dt} \\ &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 1 & 1 & 4 \end{bmatrix} \begin{bmatrix} e^{t} & & \\ & e^{-t} & \\ & & e^{2t} \end{bmatrix} \end{aligned}$$

Therefore,

$$\boldsymbol{\phi}_1(t) = e^t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\phi}_2(t) = e^{-t} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\phi}_3(t) = e^{2t} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

are three linearly independent solutions of the equivalent system. (Note that they are of the form in (6.35).) Their first components $\phi_1(t) = e^t$, $\phi_2(t) = e^{-t}$ and $\phi_3(t) = e^{3t}$ form a basis for the solution space of (6.40). Therefore,

$$y = c_1 e^t + c_2 e^{-t} + c_3 e^{2t}$$

is a general solution of (6.40).

Let $f_1, \ldots, f_n \in \mathcal{C}_{n-1}(\boldsymbol{I}, \mathbb{R})$. The matrix

$$W_{f_1,\ldots,f_n}(t) = \begin{bmatrix} f_1(t) & \cdots & f_n(t) \\ \vdots & & \vdots \\ f_1^{(n-1)}(t) & \cdots & f_n^{(n-1)}(t) \end{bmatrix}$$

is called the ***Wronski matrix*** of $f_1, \ldots, f_n$, and its determinant

$$w_{f_1,\ldots,f_n}(t) = \det\ W_{f_1,\ldots,f_n}(t)$$

is called the ***Wronskian*** of these functions. Since every element of the Wronski matrix is continuous, $w_{f_1,\ldots,f_n} \in \mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$.

Following the argument used in proving linear independence of the basis solutions of (6.36), we can show that

$$c_1 f_1 + \cdots + c_n f_n = 0$$

if and only if

$$W_{f_1,\ldots,f_n}(t)\boldsymbol{c} = \mathbf{0} \quad \text{for all } \ t \in \boldsymbol{I}$$

where $\boldsymbol{c} = \text{col}\,[\,c_1, \ldots, c_n\,]$. This shows that if $w_{f_1,\ldots,f_n}(t_0) \neq 0$ for at least one $t_0 \in \boldsymbol{I}$, then $f_1, \ldots, f_n$ are linearly independent. The converse of this result is not true in general.[4] However, if f_i are solutions of an nth order homogeneous LDE, then the converse is also true. This follows from the fact that if $f_i = \phi_i, i = 1, \ldots, n$, are linearly independent solutions of (6.36), then their Wronski matrix is a fundamental matrix of the equivalent system of first order differential equations in (6.37), and therefore, it is nonsingular for all $t \in \boldsymbol{I}$.

In summary, if $\phi_1, \ldots, \phi_n$ are solutions of (6.36), then either $w_{\phi_1,\ldots,\phi_n}(t) \neq 0$ for all $t \in \boldsymbol{I}$, in which case $\phi_1, \ldots, \phi_n$ are linearly independent, or $w_{\phi_1,\ldots,\phi_n}(t) = 0$ for all $t \in \boldsymbol{I}$, in which case $\phi_1, \ldots, \phi_n$ are linearly dependent. In other words, there is no possibility of having $w_{\phi_1,\ldots,\phi_n}(t_1) \neq 0$ at some t_1 and $w_{\phi_1,\ldots,\phi_n}(t_2) = 0$ at another t_2.

[4] For example, the functions

$$f_1(t) = t^3 \quad \text{and} \quad f_2(t) = |\,t\,|^3$$

are linearly independent on the interval $\boldsymbol{I} = (-1, 1)$, but their Wronskian vanishes identically on $\boldsymbol{I}$, that is, $w_{f_1,f_2}(t) = 0$ for all $-1 < t < 1$ (see Exercise 6.20 for details).

6.3.2 Non-Homogeneous Linear Differential Equations

Since the non-homogeneous LDE in (6.32) is a linear equation involving a linear differential operator, it has a general solution of the form

$$y = \phi_p(t) + \phi_c(t)$$

where $\phi_p(t)$ is a particular solution, and

$$\phi_c(t) = c_1\phi_1(t) + \cdots + c_n\phi_n(t)$$

is the complementary solution consisting of the basis solutions of the associated homogeneous LDE in (6.36).

In Section 6.1.2 we have seen how to obtain a particular solution of a SLDE from its complementary solution by the method of variation of parameters. Since a LDE can always be transformed into a SLDE, we conclude that the method should also be applicable to an nth order LDE. (In Chapter 2, we have already used the method to solve first and second order LDE's.) The details of the method are simple and are summarized below.

Suppose that $\phi_1(t), \ldots, \phi_n(t)$ are linearly independent basis solutions of (6.36). We assume that a particular solution of (6.32) is of the form

$$\phi_p(t) = v_1(t)\phi_1(t) + \cdots + v_n(t)\phi_n(t)$$

and impose the restrictions

$$\begin{bmatrix} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \\ \phi_1'(t) & \phi_2'(t) & \cdots & \phi_n'(t) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(n-1)}(t) & \phi_2^{(n-1)}(t) & \cdots & \phi_n^{(n-1)}(t) \end{bmatrix} \begin{bmatrix} v_1'(t) \\ v_2'(t) \\ \vdots \\ v_n'(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ u(t) \end{bmatrix} \tag{6.41}$$

on the derivatives of v_i's. Since the coefficient matrix above is a fundamental matrix of (6.37), it is nonsingular for all t, and $v_i'(t)$ can be solved uniquely. Integrating each $v_i'(t)$ we get

$$v_i(t) = \int v_i'(t)\, dt = V_i(t) + c_i$$

resulting in a general solution of (6.32)

$$y = \sum_{i=1}^{n} v_i(t)\phi_i(t) = \sum_{i=1}^{n} V_i(t)\phi_i(t) + \sum_{i=1}^{n} c_i\phi_i(t) = \phi_p(t) + \phi_c(t)$$

where $\phi_p(t)$ and $\phi_c(t)$ are a particular and the complementary solutions.

Note that the restrictions in (6.41) are equivalent to (6.12). However, to apply the method of variation of parameters to a LDE we need not transform it into an equivalent SLDE. All we need is a set of basis solutions of the associated homogeneous LDE.

Example 6.9

Consider the LDE

$$y''' - 2y'' - y' + 2y = -12e^t$$

The basis solution of the associated homogeneous LDE are obtained in Example 6.8 as

$$\phi_1(t) = e^t\,, \quad \phi_2(t) = e^{-t}\,, \quad \phi_3(t) = e^{2t}$$

Hence (6.41) becomes

$$\begin{bmatrix} e^t & e^{-t} & e^{2t} \\ e^t & -e^{-t} & 2e^{2t} \\ e^t & e^{-t} & 4e^{2t} \end{bmatrix} \begin{bmatrix} v_1'(t) \\ v_2'(t) \\ v_3'(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -12e^t \end{bmatrix}$$

Solving for v_i' we get

$$v_1'(t) = 6\,, \quad v_2'(t) = -2e^{2t}\,, \quad v_3'(t) = -4e^{-t}$$

and integrating v_i's

$$v_1(t) = 6t + c_1'\,, \quad v_2'(t) = -e^{2t} + c_2'\,, \quad v_3'(t) = 4e^{-t} + c_3'$$

Thus a general solution is obtained as

$$\begin{aligned} y &= (6t + c_1')e^t + (-e^{2t} + c_2')e^{-t} + (4e^{-t} + c_3')e^{2t} \\ &= 6te^t + c_1e^t + c_2e^{-t} + c_3e^{2t} \end{aligned}$$

Example 6.10

Let us find the general solution of

$$y'' - \frac{2}{t}\,y' + \frac{2}{t^2}\,y = -\frac{1}{t}\,, \quad t > 0$$

given that $\phi_1(t) = t$ and $\phi_2(t) = t^2$ are two linearly independent solutions of the associated homogeneous equation.

Writing (6.41) as

$$\begin{bmatrix} t & t^2 \\ 1 & 2t \end{bmatrix} \begin{bmatrix} v_1'(t) \\ v_2'(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -1/t \end{bmatrix}$$

and solving for v_1' and v_2' we obtain

$$v_1'(t) = \frac{1}{t}\,, \quad v_2'(t) = -\frac{1}{t^2}$$

Hence

$$v_1(t) = \ln t + c_1'\,, \quad v_2(t) = \frac{1}{t} + c_2'$$

and a general solution is

$$y = (\ln t + c_1')t + (\frac{1}{t} + c_2')t^2 = t\ln t + c_1t + c_2t^2$$

6.4 Homogeneous LDE With Constant Coefficients

When the coefficients of the LDE (6.36) are constant, it takes the form

$$L(\mathcal{D})(y) = (\mathcal{D}^n + a_1\mathcal{D}^{n-1} + \cdots + a_{n-1}\mathcal{D} + a_n)(y) = 0 \tag{6.42}$$

where the identity operator $\mathcal{I}$ is dropped from the term $a_n\mathcal{I}$ for convenience. In this case, the equivalent SLDE in (6.37) becomes

$$\boldsymbol{x}' = A\boldsymbol{x} \tag{6.43}$$

where A is a constant matrix having the companion form in (5.29). We immediately observe that the characteristic polynomial of A is

$$d(s) = s^n + a_1 s^{n-1} + \cdots + a_{n-1}s + a_n = L(s)$$

Incidentally, $L(s)$ is called the ***characteristic polynomial*** associated with the linear differential operator $L(\mathcal{D})$, and the equation

$$L(s) = 0 \tag{6.44}$$

is called the ***characteristic equation*** of the LDE in (6.42).

Recall from Section 6.2 that if the characteristic polynomial of A is factored as

$$L(s) = d(s) = \prod_{i=1}^{k}(s - \lambda_i)^{n_i} \tag{6.45}$$

then any solution of (6.43) can be decomposed into complex modes as

$$\boldsymbol{x} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} t^{j-1}e^{\lambda_i t}\boldsymbol{\beta}_{ij}$$

for some $\boldsymbol{\beta}_{ij} \in \mathbb{C}^{n\times 1}$. Then, as we discussed in Section 6.3.1, the complex-valued functions

$$\psi_{ij}(t) = t^{j-1}e^{\lambda_i t}\,, \quad i = 1, \ldots, k\,;\ j = 1, \ldots, n_i \tag{6.46}$$

are linearly independent and form a basis for the complex solution space of (6.42).

If λ_i are all real, then so are $\phi_{ij} = \psi_{ij}$, and therefore, they form a basis for the real solution space as well. To find a basis for the real solution space when some λ_i are complex, suppose that $L(s)$ has m distinct pairs of complex conjugate zeros $\lambda_{2i-1,2i} = \sigma_i \mp i\omega_i, i = 1, \ldots, m$, and $k - 2m$ distinct real zeros $\lambda_i, i = 2m+1, \ldots, k$, with multiplicities n_i. Then in addition to the real solutions

$$\phi_{ij}(t) = t^{j-1}e^{\lambda_i t}\,, \quad i = 2m+1, \ldots, k$$

the real and imaginary parts

$$\phi_{2i-1,j}(t) = t^{j-1}e^{\sigma_i t}\cos\omega_i t\,, \quad \phi_{2i,j}(t) = t^{j-1}e^{\sigma_i t}\sin\omega_i t\,, \quad i = 1, \ldots, m$$

of the complex solutions $\psi_{ij} = t^{j-1}e^{\lambda_i t}$ are also real solutions. Since ψ_{ij} are linearly independent then so are ϕ_{ij} (see Example 3.15), and constitute a basis for the real solution space.

The result obtained above can also be reached without reference to the equivalent SLDE in (6.43). Noting that a factorization of $L(s)$ as in (6.45) corresponds to a factorization of $L(\mathcal{D})$ so that (6.42) can formally be written as

$$[\prod_{p=1}^{k}(\mathcal{D}-\lambda_p)^{n_p}](y)=0$$

It can be shown (see Exercise 6.23) that if $p>q$ then

$$(\mathcal{D}-\lambda)^p(t^q e^{\lambda t})=0$$

Thus for $j\le n_i$

$$[\prod_{p=1}^{k}(\mathcal{D}-\lambda_p)^{n_p}](t^{j-1}e^{\lambda_i t})=[\prod_{\substack{p=1\\p\neq i}}^{k}(\mathcal{D}-\lambda_p)^{n_p}](\mathcal{D}-\lambda_i)^{n_i}(t^{j-1}e^{\lambda_i t})=0$$

that is, each ψ_{ij} in (6.46) is a complex solution of (6.42).

Example 6.11

The second order homogeneous LDE

$$y''+2y'+5y=0$$

has the characteristic equation

$$s^2-2s+5=0$$

with a simple pair of complex conjugate roots $\lambda_{1,2}=-1\mp 2i$. Then

$$\phi_1(t)=e^{-t}\cos 2t \quad \text{and} \quad \phi_2(t)=e^{-t}\sin 2t$$

form a basis for the solution space, and a general solution is obtained as

$$y=c_1e^{-t}\cos 2t+c_2e^{-t}\sin 2t$$

Example 6.12

Let us find a general solution of the homogeneous LDE

$$(\mathcal{D}-2)^2(\mathcal{D}^2-2\mathcal{D}+2)^2(y)=0$$

Since the characteristic polynomial $L(s)$ is already in factored form, having a real zero $\lambda_1=2$ with multiplicity $n_1=2$, and a pair of complex conjugate zeros $\lambda_{2,3}=1\mp i$, with multiplicities $n_{2,3}=2$, we write down a general solution as

$$y=c_1e^{2t}+c_2te^{2t}+c_3e^{t}\cos t+c_4te^{t}\cos t+c_5e^{t}\sin t+c_6te^{t}\sin t$$

6.5 The Method of Undetermined Coefficients

Consider a non-homogeneous LDE with constant coefficients

$$L(\mathcal{D})(y) = u(t) \tag{6.47}$$

We already know that once a basis for the solution space of the associated homogeneous LDE is obtained, then a particular solution of (6.47) can be found by the method of variation of parameters. The ***method of undetermined coefficients*** is an alternative and practical method to obtain a particular solution when u is a linear combination of functions of the form

$$u(t) = (p_0 t^r + \cdots + p_r)e^{\sigma t} \tag{6.48}$$

or of the form

$$u(t) = (p_0 t^r + \cdots + p_r)e^{\sigma t}\cos\omega t + (q_0 t^r + \cdots + q_r)e^{\sigma t}\sin\omega t \tag{6.49}$$

where $p_0, \ldots, p_r, q_0, \ldots, q_r, \sigma, \omega$ are real constants. Such functions include polynomials, exponential functions, trigonometric functions, and their products. For example, $u(t) = t^2 - 5$ is of the form of (6.48) with $\sigma = 0$; $u(t) = te^{2t}$ is of the same form with $\sigma = 2$; $u(t) = -2\sin 5t$ is of the form of (6.49) with $\sigma = 0$ and $\omega = 5$; and $u(t) = te^t \cos t - 3e^t \sin t$ is of same form with $\sigma = \omega = 1$.

We first consider the case where $u(t)$ of the form in (6.48). Then it is a solution of a homogeneous LDE

$$L_1(\mathcal{D})u(t) = 0 \tag{6.50}$$

where

$$L_1(\mathcal{D}) = (\mathcal{D} - \sigma)^{r+1} = (\mathcal{D} - \lambda)^{r+1}, \quad \lambda = \sigma \tag{6.51}$$

Suppose that $y = \phi_p(t)$ is a particular solution of (6.47), that is,

$$L(\mathcal{D})\phi_p(t) = u(t)$$

Operating on both sides of this equation with $L_1(\mathcal{D})$ and using (6.50), we get

$$L_1(\mathcal{D})L(\mathcal{D})\phi_p(t) = 0 \tag{6.52}$$

which shows that ϕ_p is a solution of the homogeneous LDE in (6.52).

Let $L(s)$ have a factorization as in (6.45).

If $\lambda = \sigma$ in (6.51) is not a zero of $L(s)$, then it appears as a new zero of $L_1(s)L(s)$, that is, $L_1(s)L(s)$ has a factorization

$$L_1(s)L(s) = (s - \sigma)^{r+1} \prod_{i=1}^{k} (s - \lambda_i)^{n_i}$$

Then ϕ_p must be of the form

$$\phi_p(t) = (A_0 t^r + \cdots + A_r)e^{\sigma t} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} A_{ij}\psi_{ij}(t)$$

where ψ_{ij} are the complex solutions of the associated homogeneous LDE, and are given by (6.46). However, since they can be included in the complementary solution, ϕ_p can be assumed to have the form

$$\phi_p(t) = (A_0 t^r + \cdots + A_r)e^{\sigma t} \tag{6.53}$$

Note that ϕ_p has the same structure as u. Once we decide on the form of ϕ_p, its coefficients $A_0, \ldots, A_r$ can be determined by substituting it into (6.47), and equating the coefficients of the like terms on both sides of the resulting equation.

If $\lambda = \sigma$ in (6.51) is a zero of $L(s)$, that is, if $\sigma = \lambda_p$ for some p, then $L_1(s)L(s)$ has a factorization

$$L_1(s)L(s) = (s-\sigma)^{n_p+r+1} \prod_{\substack{i=1 \\ i\neq p}}^{k} (s-\lambda_i)^{n_i}$$

Then ϕ_p must be of the form

$$\begin{aligned} \phi_p(t) &= (A_0 t^{n_p+r} + \cdots + A_r t^{n_p} + A_{p1} t^{n_p-1} + \ldots + A_{pn_p})e^{\sigma t} \\ &\quad + \sum_{\substack{i=1 \\ i\neq p}}^{k} \sum_{j=1}^{n_i} A_{ij}\psi_{ij}(t) \\ &= t^{n_p}(A_0 t^r + \cdots + A_r)e^{\sigma t} + \sum_{i=1}^{k} \sum_{j=1}^{n_i} A_{ij}\psi_{ij}(t) \end{aligned}$$

As before, ψ_{ij} can be included in the complementary solution, and ϕ_p can be assumed to have the form

$$\phi_p(t) = t^{n_p}(A_0 t^r + \cdots + A_r)e^{\sigma t} \tag{6.54}$$

Comparing with (6.53), we observe that when σ is a root of the characteristic equation with multiplicity $m = n_p$, then the assumed solution in (6.53) is modified by multiplying it with t^m. As before, the coefficients $A_0, \ldots, A_r$ of the assumed particular solution can be determined by substituting it into (6.47), and equating the coefficients of the like terms on both sides of the resulting equation.

We now consider the case where $u(t)$ of the form in (6.49). Then it is a solution of a homogeneous LDE

$$L_2(\mathcal{D})u(t) = 0 \tag{6.55}$$

where

$$L_2(\mathcal{D}) = (\mathcal{D}^2 - 2\sigma\mathcal{D} + \sigma^2 + \omega^2)^{r+1} = (\mathcal{D}-\lambda)^{r+1}(\mathcal{D}-\lambda^*)^{r+1} \tag{6.56}$$

with $\lambda = \sigma + i\omega$. Following the same argument as in the previous case, we can show (see Exercise 6.26) that if $\lambda = \sigma + i\omega$ is not a root of the characteristic equation $L(s) = 0$, then a particular solution of (6.47) is of the form

$$\phi_p(t) = (A_0 t^r + \cdots + A_r)e^{\sigma t}\cos\omega t + (B_0 t^r + \cdots + B_r)e^{\sigma t}\sin\omega t$$

and if $\lambda = \sigma + i\omega$ is a root of the characteristic equation $L(s) = 0$ with multiplicity m, then a particular solution is of the form

$$\phi_p(t) = t^m(A_0t^r + \cdots + A_r)e^{\sigma t}\cos\omega t + t^m(B_0t^r + \cdots + B_r)e^{\sigma t}\sin\omega t$$

In either case, the coefficients $A_0, \ldots, A_r$ and $B_0, \ldots, B_r$ are found by substituting the assumed particular solution into (6.47) and equating the coefficients of the like terms.

Finally, we note that if u is the sum of functions of the form in (6.48) or (6.49) with different σ's and/or ω's, then as a result of the linearity of $L(\mathcal{D})$ the assumed particular solution will be the sum of the corresponding assumed solutions.

Example 6.13

Let us find a general solution of the second order LDE

$$y'' + 2y' + 5y = 16te^t$$

The roots of the characteristic equation have already been obtained in Example 6.11 as $\lambda_{1,2} = -1 \mp 2i$. $u(t) = 16te^t$ is of the form in (6.48) with $\sigma = 1$, which is not a root of the characteristic equation. Consequently, we assume a particular solution of the form

$$\phi_p(t) = (A_0t + A_1)e^t$$

Substituting ϕ_p and its derivatives

$$\phi'(t) = (A_0t + A_0 + A_1)e^t\,, \quad \phi''(t) = (A_0t + 2A_0 + A_1)e^t$$

into the given LDE, cancelling out the nonzero terms e^t on both sides of the equation, and arranging the terms, we get

$$8A_0t + (4A_0 + 8A_1) = 16t$$

Equating the coefficients of the like terms and solving for A_0 and A_1, we obtain $A_0 = 2$ and $A_1 = -1$. Hence

$$\phi_p(t) = (2t - 1)e^t$$

Together with the complementary solution obtained in Example 6.11, we get a general solution as

$$y = \phi_p(t) + \phi_c(t) = (2t - 1)e^t + c_1e^{-t}\cos 2t + c_2e^{-t}\sin 2t$$

Example 6.14

Consider the LDE

$$(\mathcal{D} - 1)^2(\mathcal{D} - 2)(y) = 6te^t$$

$u(t) = 6te^t$ is of the form in (6.48) with $\sigma = 1$, which is a root of the characteristic equation with multiplicty $m = 2$. Consequently, we assume a particular solution of the form

$$\phi_p(t) = t^2(A_0t + A_1)e^t = (A_0t^3 + A_1t^2)e^t$$

Evaluating

$$\begin{aligned}
(\mathcal{D} - 1)\phi_p(t) &= (\mathcal{D} - 1)[(A_0t^3 + A_1t^2)e^t] \\
&= (3A_0t^2 + 2A_1t)e^t \\
(\mathcal{D} - 1)^2\phi_p(t) &= (\mathcal{D} - 1)[(3A_0t^2 + 2A_1t)e^t] \\
&= (6A_0t + 2A_1)e^t
\end{aligned}$$

and

$$\begin{aligned} L(\mathcal{D})\phi_p(t) &= (\mathcal{D}-2)(\mathcal{D}-1)^2\phi_p(t) \\ &= (\mathcal{D}-2)[(6A_0t+2A_1)e^t] \\ &= (-6A_0t+6A_0-2A_1)e^t \quad = \quad 6te^t \end{aligned}$$

we obtain $A_0 = -1$ and $A_1 = -3$. A general solution can then be written as

$$y = t^2(-t-3)e^t + c_1te^t + c_2e^t + c_3e^{2t} = (-t^3-3t^2+c_1t+c_2)e^t + c_3e^{2t}$$

Example 6.15

Let us find a general solution of the LDE

$$y'' + 2y' + 5y = 10\cos t$$

The right-hand side of the LDE is of the form in (6.49) with $\lambda = \sigma + i\omega = i$. Since $\lambda = i$ is not a root of the characteristic equation, we assume a particular solution of the form

$$\phi_p(t) = A\cos t + B\sin t$$

Substituting $\phi_p(t)$ together with

$$\phi_p'(t) = B\cos t - A\sin t\,, \quad \phi_p''(t) = -A\cos t - B\sin t$$

into the given equation and collecting the terms, we obtain

$$(4A+2B)\cos t + (4B-2A)\sin t = 10\cos t$$

and equating the coefficients of the like terms,

$$4A + 2B = 10\,, \quad 4B - 2A = 0$$

Solving for A and B, we get $A = 2$, $B = 1$. Thus a particular solution is obtained as

$$y_p = 2\cos t + \sin t$$

Together with the complementary solution obtained in Example 6.11, a general solution can be written as

$$y = 2\cos t + \sin t + c_1e^{-t}\cos 2t + c_2e^{-t}\sin 2t$$

Let us now consider the same LDE with a different right-hand side:

$$y'' + 2y' + 5y = 16te^{-t}\sin 2t$$

Now, the right-hand side of the LDE is of the form in (6.49) with $\lambda = \sigma + i\omega = -1 + i$, which is a simple root of the characteristic equation. Then we assume a particular solution of the form

$$\phi_p(t) = t(A_0t + A_1)e^{-t}\cos 2t + t(B_0t + B_1)e^{-t}\sin 2t$$

It is left to the reader to show that after evaluating the coefficients of the assumed solution by substitution we get

$$\phi_p(t) = te^{-t}(-2t\cos 2t + \sin 2t)$$

6.6 Exercises

1. Show that

$$Y(t) = \begin{bmatrix} \cos t + \sin t & \sin t - \cos t \\ 2\cos t & 2\sin t \end{bmatrix}$$

is also a fundamental matrix of the SLDE in Example 6.1. How is it related to the fundamental matrix $X(t)$ found in the same example?

2. Find the solution of the initial value problem considered in Example 6.2 for a unit pulse input

$$u(t) = \begin{cases} 1/T\,, & 0 < t < T \\ 0 \quad , & t < 0 \text{ or } t > T \end{cases}$$

and investigate the behavior of solution as $T \to 0$.

3. (a) Show that if $A(t)$ and $A(\tau)$ commute for all $t, \tau \in \boldsymbol{I}$ then a fundamental matrix of (6.4) is given by the formula

$$X(t) = e^{B(t)}\,, \quad B(t) = \int A(t)\,dt$$

Hint: First show that if $A(t)$ and $A(\tau)$ commute then $B(t)$ and $B'(t)$ commute. Use this result to show that

$$\frac{d}{dt}\,B^m(t) = B'(t)B^{m-1}(t) = A(t)B^{m-1}(t)\,, \quad m = 1, 2, \ldots$$

Finally, consider

$$X(t) = e^{B(t)} = \sum_{m=0}^{\infty} \frac{1}{m!}\,B^m(t)$$

(b) Show that if $A(t) = a(t)A$ then $A(t)$ and $A(\tau)$ commute, so that

$$X(t) = e^{b(t)A}\,, \quad b(t) = \int a(t)\,dt$$

4. Find the state transition matrix for

$$A(t) = \begin{bmatrix} -t & 1 \\ -1 & -t \end{bmatrix}$$

Hint: Use the result of Exercise 6.3.

5. Find the solution of

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -4 & -2 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

corresponting to the initial conditions

(a) $x_1(0) = x_2(0) = 1$

(b) $x_1(1) = 1\,, \quad x_2(1) = 0$

6. Consider the SLDE

$$\begin{aligned} x_1' &= (\cos t)x_2 \\ x_2' &= -(\cos t)x_1 + u(t) \end{aligned}$$

(a) Obtain a fundamental matrix. Hint: Use the result of Exercise 6.3.

(b) Find the solution for $x_1(0) = x_2(0) = 0$ and $u(t) = \cos t$.

7. Obtain the solutions of the SLDE's in Exercises 6.5 and 6.6 numerically using the MATLAB function `ode23`. In each case plot the exact and numerical solutions on the same graph.

8. Consider the SLDE

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a(t) & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where $a(t)$ is a periodic, piecewise constant function given as

$$a(t) = \begin{cases} 1\,, & 2k\pi < t < (2k+1)\pi \\ 0\,, & (2k+1)\pi < t < 2(k+1)\pi \end{cases}$$

(a) Calculate $\mathbf{\Phi}(t, 0)$ for $0 \le t \le 4\pi$. Hint: Note that

$$A(t) = \begin{cases} A_1\,, & 0 < t < \pi \quad \text{and} \quad 2\pi < t < 3\pi \\ A_2\,, & \pi < t < 2\pi \quad \text{and} \quad 3\pi < t < 4\pi \end{cases}$$

where

$$A_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Hence, $\mathbf{\Phi}(t, 0) = e^{A_1 t}$ for $0 \le t \le \pi$, $\mathbf{\Phi}(t, 0) = \mathbf{\Phi}(t, \pi)\mathbf{\Phi}(\pi, 0) = e^{A_2(t-\pi)}e^{A_1 \pi}$ for $\pi \le t \le 2\pi$, etc.

(b) Calculate and plot $x_1(t)$ and $x_2(t)$ corresponding to $x_1(0) = 1$ and $x_2(0) = 0$ for $0 \le t \le 4\pi$.

9. Transform the second-order LDE

$$y'' + 2ty' + (t^2 + 2)y = 0\,, \quad y(0) = y_0\,, \quad y'(0) = y_1$$

into a system of two first order linear differential equations by defining $x_1(t) = y(t), x_2(t) = y'(t) + ty(t)$.

10. (a) Let A and C be $n \times n$ real matrices. Show that the solution of the matrix differential equation

$$X'(t) = AX(t) + X(t)C\,, \quad X(0) = X_0$$

is

$$X = e^{At} X_0 e^{Ct}$$

(b) Suppose that $C = A^t$ and $X_0 = \boldsymbol{v}_i \boldsymbol{v}_j^t$, where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are eigenvectors of A associated with the eigenvalues λ_i and λ_j. Find a simple expression for the solution.

11. It is given that $\boldsymbol{x} = \text{col}\,[\cos t + \sin t\,,\ \cos t + 3\sin t\,]$ is a solution of a homogeneous SLDE $\boldsymbol{x}' = A\boldsymbol{x}$.

(a) Find A.

(b) Find the solution corresponding to $x_0 = \text{col}\,[\,1\,,\ 2\,]$.

12. Consider the SLDE in (6.18), where

$$A = \begin{bmatrix} 0 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & -2 & -3 \end{bmatrix}$$

(a) Find a modal matrix and the Jordan form of A.

(b) Find the solution corresponding to $\boldsymbol{x}_0 = \text{col}\,[\,1, 0, -1\,]$ and decompose it into its modes.

(c) Find an initial condition $\boldsymbol{x}_0 \neq \boldsymbol{0}$ such that $x_1 = \phi_1(t, \boldsymbol{x}_0) = 0$ for all t, where x_1 is the first component of $\boldsymbol{x}$.

13. Obtain the modal decomposition of the solution of the SLDE

$$\boldsymbol{x}' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -5 & -4 \end{bmatrix} \boldsymbol{x}\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

corresponding to

(a) $\boldsymbol{x}_0 = \text{col}\,[\,0, -1, 3\,]$

(b) $\boldsymbol{x}_0 = \text{col}\,[\,1, -1, 2\,]$

14. Obtain the modal decomposition of the solution of the SLDE

$$\boldsymbol{x}' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -5 & -7 & -3 \end{bmatrix} \boldsymbol{x}\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

corresponding to

(a) $\boldsymbol{x}_0 = \text{col}\,[\,1, 1, -3\,]$

(b) $\boldsymbol{x}_0 = \text{col}\,[\,0, 0, 2\,]$

15. Decompose the solution of (6.18) for

$$A = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

both into complex and also into real modes.

16. Show that if the $t^q e^{\lambda_i t}$ term appears in a mode

$$\boldsymbol{\phi}(t; \boldsymbol{x}_{ij}) = P_{ij} e^{J_{ij} t} \boldsymbol{\alpha}_{ij}$$

of a SLDE, then all the $t^m e^{\lambda_i t}, m < q$, terms must also appear. Hint: Rewrite the expression for $\boldsymbol{\phi}(t, \boldsymbol{x}_{ij})$ above by partitioning P_{ij} into its columns and using the expression (6.29) for $e^{J_{ij} t}$, and show that $t^q e^{\lambda_i t}$ appears in the expression if and only if $\boldsymbol{\alpha}_{ij}$ contains a nonzero element in at least one of the positions $q + 1, \ldots, n_{ij}$.

17. Let $\boldsymbol{\mathcal{V}} \subset \mathbb{R}^{n \times 1}$ be an m-dimensional A-invariant subspace. Show that if $\boldsymbol{x}_0 \in \boldsymbol{\mathcal{V}}$, then $\boldsymbol{\phi}(t; \boldsymbol{x}_0) \in \boldsymbol{\mathcal{V}}$ for all t, where $\boldsymbol{\phi}$ denotes the solution of (6.18). Hint: Let columns of the $n \times m$ matrix V form a basis for $\boldsymbol{\mathcal{V}}$. Since $\boldsymbol{\mathcal{V}}$ is A-invariant, we have $AV = VF$ for some $n \times n$ matrix F. Also, if $\boldsymbol{x}_0 \in \boldsymbol{\mathcal{V}}$ then $\boldsymbol{x}_0 = V\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^{m \times 1}$.

18. Rewrite each of the following LDE's as an equivalent SLDE by defining new dependent variables as in (6.33).

(a) $y''' - 2y'' - y' + 2y = e^t\,, \quad y(0) = 1\,, \quad y'(0) = 0$

(b) $(\mathcal{D}^2 - 4\mathcal{D} + 5)(\mathcal{D} - 1)(\mathcal{D} - 2)(y) = u(t)$

19. Show that the following sets of functions are linearly independent.

(a) $\{\,1, t, \ldots, t^n\,\}$

(b) $\{\,te^t, \cos t, \sin t\,\}$

20. (a) Show that the functions $f_1(t) = t^3$ and $f_2(t) = |\,t\,|^3$ are linearly independent on $\boldsymbol{I} = (-1, 1)$. Hint: Suppose that

$$c_1 f_1(t) + c_2 f_2(t) = 0$$

for all $-1 < t < 1$. Evaluate the expression at $t_1 = -0.5$ and $t_2 = 0.5$

(b) Show that the Wronskian of f_1 and f_2 vanishes identically on $(-1, 1)$. Hint:

$$\frac{d}{dt}\,|\,t\,|^3 = 3t|\,t\,|$$

21. Show that if ϕ_1 and ϕ_2 are solutions of a second order homogeneous differential equation

$$y'' + a_1(t)y' + a_2(t)y = 0$$

then their Wronskian $w(t) = w_{\phi_1,\phi_2}(t)$ satisfies the first order homogeneous differential equation

$$w' + a_1(t)w = 0$$

Explain why this implies that either $w(t) = 0$ for all t, or $w(t) \neq 0$ for all t.

22. Solve the following initial value problems

(a) $y'' + y' - 2y = 0\,, \quad y(0) = 3\,, \quad y'(0) = 0$

(b) $y'' - 2y' + y = 0\,, \quad y(0) = 0\,, \quad y'(0) = 3$

(c) $t^2y'' - t(t+2)y' + (t+2)y = 0\,, \quad y(1) = 1\,, \quad y'(1) = 2$
Hint: $\phi_1(t) = t$ is a solution.

(d) $(\mathcal{D}^2 + 4\mathcal{D} + 13)(\mathcal{D} + 1)(y) = 0\,, \quad y(0) = 1\,, \quad y'(0) = 5\,, \quad y''(0) = 1$

23. Show that if $p > q$, then

(a) $\mathcal{D}^p(t^q) = 0$

(b) $(\mathcal{D} - \lambda)^p(t^q e^{\lambda t}) = 0$
Hint: First prove the result for $p = q + 1$ by using induction on q.

24. Find general solutions of the following LDE's.

(a) $y'' + y' - 2y = 2t^3$

(b) $y'' + y' - 2y = 4e^{-t}$

(c) $y'' + y' - 2y = (6t - 4)e^t$

(d) $y'' + y' - 2y = \sin t$

25. Determine the form of a particular solution of

$$(\mathcal{D} - 1)^2(\mathcal{D}^2 - 2\mathcal{D} + 2)(y) = u(t)$$

for each of the following.

(a) $u(t) = t^2 - te^{2t} + \cos t$

(b) $u(t) = te^t + e^t \cos 3t$

(c) $u(t) = (t^2 - 1)e^t \sin t$

26. Show that

(a) if $\lambda = \sigma + i\omega$ is not a root of the characteristic equation $L(s) = 0$, then a particular solution of (6.47) is of the form

$$\phi_p(t) = (A_0t^r + \cdots + A_r)e^{\sigma t}\cos\omega t + (B_0t^r + \cdots + B_r)e^{\sigma t}\sin\omega t$$

(b) if $\lambda = \sigma + i\omega$ is a root of the characteristic equation $L(s) = 0$ with multiplicity m, then the particular solution in part (a) must be modified by multiplying it with t^m.

Chapter 7
Normed and Inner Product Spaces

7.1 Normed Vector Spaces

Recall that the length of a vector $\boldsymbol{x} = (x_1, x_2, x_3)$ in $\mathbb{R}^3$ is

$$\| \boldsymbol{x} \| = \sqrt{|x_1|^2 + |x_2|^2 + |x_3|^2}$$

Norm is a generalization of the concept of length to vectors of abstract spaces.

7.1.1 Vector Norms

Let $\boldsymbol{\mathcal{X}}$ be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. A function which associates with every vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ a real value denoted $\| \boldsymbol{x} \|$ is called a ***norm*** on $\boldsymbol{\mathcal{X}}$ if it satisfies the following.

N1. $\| \boldsymbol{x} \| > 0$ for any $\boldsymbol{x} \neq \boldsymbol{0}$.

N2. $\| c\boldsymbol{x} \| = | c | \, \| \boldsymbol{x} \|$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $c \in \mathbb{F}$.

N3. $\| \boldsymbol{x} + \boldsymbol{y} \| \leq \| \boldsymbol{x} \| + \| \boldsymbol{y} \|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$.

A vector space with a norm defined on it is called a ***normed vector space***.

Note that property N2 implies that $\| \boldsymbol{0} \| = 0$. Property N3 is known as the ***triangle inequality***.

If $\boldsymbol{x}$ is a nonzero vector in $\boldsymbol{\mathcal{X}}$, then $\frac{1}{\| \boldsymbol{x} \|} \boldsymbol{x}$ has unity norm, and is called a ***unit vector***.

Example 7.1

A simple norm on $\mathbb{R}^n$ is

$$\| \boldsymbol{x} \|_1 = |x_1| + \cdots + |x_n|$$

which is called the ***uniform norm***. Obviously, it satisfies properties N1 and N2, and N3 follows from the property of the absolute value that $|a + b| \leq |a| + |b|$ for $a, b \in \mathbb{R}$.

In fact, for any real number $p \geq 1$

$$\| (x_1, \ldots, x_n) \|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p} \tag{7.1}$$

is a norm on $\mathbb{R}^n$, which reduces to the uniform norm for $p = 1$. Again, properties N1 and N2 are satisfied trivially. Proof of the triangle inequality for $p > 1$ is left to the reader (see Exercises 7.3 and 7.4).

In particular,

$$\| \boldsymbol{x} \|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$$

is a norm, called the ***Euclidean norm***. The Euclidean norm is a straightforward generalization of length in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Letting $p \to \infty$, we observe that

$$\| \boldsymbol{x} \|_\infty = \max \{ \, |x_1|, \ldots, |x_n| \, \}$$

is also a norm on $\mathbb{R}^n$, called the ***infinity norm***.

As an illustration, if $\boldsymbol{x} = (4, -12, 3)$ then

$$\begin{aligned} \| \boldsymbol{x} \|_1 &= 4 + 12 + 3 &&= 19 \\ \| \boldsymbol{x} \|_2 &= \sqrt{16 + 144 + 9} &&= 13 \\ \| \boldsymbol{x} \|_\infty &= \max \{4, 12, 3\} &&= 12 \end{aligned}$$

Corresponding norms on $\mathbb{C}^n$, $\mathbb{R}^{n \times 1}$ and $\mathbb{C}^{n \times 1}$ are defined similarly.

Example 7.2

Recall from Example 3.4 that a real n-tuple $(x_1, x_2, \ldots, x_n)$ can be viewed as a function $f : \mathbb{N}_n \to \mathbb{R}$ such that

$$f[k] = x_k \, , \quad k \in \mathbb{N}_n \tag{7.2}$$

Hence for any $p \geq 1$

$$\| f \| = (\sum_{k=1}^{n} |f(k)|^p \,)^{1/p}$$

defines a norm on the function space $\mathcal{F}(\mathbb{N}_n, \mathbb{R})$.

Now consider the function space $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$ of real-valued continuous functions defined on a closed interval $\boldsymbol{I} = [\, a, b \,]$. Replacing the summation in (7.2) with an integral, we observe that for any $p \geq 1$

$$\| f \|_p = (\int_a^b |f(t)|^p \, dt \,)^{1/p} \tag{7.3}$$

is a norm on $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$ (see Exercise 7.5).[1] In particular,

$$\begin{aligned} \| f \|_1 &= \int_a^b |f(t)| \, dt \\ \| f \|_2 &= (\int_a^b |f(t)|^2 \, dt)^{1/2} \\ \| f \|_\infty &= \max_{a \leq t \leq b} \{ \, |f(t)| \, \} \end{aligned}$$

are norms on $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$, which are also called the uniform, Euclidean and infinity norms, respectively.

[1] The reason for restricting our attention to $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$ rather than $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ is to guarantee the convergence of the integral in (7.3).

As an illustration, if

$$f(t) = \sin t\,, \quad -\pi \le t \le \pi$$

then

$$\begin{aligned} \| f \|_1 &= \int_{-\pi}^{\pi} |\sin t|\, dt = 2\int_0^{\pi} \sin t\, dt = 4 \\ \| f \|_2 &= (\int_{-\pi}^{\pi} \sin^2 t\, dt)^{1/2} = (\frac{1}{2}\int_{-\pi}^{\pi} (1-\cos 2t)\, dt\,)^{1/2} = \sqrt{\pi} \\ \| f \|_\infty &= \max_{-\pi\le t\le\pi} \{\, |\sin t| \,\} = 1 \end{aligned}$$

Example 7.3

Consider the vector space $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n\times 1})$, where $\boldsymbol{I} = [\,a, b\,]$ is a closed interval. Recall from Example 3.5 that a vector-valued function $\boldsymbol{f} \in \mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n\times 1})$ can be viewed as a stack of scalar functions $f_1, \ldots, f_n$ such that $\boldsymbol{f}(t) = \text{col}\,[\,f_1(t), \ldots, f_n(t)\,]$ for every $t \in \boldsymbol{I}$.

Let

$$\| \boldsymbol{f} \| = \max_{t\in\boldsymbol{I}} \{ \sum_{i=1}^{n} |\, f_i(t)\,| \} \tag{7.4}$$

Then $\| \cdot \|$ trivially satisfies properties N1 and N2 of a norm. Since

$$\begin{aligned} \| \boldsymbol{f} + \boldsymbol{g} \| &= \max_{t\in\boldsymbol{I}} \{ \sum_{i=1}^{n} |\, f_i(t) + g_i(t)\,| \} \\ &\le \max_{t\in\boldsymbol{I}} \{ \sum_{i=1}^{n} |\, f_i(t)\,| + \sum_{i=1}^{n} |\, g_i(t)\,| \} \\ &\le \max_{t\in\boldsymbol{I}} \{ \sum_{i=1}^{n} |\, f_i(t)\,| \} + \max_{t\in\boldsymbol{I}} \{ \sum_{i=1}^{n} |\, g_i(t)\,| \} \;=\; \| \boldsymbol{f} \| + \| \boldsymbol{g} \| \end{aligned}$$

it also satisfies property N3. Hence it is a norm on $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n\times 1})$.

The summation in (7.4) is the uniform norm of the vector $\boldsymbol{f}(t) \in \mathbb{R}^{n\times 1}$. Let us denote its value by $\nu^{\boldsymbol{f}}(t)$ to indicate its dependence on $\boldsymbol{f}$ and t:

$$\nu^{\boldsymbol{f}}(t) = \| \boldsymbol{f}(t) \|_1 \quad \text{for all} \;\; t \in \boldsymbol{I} \tag{7.5}$$

(7.5) defines a scalar continuous function $\nu^{\boldsymbol{f}} \in \mathcal{C}_0(\boldsymbol{I}, \mathbb{R})$. The maximum value of $\nu^{\boldsymbol{f}}$ on $\boldsymbol{I}$ is its infinity norm. Thus (7.4) can be rewritten as

$$\| \boldsymbol{f} \| = \| \nu^{\boldsymbol{f}} \|_\infty \tag{7.6}$$

The norms in (7.5) and (7.6) bear no special significance in defining $\| \boldsymbol{f} \|$, and they can be replaced with arbitrary norms. By letting

$$\nu_p^{\boldsymbol{f}}(t) = \| \boldsymbol{f}(t) \|_p \quad \text{for all} \;\; t \in \boldsymbol{I} \tag{7.7}$$

and

$$\| \boldsymbol{f} \|_{p,q} = \| \nu_p^{\boldsymbol{f}} \|_q \tag{7.8}$$

for arbitrary $p, q \ge 1$, we can define many different norms on $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n\times 1})$. For details, the reader is referred to Exercises 7.7 and 7.8.

7.1.2 Matrix Norms

Since $\mathcal{C}^{m\times n}$ is a vector space, we may attempt to define a norm for matrices. For example, it is rather easy to show that

$$\| A \|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} \tag{7.9}$$

is a matrix norm, called the ***Frobenius norm***. For $m = 1$ or $n = 1$, it reduces to the Euclidean norm. In fact, Frobenius norm of an $m \times n$ matrix A is the Euclidean norm of an $mn \times 1$ column vector formed by stacking the columns of A. In addition to the properties of a norm, the Frobenius norm also satisfies the consistency condition

$$\| AB \|_F \leq \| A \|_F \| B \|_F$$

which is useful and often desired in matrix operations.

For any $p \geq 1$, let

$$\| A \|_p = \max_{\boldsymbol{x}\neq\mathbf{0}} \{ \frac{\| A\boldsymbol{x} \|_p}{\| \boldsymbol{x} \|_p} \} = \max_{\| \boldsymbol{x} \|_p=1} \{ \| A\boldsymbol{x} \|_p \} \tag{7.10}$$

Then we have the following properties of $\| \cdot \|_p$.

a) Since $\| A\boldsymbol{x} \|_p \geq 0$ and $\| \boldsymbol{x} \|_p > 0$ for $\boldsymbol{x} \neq \mathbf{0}$, $\| A \|_p \geq 0$. If $A \neq O$, then there exists $\boldsymbol{x} \neq \mathbf{0}_{n\times 1}$ such that $A\boldsymbol{x} \neq \mathbf{0}_{m\times 1}$, so that

$$0 < \| A\boldsymbol{x} \|_p \leq \| A \|_p \| \boldsymbol{x} \|_p$$

that is $\| A \|_p > 0$.

b) For any scalar c

$$\| cA \|_p = \max_{\| \boldsymbol{x} \|_p=1} \{ | c | \| A\boldsymbol{x} \|_p \} = | c | \max_{\| \boldsymbol{x} \|_p=1} \{ \| A\boldsymbol{x} \|_p \} = | c | \| A \|_p$$

c) Since $\| (A + B)\boldsymbol{x} \|_p \leq \| A\boldsymbol{x} \|_p + \| B\boldsymbol{x} \|_p$, we have

$$\begin{aligned} \| A + B \|_p &= \max_{\| \boldsymbol{x} \|_p=1} \{ \| A\boldsymbol{x} + B\boldsymbol{x} \|_p \} \\ &\leq \max_{\| \boldsymbol{x} \|_p=1} \{ \| A\boldsymbol{x} \|_p + \| B\boldsymbol{x} \|_p \} \\ &\leq \max_{\| \boldsymbol{x} \|_p=1} \{ \| A\boldsymbol{x} \|_p \} + \max_{\| \boldsymbol{x} \|_p=1} \{ \| B\boldsymbol{x} \|_p \} \\ &= \| A \|_p + \| B \|_p \end{aligned}$$

Hence $\| \cdot \|_p$ is a norm on $\mathcal{C}^{m\times n}$, called the ***matrix norm subordinate to the p-vector norm***.[2]

[2] In this definition, a matrix is interpreted as a mapping between two vector spaces rather than a vector: For $\boldsymbol{x} \neq \mathbf{0}$, the ratio $\| A\boldsymbol{x} \|_p / \| \boldsymbol{x} \|_p$ is the factor by which the strength $\| \boldsymbol{x} \|_p$ of the vector $\boldsymbol{x}$ (as measured by its p-norm) changes while undergoing the transformation represented by the matrix A. Hence $\| A \|_p$ represents the maximum possible change in the strength of a vector $\boldsymbol{x}$ when it is transformed into $A\boldsymbol{x}$. This interpretation of the norm of a matrix can be generalized to linear transformations between arbitrary normed vector spaces (see Exercise 7.13).

All matrix p-norms satisfy the consistency condition. This follows from the fact that

$$\| AB\boldsymbol{x} \|_p \leq \| A \|_p \| B\boldsymbol{x} \|_p \leq \| A \|_p \| B \|_p \| \boldsymbol{x} \|_p$$

so that

$$\| AB \|_p = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \{ \frac{\| AB\boldsymbol{x} \|_p}{\| \boldsymbol{x} \|_p} \} \leq \| A \|_p \| B \|_p$$

It is left to the reader (see Exercise 7.11) to show that the matrix norm subordinate to the uniform vector norm is the maximum column sum

$$\| A \|_1 = \max_{1 \leq j \leq n} \{ \sum_{i=1}^{m} |a_{ij}| \} = \max_{1 \leq j \leq n} \{ \| \boldsymbol{a}_j \|_1 \}$$

where $\boldsymbol{a}_j$ denotes the jth column of A, and the matrix norm subordinate to the infinity vector norm is the maximum row sum

$$\| A \|_\infty = \max_{1 \leq i \leq m} \{ \sum_{j=1}^{n} |a_{ij}| \} = \max_{1 \leq i \leq m} \{ \| \boldsymbol{\alpha}_i \|_1 \}$$

where $\boldsymbol{\alpha}_i$ denotes the ith row of A.[3] Note that

$$\| A^h \|_\infty = \| A \|_1 , \quad \| A^h \|_1 = \| A \|_\infty$$

We shall consider the matrix norm subordinate to the Euclidean vector norm in Chapter 8.

MATLAB provides a built-in function to compute the uniform, Euclidean, infinity and Frobenius norms of vectors and matrices. The function `norm(X,p)` returns the p-norm of X, where $p = 1, 2, \infty$ for p-norms or 'fro' for Frobenius norm.

Example 7.4

Let

$$A = \begin{bmatrix} 2 & 3 & 5 \\ 3 & 4 & 1 \end{bmatrix}$$

Then

$$\begin{array}{lcll} \| A \|_F & = & \sqrt{4+9+25+9+16+1} & = & 8 \\ \| A \|_1 & = & \max \{ (2+3), (3+4), (5+1) \} & = & 7 \\ \| A \|_\infty & = & \max \{ (2+3+5), (3+4+1) \} & = & 10 \end{array}$$

The reader should get the same answers by MATLAB.

[3] These definitions are valid for $m \times n$ matrices with $m \geq 2$. For a row matrix $\boldsymbol{\alpha} = [\, a_1, \ \ldots \ , a_n \,]$, the definition would result in inconsistent identities $\| \boldsymbol{\alpha} \|_\infty = \| \boldsymbol{\alpha} \|_1$ and $\| \boldsymbol{\alpha} \|_1 = \| \boldsymbol{\alpha} \|_\infty$ where $\| \cdot \|$ denotes the vector norm on the right and the subordinate matrix norm on the left.

7.2 Inner Product Spaces

Let $\boldsymbol{\mathcal{X}}$ be a vector space over $\mathbb{F}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. A function which associates with every pair of vectors $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$ a complex scalar denoted $\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle$ is called an ***inner product*** on $\boldsymbol{\mathcal{X}}$ if it satisfies the following.

I1. $\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and $\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.

I2. $\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle^* = \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$.

I3. $\langle \boldsymbol{x} \,|\, a\boldsymbol{y} + b\boldsymbol{z} \rangle = a \langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle + b \langle \boldsymbol{x} \,|\, \boldsymbol{z} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{\mathcal{X}}$ and $a, b \in \mathbb{F}$

A vector space with an inner product defined on it is called an ***inner product space***. If $\boldsymbol{\mathcal{X}}$ is a real vector space, then property I2 reduces to $\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle$.

The following properties of inner product are immediate consequences of the definition.

a) $\langle \boldsymbol{x} \,|\, \boldsymbol{0} \rangle = \langle \boldsymbol{0} \,|\, \boldsymbol{x} \rangle$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

b) $\langle a\boldsymbol{x} + b\boldsymbol{y} \,|\, \boldsymbol{z} \rangle = a^* \langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle + b^* \langle \boldsymbol{x} \,|\, \boldsymbol{z} \rangle$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{\mathcal{X}}$, and $a, b \in \mathbb{F}$

Example 7.5

The standard inner product on $\mathbb{C}^{n \times 1}$ is defined as

$$\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = \boldsymbol{x}^h \boldsymbol{y} = \sum_{i=1}^{n} x_i^* y_i$$

Clearly,

$$\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle = \sum_{i=1}^{n} |x_i|^2 \geq 0$$

and $\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle = 0$ if and only if $x_i = 0$ for all i, or equivalently, $\boldsymbol{x} = \boldsymbol{0}$. Also

$$\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle^* = (\boldsymbol{x}^h \boldsymbol{y})^h = \boldsymbol{y}^h \boldsymbol{x} = \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle$$

and

$$\langle \boldsymbol{x} \,|\, a\boldsymbol{y} + b\boldsymbol{z} \rangle = \boldsymbol{x}^h (a\boldsymbol{y} + b\boldsymbol{z}) = a\, \boldsymbol{x}^h \boldsymbol{y} + b\, \boldsymbol{x}^h \boldsymbol{z} = a \langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle + b \langle \boldsymbol{x} \,|\, \boldsymbol{z} \rangle$$

Similarly, the standard inner product for $\mathbb{R}^{n \times 1}$ is

$$\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = \boldsymbol{x}^t \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

Another common example is the standard inner product

$$\langle f \,|\, g \rangle = \int_a^b f^*(t) g(t)\, dt \tag{7.11}$$

defined on $\mathcal{C}_0([a, b], \mathbb{C})$. All three properties of the inner product are immediate consequences of the properties of the definite integral.

The following theorem states an important property of inner product.

Theorem 7.1 (The Schwarz Inequality) *In an inner product space* $\boldsymbol{\mathcal{X}}$

$$|\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle| \leq \sqrt{\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle}\,\sqrt{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$*, where equality holds if and only if* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are linearly dependent.*

Proof If $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly independent, then $\boldsymbol{y} \neq \boldsymbol{0}$ and $\boldsymbol{x} - c\boldsymbol{y} \neq \boldsymbol{0}$ for any $c \in \mathbb{F}$. Then

$$0 < \langle \boldsymbol{x} - c\boldsymbol{y} \,|\, \boldsymbol{x} - c\boldsymbol{y} \rangle = \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle - c\,\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle - c^*\langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle + |c|^2\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle$$

With $c = \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle / \langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle$, we get

$$0 < \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle - \frac{|\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle|^2}{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle}$$

from which the Schwarz inequality follows.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent, then either $\boldsymbol{y} = \boldsymbol{0}$, in which case we have

$$|\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle| = 0 = \sqrt{\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle}\,\sqrt{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle}$$

or $\boldsymbol{x} = c\boldsymbol{y}$ for some $c \in \mathbb{F}$ so that

$$|\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle| = |c|\,|\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle| = |c|\,\sqrt{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle}\,\sqrt{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle} = \sqrt{\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle}\,\sqrt{\langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle}$$

An important consequence of the Schwarz inequality is that if $\boldsymbol{\mathcal{X}}$ is an inner product space then

$$\|\, \boldsymbol{x} \,\| = \sqrt{\langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle}$$

is a norm on $\boldsymbol{\mathcal{X}}$. Properties N1 and N2 follow immediately from the definition of inner product, and property N3 follows by taking the square root of both sides of the inequality

$$\begin{aligned}
\|\, \boldsymbol{x} + \boldsymbol{y} \,\|^2 &= \langle \boldsymbol{x} + \boldsymbol{y} \,|\, \boldsymbol{x} + \boldsymbol{y} \rangle = \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle + \langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle + \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle + \langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle \\
&= \|\, \boldsymbol{x} \,\|^2 + 2\,\mathrm{Re}\,\{\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle\} + \|\, \boldsymbol{y} \,\|^2 \\
&\leq \|\, \boldsymbol{x} \,\|^2 + 2\,|\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle| + \|\, \boldsymbol{y} \,\|^2 \\
&\leq \|\, \boldsymbol{x} \,\|^2 + 2\,\|\, \boldsymbol{x} \,\|\,\|\, \boldsymbol{y} \,\| + \|\, \boldsymbol{y} \,\|^2 \\
&= (\,\|\, \boldsymbol{x} \,\| + \|\, \boldsymbol{y} \,\|\,)^2
\end{aligned}$$

Example 7.6

The norm defined by the standard inner product in $\mathbb{R}^{n\times 1}$ or $\mathbb{C}^{n\times 1}$ is the Euclidean norm

$$\|\, \boldsymbol{x} \,\|_2 = \sqrt{\boldsymbol{x}^h \boldsymbol{x}} = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$$

Similarly, the standard inner product in $\mathcal{C}_0([\,a, b\,], \mathbb{C})$ defines the Euclidean norm

$$\|\, f \,\|_2 = (\int_a^b |\, f(t) \,|^2\, dt\,)^{1/2}$$

Example 7.7

Let for $A, B \in \mathbb{C}^{m\times n}$

$$\langle\, A \,|\, B \,\rangle = \mathrm{tr}\,(A^h B) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^* b_{ij}$$

Then

$$\langle A \,|\, A \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^* a_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} |\, a_{ij} \,|^2 \geq 0$$

and $\langle A \,|\, A \rangle = 0$ if and only if $A = O$. Thus property I1 of inner product is satisfied. Properties I2 and I3 are obvious from the definition. Hence $\langle A \,|\, B \rangle = \text{tr}\,(A^h B)$ is an inner product on $\mathbb{C}^{m \times n}$.

The norm

$$\| A \| = \sqrt{\text{tr}\,(A^h A)} = (\sum_{i=1}^{m} \sum_{j=1}^{n} |\, a_{ij} \,|^2)^{1/2}$$

defined by this inner product is nothing but the Frobenius norm defined previously.

7.3 Orthogonality

Let $\boldsymbol{\mathcal{X}}$ be an inner product space. Two vectors $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{\mathcal{X}}$ are said to be ***orthogonal***, denoted $\boldsymbol{x} \perp \boldsymbol{y}$, if $\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = 0$. A vector $\boldsymbol{x}$ is said to be orthogonal to a set of vectors $\boldsymbol{R}$, denoted $\boldsymbol{x} \perp \boldsymbol{R}$, if it is orthogonal to every $\boldsymbol{r} \in \boldsymbol{R}$. Two sets $\boldsymbol{R}$ and $\boldsymbol{S}$ are said to be orthogonal, denoted $\boldsymbol{R} \perp \boldsymbol{S}$, if $\boldsymbol{r} \perp \boldsymbol{s}$ for all $\boldsymbol{r} \in \boldsymbol{R}$ and $\boldsymbol{s} \in \boldsymbol{S}$. A finite set $\boldsymbol{R} = \{ \boldsymbol{r}_1, \cdots, \boldsymbol{r}_k \}$ is said to be orthogonal if $\boldsymbol{r}_i \neq \boldsymbol{0}$ for all i and $\boldsymbol{r}_i \perp \boldsymbol{r}_j$ for all $i \neq j$. If, in addition, $\| \boldsymbol{r}_i \| = \langle \boldsymbol{r}_i \,|\, \boldsymbol{r}_i \rangle^{1/2} = 1$ for all i, then $\boldsymbol{R}$ is said to be an ***orthonormal*** set.

We have the following properties concerning orthogonality.

a) $\boldsymbol{0}$ is orthogonal to every vector in $\boldsymbol{\mathcal{X}}$, and it is the only vector that is orthogonal to every vector in $\boldsymbol{\mathcal{X}}$.

b) If $\boldsymbol{x} \perp \boldsymbol{y}$ then $\| \boldsymbol{x} + \boldsymbol{y} \|^2 = \| \boldsymbol{x} \|^2 + \| \boldsymbol{y} \|^2$.

c) If $\boldsymbol{x} \perp \boldsymbol{R}$ then $\boldsymbol{x} \perp \text{span}\,(\boldsymbol{R})$.

That $\boldsymbol{0}$ is orthogonal to every vector in $\boldsymbol{\mathcal{X}}$ is obvious. If $\boldsymbol{x}$ is a vector that is orthogonal to every vector in $\boldsymbol{\mathcal{X}}$, then $\| \boldsymbol{x} \|^2 = \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle = 0$, and therefore, $\boldsymbol{x} = \boldsymbol{0}$. This proves (a). If $\boldsymbol{x} \perp \boldsymbol{y}$, then

$$\begin{aligned} \| \boldsymbol{x} + \boldsymbol{y} \|^2 &= \langle \boldsymbol{x} + \boldsymbol{y} \,|\, \boldsymbol{x} + \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle + \langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle + \langle \boldsymbol{y} \,|\, \boldsymbol{x} \rangle + \langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x} \,|\, \boldsymbol{x} \rangle + \langle \boldsymbol{y} \,|\, \boldsymbol{y} \rangle \;=\; \| \boldsymbol{x} \|^2 + \| \boldsymbol{y} \|^2 \end{aligned}$$

proving (b). Note that property (b) is a generalization of the Pythagorean theorem. Finally, if $\boldsymbol{x} \perp \boldsymbol{R}$ and $\boldsymbol{y} = c_1 \boldsymbol{r}_1 + \cdots + c_k \boldsymbol{r}_k$ for some $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k \in \boldsymbol{R}$, then

$$\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = c_1 \langle \boldsymbol{x} \,|\, \boldsymbol{r}_1 \rangle + \cdots + c_k \langle \boldsymbol{x} \,|\, \boldsymbol{r}_k \rangle = 0$$

as $\boldsymbol{x} \perp \boldsymbol{r}_i$ for all i, and therefore, $\boldsymbol{x} \perp \boldsymbol{y}$.

Example 7.8

In $\mathbb{R}^{3 \times 1}$ the vector $\boldsymbol{r}_3 = \text{col}\,[\,1, 1, 1\,]$ is orthogonal to each of the vectors

$$\boldsymbol{r}_1 = \text{col}\,[\,1, -1, 0\,] \quad \text{and} \quad \boldsymbol{r}_2 = \text{col}\,[\,0, -1, 1\,]$$

with respect to the standard inner product, as $\boldsymbol{r}_3^t\boldsymbol{r}_1 = \boldsymbol{r}_3^t\boldsymbol{r}_2 = 0$. Therefore, $\boldsymbol{r}_3 \perp \text{span}\,(\boldsymbol{r}_1, \boldsymbol{r}_2)$. Indeed, for any

$$\boldsymbol{x} = c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 = \begin{bmatrix} c_1 \\ -c_1 - c_2 \\ c_2 \end{bmatrix}$$

$\boldsymbol{r}_3^t\boldsymbol{x} = 0$.

Clearly, orthogonality of two vectors depends on the particular inner product chosen (see Exercise 7.17). In the rest of this chapter we shall use the standard inner product for $\mathbb{C}^{n\times 1}$ ($\mathbb{R}^{n\times 1}$) or $\mathcal{C}_0([\,a, b\,], \mathbb{C})$, and unless we indicate otherwise, we shall use the notation $\|\cdot\|$ to denote the Euclidean norm defined by the standard inner product.

Let $\boldsymbol{R} = (\,\boldsymbol{r}_1, \cdots, \boldsymbol{r}_k\,)$ be a finite ordered set. The matrix $G = [\,\langle\,\boldsymbol{r}_i\,|\,\boldsymbol{r}_j\,\rangle\,]_{k\times k}$ is called the ***Gram matrix*** of the vectors $\boldsymbol{r}_1, \cdots, \boldsymbol{r}_k$. We claim that

d) $\boldsymbol{R}$ is linearly independent if and only if G is nonsingular

e) if $\boldsymbol{R}$ is orthogonal then $\boldsymbol{R}$ is linearly independent.

If $\boldsymbol{R}$ is linearly dependent then there exists scalars c_i, not all zero, such that

$$\sum_{j=1}^{k} c_j\boldsymbol{r}_j = \boldsymbol{0}$$

Taking inner product of both sides with $\boldsymbol{r}_i$, we obtain

$$\sum_{j=1}^{k}\langle\,\boldsymbol{r}_i\,|\,\boldsymbol{r}_j\,\rangle c_j = 0\,, \quad i = 1, \ldots, k$$

or equivalently, $G\boldsymbol{c} = \boldsymbol{0}$, where $\boldsymbol{c} = \text{col}\,[\,c_1, \ldots, c_k\,] \neq \boldsymbol{0}$. This implies that G is singular. Conversely, if G is singular then $G\boldsymbol{c} = \boldsymbol{0}$ for some $\boldsymbol{c} \neq \boldsymbol{0}$. Then

$$0 = \boldsymbol{c}^h G\boldsymbol{c} = \sum_{i=1}^{k}\sum_{j=1}^{k} c_i^*\langle\,\boldsymbol{r}_i\,|\,\boldsymbol{r}_j\,\rangle c_j = \langle\,\sum_{i=1}^{k} c_i\boldsymbol{r}_i\,|\,\sum_{j=1}^{k} c_j\boldsymbol{r}_j\,\rangle = \|\,\sum_{i=1}^{k} c_i\boldsymbol{r}_i\,\|^2$$

and therefore

$$\sum_{i=1}^{k} c_i\boldsymbol{r}_i = \boldsymbol{0}$$

implying that $\boldsymbol{R}$ is linearly dependent. This proves (d). (e) follows from (d) on noting that if $\boldsymbol{R}$ is orthogonal then $G = \text{diag}\,[\,\|\,\boldsymbol{r}_1\,\|^2, \ldots, \|\,\boldsymbol{r}_k\,\|^2\,]$.

Let $\dim\,(\boldsymbol{\mathcal{X}}) = n$ and let $\boldsymbol{R}$ be an orthogonal (orthonormal) set containing n vectors. Then since $\boldsymbol{R}$ is linearly independent, by Corollary 3.1 it is a basis for $\boldsymbol{\mathcal{X}}$, called an ***orthogonal (orthonormal) basis***.

Example 7.9

Let $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k \in \mathbb{C}^{n\times 1}$ be arranged as the column of an $n \times k$ matrix:

$$R = [\,\boldsymbol{r}_1 \cdots \boldsymbol{r}_k\,]$$

Since $\langle \boldsymbol{r}_i \,|\, \boldsymbol{r}_j \rangle = \boldsymbol{r}_i^h \boldsymbol{r}_j$ it follows that the Gram matrix is

$$G = R^h R$$

Consider the vectors in Example 7.8. Constructing

$$R = \begin{bmatrix} 1 & 0 & 1 \\ -1 & -1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad G = R^t R = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

we observe that G is nonsingular. Hence, $(\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3)$ is linearly independent, and therefore, is a basis for $\mathbb{R}^{3\times 1}$.

7.4 The Projection Theorem and Its Applivations

7.4.1 The Projection Theorem

Let $\boldsymbol{\mathcal{X}}$ be an inner product space with $\dim(\boldsymbol{\mathcal{X}}) = n$ and let $\boldsymbol{\mathcal{U}} \subset \boldsymbol{\mathcal{X}}$ be a subspace with $\dim(\boldsymbol{\mathcal{U}}) = k$. The set

$$\boldsymbol{\mathcal{U}}^{\perp} = \{\, \boldsymbol{x} \,|\, \boldsymbol{x} \perp \boldsymbol{\mathcal{U}} \,\}$$

is called the ***orthogonal complement*** of $\boldsymbol{\mathcal{U}}$.

Let $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k)$ be an ordered basis for $\boldsymbol{\mathcal{U}}$. If $\boldsymbol{x} \perp \boldsymbol{\mathcal{U}}$ then obviously $\boldsymbol{x} \perp \boldsymbol{r}_i$ for all i. Conversely, if $\boldsymbol{x} \perp \boldsymbol{r}_i$ for all i, then $\boldsymbol{x} \perp \text{span}(\boldsymbol{R}) = \boldsymbol{\mathcal{U}}$. Thus $\boldsymbol{\mathcal{U}}^{\perp}$ can also be characterized as

$$\boldsymbol{\mathcal{U}}^{\perp} = \{\, \boldsymbol{x} \,|\, \boldsymbol{x} \perp \boldsymbol{r}_i \,,\ i = 1, \ldots, k \,\}$$

Using this characterization, it can be shown that $\boldsymbol{\mathcal{U}}^{\perp}$ is also a subspace of $\boldsymbol{\mathcal{X}}$ (see Exercise 7.19).

If $\boldsymbol{x} \in \boldsymbol{\mathcal{U}} \cap \boldsymbol{\mathcal{U}}^{\perp}$ then $\boldsymbol{x} \perp \boldsymbol{x}$, and therefore, $\boldsymbol{x} = \boldsymbol{0}$. This shows that $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{U}}^{\perp}$ are linearly independent.

Let $\boldsymbol{x}$ be an arbitrary vector in $\boldsymbol{\mathcal{X}}$, and consider the $k \times k$ linear system of equation

$$\begin{bmatrix} \langle \boldsymbol{r}_1 \,|\, \boldsymbol{r}_1 \rangle & \cdots & \langle \boldsymbol{r}_1 \,|\, \boldsymbol{r}_k \rangle \\ \vdots & & \vdots \\ \langle \boldsymbol{r}_k \,|\, \boldsymbol{r}_1 \rangle & \cdots & \langle \boldsymbol{r}_k \,|\, \boldsymbol{r}_k \rangle \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix} = \begin{bmatrix} \langle \boldsymbol{r}_1 \,|\, \boldsymbol{x} \rangle \\ \vdots \\ \langle \boldsymbol{r}_k \,|\, \boldsymbol{x} \rangle \end{bmatrix} \tag{7.12}$$

Since the coefficient matrix is the Gram matrix of $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k$, it is nonsingular and hence (7.12) has a unique solution $\boldsymbol{c} = \boldsymbol{\alpha} = \text{col}\,[\,\alpha_1, \ldots, \alpha_k\,]$. Let

$$\boldsymbol{x}_u = \sum_{j=1}^{k} \alpha_j \boldsymbol{r}_j \,, \quad \boldsymbol{x}_v = \boldsymbol{x} - \boldsymbol{x}_u$$

Then $\boldsymbol{x}_u \in \boldsymbol{\mathcal{U}}$, and since

$$\langle \boldsymbol{r}_i \,|\, \boldsymbol{x}_v \rangle = \langle \boldsymbol{r}_i \,|\, \boldsymbol{x} \rangle - \langle \boldsymbol{r}_i \,|\, \boldsymbol{x}_u \rangle = \langle \boldsymbol{r}_i \,|\, \boldsymbol{x} \rangle - \sum_{j=1}^{k} \alpha_j \langle \boldsymbol{r}_i \,|\, \boldsymbol{r}_j \rangle = 0 \,, \quad i = 1, \ldots, k$$

$\boldsymbol{x}_v \in \boldsymbol{\mathcal{U}}^{\perp}$. This shows that $\boldsymbol{\mathcal{U}} + \boldsymbol{\mathcal{U}}^{\perp} = \boldsymbol{\mathcal{X}}$. Together with linear independence of $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{U}}^{\perp}$ proved earlier, we reach the following theorem.

Theorem 7.2 (The Projection Theorem) $\mathcal{X} = \mathcal{U} \oplus \mathcal{U}^{\perp}$.

The unique vector $\boldsymbol{x}_u$ is called the ***orthogonal projection*** of $\boldsymbol{x}$ on $\mathcal{U}$. Note that $\boldsymbol{x}_v$ is the orthogonal projection of $\boldsymbol{x}$ on $\mathcal{U}^{\perp}$.

As a consequence of the projection theorem we have the following result.

Corollary 7.2.1 *Let $\boldsymbol{x}_u$ be the orthogonal projection of $\boldsymbol{x}$ on $\mathcal{U}$, and let $\boldsymbol{x}_v = \boldsymbol{x} - \boldsymbol{x}_u$. Then*

$$\min_{\boldsymbol{u} \in \mathcal{U}} \{ \| \boldsymbol{x} - \boldsymbol{u} \| \} = \| \boldsymbol{x}_v \|$$

and the minimum is achieved at $\boldsymbol{u} = \boldsymbol{x}_u$.

Proof Writing $\boldsymbol{x} - \boldsymbol{u} = \boldsymbol{x}_u - \boldsymbol{u} + \boldsymbol{x}_v$ and noting that $\boldsymbol{x}_u - \boldsymbol{u} \perp \boldsymbol{x}_v$, we have

$$\| \boldsymbol{x} - \boldsymbol{u} \|^2 = \| \boldsymbol{x}_u - \boldsymbol{u} \|^2 + \| \boldsymbol{x}_v \|^2$$

from which the result follows.

An illustration of the projection theorem and its corollary is given in Figure 7.1 for $\mathcal{X} = \mathbb{R}^{2\times 1}$.

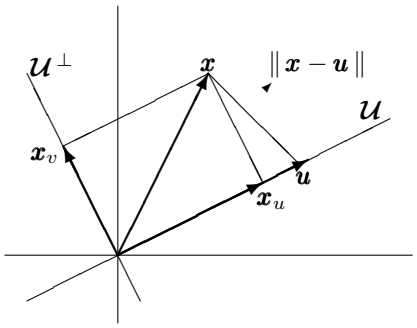

Figure 7.1: Illustration of the projection theorem

Equations (7.12) provide a computational procedure for determining $\boldsymbol{x}_u$. In particular, if $\boldsymbol{R}$ is an orthogonal basis for $\mathcal{U}$ then the solution of (7.12) is obtained as $c_j = \alpha_j = \langle \boldsymbol{r}_j \,|\, \boldsymbol{x} \rangle / \langle \boldsymbol{r}_j \,|\, \boldsymbol{r}_j \rangle$ and we have

$$\boldsymbol{x}_u = \sum_{j=1}^{k} \frac{\langle \boldsymbol{r}_j \,|\, \boldsymbol{x} \rangle}{\langle \boldsymbol{r}_j \,|\, \boldsymbol{r}_j \rangle} \boldsymbol{r}_j$$

If $\boldsymbol{R}$ is orthonormal then this expression further reduces to

$$\boldsymbol{x}_u = \sum_{j=1}^{k} \langle \boldsymbol{r}_j \,|\, \boldsymbol{x} \rangle \, \boldsymbol{r}_j \tag{7.13}$$

We can derive a compact formula for computing orthogonal projections in n-spaces. Let $\mathcal{X} = \mathbb{C}^{n\times 1}$ and

$$\mathcal{U} = \text{span}\,(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k) = \text{im}\,(\,[\boldsymbol{r}_1 \;\; \boldsymbol{r}_2 \;\cdots\; \boldsymbol{r}_k\,]\,) = \text{im}\,(R)$$

where $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k$ form a basis for $\boldsymbol{\mathcal{U}}$. For a given vector $\boldsymbol{x}$ let

$$\boldsymbol{x}_u = c_1\boldsymbol{r}_1 + \cdots + c_k\boldsymbol{r}_k = R\boldsymbol{c}\,, \quad \boldsymbol{c} = \text{col}\,[\,c_1, \ldots, c_k\,]$$

Since the Gram matrix of $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_k$ is $G = R^hR$, (7.12) becomes

$$R^hR\boldsymbol{c} = R^h\boldsymbol{x}$$

from which we obtain

$$\boldsymbol{c} = (R^hR)^{-1}R^h\boldsymbol{x} \quad \text{and} \quad \boldsymbol{x}_u = R(R^hR)^{-1}R^h\boldsymbol{x}$$

Example 7.10

In $\mathbb{R}^{3\times 1}$ the orthogonal projection of $\boldsymbol{x} = \text{col}\,[\,x_1, x_2, x_3\,]$

a) on $\boldsymbol{\mathcal{E}}_1 = \text{span}\,(\boldsymbol{e}_1)$ is $\boldsymbol{x}_1 = x_1\boldsymbol{e}_1 = \text{col}\,[\,x_1, 0, 0\,]$

b) on $\boldsymbol{\mathcal{E}}_2 = \text{span}\,(\boldsymbol{e}_2)$ is $\boldsymbol{x}_2 = x_2\boldsymbol{e}_2 = \text{col}\,[\,0, x_2, 0\,]$

c) on $\boldsymbol{\mathcal{E}}_{12} = \text{span}\,(\boldsymbol{e}_1, \boldsymbol{e}_2)$ is $\boldsymbol{x}_{12} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 = \text{col}\,[\,x_1, x_2, 0\,]$

The reader can interpret $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $\boldsymbol{x}_{12}$ as the components of the vector $\boldsymbol{x}$ on the x_1 axis, on the x_2 axis, and on the x_1x_2 plane, respectively.

Now let $\boldsymbol{u} = \text{col}\,[\,1, 1, 1\,]$ and $\boldsymbol{\mathcal{U}} = \text{span}\,(\boldsymbol{u})$. Then the orthogonal projection of $\boldsymbol{x}$ on $\boldsymbol{\mathcal{U}}$ is

$$\boldsymbol{x}_u = \boldsymbol{u}(\boldsymbol{u}^t\boldsymbol{u})^{-1}\boldsymbol{u}^t\boldsymbol{x} = \frac{\boldsymbol{u}^t\boldsymbol{x}}{\boldsymbol{u}^t\boldsymbol{u}}\,\boldsymbol{u} = \frac{x_1 + x_2 + x_3}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

The reader can easily verify that $\boldsymbol{x}_u \perp \boldsymbol{x} - \boldsymbol{x}_u$.

Orthogonal projections are not restricted to finite dimensional vector spaces as we illustrate by the following example.

* **Example 7.11**

In $\mathcal{C}_0([\,0, 1\,], \mathbb{R})$, let

$$f_1(t) = 1\,, \quad f_2(t) = 2\sqrt{3}(t - 1/2)$$

Since

$$\langle\, f_1 \,|\, f_1 \,\rangle = \int_0^1 dt = 1\,, \quad \langle\, f_2 \,|\, f_2 \,\rangle = \int_0^1 12(t - 1/2)^2\, dt = 1$$

and

$$\langle\, f_1 \,|\, f_2 \,\rangle = \langle\, f_2 \,|\, f_1 \,\rangle = \int_0^1 2\sqrt{3}(t - 1/2)\, dt = 0$$

(f_1, f_2) is an orthonormal set.

Let $\boldsymbol{\mathcal{U}} = \text{span}\,(\,f_1, f_2\,)$ and $f(t) = t^2$. The orthogonal projection of f on $\boldsymbol{\mathcal{U}}$ is $f_u = \alpha_1 f_1 + \alpha_2 f_2$, where

$$\alpha_1 = \langle\, f_1 \,|\, f \,\rangle = \int_0^1 t^2\, dt = \frac{1}{3}$$

and

$$\alpha_2 = \langle\, f_2 \,|\, f \,\rangle = \int_0^1 2\sqrt{3}(t - 1/2)t^2\, dt = \frac{1}{2\sqrt{3}}$$

Thus

$$f_u(t) = (1/3) + (t - 1/2) = t - 1/6$$

Referring to Figure 7.1, we observe that the angle between the straight lines that contain the vectors $\boldsymbol{x}$ and $\boldsymbol{u}$ is given by

$$\theta(\boldsymbol{x},\boldsymbol{u}) = \cos^{-1}\frac{\|\,\boldsymbol{x}_u\,\|}{\|\,\boldsymbol{x}\,\|} = \cos^{-1}\frac{|\langle\,\boldsymbol{u}\,|\,\boldsymbol{x}\,\rangle|}{\|\,\boldsymbol{x}\,\|\,\|\,\boldsymbol{u}\,\|}\,, \quad 0 \le \theta \le \pi/2$$

where the second equality follows from

$$\|\,\boldsymbol{x}_u\,\| = \|\,\frac{\langle\,\boldsymbol{u}\,|\,\boldsymbol{x}\,\rangle}{\langle\,\boldsymbol{u}\,|\,\boldsymbol{u}\,\rangle}\,\| = \frac{|\langle\,\boldsymbol{u}\,|\,\boldsymbol{x}\,\rangle|}{\|\,\boldsymbol{u}\,\|^2}\,\|\,\boldsymbol{u}\,\|$$

Note that since $\|\,\boldsymbol{x}_u\,\| \le \|\,\boldsymbol{x}\,\|$, $\theta(\boldsymbol{x},\boldsymbol{u})$ is well-defined. This definition of angle between two vectors in the plane can readily be generalized to vectors of any inner product space. It can be used as a measure of alignment, and therefore, linear independence of two vectors: The larger the angle between two vectors the more linearly independent they are. Thus orthogonal vectors are maximally linearly independent. This explains why the vectors $\boldsymbol{u}_1 = (2.0, 1.0)$ and $\boldsymbol{u}_2 = (1.0, 2.0)$ in Example 3.20 can be considered to be more linearly independent than the vectors $\boldsymbol{v}_1 = (1.1, 1.0)$ and $\boldsymbol{v}_2 = (1.0, 1.1)$ of the same example: Comparing the angles between vectors of each pair, we see that

$$\theta(\boldsymbol{u}_1,\boldsymbol{u}_2) = \cos^{-1}\frac{4.0}{5.0} = 0.6435 \;\; (\approx 36.9^\circ)$$

whereas

$$\theta(\boldsymbol{v}_1,\boldsymbol{v}_2) = \cos^{-1}\frac{2.20}{2.21} = 0.0952 \;\; (\approx 5.5^\circ)$$

7.4.2 The Gram-Schmidt Orthogonalization Process

Let $\boldsymbol{R}_m = (\,\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_m\,)$ be an ordered linearly independent set in an inner product space $\boldsymbol{\mathcal{X}}$. Define vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m$ successively as

$$\begin{aligned} \boldsymbol{u}_1 &= \boldsymbol{r}_1 \\ \boldsymbol{u}_i &= \boldsymbol{r}_i - \sum_{j=1}^{i-1} \frac{\langle\,\boldsymbol{r}_i\,|\,\boldsymbol{u}_j\,\rangle}{\langle\,\boldsymbol{u}_j\,|\,\boldsymbol{u}_j\,\rangle}\,\boldsymbol{u}_j\,, \quad i = 2, \ldots, m \end{aligned} \tag{7.14}$$

We claim that

a) $\boldsymbol{u}_i \neq \boldsymbol{0}, i = 1, \ldots, m$, so that the process continues to the end without encountering a problem of division by zero

b) the set $\boldsymbol{U}_m = (\,\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m\,)$ is orthogonal

c) span $(\boldsymbol{U}_m)$ = span $(\boldsymbol{R}_m)$

We prove the claims by induction on m. Since $\boldsymbol{u}_1 = \boldsymbol{r}_1 \neq \boldsymbol{0}$, they are obviously true for $m = 1$. Suppose they are true for $m = k$, and consider the case $m = k + 1$. Since

$$\boldsymbol{v}_{k+1} = \sum_{j=1}^{k} \frac{\langle\,\boldsymbol{r}_{k+1}\,|\,\boldsymbol{u}_j\,\rangle}{\langle\,\boldsymbol{u}_j\,|\,\boldsymbol{u}_j\,\rangle}\,\boldsymbol{u}_j$$

is the orthogonal projection of $\boldsymbol{r}_{k+1}$ on $\text{span}\,(\boldsymbol{U}_k) = \text{span}\,(\boldsymbol{R}_k)$ we have

$$\boldsymbol{u}_{k+1} = \boldsymbol{r}_{k+1} - \boldsymbol{v}_{k+1} \neq \boldsymbol{0}$$

for otherwise, $\boldsymbol{r}_{k+1} = \boldsymbol{v}_{k+1} \in \text{span}\,(\boldsymbol{R}_k)$ contradicting linear independence of $\boldsymbol{R}_{k+1}$. Also, since $\boldsymbol{u}_{k+1} \perp \boldsymbol{U}_k$ and $\boldsymbol{U}_k$ is orthogonal by induction hypothesis, $\boldsymbol{U}_{k+1}$ is also orthogonal. Finally,

$$\begin{aligned} \text{span}\,(\boldsymbol{U}_{k+1}) &= \text{span}\,(\boldsymbol{U}_k) \oplus \text{span}\,(\boldsymbol{u}_{k+1}) \\ &= \text{span}\,(\boldsymbol{R}_k) \oplus \text{span}\,(\boldsymbol{u}_{k+1}) \\ &= \text{span}\,(\boldsymbol{R}_k) \oplus \text{span}\,(\boldsymbol{r}_{k+1}) = \text{span}\,(\boldsymbol{R}_{k+1}) \end{aligned}$$

The process described above, which generates an orthogonal set from a linearly independent set, is known as the ***Gram-Schmidt orthogonalization process*** (GSOP).

The GSOP can also be used to check if a given set is linearly independent: Suppose that the subset $\boldsymbol{R}_k$ is linearly independent, but $\boldsymbol{R}_{k+1}$ is not for some $k \leq m$. Then, since $\boldsymbol{r}_{k+1} \in \text{span}\,(\boldsymbol{R}_k) = \text{span}\,(\boldsymbol{U}_k)$, the process gives $\boldsymbol{u}_{k+1} = \boldsymbol{0}$ at the $(k+1)$st step. Conversely, if the process continues up to the kth step, and gives $\boldsymbol{u}_{k+1} = \boldsymbol{0}$, then we conclude that $\boldsymbol{R}_k$ is linearly independent, but $\boldsymbol{R}_{k+1}$ is not.

Example 7.12

Let us complete $\{\boldsymbol{r}\}$ to an orthogonal basis for $\mathbb{R}^{3\times 1}$, where $\boldsymbol{r} = \text{col}\,[\,1, 1, 0\,]$.

Referring to Corollary 3.2, all we have to do is to apply the GSOP to the set $(\boldsymbol{r}, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ and obtain an orthogonal set while eliminating the vectors that are linearly dependent on the previous ones. The process continues as follows.

$$\begin{aligned} \boldsymbol{u}_1 &= \boldsymbol{r} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ \boldsymbol{u}_2 &= \boldsymbol{e}_1 - \frac{\boldsymbol{e}_1^t \boldsymbol{u}_1}{\boldsymbol{u}_1^t \boldsymbol{u}_1}\,\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} \\ \boldsymbol{u}_3 &= \boldsymbol{e}_2 - \frac{\boldsymbol{e}_2^t \boldsymbol{u}_1}{\boldsymbol{u}_1^t \boldsymbol{u}_1}\,\boldsymbol{u}_1 - \frac{\boldsymbol{e}_2^t \boldsymbol{u}_2}{\boldsymbol{u}_2^t \boldsymbol{u}_2}\,\boldsymbol{u}_2 \\ &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \frac{-1/2}{1/2}\begin{bmatrix} 1/2 \\ -1/2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

Since $\boldsymbol{u}_3 = \boldsymbol{0}$, we conclude that $\boldsymbol{e}_2$ is linearly dependent on $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ (equivalently, on $\boldsymbol{r}$ and $\boldsymbol{e}_1$), discard $\boldsymbol{e}_2$, and continue with $\boldsymbol{e}_3$. Observing that $\boldsymbol{e}_3 \perp \{\boldsymbol{u}_1, \boldsymbol{u}_2\}$, we immediately take

$$\boldsymbol{u}_3 = \boldsymbol{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

7.4.3 The Least-Squares Problem

Let $A \in \mathbb{F}^{m\times n}$ and $\boldsymbol{y} \in \mathbb{F}^{n\times 1}$. Recall from Section 3.5 that if $\boldsymbol{y} \notin \text{im}\,(A)$ then the linear equation

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

has no solution. In such a case, we might be interested in finding an approximate solution $\boldsymbol{x} = \boldsymbol{\phi}$ such that $A\boldsymbol{\phi}$ is as close to $\boldsymbol{y}$ as possible with respect to a suitable measure. If we use the Euclidean norm on $\mathbb{F}^{m\times 1}$ as a measure of closeness, the problem can be formulated as

$$\min_{\boldsymbol{x}\in\mathbb{F}^{n\times 1}} \{ \| \boldsymbol{y} - A\boldsymbol{x} \| \} \tag{7.15}$$

Since

$$\{ A\boldsymbol{x} \,|\, \boldsymbol{x} \in \mathbb{F}^{n\times 1} \} = \text{im}\,(A)$$

problem (7.15) is a matter of finding the orthogonal projection of $\boldsymbol{y}$ on im (A). Let $\boldsymbol{y}_A$ be the orthogonal projection of $\boldsymbol{y}$ on im (A). Then

$$\min_{\boldsymbol{x}\in\mathbb{F}^{n\times 1}} \{ \| \boldsymbol{y} - A\boldsymbol{x} \| \} = \| \boldsymbol{y} - \boldsymbol{y}_A \|$$

and the minimum is achieved at a solution $\boldsymbol{x} = \boldsymbol{\phi}_{LS}$ of the consistent equation

$$A\boldsymbol{x} = \boldsymbol{y}_A \tag{7.16}$$

Such a solution is called a ***least-squares solution*** of the equation $A\boldsymbol{x} = \boldsymbol{y}$, for the reason that it minimizes the sum of the squares of the differences between the elements of $\boldsymbol{y}$ and those of $A\boldsymbol{x}$.

Let $r(A) = r$ and let the columns of $R = [\,\boldsymbol{r}_1 \;\cdots\; \boldsymbol{r}_r\,]$ be a basis for im (A). Then the orthogonal projection of $\boldsymbol{y}$ on im (A) is

$$\boldsymbol{y}_A = R\boldsymbol{\alpha} = R(R^hR)^{-1}R^h\boldsymbol{y}$$

Once $\boldsymbol{y}_A$ is found, a least-squares solution can be obtained by solving (7.16).

In the special case when $r(A) = n$, we can choose $R = A$. Then

$$\boldsymbol{y}_A = A(A^hA)^{-1}A^h\boldsymbol{y}$$

and (7.16) becomes

$$A\boldsymbol{x} = A(A^hA)^{-1}A^h\boldsymbol{y}$$

Clearly, the formula

$$\boldsymbol{x} = \boldsymbol{\phi}_{LS} = (A^hA)^{-1}A^h\boldsymbol{y} \tag{7.17}$$

gives a least-squares solution.

Example 7.13

A constant quantity x is measured three times, and the values $x_1 = 14.6, x_2 = 15.6, x_3 = 14.8$ are obtained. What is the best estimate of x based on the measurements? In what sense?

The problem can be formulated as a linear equation as

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} x = \begin{bmatrix} 14.6 \\ 15.6 \\ 14.8 \end{bmatrix}$$

which is obviously inconsistent. A least-squares solution can be obtained from (7.17) as

$$x = \phi_{LS} = \frac{1}{3}\,[\,1 \quad 1 \quad 1\,] \begin{bmatrix} 14.6 \\ 15.6 \\ 14.8 \end{bmatrix} = \frac{14.6 + 15.6 + 14.8}{3} = 15.0$$

Observe that the least-squares solution $x = 15.0$ is simply the average of x_1, x_2 and x_3. It is the best estimate in Euclidean norm in the sense that it minimizes the sum-square error

$$e^2 = (x - 14.6)^2 + (x - 15.6)^2 + (x - 14.8)^2$$

If the error were measured with infinity norm, then we would try to minimize the absolute error

$$e = \max\{\,|x - x_1|, |x - x_2|, |x - x_3|\,\}$$

In this case the best estimate would be $x = 15.1$.

* **Example 7.14**

An operation analyst conducts a study to analyze the relationship between production volume and manufacturing expenses in the auto tyre industry. He assumes a linear relation

$$y = ax + b$$

between the number of tyres produced per day (x) and the daily manufacturing cost (y), and collects data (x_i, y_i) from N companies. His problem is to compute a and b such that the assumed model fits best the collected data.

Clearly, he is faced with a least-squares problem involving the linear system

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

the solution of which is given by (7.17). Straightforward computations yield

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{N\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} N\sum x_i y_i - (\sum x_i)(\sum y_i) \\ (\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i) \end{bmatrix}$$

Letting

$$\mu_x = \frac{1}{N}\sum x_i\,, \qquad \sigma_x^2 = \frac{1}{N}\sum (x_i - \mu_x)^2$$
$$\mu_y = \frac{1}{N}\sum y_i\,, \qquad \sigma_{xy} = \frac{1}{N}\sum (x_i - \mu_x)(y_i - \mu_y)$$

the least-squares solution above can be manipulated into

$$a = \frac{\sigma_{xy}}{\sigma_x^2}\,, \quad b = \mu_y - a\mu_x$$

As a numerical example, suppose that the operation analyst collects the following data from $N = 10$ selected firms, where y is in thousands of dollars:

x :	600	700	825	925	1050	1125	1200	1275	1400	1500
y :	14.8	15.8	16.9	18.0	19.5	19.9	22.4	25.0	26.3	28.7

The solution of the least-squares problem yields a linear model

$$y = 0.01392\,x + 5.975$$

the graph of which is shown in Figure 7.2 together with the data points. The parameters a and b of the linear model are chosen so as to minimize the total sum-square-error

$$e^2 = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (ax_i + b - y_i)^2 = \| A\boldsymbol{x} - \boldsymbol{y} \|^2$$

Based on this model, the analyst expects the total cost of a firm that produces $x = 1000$ tyres to be

$$y = (0.01392)(1000) + 5.975 = 19.895$$

thousand dollars.

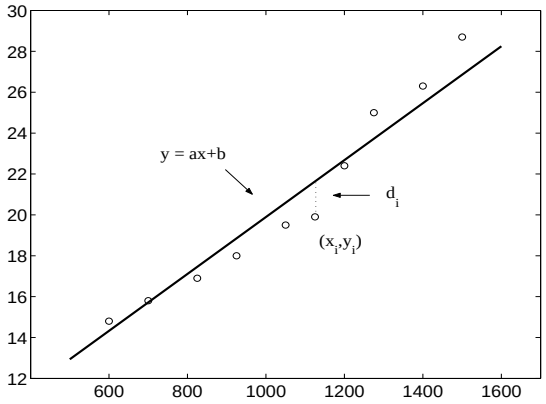

Figure 7.2: Data points and the least-squares linear model

7.4.4 Fourier Series

Let $\boldsymbol{\mathcal{X}}$ be an n-dimensional inner product space, let $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n)$ be an orthonormal basis for $\boldsymbol{\mathcal{X}}$, and let the subspaces $\boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2, \ldots, \boldsymbol{\mathcal{V}}_n$ be defined as

$$\begin{aligned}
\boldsymbol{\mathcal{V}}_1 &= \text{span}\,(\boldsymbol{u}_1) \\
\boldsymbol{\mathcal{V}}_2 &= \text{span}\,(\boldsymbol{u}_1, \boldsymbol{u}_2) \\
&\vdots \\
\boldsymbol{\mathcal{V}}_n &= \text{span}\,(\boldsymbol{u}_1, \boldsymbol{u}_2, \ \ldots\ , \boldsymbol{u}_n)
\end{aligned}$$

Consider an arbitrary vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. The orthogonal projections of $\boldsymbol{x}$ on $\boldsymbol{\mathcal{V}}_1, \ldots, \boldsymbol{\mathcal{V}}_n$ are computed as

$$\begin{aligned}
\boldsymbol{x}_1 &= c_1\boldsymbol{u}_1 \\
\boldsymbol{x}_2 &= c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 \\
&\vdots \\
\boldsymbol{x}_n &= c_1\boldsymbol{u}_1 + c_2\boldsymbol{u}_2 + \cdots + c_n\boldsymbol{u}_n
\end{aligned} \tag{7.18}$$

where

$$c_p = \langle \boldsymbol{u}_p \,|\, \boldsymbol{x} \rangle, \quad p = 1, 2, \ldots, n$$

Note that each $c_p \boldsymbol{u}_p$ is the orthogonal projection of $\boldsymbol{x}$ on the one dimensional subspace $\mathcal{U}_p = \text{span}\,(\boldsymbol{u}_p)$. To construct the orthogonal projection on $\mathcal{V}_2 = \mathcal{U}_1 \oplus \mathcal{U}_2$, we simply add the orthogonal projections on $\mathcal{U}_1$ and $\mathcal{U}_2$; to construct the orthogonal projection on $\mathcal{V}_3 = \mathcal{U}_1 \oplus \mathcal{U}_2 \oplus \mathcal{U}_3$, we add the orthogonal projections on $\mathcal{U}_1$, $\mathcal{U}_2$ and $\mathcal{U}_3$; and so on. This is a consequence of $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n)$ being orthogonal. Otherwise, the orthogonal projection on $\mathcal{U}_1 \oplus \mathcal{U}_2$ would be different from the sum of the orthogonal projections on the individual subspaces (see Exercise 7.34).

From the discussion in the previous section we know that each $\boldsymbol{x}_q$ is the best approximation to $\boldsymbol{x}$ in terms of the vectors of $\mathcal{V}_q$, $q = 1, 2, \ldots, n$. Since

$$\mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n$$

$\boldsymbol{x}_2$ is a better approximation to $\boldsymbol{x}$ than $\boldsymbol{x}_1$ is, $\boldsymbol{x}_3$ is better than $\boldsymbol{x}_2$ is, and so on. That is,

$$\| \boldsymbol{x} - \boldsymbol{x}_1 \| \geq \| \boldsymbol{x} - \boldsymbol{x}_2 \| \geq \cdots \geq \| \boldsymbol{x} - \boldsymbol{x}_n \|$$

In fact, since $\mathcal{V}_n = \mathcal{X}$, we have $\boldsymbol{x}_n = \boldsymbol{x}$ so that $\| \boldsymbol{x} - \boldsymbol{x}_n \| = 0$. The situation, which is illustrated in Figure 7.3 for $\mathcal{X} = \mathbb{R}^3$, is easy to understand when $\mathcal{X}$ is finite dimensional. The infinite dimensional case is more interesting.

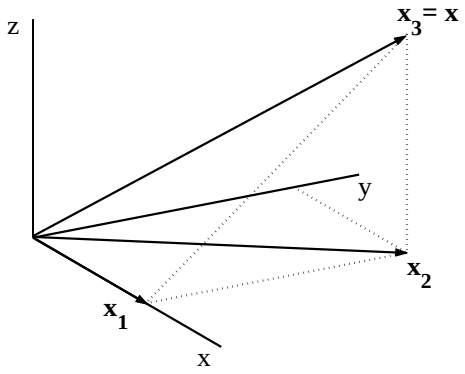

Figure 7.3: Orthogonal projections of a vector

Suppose $\dim(\mathcal{X}) = \infty$, and suppose $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n, \ldots)$ is an infinite sequence of orthonormal vectors in $\mathcal{X}$.[4] For an arbitrary $\boldsymbol{x} \in \mathcal{X}$, let us define the subspaces $\mathcal{V}_q$ and the vectors $\boldsymbol{x}_q$ as in (7.18) and (7.18) with the index q running not just up to n but up to infinity. Then again each $\boldsymbol{x}_q$ is the orthogonal projection of $\boldsymbol{x}$ on the q-dimendional subspace $\mathcal{V}_q$, and hence it is the best approximation to $\boldsymbol{x}$ in terms of the vectors in $\mathcal{V}_q$. That is,

$$\min_{\boldsymbol{u} \in \mathcal{V}_q} \| \boldsymbol{x} - \boldsymbol{u} \| = \| \boldsymbol{x} - \boldsymbol{x}_q \|, \quad q = 1, 2, \ldots$$

[4]Here we assume that such an infinite orthonormal set exists. Although with our present knowledge we cannot guarantee the existence of such a set, we may try to construct one.

Since $\boldsymbol{\mathcal{V}}_q \subset \boldsymbol{\mathcal{V}}_{q+1}$, we also have

$$\| \boldsymbol{x} - \boldsymbol{x}_{q+1} \| = \min_{\boldsymbol{u} \in \boldsymbol{\mathcal{V}}_{q+1}} \| \boldsymbol{x} - \boldsymbol{u} \| \leq \min_{\boldsymbol{u} \in \boldsymbol{\mathcal{V}}_q} \| \boldsymbol{x} - \boldsymbol{u} \| = \| \boldsymbol{x} - \boldsymbol{x}_q \|$$

that is,

$$\| \boldsymbol{x} - \boldsymbol{x}_1 \| \geq \| \boldsymbol{x} - \boldsymbol{x}_2 \| \geq \cdots \geq \| \boldsymbol{x} - \boldsymbol{x}_q \| \geq \cdots$$

The interesting point is that although $\boldsymbol{x}_q$ approximate $\boldsymbol{x}$ better and better as q increases, without further information about the set $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n, \ldots)$ we cannot say that

$$\lim_{q \to \infty} \| \boldsymbol{x} - \boldsymbol{x}_q \| = 0 \tag{7.19}$$

However, if (7.19) is true, we formally write

$$\boldsymbol{x} = \sum_{p=1}^{\infty} c_p \boldsymbol{u}_p \tag{7.20}$$

The expression on the right-hand-side of (7.20) is known as the ***Fourier series*** of $\boldsymbol{x}$ in terms of $(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n, \ldots)$.

Example 7.15

Refer to Example 3.26. Let $\boldsymbol{\mathcal{X}} = \mathcal{F}(\mathbb{D}_N, C)$, and define

$$\langle f \,|\, g \rangle = \frac{1}{N} \sum_{k=0}^{N-1} f^*[k] g[k]$$

which is the standard inner product on $\mathbb{C}^{N \times 1}$ scaled by $1/N$. (Recall that $\mathcal{F}(\mathbb{D}_N, C)$ is essentially the same as $\mathbb{C}^{N \times 1}$.)

Consider the set of functions $\phi_p, p = 0, \ldots, N-1$ defined in Example 3.26. Using the hint in Exercise 3.20, it can easily be shown that $(\, \phi_p \,)$ is an orthonormal set in $\mathcal{F}(\mathbb{D}_N, C)$. Then for a given $f \in \mathcal{F}(\mathbb{D}_N, C)$, (7.20) reduces to the (discrete) Fourier series in (3.7), where

$$c_p = \langle \, \phi_p \,|\, f \, \rangle = \frac{1}{N} \sum_{k=0}^{N-1} \phi_p^*[k] f[k]$$

and each $c_p \phi_p$ term is the orthogonal projection of f on span (ϕ_p).

Example 7.16

Let $\boldsymbol{\mathcal{X}}$ be the vector space of piece-wise continuous complex-valued functions defined on a real interval $(\, 0, T \,)$. Let us define an inner product on $\boldsymbol{\mathcal{X}}$ as

$$\langle f \,|\, g \rangle = \frac{1}{T} \int_0^T f^*(t) g(t) \, dt \tag{7.21}$$

which is the familiar standard inner product scaled with $1/T$. Consider the following set of functions:

$$\phi_k(t) = e^{ik\frac{2\pi}{T}t}, \quad k = \ldots, -1, 0, 1, \ldots$$

It is left to the reader to prove that (ϕ_k) is an orthonormal set with respect to the inner product in (7.21). Hence the Fourier coefficients of a given function f are computed as

$$\alpha_k = \langle\, \phi_k \,|\, f \,\rangle = \frac{1}{T} \int_0^T f(t) e^{-ik\frac{2\pi}{T}t}\, dt$$

As a specific example consider the piece-wise continuous function

$$f(t) = \begin{cases} 1\,, & 0 < t < 0.5 \\ 0\,, & 0.5 < t < 1 \end{cases}$$

defined on the interval $(0, 1)$. Then its Fourier coefficients are computed as

$$\alpha_k = \int_0^{0.5} e^{-i2k\pi t}\, dt = \begin{cases} \frac{1}{2}\ , & k = 0 \\ \frac{1}{k\pi i}\,, & k \text{ odd} \\ 0\ \ , & k \neq 0, k \text{ even} \end{cases}$$

Hence the orthogonal projection of f on the subspace

$$\boldsymbol{\mathcal{U}}_q = \text{span}\,(\,\phi_{-q}, \ldots, \phi_{-1}, \phi_0, \phi_1, \ldots, \phi_q\,)$$

is given as

$$f_q(t) = \frac{1}{2} + \sum_{\substack{k=1 \\ k \text{ odd}}}^{q} (\frac{1}{k\pi i}\, e^{i2k\pi t} - \frac{1}{k\pi i}\, e^{-i2k\pi t}) = \frac{1}{2} + \sum_{\substack{k=1 \\ k \text{ odd}}}^{q} \frac{2}{k\pi}\, \sin 2k\pi t$$

Plots of f and f_q for $q = 0, 1, 3, 9$ are shown in Figure 7.4.

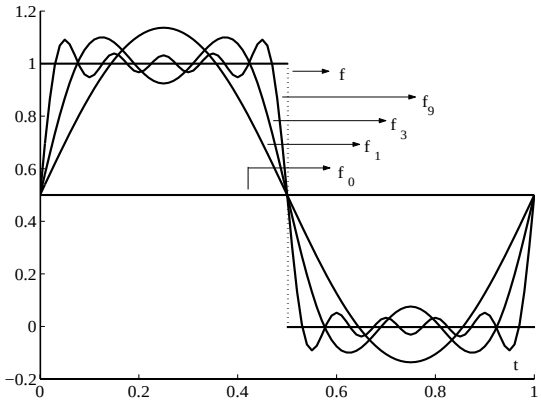

Figure 7.4: Fourier approximations of a function

7.5 Exercises

1. (a) Find the uniform, Euclidean and infinity norms of the following vectors.

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ -3 \\ 5 \end{bmatrix}\,, \quad \boldsymbol{y} = \begin{bmatrix} 1+i \\ -3i \\ 2 \end{bmatrix}$$

(b) Repeat part (a) by using MATLAB command `norm`.

2. In $\mathbb{R}^2$ plot the locus of points $\boldsymbol{x}$ for which $\| \boldsymbol{x} \|_p = 1$ for $p = 1, 2, \infty$.

3. Let $p > 1$ be a real number and $q > 1$ be such that $p^{-1} + q^{-1} = 1$.

(a) Show that

$$u^{1/p} v^{1/q} \leq u/p + v/q \quad \text{for all} \;\; u \geq 0\,, \quad v \geq 0$$

Hint: First show that

$$(1 + x)^p \geq 1 + px \quad \text{for all} \;\; x \geq -1\,, \quad p > 1$$

and let $1 + x = (u/v)^{1/p}$.

(b) Prove ***Hölder's inequality***

$$\sum_{i=1}^{n} |x_i y_i| \leq \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p} \Big(\sum_{i=1}^{n} |y_i|^q \Big)^{1/q}$$

for $\boldsymbol{x} = \text{col}\,[\, x_1, \ldots, x_n \,]$ and $\boldsymbol{y} = \text{col}\,[\, y_1, \ldots, y_n \,] \in \mathbb{C}^{n \times 1}$. Hint: Apply the inequality in (a) to

$$u = u_i = \frac{|x_i|^p}{\sum\limits_{i=1}^{n} |x_i|^p}\,, \quad v = v_i = \frac{|y_i|^q}{\sum\limits_{i=1}^{n} |y_i|^q}$$

and then take summation on i.

4. Prove Minkowski's inequality

$$\Big(\sum_{i=1}^{n} |x_i + y_i|^p \Big)^{1/p} \leq \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p} + \Big(\sum_{i=1}^{n} |y_i|^p \Big)^{1/p}$$

Hint: Take summation of both sides of the inequalities

$$|x_i + y_i|^p \leq |x_i|\, |x_i + y_i|^{p-1} + |y_i|\, |x_i + y_i|^{p-1}\,, \quad i = 1, \ldots, n$$

and use Hölder's inequality and the identity $q(p - 1) = p$.

5. Show that

$$\| f \|_p = \Big(\int_a^b |f(t)|^p \, dt \Big)^{1/p}$$

is a norm on $\mathcal{C}_0([a, b], \mathbb{R})$. Hint: Derive integral counterparts of Hölder's and Minkowski's inequalities.

6. Find the uniform, Euclidean and infinity norms of the following functions.

(a) $f(t) = t - 1\,, \quad 0 \leq t \leq 2$

(b) $g(t) = e^{i\omega t}\,, \quad -\pi/\omega \leq t \leq \pi/\omega$

(c) $h(t) = 1/t\,, \quad 1 \leq t \leq T\,, \quad T \to \infty$

7. Refer to Example 7.3. Since the components of any $\boldsymbol{f} \in \mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n \times 1})$ are continuous, the function $\nu_p^{\boldsymbol{f}}$ defined in (7.7) is a continuous function for any $p \geq 1$. Hence, $\| \boldsymbol{f} \|_{p,q}$ in (7.8) is a well-defined quantity for any $q \geq 1$. Now, $\boldsymbol{f} \neq \boldsymbol{0}$ implies $\nu_p^{\boldsymbol{f}} \neq 0$, which in turn implies $\| \boldsymbol{f} \|_{p,q} > 0$. Also, for any $c \in \mathbb{R}$,

$$\nu_p^{c\boldsymbol{f}}(t) = \| \, (c\boldsymbol{f})(t) \, \|_p = \| \, c\boldsymbol{f}(t) \, \|_p = |\, c \,|\, \| \, \boldsymbol{f}(t) \, \|_p = |\, c \,|\, \nu_p^{\boldsymbol{f}}(t) \quad \text{for all} \;\; t \in \boldsymbol{I}$$

so that

$$\| c\boldsymbol{f} \|_{p,q} = \| \nu_p^{c\boldsymbol{f}} \|_q = \| \, | \, c \, | \, \nu_p^{\boldsymbol{f}} \|_q = | \, c \, | \, \| \nu_p^{\boldsymbol{f}} \|_q = | \, c \, | \, \| \boldsymbol{f} \|_{p,q}$$

Thus $\| \boldsymbol{f} \|_{p,q}$ satisfies the first two properties of a norm. Show that it also satisfies the triangle inequality, so that it is a norm on $\mathcal{C}_0(\boldsymbol{I}, \mathbb{R}^{n \times 1})$.

8. Refer to Example 7.3.

   (a) Let $\boldsymbol{f}(t) = \text{col}\,[\,1, t\,], \quad 0 \leq t \leq 1$. Find $\| \boldsymbol{f} \|_{1,2}$ and $\| \boldsymbol{f} \|_{2,\infty}$

   (b) Let $\boldsymbol{h}(t) = \text{col}\,[\,t, t-1\,], \quad 0 \leq t \leq 1$. Find $\| \boldsymbol{h} \|_{1,\infty}$ and $\| \boldsymbol{h} \|_{\infty,1}$

9. Two norms $\| \cdot \|$ and $\| \cdot \|'$ on $\boldsymbol{\mathcal{X}}$ are said to be equivalent if there exist $0 < c_1 \leq c_2$ such that

$$c_1 \| \boldsymbol{x} \| \leq \| \boldsymbol{x} \|' \leq c_2 \| \boldsymbol{x} \| \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

in which case

$$\frac{1}{c_2} \| \boldsymbol{x} \|' \leq \| \boldsymbol{x} \| \leq \frac{1}{c_1} \| \boldsymbol{x} \|' \quad \text{for all } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

   (a) Show that all p-norms on $\mathbb{C}^{n \times 1}$ (including $p = \infty$) are equivalent. Hint: First show that all p-norms are equivalent to the ∞-norm.

   (b) Does a corresponding result hold for the p-norms on $\mathcal{C}_0([\,0, 1\,], \mathbb{R})$? Hint: Suppose that there exist $0 < c_1 \leq c_2$ such that

$$c_1 \| f \|_1 \leq \| f \|_\infty \leq c_2 \| f \|_1 \quad \text{for all } f \in \mathcal{C}_0([\,0, 1\,], \mathbb{R})$$

   Let

$$f(t) = \begin{cases} 1 - nt\,, & 0 \leq t \leq 1/n \\ 0 \quad , & 1/n \leq t \leq 1 \end{cases}$$

   and show that the second equality is violated for sufficiently large n.

10. Show that the following are norms on $\mathbb{F}^{m \times n}$.

   (a)

$$\| A \| = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|$$

   (b)

$$\| A \| = \max_{i,j} \{\, |a_{ij}| \,\}$$

11. (a) Show that the matrix norm subordinate to the uniform vector norm is

$$\| A \|_1 = \max_{1 \leq j \leq n} \{ \sum_{i=1}^{m} |a_{ij}| \} = \max_{1 \leq j \leq n} \{ \| \boldsymbol{a}_j \|_1 \}$$

   where $\boldsymbol{a}_j$ denotes the jth column of A Hint: Suppose that the maximum of the right-hand side is achieved for $j = q$. Show that for arbitrary $\boldsymbol{x} = \text{col}\,[\,x_1, \ldots, x_n\,]$

$$\| A\boldsymbol{x} \|_1 \leq (|x_1| + \cdots + |x_n|) \cdot \| \boldsymbol{a}_q \|_1 = \| \boldsymbol{a}_q \|_1 \| \boldsymbol{x} \|_1$$

   with equality holding for $\boldsymbol{x} = \boldsymbol{e}_q$.

(b) Show that the matrix norm subordinate to the infinity vector norm is

$$\| A \|_\infty = \max_{1 \le i \le m} \{ \sum_{j=1}^{n} |a_{ij}| \} = \max_{1 \le i \le m} \{ \| \boldsymbol{\alpha}_i \|_1 \}$$

where $\boldsymbol{\alpha}_i$ denotes the ith row of A. Hint: Suppose that the maximum of the right-hand side is achieved for $i = p$. Show that for arbitrary $\boldsymbol{x} = \text{col}\,[\, x_1, \ldots, x_n \,]$

$$\| A\boldsymbol{x} \|_\infty \le \max_i \{ \sum_{j=1}^{n} |a_{ij}| \} \cdot \max_j \{ |x_j| \} = \| \boldsymbol{\alpha}_p \|_1 \| \boldsymbol{x} \|_\infty$$

with equality holding for

$$\boldsymbol{x} = \text{col}\,[\, \text{sign}(a_{p1}), \ldots, \text{sign}(a_{pn}) \,]$$

12. (a) Find the uniform and infinity norms of the following matrices.

(b) Use MATLAB command `norm` to verify your results.

$$A = \begin{bmatrix} -1 & 1 \\ 0 & 3 \\ 2 & -1 \end{bmatrix}, \quad B = A^t$$

13. Let $\mathcal{A} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ be defined as $\mathcal{A}(X) = X^t$. Clearly, $\mathcal{A}$ is a linear transformation. Find

$$\| \mathcal{A} \|_1 = \max_{X \ne O} \frac{\| X^t \|_1}{\| X \|_1}$$

14. Let $\boldsymbol{\mathcal{X}}$ be a normed vector space with a norm $\| \cdot \|$. An infinite sequence of vectors $(\boldsymbol{x}_n)$ is said to ***converge*** (in the norm $\| \cdot \|$) to a limit vector $\boldsymbol{x}$, denoted as

$$\lim_{n \to \infty} \boldsymbol{x}_n = \boldsymbol{x}$$

if the sequence of real numbers $(\| \boldsymbol{x}_n - \boldsymbol{x} \|)$ converges to 0. Equivalently, $(\boldsymbol{x}_n)$ converges to $\boldsymbol{x}$ if for any $\epsilon > 0$ there exists an integer $N > 0$ such that

$$\| \boldsymbol{x}_n - \boldsymbol{x} \| < \epsilon \quad \text{for all} \;\; n \ge N$$

Check if the sequence $(\boldsymbol{x}_n)$, where

$$\boldsymbol{x}_n = \begin{bmatrix} 1 - \frac{1}{(-2)^n} \\ \\ \frac{1}{3^n} \end{bmatrix}$$

converges in $\mathbb{R}^{2\times 1}$, and find its limit if it does. Does your answer depend on the particular norm you choose for $\mathbb{R}^{2\times 1}$?

15. A sequence $(\boldsymbol{x}_n)$ in a normed vector space $\boldsymbol{\mathcal{X}}$ is called a ***Cauchy sequence*** if

$$\lim_{m,n \to \infty} \| \boldsymbol{x}_n - \boldsymbol{x}_m \| = 0$$

Show that every convergent sequence in $\boldsymbol{\mathcal{X}}$ is a Cauchy sequence.

16. (a) Find all orthogonal pairs of the following vectors in $\mathbb{R}^{3\times 1}$ w.r.t. the standard inner product:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \boldsymbol{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \boldsymbol{x}_4 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

(b) Repeat (a) for the following vectors in $\mathbb{C}^{3\times 1}$:

$$\boldsymbol{z}_1 = \begin{bmatrix} 1 \\ i \\ 1+i \end{bmatrix}, \boldsymbol{z}_2 = \begin{bmatrix} 3i \\ 1 \\ 1-i \end{bmatrix}, \boldsymbol{z}_3 = \begin{bmatrix} 0 \\ 1+i \\ -1 \end{bmatrix}, \boldsymbol{z}_4 = \begin{bmatrix} -1 \\ -i \\ 1+i \end{bmatrix}$$

(c) Repeat (a) for the following vectors in $\mathcal{F}([\,0,1\,],\mathbb{R})$:

$$f_1 = 1\,,\ f_2(t) = t+1\,,\ f_3(t) = 2t-1\,,\ f_4(t) = 6t^2 - 6t + 1$$

17. For $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^{2\times 1}$ let

$$\langle\, \boldsymbol{x} \,|\, \boldsymbol{y} \,\rangle_Q = x_1y_1 + x_1y_2 + x_2y_1 + 2x_2y_2 = \boldsymbol{x}^t Q \boldsymbol{y}$$

where

$$Q = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Show that $\langle \cdot \,|\, \cdot \rangle_Q$ is an inner product. Hint:

$$\langle\, \boldsymbol{x} \,|\, \boldsymbol{x} \,\rangle_Q = (x_1 + x_2)^2 + x_2^2$$

(b) Are $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ orthogonal with respect to this inner product? Find a vector that is orthogonal to $\boldsymbol{e}_1$ and a vector orthogonal to $\boldsymbol{e}_2$.

(c) The norm defined by this inner product is

$$\|\, \boldsymbol{x} \,\|_Q = \sqrt{\boldsymbol{x}^t Q \boldsymbol{y}} = \sqrt{(x_1 + x_2)^2 + x_2^2}$$

Compute $\|\, \boldsymbol{e}_1 \,\|_Q$ and $\|\, \boldsymbol{e}_2 \,\|_Q$.

18. Show that Schwarz inequality for $\mathbb{R}^{n\times 1}$ and $\mathcal{F}(\boldsymbol{I}, \mathbb{R})$ is a special case of Hölder's inequality.

19. Let $\boldsymbol{\mathcal{U}}$ be a subset of an inner product space $\boldsymbol{\mathcal{X}}$. Prove that $\boldsymbol{\mathcal{U}}^{\perp}$ is a subspace of $\boldsymbol{\mathcal{X}}$. Hint: Show that $\boldsymbol{\mathcal{U}}^{\perp}$ is closed under vector addition and scalar multiplication.

20. In $\mathbb{R}^{3\times 1}$, let

$$\boldsymbol{\mathcal{U}} = \text{span}\,(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix})$$

Find bases for $\boldsymbol{\mathcal{U}}^{\perp}$ and $(\boldsymbol{\mathcal{U}}^{\perp})^{\perp}$.

21. In $\mathcal{C}_0([\,-T,T\,],\mathbb{R})$, let

$$\boldsymbol{\mathcal{U}} = \{\, f \,|\, f(-t) = f(t) \,\}$$

Characterize the orthogonal complement of $\boldsymbol{\mathcal{U}}$ with respect to the inner product in (7.11).

22. Apply GSOP to

$$\boldsymbol{u}_1 = \text{col}\,[\,1,1,0\,]\,, \quad \boldsymbol{u}_2 = \text{col}\,[\,0,2,1\,]\,, \quad \boldsymbol{u}_3 = \text{col}\,[\,4,0,1\,]$$

to generate an orthogonal basis for $\mathbb{R}^{3\times 1}$.

23. The MATLAB command `orth(A)` finds an orthonormal basis for $\text{im}\,(A)$. Thus if

$$A = [\,\boldsymbol{a}_1 \ \cdots\ \boldsymbol{a}_n\,] \quad \text{and} \quad B = \texttt{orth}(A) = [\,\boldsymbol{b}_1 \ \cdots\ \boldsymbol{b}_q\,]$$

then $(\,\boldsymbol{b}_1,\ldots,\boldsymbol{b}_q\,)$ is an orthonormal set generated by $(\,\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n\,)$.

(a) Use `orth` to generate an orthogonal basis for $\mathbb{R}^{3\times 1}$ from the vectors in Exercise 7.22.

(b) Use `orth` to compute the rank of the matrix

$$A = \begin{bmatrix} 1 & 2 & 1 & -3 \\ -1 & 1 & 2 & -3 \\ 2 & -1 & -3 & 4 \end{bmatrix}$$

24. Let $\boldsymbol{x}_1 = \text{col}\,[\,1, 1, 0\,]$ and $\boldsymbol{x}_2 = \text{col}\,[\,1, 1, 1\,]$.

(a) Apply GSOP to $(\,\boldsymbol{x}_1, \boldsymbol{x}_2\,)$ to generate an orthonormal set $(\,\boldsymbol{v}_1, \boldsymbol{v}_2\,)$.

(b) Find orthogonal projections of $\boldsymbol{x} = \text{col}\,[\,0, 1, 1\,]$ on $\boldsymbol{\mathcal{S}}_1 = \text{span}\,(\,\boldsymbol{x}_1\,)$ and on $\boldsymbol{\mathcal{S}}_2 = \text{span}\,(\,\boldsymbol{x}_1, \boldsymbol{x}_2\,)$.

(c) Repeat (a) and (b) for $\boldsymbol{x}_1 = \text{col}\,[\,1, -1, 0\,]$, $\boldsymbol{x}_2 = \text{col}\,[\,0, 1, 1\,]$ and $\boldsymbol{x} = \text{col}\,[\,1, 1, 1\,]$

25. In $\mathbb{R}^3$ find the minimum distance from the origin to the plane

$$2x_1 + 3x_2 - x_3 = -5$$

and also find the point on the plane closest to the origin. Hint: The given plane has a normal $\boldsymbol{n} = (2, 3, -1)$.

26. A mirror lies in the plane defined by

$$-2x_1 + 3x_2 + x_3 = 0$$

which defines a subspace of $\mathbb{R}^3$. Find the reflected image of the vector $\boldsymbol{x} = \text{col}\,[\,5, 2, -3\,]$.

27. Consider $\mathcal{C}_0([\,-1, 1\,], \mathbb{R})$ with the inner product given in (7.11).

(a) Apply GSOP to the set $(1, t, t^2)$ to generate an orthonormal set of functions.

(b) Find the orthogonal projection of $g(t) = t^3$ on span $(\,1, t, t^2\,)$.

28. Let $A \in \mathcal{C}^{n\times n}$, and $\boldsymbol{\mathcal{U}} \in \mathcal{C}^{n\times 1}$ be a proper subspace. Show that if $\boldsymbol{\mathcal{U}}$ is A-invariant, then $\boldsymbol{\mathcal{U}}^{\perp}$ is A^h-invariant.

29. (a) Find a least squares solution to

$$\begin{bmatrix} 1 & 2 \\ -1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \\ 6 \end{bmatrix}$$

(b) Use MATLAB command `x=pinv(A)*b` to verify your result.

30. Consider the linear equation

$$\begin{bmatrix} 1 & 1 & 2 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}$$

(a) Characterize all least squares solutions $\boldsymbol{x}_{LS}$ of the given equation.

(b) Among all least squares solutions, find the one with minimum norm.

31. (a) In the xy plane plot the locus of all vectors (points) of the form $\boldsymbol{x} = \boldsymbol{p} + c\,\boldsymbol{q}$, $c \in \mathbb{R}$, where

$$\boldsymbol{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{q} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

(b) Determine geometrically the value of c such that $\|\,\boldsymbol{p} + c\,\boldsymbol{q}\,\|$ is minimum.

(c) Formulate and solve the problem in part (b) as a least-squares problem.

32. (a) Find a, b, c which minimize

$$\int_{-1}^{1} (t^3 - a - bt - ct^2)^2 \, dt$$

(b) Formulate and solve the problem as a least-squares problem in $\mathcal{C}_0([-1, 1], \mathbb{R})$.

33. Let $f(t)$ be a continuous function defined on an interval $0 \leq t \leq 1$. Consider the problem of approximating f by an $(n-1)$st degree polynomial

$$p(t) = \sum_{k=0}^{n-1} p_k t^k$$

whose coefficients $p_k, k = 0, \ldots, n-1$, are to be determined such that the error

$$E = \int_0^1 [p(t) - f(t)]^2 \, dt$$

is minimized.

(a) Obtain a system of n linear equations in the n unknowns $p_k, k = 0, \ldots, n-1$, by setting

$$\frac{\partial E}{\partial p_k} = 0\,, \quad k = 0, \ldots, n-1$$

Show that the coefficient matrix of the resulting linear system $H\boldsymbol{p} = \boldsymbol{b}$ is a Hilbert matrix of order n.

(b) Interpret the problem as a least-squares problem.

34. Let

$$\boldsymbol{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}\,, \quad \boldsymbol{u}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}\,, \quad \boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

(a) Find the orthogonal projections $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of $\boldsymbol{x}$ on span $(\boldsymbol{u}_1)$ and span $(\boldsymbol{u}_2)$.

(b) Find the orthogonal projection $\boldsymbol{x}_{12}$ of $\boldsymbol{x}$ on span $(\boldsymbol{u}_1, \boldsymbol{u}_2)$. Is $\boldsymbol{x}_{12} = \boldsymbol{x}_1 + \boldsymbol{x}_2$? Explain.

35. Refer to Example 7.16. Obtain the Fourier series of the function

$$f(t) = t\,, \quad 0 \leq t \leq 1$$

Use MATLAB to compute and plot the Fourier series truncated at $k = 0, 1, 5, 10$.

36. Let $(\boldsymbol{x}_i, i = 1, \ldots, k)$ be an orthonormal set in an inner product space $\boldsymbol{\mathcal{X}}$. Show that for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$

$$\sum_{i=1}^{k} |\langle \boldsymbol{x} \,|\, \boldsymbol{x}_i \rangle|^2 \leq \|\boldsymbol{x}\|^2$$

The inequality above is known as the ***Bessel's inequality*** and is true whether $\boldsymbol{\mathcal{X}}$ is finite or infinite dimensional.

Chapter 8
Unitary and Hermitian Matrices

8.1 Unitary Matrices

A complex square matrix U that satisfies

$$U^h U = UU^h = I$$

is called ***unitary***. If U is a real unitary matrix then

$$U^t U = UU^t = I$$

and U is called ***orthogonal***. Equivalently, a complex matrix U is unitary if $U^{-1} = U^h$, and a real matrix is orthogonal if $U^{-1} = U^t$. Note that the columns of an $n \times n$ unitary (orthogonal) matrix form an orthonormal basis for $\mathbb{C}^{n\times 1}$ ($\mathbb{R}^{n\times 1}$).

If U is unitary, then

$$\langle\, U\boldsymbol{x} \,|\, U\boldsymbol{y} \,\rangle = \boldsymbol{x}^h U^h U\boldsymbol{y} = \boldsymbol{x}^h\boldsymbol{y} = \langle\, \boldsymbol{x} \,|\, \boldsymbol{y} \,\rangle \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{n\times 1}$$

Consequently, $\| U\boldsymbol{x} \| = \| \boldsymbol{x} \|$ for all $\boldsymbol{x} \in \mathbb{C}^{n\times 1}$, and if $\{ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \}$ is an orthonormal set, then so is $\{ U\boldsymbol{x}_1, \ldots, U\boldsymbol{x}_k \}$. Also,

$$\theta(U\boldsymbol{x}, U\boldsymbol{y}) = \theta(\boldsymbol{x}, \boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{n\times 1}$. In other words, a mapping by a unitary transformation preserves norms and angles.

Example 8.1

It can easily be verified that

$$R = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

is an orthogonal matrix.

Let

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \boldsymbol{y} = R\boldsymbol{x} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} x_1 - x_2 \\ x_1 + x_2 \end{bmatrix}$$

Then

$$\| \boldsymbol{y} \|^2 = \frac{(x_1 - x_2)^2 + (x_1 + x_2)^2}{2} = x_1^2 + x_2^2 = \| \boldsymbol{x} \|^2$$

Expressing R as

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \theta = \frac{\pi}{4}$$

we observe that the transformation $\boldsymbol{y} = R\boldsymbol{x}$ corresponds to a counterclockwise rotation in the plane by an angle of $\theta = \pi/4$ (see Example 5.3). If

$$R_1 = \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix} \quad \text{and} \quad R_2 = \begin{bmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix}$$

are two such rotation matrices corresponding to conterclockwise rotations by θ_1 and θ_2, then we expect that $R = R_2R_1$ should also be a rotation matrix corresponding to a conterclockwise rotation by $\theta = \theta_1 + \theta_2$. Indeed, simple trigonometric identities give

$$\begin{aligned} R &= \begin{bmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{bmatrix} \begin{bmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{bmatrix} \\ &= \begin{bmatrix} \cos\theta_2\cos\theta_1 - \sin\theta_2\sin\theta_1 & -\cos\theta_2\sin\theta_1 - \sin\theta_2\cos\theta_1 \\ \cos\theta_2\sin\theta_1 + \sin\theta_2\cos\theta_1 & \cos\theta_2\cos\theta_1 - \sin\theta_2\sin\theta_1 \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta_1+\theta_2) & -\sin(\theta_1+\theta_2) \\ \sin(\theta_1+\theta_2) & \cos(\theta_1+\theta_2) \end{bmatrix} \end{aligned}$$

Note that R is also an orthogonal matrix.

Rotation matrices in the 3-space can be defined similarly (see Exercise 8.3).

If U is a unitary matrix, then

$$1 = \det(U^hU) = (\det U^h)(\det U) = (\det U)^*(\det U) = |\det U|^2$$

so that $|\det U| = 1$. If U is orthogonal then $\det U$ is real, and therefore

$$\det U = \mp 1$$

As a simple example, the reader can verify that $\det U = 1$ for the rotation matrix in Example 8.1.

Structure of unitary matrices is characterized by the following theorem.

Theorem 8.1 *Let $U \in \mathbb{C}^{n\times n}$ be a unitary matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

a) $|\lambda_i| = 1, i = 1, \ldots, n$

b) there exists a unitary matrix $P \in \mathbb{C}^{n\times n}$ such that

$$P^hUP = D = \mathit{diag}\,[\lambda_1, \ldots, \lambda_n]$$

Proof We use induction on n.

For $n = 1$, $U = u$ (a scalar) with $\lambda = u$. Then $U^hU = u^*u = |u|^2 = 1$, and the result is trivially true with $P = 1$ and $D = u$.

Suppose that (a) and (b) are true for all unitary matrices of order $n-1$, and consider a unitary matrix $U = U_1$ of order n. Let λ_1 be an eigenvalue of U, and let $\boldsymbol{v}_1$ be a unit eigenvector (scaled to have unity norm) associated with λ_1. Choose V_1 such that

$$P_1 = [\boldsymbol{v}_1 \;\; V_1]$$

is a unitary matrix. (Columns of V_1 complete $\{\boldsymbol{v}_1\}$ to an orthonormal basis for $\mathbb{C}^{n\times 1}$.) Then

$$\boldsymbol{v}_1^h V_1 = \mathbf{0}_{1\times(n-1)} \quad \text{and} \quad V_1^h \boldsymbol{v}_1 = \mathbf{0}_{(n-1)\times 1}$$

from which we obtain

$$\begin{aligned} P_1^h U_1 P_1 &= \begin{bmatrix} \boldsymbol{v}_1^h \\ V_1^h \end{bmatrix} U_1 \begin{bmatrix} \boldsymbol{v}_1 & V_1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^h U_1 \boldsymbol{v}_1 & \boldsymbol{v}_1^h U_1 V_1 \\ V_1^h U_1 \boldsymbol{v}_1 & V_1^h U_1 V_1 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 \boldsymbol{v}_1^h \boldsymbol{v}_1 & \lambda_1^* \boldsymbol{v}_1^h V_1 \\ \lambda_1 V_1^h \boldsymbol{v}_1 & V_1^h U_1 V_1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & U_2 \end{bmatrix} \end{aligned}$$

Since

$$(P_1^h U_1 P_1)^h (P_1^h U_1 P_1) = P_1^h U_1^h P_1 P_1^h U_1 P_1 = I$$

we have

$$\begin{bmatrix} \lambda_1^* \lambda_1 & \mathbf{0} \\ \mathbf{0} & U_2^h U_2 \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix}$$

which implies that $\lambda_1^* \lambda_1 = 1$, that is, $|\,\lambda_1\,|^2 = 1$, and that U_2 is unitary. Let U_2 have the eigenvalues $\lambda_2, \ldots, \lambda_n$. Then they are also eigenvalues of $U = U_1$. Since U_2 is of order $n-1$, by induction hypothesis $|\,\lambda_2\,|^2 = \cdots = |\,\lambda_n\,|^2 = 1$ and there exists a unitary matrix P_2 such that

$$P_2^h U_2 P_2 = D_2 = \text{diag}\,[\,\lambda_2, \ldots, \lambda_n\,]$$

Let

$$P = P_1 \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix}$$

Then

$$\begin{aligned} P^h U P &= \begin{bmatrix} 1 & \\ & P_2^h \end{bmatrix} P_1^h U_1 P_1 \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \\ & P_2^h \end{bmatrix} \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & U_2 \end{bmatrix} \begin{bmatrix} 1 & \\ & P_2 \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & \mathbf{0} \\ \mathbf{0} & D_2 \end{bmatrix} = D_1 \end{aligned}$$

Theorem 8.1 simply states that eigenvalues of a unitary (orthogonal) matrix are located on the unit circle in the complex plane, that such a matrix can always be diagonalized (even if it has multiple eigenvalues), and that a modal matrix can be chosen to be unitary (orthogonal).

Example 8.2

The matrix

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

is unitary as

$$U^h U = \frac{1}{2} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Its characteristic equation

$$s^2 - \sqrt{2}s + 1 = 0$$

gives the eigenvalues $\lambda_{1,2} = (1 \mp i)/\sqrt{2}$. We observe that $|\lambda_{1,2}| = 1$ as expected. An eigenvector associated with λ_1 is found by solving

$$(U - \lambda_1 I)\boldsymbol{v} = \frac{1}{\sqrt{2}} \begin{bmatrix} -i & i \\ i & -i \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

as

$$\boldsymbol{v}_1 = \text{col}\,[1, 1]$$

Similarly, an eigenvector associated with λ_2 is found by solving

$$(U - \lambda_2 I)\boldsymbol{v} = \frac{1}{\sqrt{2}} \begin{bmatrix} i & i \\ i & i \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

as

$$\boldsymbol{v}_1 = \text{col}\,[-1, 1]$$

Note that we need not look specifically for an eigenvector $\boldsymbol{v}_2$ that is orthogonal to $\boldsymbol{v}_1$; eigenvectors of a unitary matrix associated with distinct eigenvalues are orthogonal (see Exercise 8.11).

Normalizing the eigenvectors, we obtain a unitary modal matrix

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

The reader can easily verify that

$$P^h U P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1+i & \\ & 1-i \end{bmatrix}$$

8.2 Hermitian Matrices

Recall that a matrix $H \in \mathbb{C}^{n\times n}$ is called Hermitian if $H^h = H$, and that a real Hermitian matrix is symmetric.

The following theorem characterizes structure of Hermitian matrices.

Theorem 8.2 *Let $H \in \mathbb{C}^{n\times n}$ be a Hermitian matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

a) $\lambda_i^* = \lambda_i, i = 1, \ldots, n$, that is, eigenvalues of A are real

b) there exists a unitary matrix $P \in \mathbb{C}^{n\times n}$ such that

$$P^h H P = D = \textit{diag}\,[\lambda_1, \ldots, \lambda_n]$$

Proof If $\boldsymbol{v}$ is a unit eigenvector of A associated with an eigenvalue λ, then

$$H\boldsymbol{v} = \lambda\boldsymbol{v}$$

and

$$\boldsymbol{v}^h H = \boldsymbol{v}^h H^h = (H\boldsymbol{v})^h = (\lambda\boldsymbol{v})^h = \lambda^*\boldsymbol{v}^h$$

Premultiplying both sides of the first equality by $\boldsymbol{v}^h$, postmultiplying both sides of the second equality by $\boldsymbol{v}$, and noting that $\boldsymbol{v}^h\boldsymbol{v} = \| \boldsymbol{v} \|^2 = 1$, we get

$$\boldsymbol{v}^h H \boldsymbol{v} = \lambda = \lambda^*$$

Hence all eigenvalues of H are real.

The existence of a unitary modal matrix P that diagonalizes H can be shown by following almost the same lines as in the proof of Theorem 8.1, and is left to the reader as an exercise.

Hence, like unitary matrices, Hermitian (symmetric) matrices can always be diagonalized by means of a unitary (orthogonal) modal matrix.

Example 8.3

The real symmetrix matrix

$$S = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

has the characteristic polynomial $d(s) = (s-1)^2(s-7)$. We observe that the eigenvalues are real.

Two linearly independent eigenvectors associated with the multiple eigenvalue $\lambda_1 = 1$ can be found by solving

$$(S - \lambda_1 I)\boldsymbol{v} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \boldsymbol{v} = \boldsymbol{0}$$

as

$$\boldsymbol{v}_{11} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}, \quad \boldsymbol{v}_{22} = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$$

Applying the Gram-Schmidt process to $\{\boldsymbol{v}_{11}, \boldsymbol{v}_{12}\}$, and normalizing the orthogonal eigenvector generated by the process, we obtain two orthonormal eigenvectors associated with $\lambda_1 = 1$ as

$$\boldsymbol{u}_{11} = \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_{12} = \frac{1}{\sqrt{30}} \begin{bmatrix} -2 \\ -1 \\ 5 \end{bmatrix}$$

An eigenvector associated with $\lambda_2 = 7$ is found as

$$\boldsymbol{v}_2 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

Like the eigenvectors of a unitary matrix, eigenvectors of a Hermitian matrix associated with distinct eigenvalues are also orthogonal (see Exercise 8.11). Therefore, we need not specifically look for an eigenvector $\boldsymbol{v}_2$ that is orthogonal to $\boldsymbol{v}_{11}$ and $\boldsymbol{v}_{12}$. After normalizing $\boldsymbol{v}_2$, we obtain a unit eigenvector associated with $\lambda_2 = 7$ as

$$\boldsymbol{u}_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$$

The reader can verify that the modal matrix

$$P = [\, \boldsymbol{u}_{11} \;\; \boldsymbol{u}_{12} \;\; \boldsymbol{u}_2 \,] = \begin{bmatrix} -\frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{30}} & \frac{2}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & -\frac{1}{\sqrt{30}} & \frac{1}{\sqrt{6}} \\ 0 & \frac{5}{\sqrt{30}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

is orthogonal and that

$$P^t S P = \text{diag}\,[\,1, 1, 7\,]$$

8.3 Quadratic Forms

8.3.1 Real Quadratic Forms

Let $S \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$. An expression of the form

$$q(\boldsymbol{x}) = \boldsymbol{x}^t S \boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} x_i x_j \tag{8.1}$$

is called a ***quadratic form*** in $\boldsymbol{x}$. Note that $q(\boldsymbol{x})$ is a scalar for every $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$.

Clearly $q(\boldsymbol{0}) = 0$. If $q(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$, then $q(\boldsymbol{x})$ is said to be ***positive definite***. If $q(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$, and $q(\boldsymbol{y}) = 0$ for at least one $\boldsymbol{y} \neq \boldsymbol{0}$, then $q(\boldsymbol{x})$ is said to be ***positive semi-definite***. $q(\boldsymbol{x})$ is said to be negative definite (negative semi-definite) if $-q(\boldsymbol{x})$ is positive definite (positive semi-definite), and indefinite if it is neither positive definite nor negative definite.

A real symmetric matrix S is said to be positive definite (positive semi-definite, negative definite, negative semi-definite, indefinite) if the associated quadratic form $q(\boldsymbol{x}) = \boldsymbol{x}^t S \boldsymbol{x}$ is positive definite (positive semi-definite, negative definite, negative semi-definite, indefinite).

Example 8.4

The quadratic form

$$q_1(x_1, x_2) = x_1^2 + 2x_1x_2 + 4x_2^2$$

involving the real variables x_1 and x_2 can be written as

$$q_1(\boldsymbol{x}) = [\, x_1 \;\; x_2 \,] \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \boldsymbol{x}^t S_1 \boldsymbol{x}$$

Note that the diagonal elements of S are the coefficients of the square terms x_1^2 and x_2^2, and the symmetrically located off-diagonal elements are one half of the coefficient of the cross-product term x_1x_2. Since

$$q_1(x_1, x_2) = (x_1 + x_2)^2 + 3x_2^2$$

q_1 is positive definite ($q_1 \geq 0$, and $q_1 = 0$ implies $x_1 = x_2 = 0$). Therefore, the symmetric matrix

$$S_1 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

is also positive definite.

The quadratic form

$$q_2(x_1, x_2) = x_1^2 + 4x_1x_2 + 4x_2^2 = (x_1 + 2x_2)^2$$

is positive semi-definite, because $q_2 \geq 0$ and $q_2 = 0$ for any $x_1 = -2x_2 \neq 0$. Hence the matrix of q_2

$$S_2 = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

is also positive semi-definite.

The quadratic form

$$q_3(x_1, x_2) = x_1^2 + 6x_1x_2 + 4x_2^2$$

is indefinite, because $q_3(1, 0) = 1 > 0$ and $q_3(1, -1) = -1 < 0$. Thus its matrix

$$S_3 = \begin{bmatrix} 1 & 3 \\ 3 & 4 \end{bmatrix}$$

is indefinite.

In Example 8.4 we established positive definiteness of a quadratic form by expressing it as a linear combination of square terms, which is not always as easy as it was in this example. A systematic way of testing sign properties of a quadratic form is based on Theorem 8.2, and is described below.

Since $P^tSP = D$ for some orthogonal matrix P, a change of the variables as

$$\boldsymbol{x} = P\tilde{\boldsymbol{x}} \tag{8.2}$$

transforms the quadratic form in (8.1) into

$$q(\boldsymbol{x}) = \tilde{\boldsymbol{x}}^t P^t S P \tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}^t D \tilde{\boldsymbol{x}} = \tilde{q}(\tilde{\boldsymbol{x}}) \tag{8.3}$$

Since P is nonsingular, it represents a change of basis in $\mathbb{R}^{n\times 1}$. Therefore, q and $\tilde{q}$ are equivalent and thus have the same sign property. Also, since

$$\tilde{q}(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{x}}^t D \tilde{\boldsymbol{x}} = \sum_{i=1}^{n} \lambda_i \tilde{x}_i^2$$

sign of $\tilde{q}$ is completely determined by the eigenvalues λ_i of S. We conclude that a symmetric matrix (whose eigenvalues are real) is positive (negative) definite if and only if all eigenvalues are positive (negative), positive (negative) semi-definite if and only if all eigenvalues are nonnegative (nonpositive) and at least one eigenvalue is zero, and indefinite if and only if it has both positive and negative eigenvalues.

Example 8.5

The matrix S_1 in Example 8.4 has the eigenvalues

$$\lambda_1 = (5 + \sqrt{13})/2 \approx 4.3028\,, \quad \lambda_2 = (5 - \sqrt{13})/2 \approx 0.6972$$

Since both eigenvalues are positive, S_1 and hence q_1 are positive definite.

S_2 has the eigenvalues

$$\lambda_1 = 4\,, \quad \lambda_2 = 0$$

and therefore it is positive semi-definite.

The indefinite matrix S_3 has a positive and a negative eigenvalue:

$$\lambda_1 = (5 + \sqrt{45})/2 \approx 5.8541\,, \quad \lambda_2 = (5 - \sqrt{45})/2 \approx -0.8541$$

8.3.2 Bounds of Quadratic Forms

Let $S \in \mathbb{R}^{n\times n}$ be a symmetric matrix with real eigenvalues $\lambda_1, \ldots, \lambda_n$, and associated orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ that form a basis for $\mathbb{R}^{n\times 1}$. For an arbitrary $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ expressed as

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i$$

we have

$$\| \boldsymbol{x} \|^2 = \sum_{i=1}^{n} \| \alpha_i \boldsymbol{v}_i \|^2 = \sum_{i=1}^{n} | \alpha_i |^2$$

Considering

$$\boldsymbol{x}^t S \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \boldsymbol{v}_i^t S \boldsymbol{v}_j = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \lambda_i \boldsymbol{v}_i^t \boldsymbol{v}_j = \sum_{i=1}^{n} \lambda_i | \alpha_i |^2$$

we find that

$$\lambda_{\min} \sum_{i=1}^{n} | \alpha_i |^2 \leq \boldsymbol{x}^t S \boldsymbol{x} \leq \lambda_{\max} \sum_{i=1}^{n} | \alpha_i |^2$$

or equivalently

$$\lambda_{\min} \| \boldsymbol{x} \|^2 \leq \boldsymbol{x}^t S \boldsymbol{x} \leq \lambda_{\max} \| \boldsymbol{x} \|^2 \tag{8.4}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and and maximum eigenvalues of S. Clearly, equality on either side holds if $\boldsymbol{x}$ equals the corresponding eigenvector. (8.4) establishes bounds on a quadratic form $q(\boldsymbol{x}) = \boldsymbol{x}^t S \boldsymbol{x}$. It also provides an alternative explanation to the relation between sign-definiteness of a quadratic form and the eigenvalues of its symmetric matrix.

8.3.3 Quadratic Forms in Complex Variables

A quadratic form can also be formed by a complex vector $\boldsymbol{z} \in \mathbb{C}^{n\times 1}$ just by replacing the real symmetric matrix S in (8.1) by a complex Hermitian matrix H. Thus a quadratic form in $\boldsymbol{z} \in \mathbb{C}^{n\times 1}$ is

$$q(\boldsymbol{z}) = \boldsymbol{z}^h H \boldsymbol{z} = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} z_i^* z_j \tag{8.5}$$

where $H \in \mathbb{C}^{n\times n}$ is Hermitian. Although $\boldsymbol{z}$ is complex, since

$$q^*(\boldsymbol{z}) = (\boldsymbol{z}^h H \boldsymbol{z})^h = \boldsymbol{z}^h H^h \boldsymbol{z} = \boldsymbol{z}^h H \boldsymbol{z} = q(\boldsymbol{z})$$

the quadratic form in (8.5) is real. This allows us to extend the definitions of definite and semi-definite quadratic forms in real variables and real symmetric matrices to quadratic forms in complex variables and complex Hermitian matrices.

Example 8.6

Consider the quadratic form

$$q(z_1, z_2) = \boldsymbol{z}^h H \boldsymbol{z} = [\, z_1^* \;\; z_2^* \,] \begin{bmatrix} 1 & i \\ -i & 2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1^* z_1 + i z_1^* z_2 - i z_1 z_2^* + 2 z_2^* z_2$$

in the complex variables z_1 and z_2.

Rewriting the quadratic form as

$$q(z_1, z_2) = (z_1 + i z_2)^* (z_1 + i z_2) + z_2^* z_2 = |\, z_1 + i z_2 \,|^2 + |\, z_2 \,|^2$$

we observe that q is positive definite. Thus the Hermitian matrix

$$H = \begin{bmatrix} 1 & i \\ -i & 2 \end{bmatrix}$$

of the quadratic form is also positive definite.

Note that by letting $z_1 = x_1 + i y_1$ and $z_2 = x_2 + i y_2$, we can express q as

$$\begin{aligned} q(z_1, z_2) &= |\, (x_1 - y_2) + i(x_2 + y_1) \,|^2 + |\, z_2 \,|^2 \\ &= (x_1 - y_2)^2 + (x_2 + y_1)^2 + x_2^2 + y_2^2 \\ &= x_1^2 + y_1^2 - 2x_1 y_2 + 2x_2 y_1 + 2x_2^2 + 2y_2^2 \\ &= Q(x_1, x_2, y_1, y_2) \end{aligned}$$

which is a quadratic form in real variables.

Expressions similar to (8.3) and (8.4) can easily be derived for a quadratic form in complex variables. With $\boldsymbol{z} = P\tilde{\boldsymbol{z}}$, where P is a unitary modal matrix of H, the quadratic form $q(\boldsymbol{z}) = \boldsymbol{z}^h H \boldsymbol{z}$ is transformed into

$$q(\boldsymbol{z}) = \tilde{\boldsymbol{z}}^h P^h H P \tilde{\boldsymbol{z}} = \tilde{\boldsymbol{z}}^h D \tilde{\boldsymbol{z}} = \tilde{q}(\tilde{\boldsymbol{z}}) = \sum_{i=1}^{n} \lambda_i |\, \tilde{z}_i \,|^2 \tag{8.6}$$

Again, the sign properties of $q(\boldsymbol{z})$ can be deduced from the eigenvalues of H. For example, the Hermitian matrix of the quadratic form in Example 8.6 has the real eigenvalues

$$\lambda_1 = (3 + \sqrt{5})/2 \approx 2.6180\,, \quad \lambda_1 = (3 - \sqrt{5})/2 \approx 0.3280$$

Since both eigenvalues are positive we conclude that the quadratic form (and hence its matrix) are positive definite.

Similarly, for $q(\boldsymbol{z}) = \boldsymbol{z}^h H \boldsymbol{z}$ (8.4) becomes

$$\lambda_{\min} \| \boldsymbol{z} \|^2 \leq \boldsymbol{z}^h H \boldsymbol{z} \leq \lambda_{\max} \| \boldsymbol{z} \|^2 \tag{8.7}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and and maximum eigenvalues of H.

Finally, we observe from Example 8.6 that a quadratic form in complex variables is equivalent to a quadratic form in real variables (which are the real and imaginary parts of the complex variables). To prove this statement in general, let

$$\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}\,, \quad H = S + iK$$

in (8.5), where $\boldsymbol{x}$ and $\boldsymbol{y}$ are real, S is a real symmetric matrix, and K is a real skew-symmetric matrix. Then, noting that

$$\boldsymbol{x}^t K \boldsymbol{x} = \boldsymbol{y}^t K \boldsymbol{y} = 0\,, \quad \boldsymbol{x}^t S \boldsymbol{y} = \boldsymbol{y}^t S \boldsymbol{x}$$

$q(\boldsymbol{z})$ can be expressed as

$$\begin{aligned} q(\boldsymbol{z}) &= (\boldsymbol{x}^t - i\boldsymbol{y}^t)(S + iK)(\boldsymbol{x} + i\boldsymbol{y}) \\ &= \boldsymbol{x}^t S \boldsymbol{x} - \boldsymbol{x}^t K \boldsymbol{y} + \boldsymbol{y}^t K \boldsymbol{x} + \boldsymbol{y}^t S \boldsymbol{y} \\ &= [\,\boldsymbol{x}^t \;\; \boldsymbol{y}^t\,] \begin{bmatrix} S & -K \\ K & S \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \;=\; Q(\boldsymbol{x}, \boldsymbol{y}) \end{aligned} \tag{8.8}$$

involving real quantities only.

Since the quadratic forms $q(\boldsymbol{x})$ and $Q(\boldsymbol{y}, \boldsymbol{z})$ in (8.8) are equivalent, the eigenvalues of their matrices $H = S + iK$ and

$$\tilde{H} = \begin{bmatrix} S & -K \\ K & S \end{bmatrix} \tag{8.9}$$

must be related. This relation is studied in Exercise 8.12.

8.3.4 Conic Sections and Quadric Surfaces

Recall from analytic geometry that an equation of the form

$$s_{11}x_1^2 + 2s_{12}x_1x_2 + s_{22}x_2^2 = 1$$

where not all coefficients are zero, defines a central conic in the x_1x_2 plane. A suitable way to investigate the properties of such a conic is to rewrite the defining equation in compact form as

$$\boldsymbol{x}^t S \boldsymbol{x} = 1 \tag{8.10}$$

with the obvious definitions of $\boldsymbol{x}$ and S.

Let S have the eigenvalues $\lambda_1 \geq \lambda_2$ and an orthogonal modal matrix P such that

$$P^t S P = D = \text{diag}\,[\,\lambda_1, \lambda_2\,]$$

Then a change of coordinate system as $\boldsymbol{x} = P\tilde{\boldsymbol{x}}$ transforms the equation of the conic into

$$\tilde{\boldsymbol{x}}^t D \tilde{\boldsymbol{x}} = \lambda_1 \tilde{x}_1^2 + \lambda_2 \tilde{x}_2^2 = 1$$

Depending on the signs of the (real) eigenvalues λ_1 and λ_2, we consider the following distinct cases:

a) $\lambda_1 \geq \lambda_2 > 0$ (S positive definite): Letting $a_1 = 1/\sqrt{\lambda_1},\ a_2 = 1/\sqrt{\lambda_2}$, the equation of the conic takes the form

$$\frac{\tilde{x}_1^2}{a_1^2} + \frac{\tilde{x}_2^2}{a_2^2} = 1$$

which represents an ellipse in the $\tilde{x}_1\tilde{x}_2$ plane with axes of length a_1 and a_2.

b) $\lambda_1 > \lambda_2 = 0$ (S positive semi-definite): Again, letting $a_1 = 1/\sqrt{\lambda_1}$, the equation becomes

$$\tilde{x}_1^2 = a_1^2$$

which represents two straight lines $\tilde{x}_1 = \mp a_1$.

c) $\lambda_1 > 0 > \lambda_2$ (S indefinite): With $a_1 = 1/\sqrt{\lambda_1},\ a_2 = 1/\sqrt{-\lambda_2}$, we have

$$\frac{\tilde{x}_1^2}{a_1^2} - \frac{\tilde{x}_2^2}{a_2^2} = 1$$

which represents a hyperbola.

d) $0 \geq \lambda_1 \geq \lambda_2$ (S negative definite or negative semi-definite): In these cases no point in the $\tilde{x}_1\tilde{x}_2$ plane (and therefore, no point in the x_1x_2 plane) satisfies the equation.

Note that the parabola, which is another conic, does not occur in any of the cases considered above, because it is not a central conic. A more general equation, which also includes the parabola, is considered in Exercise 8.28.

Example 8.7

Consider the equation

$$(10-c)x_1^2 + 2(6+c)x_1x_2 + (10-c)x_2^2 = \boldsymbol{x}^t S\boldsymbol{x} = 1$$

where c is a real parameter, and

$$S = \begin{bmatrix} 10-c & 6+c \\ 6+c & 10-c \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The matrix S has the eigenvalues

$$\lambda_1 = 16\,, \quad \lambda_2 = 4 - 2c$$

and an orthogonal modal matrix

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

which is independent of the parameter c. Thus the change of basis as in (8.2) transforms the equation of the conic into

$$16\tilde{x}_1^2 + (4-2c)\tilde{x}_2^2 = 1$$

For $c = 0, 2, 4$, the equation becomes and represents

$$\begin{aligned} c = 0: &\quad 16\tilde{x}_1^2 + 4\tilde{x}_2^2 = 1 &:& \text{ an elipse} \\ c = 2: &\quad 16\tilde{x}_1^2 = 1 &:& \text{ two parallel lines} \\ c = 4: &\quad 16\tilde{x}_1^2 - 4\tilde{x}_2^2 = 1 &:& \text{ a hyperbola} \end{aligned}$$

The loci of the points that satisfy the given equation for each case are shown in Figure 8.1. Note that the transformation in (8.2) corresponds to a counter-clock-wise rotation of the coordinate axes by 45°.

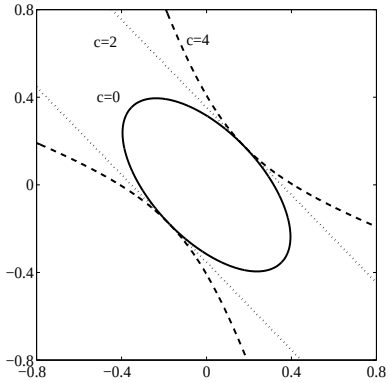

Figure 8.1: Conics in Example 8.7

A three-dimensional version of (8.10) defines a ***quadric surface*** in the $x_1x_2x_3$ space. As in the two-dimensional case, a change of coordinate system as $\boldsymbol{x} = P\tilde{\boldsymbol{x}}$ transforms the equation of the quadric into

$$\tilde{\boldsymbol{x}}^t D\tilde{\boldsymbol{x}} = \lambda_1\tilde{x}_1^2 + \lambda_2\tilde{x}_2^2 + \lambda_3\tilde{x}_3^2 = 1$$

where, without loss of generality, we may assume $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Then we have the following distinct cases:

a) $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$ (S positive definite): The quadric equation can be expressed as

$$\frac{\tilde{x}_1^2}{a_1^2} + \frac{\tilde{x}_2^2}{a_2^2} + \frac{\tilde{x}_3^2}{a_3^2} = 1$$

which represents an ellipsoid with axes of length a_1, a_2 and a_3.

b) $\lambda_1 \geq \lambda_2 > \lambda_3 = 0$ (S positive semi-definite): Now the equation becomes

$$\frac{\tilde{x}_1^2}{a_1^2} + \frac{\tilde{x}_2^2}{a_2^2} = 1$$

which represents an elliptic cylinder.

c) $\lambda_1 \geq \lambda_2 > 0 > \lambda_3$ (S indefinite): We have a hyperboloid of one sheet described as

$$\frac{\tilde{x}_1^2}{a_1^2} + \frac{\tilde{x}_2^2}{a_2^2} - \frac{\tilde{x}_3^2}{a_3^2} = 1$$

d) $\lambda_1 > \lambda_2 = \lambda_3 = 0$ (S positive semi-definite): The equation becomes

$$\frac{\tilde{x}_1^2}{a_1^2} = 1$$

which represents two parallel planes.

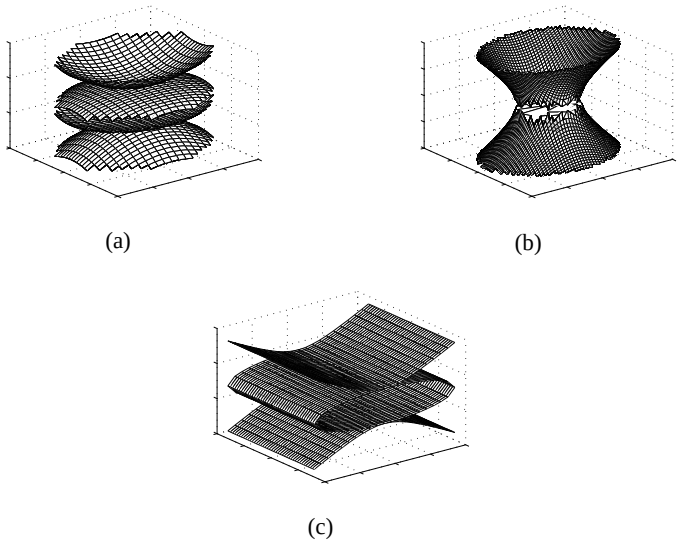

Figure 8.2: Central quadric surfaces

e) $\lambda_1 > \lambda_2 = 0 > \lambda_3$ (S indefinite): We have a hyperbolic cylinder described as

$$\frac{\tilde{x}_1^2}{a_1^2} - \frac{\tilde{x}_3^2}{a_3^2} = 1$$

f) $\lambda_1 > 0 > \lambda_2 \geq \lambda_3$ (S indefinite): The equation

$$\frac{\tilde{x}_1^2}{a_1^2} - \frac{\tilde{x}_2^2}{a_2^2} - \frac{\tilde{x}_3^2}{a_3^2} = 1$$

represents a hyperboloid of two sheets.

g) $0 \geq \lambda_1 \geq \lambda_2 \geq \lambda_3$ (S negative definite or negative semi-definite): The equation is never satisfied.

The quadric surfaces corresponding to the above cases are illustrated in Figure 8.2, where (a) contains an ellipsoid and a hyperboloid of two sheets, (b) a hyperboloid of one sheet, and (c) an elliptic and a hyperbolic cylinder.

8.4 The Singular Value Decomposition

Many applications of linear algebra require knowledge of the rank of a matrix, construction of bases for its row and column spaces or their orthogonal complements, or computation of projections onto these subspaces. Such applications usually involve matrices whose elements are given only approximately (e.g., as a result of some measurement). In such cases, approximate answers, together with a measure of approximation, make more sense than exact

answers. For example, determining whether a given matrix is close (according to a specified measure) to a matrix of defective rank may be more significant than computing the rank of the given matrix itself. In theory, the rank of a matrix can easily be computed using the Gaussian elimination. However, as we noted in Example 4.4, it is not reliable when the matrix has nearly linearly dependent rows (or columns). This may pose serious problems in practical situations.

Singular value decomposition (SVD) is a computationally reliable method of transforming a given matrix into a simple form, from which its rank, bases for its column and row spaces and their orthogonal complements and projections onto these subspaces can be computed easily. In the following we shall first prove the SVD Theorem and then study its uses.

8.4.1 The Singular Value Decomposition Theorem

Theorem 8.3 (The Singular Value Decomposition) *Given $A \in \mathbb{C}^{m\times n}$, there exist unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that*

$$U^h A V = \Sigma = \begin{bmatrix} \Sigma_1 & O \\ O & O \end{bmatrix} \tag{8.11}$$

where

$$\Sigma_1 = \mathit{diag}\,[\,\sigma_1, \ldots, \sigma_k\,]$$

with $\sigma_1 \geq \cdots \geq \sigma_k > 0$ for some $k \leq m, n$.

Proof $A^h A \in \mathbb{C}^{n\times n}$ is Hermitian, and since

$$\boldsymbol{x}^h A^h A \boldsymbol{x} = (A\boldsymbol{x})^h (A\boldsymbol{x}) = \| A\boldsymbol{x} \|^2 \geq 0 \quad \text{for all } \boldsymbol{x}$$

it is at least positive semi-definite. Let the eigenvalues of $A^h A$ be

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_k^2 > 0 = \sigma_{k+1}^2 = \cdots = \sigma_n^2$$

and let

$$V = [\,V_1 \;\; V_2\,]$$

be a unitary modal matrix of $A^h A$, where V_1 and V_2 consist of the orthonormal eigenvectors associated with the nonzero and zero eigenvalues. (If $A^h A$ is positive definite then $k = n$, and $V = V_1$.) Then

$$A^h A V_1 = V_1 \Sigma_1^2 \implies V_1^h A^h A V_1 = \Sigma_1^2$$

and

$$A^h A V_2 = O \implies V_2^h A^h A V_2 = O \implies A V_2 = O$$

Let

$$U_1 = A V_1 \Sigma_1^{-1}$$

Since

$$U_1^h U_1 = \Sigma_1^{-1} V_1^h A^h A V_1 \Sigma_1^{-1} = \Sigma_1^{-1} \Sigma_1^2 \Sigma_1^{-1} = I$$

columns of U_1 are orthonormal. Choose U_2 such that

$$U = [\, U_1 \;\; U_2 \,]$$

is unitary. (Columns of U_2 complete the columns of U_1 to an orthonormal basis for $\mathbb{C}^{m\times 1}$.) Then

$$\begin{aligned} U^h AV &= \begin{bmatrix} U_1^h AV_1 & U_1^h AV_2 \\ U_2^h AV_1 & U_2^h AV_2 \end{bmatrix} = \begin{bmatrix} \Sigma_1^{-1} V_1^h A^h AV_1 & O \\ U_2^h U_1 \Sigma_1 & O \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_1 & O \\ O & O \end{bmatrix} = \Sigma \end{aligned}$$

The non-negative scalars $\sigma_i, i = 1, \ldots, n$, are called the ***singular values*** of A, and the columns of V and U are called the right and left ***singular vectors*** of A, respectively. Although the proof of Theorem 8.3 provides a constructive method, in practice, the singular value decomposition of A is obtained by means of a different algorithm which involves unitary transformations that do not require computation of the eigenvalues and eigenvectors of $A^h A$. The MATLAB command `[U,S,V]=svd(A)` that computes the singular value decomposition of A uses such an algoritm.

The following results are immediate consequences of Theorem 8.3:

a) $k = r$, that is, the number of nonzero singular values is the rank of A.

b) $A \in \mathbb{C}^{n\times n}$ is nonsingular if and only if all its singular values are positive.

c) The right singular vectors of A are the eigenvectors of $A^h A$.

d) The left singular vectors of A are the eigenvectors of AA^h.

e) $A^h A$ and AA^h have the same nonzero eigenvalues, $\sigma_1, \ldots, \sigma_k$.

f) If $A = U\Sigma V^h$ then $A^h = V\Sigma^h U^h$. Thus A and A^h have the same nonzero singular values.

g) If $A \in \mathbb{C}^{n\times n}$ is Hermitian with (real) eigenvalues $\lambda_i, i = 1, \ldots, n$, then its singular values are $\sigma_i = |\, \lambda_i \,|$, $i = 1, \ldots, n$.

(a) follows from (8.11) on noting that U and V are nonsingular, and (b) is a direct consequence of (a). (c) is the definition used in the proof of Theorem 8.3. (d) follows from

$$AA^h U = U\Sigma V^h V\Sigma^h U^h U = U\Sigma\Sigma^h$$

(e) and (f) are obvious. Finally, (g) is a result of the fact that if A is Hermitian with eigenvalues λ_i then $A^h A = A^2$ has the eigenvalues λ_i^2, so that singular values of A are $\sigma_i = \sqrt{\lambda_i^2} = |\, \lambda_i \,|$.

The following corollary of Theorem 8.3 characterizes various subspaces associated with a matrix. The proof is left to the reader (see Exercise 8.20).

Corollary 8.3.1 *Let $A \in \mathbb{C}^{m\times n}$ have the singular value decomposition*

$$A = U\Sigma V^h = [\, U_1 \;\; U_2 \,] \begin{bmatrix} \Sigma_1 & O \\ O & O \end{bmatrix} \begin{bmatrix} V_1^h \\ V_2^h \end{bmatrix} = U_1 \Sigma_1 V_1^h$$

Then

a) $im\,(U_1) = im\,(A)$

b) $im\,(U_2) = ker\,(A^h)$

c) $im\,(V_1) = im\,(A^h)$

d) $im\,(V_2) = ker\,(A)$

Thus $U_1U_1^h$ is the orthogonal projection matrix on $\text{im}\,(A)$, $U_2U_2^h$ is the orthogonal projection matrix on $\ker(A^h)$, etc.

As a consequence of Corollary 8.3.1 we also have

$$\mathbb{C}^{m\times 1} = \text{im}\,(U_1) \overset{\perp}{\oplus} \text{im}\,(U_2) = \text{im}\,(A) \overset{\perp}{\oplus} \ker\,(A^h)$$

and the dual relation

$$\mathbb{C}^{n\times 1} = \text{im}\,(V_1) \overset{\perp}{\oplus} \text{im}\,(V_2) = \text{im}\,(A^h) \overset{\perp}{\oplus} \ker\,(A)$$

where the symbol $\overset{\perp}{\oplus}$ denotes a direct sum decomposition into orthogonal subspaces.

Example 8.8

Let us find the singular value decomposition of

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 0 & 2 \end{bmatrix}$$

Eigenvalues of

$$A^tA = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$$

are $\lambda_1 = 6$ and $\lambda_2 = 4$. Hence

$$\sigma_1 = \sqrt{6}\,, \quad \sigma_2 = 2$$

An orthogonal modal matrix of A^tA can be found as

$$V = V_1 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Letting

$$U_1 = AV_1\Sigma_1^{-1} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} \\ 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & -1/\sqrt{2} \end{bmatrix}$$

and completing the columns of U_1 to an orthonormal basis for $\mathbb{R}^{3\times 1}$, the singular value decomposition of A is obtained as

$$\begin{aligned} A &= U\Sigma V^t \\ &= \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \end{aligned}$$

The reader can verify that

$$AA^t = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & 2 \\ 0 & 2 & 4 \end{bmatrix}$$

has the eigenvalues $\lambda_1 = 6, \lambda_2 = 4, \lambda_3 = 0$, and that columns of U are orthonormal eigenvectors of AA^t.

Finally, from

$$A \xrightarrow{\text{e.c.o.}} \begin{bmatrix} 2 & -2 \\ 2 & 0 \\ 2 & 2 \end{bmatrix} \xrightarrow{\text{e.c.o.}} U_1$$

we observe that $\text{im}\,(A) = \text{im}\,(U_1)$, verifying the result of Corollary 8.3.1(a). Other results of the corollary can be verified similarly.

8.4.2 The Least-Squares Problem and The Pseudoinverse

Corollary 8.1 is especially useful in solving least-squares problems. Recall that the linear system

$$A\boldsymbol{x} = \boldsymbol{y} \tag{8.12}$$

has a solution if and only if $\boldsymbol{y} \in \text{im}\,(A)$. If not, we look for solution(s) of the consistent system

$$A\boldsymbol{x} = \boldsymbol{y}_A \tag{8.13}$$

where $\boldsymbol{y}_A$ is the orthogonal projection of $\boldsymbol{y}$ on $\text{im}\,(A)$. Such solutions are called least-square solutions as they minimize the squared error $\| A\boldsymbol{x} - \boldsymbol{y} \|_2^2$.

Let A have the singular value decomposition

$$A = U\Sigma V^h = U_1\Sigma_1 V_1^h$$

Then the orthogonal projection of $\boldsymbol{y}$ on $\text{im}\,(A)$ is

$$\boldsymbol{y}_A = U_1 U_1^h \boldsymbol{y}$$

and (8.13) becomes

$$U_1\Sigma_1 V_1^h \boldsymbol{x} = U_1 U_1^h \boldsymbol{y} \tag{8.14}$$

It is left to the reader (Exercise 8.22) to show that the above $m \times n$ system is equivalent to the $r \times n$ system

$$V_1^h \boldsymbol{x} = \Sigma_1^{-1} U_1^h \boldsymbol{y} \tag{8.15}$$

In other words, the least-squares solutions of (8.12) are precisely the solutions of (8.15). In particular,

$$\boldsymbol{x} = \boldsymbol{\phi}_{LS} = V_1\Sigma_1^{-1}U_1^h\boldsymbol{y} = A^\dagger \boldsymbol{y} \tag{8.16}$$

is a least-squares solution with minimum 2-norm (Exercise 8.22). Note that $\boldsymbol{\phi}_{LS}$ can be written in open form as

$$\boldsymbol{\phi}_{LS} = \frac{\boldsymbol{u}_1^h \boldsymbol{y}}{\sigma_1}\boldsymbol{v}_1 + \cdots + \frac{\boldsymbol{u}_r^h \boldsymbol{y}}{\sigma_r}\boldsymbol{v}_r \tag{8.17}$$

Consider the matrix

$$A^\dagger = V_1 \Sigma_1^{-1} U_1^h \tag{8.18}$$

that appears in (8.16). Since

$$\begin{aligned} AA^\dagger A &= U_1\Sigma_1 V_1^h V_1 \Sigma_1^{-1} U_1^h U_1 \Sigma_1 V_1^h = U_1 \Sigma_1 V_1^h = A \\ A^\dagger A A^\dagger &= V_1 \Sigma_1^{-1} U_1^h U_1 \Sigma_1 V_1^h V_1 \Sigma_1^{-1} U_1^h = V_1 \Sigma_1^{-1} U_1^h = A^\dagger \end{aligned}$$

it follows that $A^\dagger$ is a generalized inverse of A, called the ***Moore-Penrose generalized inverse*** or the ***pseudoinverse*** of A. An interesting property of $A^\dagger$ is that since

$$\begin{aligned} A^\dagger A &= V_1 \Sigma_1^{-1} U_1^h U_1 \Sigma_1 V_1^h = V_1 V_1^h \\ AA^\dagger &= U_1 \Sigma_1 V_1^h V_1 \Sigma_1^{-1} U_1^h = U_1 U_1^h \end{aligned}$$

$A^\dagger A$ and $AA^\dagger$ are both Hermitian.[1] Moreover, $A^\dagger$ reduces to a left inverse when $r(A) = n$ (in which case $V_1 V_1^h = I_n$), to a right inverse when $r(A) = m$ (in which case $U_1 U_1^h = I_m$), and to A^{-1} when A is square and nonsingular.

Example 8.9

Consider a linear system with

$$A = \begin{bmatrix} 5 & 0 & 5 \\ 1 & 1 & 2 \\ 0 & 5 & 5 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 4 \\ -15 \\ 2 \end{bmatrix}$$

Following the procedure of Section 7.4.2, the orthogonal projection of $\boldsymbol{y}$ on im (A) is computed as

$$R = \begin{bmatrix} 5 & 0 \\ 1 & 1 \\ 0 & 5 \end{bmatrix}, \quad \boldsymbol{y}_A = R(R^t R)^{-1} R^t = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

The general solution of $A\boldsymbol{x} = \boldsymbol{y}_A$ is then obtained as

$$\boldsymbol{x} = \begin{bmatrix} 0.2 \\ -0.2 \\ 0.0 \end{bmatrix} + c \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

which characterizes all least-squares solutions. Among these least-squares solutions

$$\boldsymbol{x} = \boldsymbol{\phi}_{LS} = \begin{bmatrix} 0.2 \\ -0.2 \\ 0.0 \end{bmatrix}$$

[1] In fact, $\hat{A}_G = A^\dagger$ is the unique generalized inverse such that $\hat{A}_G A$ and $A\hat{A}_G$ are both Hermitian. We will not prove this fact.

(corresponding to $c = 0$) has the minimum 2-norm.

The singular value decomposition of A produces

$$\Sigma = \text{diag}\,[\,9, 5, 0\,]$$

$$U = \frac{1}{3\sqrt{6}} \begin{bmatrix} 5 & 3\sqrt{3} & \sqrt{2} \\ 2 & 0 & -5\sqrt{2} \\ 5 & -3\sqrt{3} & \sqrt{2} \end{bmatrix}, \quad V = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & \sqrt{3} & \sqrt{2} \\ 1 & -\sqrt{3} & \sqrt{2} \\ 2 & 0 & -\sqrt{2} \end{bmatrix}$$

Then the pseudoinverse of A is obtained as

$$\begin{aligned} A^\dagger &= \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & \sqrt{3} \\ 1 & -\sqrt{3} \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \frac{1}{3\sqrt{6}} \begin{bmatrix} 5 & 2 & 5 \\ 3\sqrt{3} & 0 & -3\sqrt{3} \end{bmatrix} \\ &= \frac{1}{810} \begin{bmatrix} 106 & 10 & -56 \\ -56 & 10 & 106 \\ 50 & 20 & 50 \end{bmatrix} \end{aligned}$$

and the expression in (8.16) gives the same minimum norm least-squares solution as obtained above. (Alternatively, (8.17) produces the same solution.)

The results can be checked using the MATLAB commands `[U,S,V]=svd(A)`, which produces the singular value decomposition of A, and `API=pinv(A)`, which computes the pseudoinverse of A. The reader is also urged to verify that $A^\dagger A$ and $AA^\dagger$ are both symmetric.

8.4.3 The SVD and Matrix Norms

Recall that

$$\| A \|_p = \max_{\| \boldsymbol{x} \|_p = 1} \{ \| A\boldsymbol{x} \|_p \}$$

is the matrix norm subordinate to the p-vector norm. Since

$$\| A\boldsymbol{x} \|_2^2 = \boldsymbol{x}^h A^h A \boldsymbol{x} \leq \lambda_{\max}(A^h A) \| \boldsymbol{x} \|^2$$

with equality holding for the eigenvector corresponding to $\lambda_{\max}(A^h A)$, we find that

$$\| A \|_2 = \sqrt{\lambda_{\max}(A^h A)} = \sigma_1 \tag{8.19}$$

This is a significant result, which states that the matrix norm subordinate to the Euclidean vector norm is its largest singular value.

An equally significant result, which is left to the reader to prove (Exercise 8.25), is that

$$\| A \|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2} \tag{8.20}$$

To appreciate the significance of (8.20), let

$$A' = U\Sigma' V^h$$

where

$$\Sigma' = \begin{bmatrix} \Sigma_1' & O \\ O & O \end{bmatrix}$$

with $\Sigma_1' = \text{diag}\,[\,\sigma_1, \ldots, \sigma_q\,]$ and $q < r$. Then $r(A') = q$ and

$$\| A - A' \|_F = \sqrt{\sigma_{q+1}^2 + \cdots + \sigma_r^2} \tag{8.21}$$

Moreover, if B is any other matrix with $r(B) = q$ then

$$\| A - B \|_F \geq \| A - A' \|_F$$

Thus A' is the best rank-q approximation to A in Frobenius norm.[2]

Example 8.10

Consider the following matrix generated by the MATLAB command `A=rand(3,3)`.

$$A = \begin{bmatrix} 0.4103 & 0.3529 & 0.1389 \\ 0.8936 & 0.8132 & 0.2028 \\ 0.0579 & 0.0099 & 0.1987 \end{bmatrix}$$

The command `[U,S,V]=svd(A)` produces

$$S = \begin{bmatrix} 1.3485 & 0 & 0 \\ 0 & 0.1941 & 0 \\ 0 & 0 & 0.0063 \end{bmatrix}$$

in addition to U and V that are not shown here. Since all singular values of A are nonzero, we conclude that A is nonsingular, that is, it has rank $r(A) = 3$. However, the third singular value is quite small compared to the others, which suggests that A is close to a rank-2 (i.e., singular) matrix.

Let

$$Q = \begin{bmatrix} 1.3485 & 0 & 0 \\ 0 & 0.1941 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$B = UQV^t = \begin{bmatrix} 0.4065 & 0.3568 & 0.1398 \\ 0.8953 & 0.8114 & 0.2023 \\ 0.0589 & 0.0088 & 0.1985 \end{bmatrix}$$

Then B has rank $r(B) = 2$ (which can be verified by MATLAB), and is the closest rank-2 matrix to A in Frobenius norm. The command `norm(A-B,'fro')` computes the Frobenius norm of the difference of A and B as

$$\| A - B \|_F = 0.0063$$

as expected.

Let A be a nonsingular matrix of order n having a singular value decomposition

$$A = U\Sigma V^h$$

where $\Sigma = \text{diag}\,[\,\sigma_1, \ldots, \sigma_n\,]$ with $\sigma_1 \geq \cdots \geq \sigma_n > 0$. Then

$$A^{-1} = V\Sigma^{-1}U^h \tag{8.22}$$

[2]The proof of this result is beyond the scope of this book.

which shows that A^{-1} has the singular values $\sigma_n^{-1} \geq \cdots \geq \sigma_1^{-1} > 0$.[3] Thus

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

The ratio

$$\mu = \frac{\sigma_1}{\sigma_n} = \| A \|_2 \| A^{-1} \|_2 \geq 1$$

is called the ***condition number*** of A, and is a measure of linear independence of the columns of A. The larger the condition number of a matrix, the closer it is to being singular.

Example 8.11

The matrix

$$A = \begin{bmatrix} 0.2676 & 0.5111 & 0.7627 \\ 0.6467 & 0.6931 & 0.5241 \\ 0.7371 & 0.6137 & 0.1690 \end{bmatrix}$$

has the singular values

$$\sigma_1 = 1.6564\,, \quad \sigma_2 = 0.5411\,, \quad \sigma_3 = 0.0001$$

as calculated and displayed up to the fourth decimal digit by MATLAB.

The wide separation of the singular values indicate that A is nearly singular. Indeed, the condition number of A, computed by the MATLAB command `cond(A)` as [4]

$$\mu = 1.6305e + 004$$

implies that A is badly conditioned, and that A^{-1} has large element values of the order of 10^4. The reader can verify this observation by computing A^{-1} using MATLAB.

8.5 Exercises

1. (a) Verify that the following matrices are unitary.
   (b) Find a unitary modal matrix for each of these matrices and diagonalize them.
   (c) Use MATLAB command `eig` to verify your results in part (b).

$$A = \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

2. Prove that the product of two unitary matrices is unitary. Verify this result for the matrices A an B in Exercise 8.1.

3. In the xyz space a counterclockwise rotation about the z axis by an angle of θ_z is represented by the matrix

$$R_z = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 \\ \sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

[3] The expression in (8.22) is similar to the singular value decomposition of A^{-1} except that its singular values are in ascending order.

[4] The difference between μ and σ_1/σ_3 is due to the limited number of digits used to display the values.

(a) Determine the structure of the rotation matrices R_y and R_x about the y and x axes.

(b) Show that R_x, R_y and R_z are orthogonal.

(c) Characterize an invariant subspace for each of R_x, R_y and R_z.

4. Let

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n \times n}$$

(a) Show that A is orthogonal.

(b) Find the eigenvalues and eigenvectors of A.

(c) Find a unitary modal matrix and the Jordan form of A.

5. Use MATLAB command `eig` to verify your answer to Exercise 8.4 for $n = 2, 3, 4$.

6. Let $\boldsymbol{b}$ be a unit vector. Show that $A = I - \boldsymbol{b}\boldsymbol{b}^t$ is a symmetric projection matrix.

7. Prove the ***Schur's theorem***: *Let A be an $n \times n$ complex matrix. There exists a unitary matrix U such that $U^h AU$ is an upper triangular matrix with diagonal elements being the eigenvalues of A.* Hint: Refer to Exercise 5.13.

8. $A \in \mathcal{C}^{n \times n}$ is said to be an involution if $A^2 = I$. Show that any two of the following imply the third.

(a) A is Hermitian

(b) A is unitary

(c) A is an involution.

9. Verify the result of Exercise 8.8 for the matrices B and C in Exercise 8.1.

10. Prove that a Hermitian matrix A has a unitary modal matrix, and thus complete the proof of Theorem 8.2. Hint: Let $\boldsymbol{v}_1$ be a unit eigenvector of $A = A_1$ associated with some eigenvalue λ_1. Choose V_1 such that $P_1 = [\, \boldsymbol{v}_1 \;\; V_1 \,]$ is unitary, and consider $P_1^h A P_1$.

11. (a) Show that eigenvectors of a unitary matrix associated with distinct eigenvalues are orthogonal.

(b) Show that eigenvectors of a Hermitian matrix associated with distinct eigenvalues are orthogonal.

12. Show that if λ is an eigenvalue of the Hermitian matrix $H = S + iK$ then it is also an eigenvalue of $\tilde{H}$ in (8.9) and vice versa. Hint: Let $\boldsymbol{v} = \boldsymbol{u} + i\boldsymbol{w}$ be an eigenvector of H associated with λ and consider the real and imaginary parts of the expression

$$(\lambda I - H)\boldsymbol{v} = \boldsymbol{0}$$

13. Let S be an $n \times n$ real symmetric matrix with an orthogonal modal matrix Q and the diagonal Jordan form D. Find a modal matrix and the Jordan form of

$$H = \begin{bmatrix} 0 & jS \\ -jS & 0 \end{bmatrix}$$

in terms of Q and D. Hint: Let

$$\tilde{H} = \begin{bmatrix} Q^t & O \\ O & Q^t \end{bmatrix} \begin{bmatrix} 0 & jS \\ -jS & 0 \end{bmatrix} \begin{bmatrix} Q & O \\ O & Q \end{bmatrix} = \begin{bmatrix} 0 & jD \\ -jD & 0 \end{bmatrix}$$

and find eigenvalues and eigenvectors of $\tilde{H}$.

14. An $n \times n$ complex matrix A is said to be ***normal*** if it satisfies $A^h A = AA^h$. Clearly, unitary and Hermitian matrices are normal.
    (a) Show that a normal triangular matrix must be diagonal.
    (b) Prove that A can be diagonalized by a unitary similarity transformation if and only if it is normal. Hint: To prove sufficiency, use the Schur's theorem and the result of part (a).

15. Verify the result of Exercise 8.14 for the matrix

$$A = \begin{bmatrix} 2+i & 1 \\ -1 & 2+i \end{bmatrix}$$

which is neither unitary nor Hermitian.

16. Investigate the sign properties of the following quadratic forms.
    (a) $q(x, y) = 2x^2 + 8xy + 2y^2$
    (b) $q(x_1, x_2, x_3) = 2x_1^2 + x_2^2 + 7x_3^2 - 2x_1x_2 + 2x_1x_3 - 4x_2x_3$
    (c) $q(z_1, z_2) = 2|z_1|^2 + |z_2|^2 + 2\,\mathrm{Im}\,(z_1 z_2^*)$
    (d) $q(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3$

17. Let $A \in \mathbb{C}^{m\times n}$. Prove the following.
    (a) $A^h A$ and AA^h are non-negative-definite.
    (b) $A^h A$ is positive definite if and only if $r(A) = n$.
    (c) AA^h is positive definite if and only if $r(A) = m$.

18. (a) Show that

$$Q = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$$

    is positive definite.
    (b) Show that $\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle_Q = \boldsymbol{x}^t Q \boldsymbol{y}$ is an inner product in $\mathbb{R}^{3\times 1}$.
    (c) Apply GSOP to $(\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3)$ to generate an orthogonal basis for $\mathbb{R}^{3\times 1}$, where orthogonality is defined with respect to $\langle \cdot \,|\, \cdot \rangle_Q$.
    (d) Find the orthogonal projection of $\boldsymbol{e}_3$ on span $(\boldsymbol{e}_1, \boldsymbol{e}_2)$.

19. Obtain singular value decompositions of

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

    (a) by hand,
    (b) by using MATLAB.

20. Prove Corollary 8.1. Hint: $\mathbb{C}^{m\times 1} = \mathrm{im}\,(U_1) \overset{\perp}{\oplus} \mathrm{im}\,(U_2)$ and $\mathbb{C}^{n\times 1} = \mathrm{im}\,(V_1) \overset{\perp}{\oplus} \mathrm{im}\,(V_2)$.

21. Let

$$A = \begin{bmatrix} -1 & i \\ -i & -1 \end{bmatrix}$$

    Obtain the singular value decompositions of A and A^{100}.

22. (a) Show that the systems in (8.14) and (8.15) have the same solution(s).

    (b) Let $\boldsymbol{x} = \boldsymbol{\phi}$ be any solution of a consistent linear system $A\boldsymbol{x} = \boldsymbol{y}$, and let $\boldsymbol{\phi}_0$ be the orthogonal projection of $\boldsymbol{\phi}$ on $\ker(A)$. Prove that $\boldsymbol{x} = \boldsymbol{\phi} - \boldsymbol{\phi}_0$ is the unique minimum 2-norm solution of $A\boldsymbol{x} = \boldsymbol{y}$.

    (c) Prove that $\boldsymbol{x} = \boldsymbol{\phi}_{LS} = V_1\Sigma_1^{-1}U_1^h\boldsymbol{y}$ is the minimum 2-norm least-squares solution of $A\boldsymbol{x} = \boldsymbol{y}$.

23. (a) Find all least-squares solutions of $A\boldsymbol{x} = \boldsymbol{y}$, where

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \\ 0 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 9 \\ -9 \\ 9 \\ -9 \end{bmatrix}$$

    (b) Among all least-squares solutions find the one with minimum Euclidean norm.

    (c) Use the MATLAB command `x=pinv(C)*y` to verify your answer in part (b).

24. Show that the minimum 2-norm least squares solution to $C\boldsymbol{x} = \boldsymbol{y}$, where C is as in Exercise 8.19 and $\boldsymbol{y} = \text{col}\,[\,y_1, y_2, y_3, y_4\,]$ is given by

$$\boldsymbol{x} = \boldsymbol{\phi}_{LS} = \frac{1}{6}\begin{bmatrix} 2y_1 - y_2 + 2y_3 - y_4 \\ y_1 + y_2 + y_3 + y_4 \\ -y_1 + 2y_2 - y_3 + 2y_4 \end{bmatrix}$$

    Verify your result by computing the minimum norm least-squares solution of $C\boldsymbol{x} = \boldsymbol{y}$ by using the MATLAB command `x=pinv(C)*y` for several randomly generated $\boldsymbol{y}$ vectors.

25. (a) Prove (8.20). Hint: From

$$A = U_1\Sigma_1V_1^h = \sum_{k=1}^{n}\sigma_k\boldsymbol{u}_k\boldsymbol{v}_k^h$$

    it follows that

$$a_{ij} = \sum_{k=1}^{r}\sigma_k u_{ik}v_{jk}^*$$

    where

$$\boldsymbol{u}_k = \begin{bmatrix} u_{1k} \\ \vdots \\ u_{mk} \end{bmatrix} \quad \text{and} \quad \boldsymbol{v}_k = \begin{bmatrix} v_{1k} \\ \vdots \\ v_{nk} \end{bmatrix}$$

    are the kth left and right singular vectors of A. Manipulate the expression

$$\|\,A\,\|_F = \sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}a_{ij}^* = \sum_{i=1}^{m}\sum_{j=1}^{n}(\sum_{k=1}^{r}\sigma_k u_{ik}v_{jk}^*)(\sum_{l=1}^{r}\sigma_l u_{il}v_{jl}^*)$$

    (b) Prove (8.21).

26. Generate a 4×3 random matrix D using the MATLAB command `rand`, and let $E = C + 0.001D$, where C is the matrix in Exercise 8.19. Using MATLAB

    (a) obtain the singular value decomposition of E,

    (b) compute a rank-2 matrix F which is closest to E in Frobenius norm,

(c) compute $\| E - F \|_F$ and $\| E - C \|_F$.

27. Let

$$A = \begin{bmatrix} 0.5055 & 0.6412 & 0.8035 \\ 0.1693 & 0.0162 & 0.6978 \\ 0.5247 & 0.8369 & 0.4619 \end{bmatrix}$$

Using MATLAB

(a) Compute the condition number and the inverse of A.

(b) Find the best rank-1 and rank-2 approximations A_1 and A_2 of A in Frobenius norm. Compute also $\| A - A_1 \|_F$ and $\| A - A_2 \|_F$, and comment on the results.

28. (Application) The most general expression of a conic in the x_1x_2 plane is

$$s_{11}x_1^2 + 2s_{12}x_1x_2 + s_{22}x_2^2 + 2r_1x_1 + 2r_2x_2 = 1$$

Let the equation be expressed in compact form as

$$\boldsymbol{x}^t S\boldsymbol{x} + 2\boldsymbol{r}^t\boldsymbol{x} = 1$$

If $S = O$ and $\boldsymbol{r} = \boldsymbol{0}$, then the solution set is empty. If $S = O$ and $\boldsymbol{r} \neq \boldsymbol{0}$, then the conic degenerates into a straight line.
Suppose $S \neq O$. Let S have the eigenvalues $\lambda_1 \geq \lambda_2$ and an orthogonal modal matrix P such that $P^tSP = D = \text{diag}\,[\,\lambda_1, \lambda_2\,]$. Then a change of the coordinate system as

$$\boldsymbol{x} = P\tilde{\boldsymbol{x}} - \boldsymbol{x}_o\,, \quad \tilde{\boldsymbol{x}} = P^t(\boldsymbol{x} + \boldsymbol{x}_o) \tag{8.23}$$

transforms the equation of the conic into

$$\tilde{\boldsymbol{x}}^t D\tilde{\boldsymbol{x}} + 2(\boldsymbol{r} - S\boldsymbol{x}_o)^t P\tilde{\boldsymbol{x}} = 1 + 2\boldsymbol{r}^t\boldsymbol{x}_o - \boldsymbol{x}_o^t S\boldsymbol{x}_o$$

The transformation in (8.23) corresponds to a rotation of the coordinate axes, followed by a shift of the origin as illustrated in Figure 8.3. The purpose of shifting the origin of the coordinate system after the rotation is to eliminate, if posible, the linear term $2(\boldsymbol{r} - S\boldsymbol{x}_o)^t P\tilde{\boldsymbol{x}}$ so that the equation takes the form of that of a central conic.

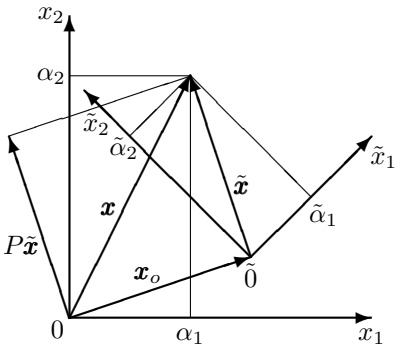

Figure 8.3: The coordinate transformation in (8.23)

The following cases need to be considered:

(a) $\lambda_1 \geq \lambda_2 > 0$. In this case D is nonsingular. Show that, a choice of $\boldsymbol{x}_o = S^{-1}\boldsymbol{r}$, reduces the equation of the conic into

$$\frac{\tilde{x}_1^2}{a_1^2} + \frac{\tilde{x}_2^2}{a_2^2} = 1$$

which represents an ellipse in the $\tilde{x}_1\tilde{x}_2$ plane.

(b) $\lambda_1 > 0 = \lambda_2$. This case is more difficult to analyze than the previous one, because S is singular now. Show that a suitable choice of $\boldsymbol{x}_o$ reduces the equation of the conic into either

$$\tilde{x}_1^2 = c^2$$

which represents two parallel lines, or into

$$c\tilde{x}_1^2 + \tilde{x}_2 = 0$$

which represents a parabola.

(c) $\lambda_1 > 0 > \lambda_2$. As in case (a), choose $\boldsymbol{x}_o = S^{-1}\boldsymbol{r}$. Work out the details to show that, depending on the value of $1 + \boldsymbol{r}^t S^{-1}\boldsymbol{r}$, the equation either represents two intersecting straight lines or a hyperbola.

(d) $\lambda_1 = 0 > \lambda_2$. This case is similar to case (b). Show that the solution set may be empty, consist of a single point, two parallel straight lines, or a parabola.

(e) $0 > \lambda_1 \geq \lambda_2$. This case is similar to case (a). Show that in addition to being an ellipse, the solution set may also be empty or consist of a single point.

29. (a) Use MATLAB to plot the graph of the conic described by

$$x_1^2 + px_1x_2 + 2x_2^2 - 3x_2 - 6 = 0$$

in a rectangular region $-5 \leq x_1, x_2 \leq 5$ of the x_1x_2 plane for each of the values $p = -1$, $p = 2\sqrt{2}$ and $p = 4$.

(b) Transform the equation of the conic into one of the standard forms in Exercise 8.28 for each of the given values of the parameter p.

Appendix A
Complex Numbers

A.1 Fields

A field is a set $\mathbb{F}$ together with two operations, called addition and multiplication and denoted by the usual symbolism, which satisfy the following conditions.

A1. $a + b = b + a$ for all $a, b \in \mathbb{F}$

A2. $(a + b) + c = a + (b + c)$ for all $a, b, c \in \mathbb{F}$

A3. There exists an element denoted by $0 \in \mathbb{F}$ such that $a + 0 = a$ for all $a \in \mathbb{F}$

A4. For each $a \in \mathbb{F}$ there exists an element $-a \in \mathbb{F}$ such that $a + (-a) = 0$

M1. $ab = ba$ for all $a, b \in \mathbb{F}$

M2. $(ab)c = a(bc)$ for all $a, b, c \in \mathbb{F}$

M3. There exists an element denoted by $1 \in \mathbb{F}$ such that $1a = a$ for all $a \in \mathbb{F}$

M4. For each $0 \neq a \in \mathbb{F}$ there exists an element $a^{-1} \in \mathbb{F}$ such that $aa^{-1} = 1$

D. $a(b + c) = ab + ac$ for all $a, b, c \in \mathbb{F}$

It can be shown that the additive identity 0 and the multiplicative identity 1 are unique in $\mathbb{F}$. Also each element a has a unique additive inverse $-a$, and each $a \neq 0$ has a unique multiplicative inverse a^{-1}. The subtraction operation in $\mathbb{F}$ is defined in terms of addition as

$$a - b = a + (-b)$$

and the division operation is defined in terms of multiplication as

$$a/b = ab^{-1}, \quad b \neq 0$$

Familiar examples of fields are the field of rational numbers and the field of real numbers (denoted $\mathbb{R}$). Another common one is the field of complex numbers explained next.

A.2 Complex Numbers

A complex number is of the form

$$z = a + ib$$

where $a, b \in \mathbb{R}$ and

$$i^2 = -1$$

a and b are called the real and imaginary parts of z, respectively, denoted

$$a = \text{Re}\, z\,, \quad b = \text{Im}\, z$$

Two complex numbers $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$ are called equal if $a_1 = a_2$ and $b_1 = b_2$. The addition and multiplication of z_1 and z_2 are defined as

$$z_1 + z_2 = (a_1 + a_2) + i(b_1 + b_2)$$

and

$$z_1 z_2 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + a_2 b_1)$$

Note that multiplication of two complex numbers is performed by the usual rules for algebraic multiplication with $i^2 = -1$.

Defining additive and multiplicative identities as

$$0 = 0 + i0\,, \quad 1 = 1 + i0$$

additive inverse of $z = a + ib$ as

$$-z = (-a) + i(-b)$$

and the multiplicative inverse as

$$z^{-1} = a/(a^2 + b^2) - ib/(a^2 + b^2)$$

it can be shown that the set of all complex numbers $\mathbb{C}$ together with the above addition and multiplication, is a field. Every real number can be considered as a complex number with imaginary part equal to 0, that is $a = a + i0$. Its additive inverse and multiplicative inverse (if $a \neq 0$) as a complex number are the same as its additive and multiplicative inverses as a real number. Thus $\mathbb{R}$, which is itself a field, is a subfield of $\mathbb{C}$ with respect to the same addition and multiplication operations.

The complex conjugate of $z = a + ib$ is defined to be

$$z^* = a - ib$$

Note that

$$zz^* = (a + ib)(a - ib) = a^2 + b^2$$

There is a geometrical representation of complex numbers. To a given complex number $z = a + ib$ we associate the point in a plane with abscissa a and ordinate b, relative to a rectangular coordinate system in the plane, as shown in Figure A.1. In this way there is a one-to-one correspondence between the set of all complex numbers and the set of all points in the plane. The absolute value or modulus of z, is defined as

$$|z| = \sqrt{zz^*} = (a^2 + b^2)^{1/2}$$

Geometrically, this is the polar distance r of the point (a, b) from the origin $(0, 0)$, that is, $|z| = r$. We also define the argument of $z \neq 0$, denoted arg z, to be the polar angle θ shown in the figure, that is

$$\text{arg}\, z = \theta = \tan^{-1}(b/a)$$

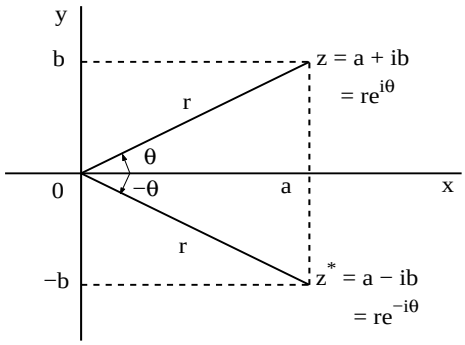

Figure A.1: Representation of a complex number

Note that

$$z = r(\cos\theta + i\sin\theta)$$

Using the series representations

$$\begin{aligned}\cos\theta &= 1 - \theta^2/2! + \theta^4/4! - \cdots \\ \sin\theta &= \theta - \theta^3/3! + \theta^5/5! - \cdots\end{aligned}$$

and rearranging the terms we observe that

$$\cos\theta + i\sin\theta = 1 + (i\theta) + (i\theta)^2/2! + (i\theta)^3/3! + \cdots$$

By analogy to the series representation of the real quantity

$$e^x = 1 + x + x^2/2! + x^3/3! + \cdots$$

the above series can conveniently be defined as $e^{i\theta}$. Thus we obtain

$$z = r(\cos\theta + i\sin\theta) = re^{i\theta}$$

which is called the polar representation of the complex number z. Polar representation provides simplicity in multiplication and division of complex numbers. If $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$, then

$$z_1 z_2 = r_1 r_2 e^{i(\theta_1+\theta_2)}$$

and if $z_2 \neq 0$ ($r_2 \neq 0$), then

$$z_1/z_2 = (r_1/r_2)e^{i(\theta_1-\theta_2)}$$

A.3 Complex-Valued Functions

If f and g are real-valued functions of a real variable t, then

$$h(t) = f(t) + ig(t)$$

defines a complex-valued function h of t. If f and g are differentiable on an interval $a < t < b$, then h is also differentiable, and its derivative is given by

$$h'(t) = f'(t) + ig'(t)$$

A useful complex-valued function is e^{zt}, where $z = a + ib$ is a complex number and t is a real variable. Using the polar representation, e^{zt} can be expressed as

$$e^{zt} = e^{(a+ib)t} = e^{at}e^{ibt} = e^{at}(\cos bt + i \sin bt)$$

Differentiating real and imaginary parts of e^{zt}, and rearranging the terms we get

$$\begin{aligned}
\frac{d}{dt} e^{zt} &= ae^{at}(\cos bt + i \sin bt) + e^{at}(-b \sin bt + ib \cos bt) \\
&= e^{at}(a \cos bt - b \sin bt) + ie^{at}(a \sin bt + b \cos bt) \\
&= (a + ib)e^{at}(\cos bt + i \sin bt) \\
&= ze^{zt}
\end{aligned}$$

Thus the usual differentiation property of the real-valued exponential function is generalized to the complex-valued exponential function.

Appendix B
Existence and Uniqueness Theorems

Consider a system of first order ordinary differential equations together with a set of initial conditions:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}, t)\,, \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_o \tag{B.1}$$

where $\boldsymbol{f} : \mathbb{R}^{n\times 1} \times \mathbb{R} \to \mathbb{R}^{n\times 1}$ is a vector-valued function defined on some interval $\boldsymbol{I} \subset \mathbb{R}$ containing t_0. We assume that

a) for every fixed $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, the function $\boldsymbol{f}(\boldsymbol{x}, \cdot) : t \to \boldsymbol{f}(\boldsymbol{x}, t)$ is piecewise continuous on $\boldsymbol{I}$, and

b) $\boldsymbol{f}$[1] satisfies a ***Lipschitz condition*** on $\boldsymbol{I}$, that is, there exists a constant $K > 0$ such that

$$\| \boldsymbol{f}(\boldsymbol{x}_1, t) - \boldsymbol{f}(\boldsymbol{x}_2, t) \| \leq K \| \boldsymbol{x}_1 - \boldsymbol{x}_2 \| \tag{B.2}$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^{n\times 1}$ and $t \in \boldsymbol{I}$.

Recall that a vector-valued function $\boldsymbol{\phi}$ defined on $\boldsymbol{I}$ is called a solution of (B.1) if $\boldsymbol{\phi}(t_0) = \boldsymbol{x}_o$ and

$$\boldsymbol{\phi}'(t) = \boldsymbol{f}(\boldsymbol{\phi}(t), t) \tag{B.3}$$

at all continuity points of $\boldsymbol{f}$. Clearly, if $\boldsymbol{\phi}$ is a solution, then integrating both sides of (B.3) from t_0 to t, we obtain

$$\boldsymbol{\phi}(t) = \boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}(\tau), \tau)\, d\tau \tag{B.4}$$

Conversely, if $\boldsymbol{\phi}$ satisfies the integral equation in (B.4), then it is a solution of (B.1). We will use this fact in the proof of the following existence and uniqueness theorem.

Theorem B.1 *Under the assumptions (a) and (b) above, the initial-value problem in (B.1) has a unique, continuous solution on* $\boldsymbol{I}$.

[1] The Lipschitz condition in (B.2) is stronger than continuity of $\boldsymbol{f}$ in $\boldsymbol{x}$. For example, the scalar function $f(x, t) = \sqrt{x}$ defined on $\boldsymbol{I} = [0, \infty)$ is continuous everywhere on $\boldsymbol{I}$, but does not satisfy a Lipschitz condition. With $x_1 = x$ and $x_2 = 0$, there exists no K that satisfies

$$\sqrt{x} \leq Kx$$

for all $x \geq 0$.

Proof Define a sequence of continuous functions recursively as

$$\begin{aligned} \boldsymbol{\phi}_0(t) &= \boldsymbol{x}_0 \\ \boldsymbol{\phi}_m(t) &= \boldsymbol{x}_0 + \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}_{m-1}(\tau), \tau)\, d\tau\,, \quad m = 1, 2, \ldots \end{aligned} \tag{B.5}$$

Fix $T > 0$ such that $\boldsymbol{J} = [t_0, t_0 + T] \subset \boldsymbol{I}$. Since $\boldsymbol{f}(\boldsymbol{x}_0, t)$ is piecewise continuous, it is bounded on $\boldsymbol{J}$. Let

$$B = \max_{t \in \boldsymbol{J}} \{\, \boldsymbol{f}(\boldsymbol{x}_0, t) \,\}$$

We claim that

$$\| \boldsymbol{\phi}_m(t) - \boldsymbol{\phi}_{m-1}(t) \| \leq \frac{B}{K} \frac{K^m (t - t_0)^m}{m!}\,, \quad m = 1, 2, \ldots \tag{B.6}$$

for all $t \in \boldsymbol{J}$. The claim is true for $m = 1$ as

$$\| \boldsymbol{\phi}_1(t) - \boldsymbol{\phi}_0(t) \| \leq \| \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}_0(\tau), \tau)\, d\tau \| \leq \int_{t_0}^{t} \| \boldsymbol{f}(\boldsymbol{x}_0, \tau) \|\, d\tau \leq B(t - t_0)$$

Suppose it is true for $m = k$. Then for $m = k + 1$

$$\begin{aligned} \| \boldsymbol{\phi}_{k+1}(t) - \boldsymbol{\phi}_k(t) \| &\leq \| \int_{t_0}^{t} [\, \boldsymbol{f}(\boldsymbol{\phi}_k(\tau), \tau) - \boldsymbol{f}(\boldsymbol{\phi}_{k-1}(\tau), \tau) \,]\, d\tau \| \\ &\leq \int_{t_0}^{t} K \| \boldsymbol{\phi}_k(\tau) - \boldsymbol{\phi}_{k-1}(\tau) \|\, d\tau \\ &\leq \int_{t_0}^{t} K \frac{B}{K} \frac{K^k (\tau - t_0)^k}{k!}\, d\tau \\ &\leq \frac{B}{K} \frac{K^{k+1} (t - t_0)^{k+1}}{(k+1)!} \end{aligned}$$

so that it is also true for $m = k + 1$. Hence it is true for all $m \geq 1$. Since $t - t_0 \leq T$ for all $t \in \boldsymbol{J}$, (B.6) further implies that

$$\| \boldsymbol{\phi}_m(t) - \boldsymbol{\phi}_{m-1}(t) \| \leq \frac{B}{K} \frac{(KT)^m}{m!}\,, \quad m = 1, 2, \ldots$$

Define

$$\phi_m(t) = \| \boldsymbol{\phi}_m(t) - \boldsymbol{\phi}_0(t) \|$$

Then

$$\begin{aligned} \phi_m(t) &= \| \sum_{k=1}^{m} [\, \boldsymbol{\phi}_k(t) - \boldsymbol{\phi}_{k-1}(t) \,] \| \leq \sum_{k=1}^{m} \| \boldsymbol{\phi}_k(t) - \boldsymbol{\phi}_{k-1}(t) \| \\ &\leq \frac{B}{K} \sum_{k=1}^{m} \frac{(KT)^k}{k!} \leq \frac{B}{K} (e^{KT} - 1)\,, \quad m = 1, 2, \ldots \end{aligned}$$

for all $t \in \boldsymbol{J}$. This implies that the sequence of nonnegative-valued continuous functions $\{\, \phi_m \,\}$ converges uniformly on $\boldsymbol{J}$.[2] Consequently, the sequence of vector-valued continuous functions $\{\, \boldsymbol{\phi}_m \,\}$ converges uniformly to a continuous limit function $\boldsymbol{\phi}$.[3]

Uniform convergence of $\{\, \boldsymbol{\phi}_m \,\}$, together with the Lipschitz condition on $\boldsymbol{f}$ further imply that

[2] This is a direct consequence of the comparison test. For details the reader is referred to a book on advanced calculus.

[3] That is, given any $\epsilon > 0$, there exist $M > 0$ such that

$$\| \boldsymbol{\phi}(t) - \boldsymbol{\phi}_m(t) \| \leq \epsilon$$

for all $m \geq M$ and for all $t \in \boldsymbol{J}$.

a)

$$\lim_{m \to \infty} \boldsymbol{f}(\boldsymbol{\phi}_m(t), t) = \boldsymbol{f}(\boldsymbol{\phi}(t), t)$$

b)

$$\lim_{m \to \infty} \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}_m(\tau), \tau)\, d\tau = \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}(\tau), \tau)\, d\tau$$

Thus

$$\begin{aligned} \boldsymbol{\phi}(t) &= \lim_{m \to \infty} \boldsymbol{\phi}_m(t) \\ &= \boldsymbol{x}_o + \lim_{m \to \infty} \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}_{m-1}(\tau), \tau)\, d\tau \\ &= \boldsymbol{x}_o + \int_{t_0}^{t} \boldsymbol{f}(\boldsymbol{\phi}(\tau), \tau)\, d\tau \end{aligned}$$

for all $t \in \boldsymbol{J}$, proving that $\boldsymbol{\phi}$ is a solution of (B.1) on $\boldsymbol{J}$.

To prove uniqueness of $\boldsymbol{\phi}$, suppose (B.1) has another solution $\boldsymbol{\psi}$ on $\boldsymbol{J}$. Define

$$g(t) = \| \boldsymbol{\phi}(t) - \boldsymbol{\psi}(t) \| = \| \int_{t_0}^{t} [\, \boldsymbol{f}(\boldsymbol{\phi}(\tau), \tau) - \boldsymbol{f}(\boldsymbol{\psi}(\tau), \tau)\,]\, d\tau \|$$

Then

$$g(t) \le \int_{t_0}^{t} \| \boldsymbol{f}(\boldsymbol{\phi}(\tau), \tau) - \boldsymbol{f}(\boldsymbol{\psi}(\tau), \tau) \|\, d\tau \le \int_{t_0}^{t} K g(\tau)\, d\tau$$

for all $t \in \boldsymbol{J}$. Let

$$h(t) = e^{-K(t-t_0)} \int_{t_0}^{t} K g(\tau)\, d\tau$$

Then $h(t_0) = 0$ and

$$h'(t) = K e^{-K(t-t_0)} [\, g(t) - \int_{t_0}^{t} K g(\tau)\, d\tau\,] \le 0$$

so that

$$h(t) \le 0 \quad \text{for all} \;\; t \in \boldsymbol{J}$$

Hence

$$0 \le g(t) \le \int_{t_0}^{t} K g(\tau)\, d\tau \le 0 \quad \text{for all} \;\; t \in \boldsymbol{J}$$

This implies $g(t) = 0$ for all $t \in \boldsymbol{J}$, or equivalently,

$$\boldsymbol{\phi}(t) = \boldsymbol{\psi}(t) \quad \text{for all} \;\; t \in \boldsymbol{J}$$

contradicting the assumption that $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are two different solutions on $\boldsymbol{J}$. In conclusion, (B.1) has a unique solution on $\boldsymbol{J}$.

The case $t < t_0$ can be analyzed similarly by considering a closed interval $\boldsymbol{J} = [t_0 - T, T] \subset \boldsymbol{I}$. Since T is arbitrary in both cases, it follows that (B.1) has a unique, continuous solution on $\boldsymbol{I}$.

The functions in (B.5) that converge to the solution of (B.1) are known as the ***Picard iterates***, and provide a constructive method for the proof of the existence of a solution. The proof of the uniqueness part of the theorem is a variation of the well-known ***Bellman-Gronwal Lemma***.

Proof of Theorem 6.1

The proof follows immediately from Theorem B.1 on noting that

$$\boldsymbol{f}(\boldsymbol{x}, t) = A(t)\boldsymbol{x} + \boldsymbol{u}(t)$$

is piecewise continuous for every fixed $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$, and satisfies a Lipschitz condition with

$$K = \sup_{t\in \boldsymbol{I}} \| A(t) \|$$

Appendix C
The Laplace Transform

C.1 Definition and Properties

The one-sided (or unilateral) ***Laplace transform*** of a real- or complex-valued function f of a real variable t is a complex-valued function F of a complex variable s, defined by

$$F(s) = \int_0^\infty f(t)e^{-st}\,dt \tag{C.1}$$

provided the integral converges. For convenience, the Laplace transform of f is also denoted by $\mathcal{L}\{f\}$.

Let f be a piece-wise continuous function that is bounded by an exponential, that is, there exist $M, \alpha \in \mathbb{R}$ such that

$$|f(t)| \leq Me^{\alpha t}$$

holds for all t. Such a function is said to be of ***exponential order*** α. Then for any $s = \sigma + i\omega \in \mathbb{C}$ with $\sigma > \alpha$

$$\begin{aligned} |\int_0^\infty f(t)e^{-st}\,dt\,| &\leq \int_0^\infty |f(t)e^{-st}|\,dt \\ &\leq \int_0^\infty Me^{(\alpha-\sigma)t}\,dt \\ &\leq \frac{M}{\sigma-\alpha} \end{aligned}$$

and thus the integral in (C.1) converges. The region

$$\mathcal{C}_\alpha = \{\, s = \sigma + i\omega \,|\, \sigma > \alpha \,\}$$

is called the ***region of convergence*** of F.

Let f be a function of exponential order α with a Laplace transform $F(s)$ that exists in a region $\mathcal{C}_\alpha$, and suppose that $f(t) = 0$ for $t < 0$. Then f can be obtained uniquely from F by means of a line integral as

$$f(t) = \lim_{\omega \to \infty} \int_\Gamma F(s)e^{st}\,ds \tag{C.2}$$

where Γ is a vertical straight line in $\mathcal{C}_\alpha$ extending from $s = \sigma - i\omega$ to $s = \sigma + i\omega$. The integral on the right-hand side of (C.2) is called the ***inverse Laplace transform*** of F, denoted

by $\mathcal{L}^{-1}(F)$.[1] We use the notation

$$f(t) \longleftrightarrow F(s)$$

to indicate that f and F are a Laplace transform-inverse Laplace transform pair.

Some useful properties of the Laplace transform are stated below, where it is assumed that the Laplace transforms involved exist.

a) Linearity

$$af(t) + bg(t) \longleftrightarrow aF(s) + bG(s)$$

b) Shift

$$f(t - t_0) \longleftrightarrow e^{-st_0}F(s), \quad t_0 > 0$$

$$e^{s_0 t}f(t) \longleftrightarrow F(s - s_0), \quad s_0 \in \mathbb{C}$$

c) Scaling

$$f(at) \longleftrightarrow \frac{1}{a}F(\frac{s}{a}), \quad a > 0$$

d) Differentiation

$$f^{(n)}(t) \longleftrightarrow s^n F(s) - s^{n-1}f(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0)$$

$$t^n f(t) \longleftrightarrow (-1)^n \frac{d^n}{ds^n}F(s)$$

The first three of the properties above are direct consequences of the definition. For example, the Laplace transform of the shifted function $f(t - t_0)$ is

$$\begin{aligned}\int_0^\infty f(t - t_0)e^{-st}\,dt &= \int_{-t_0}^\infty f(\tau)e^{-s(\tau+t_0)}\,d\tau \\ &= e^{-st_0}\int_0^\infty f(\tau)e^{-s\tau}\,d\tau = e^{-st_0}F(s)\end{aligned}$$

proving the first property in (b).[2] Proofs of the properties in (d) require some manipulations. Evaluating the integral in (C.1) written for f' by parts, we obtain

$$\begin{aligned}\mathcal{L}\{f'\} &= \int_0^\infty f'(t)e^{-st}\,dt \\ &= [\,f(t)e^{-st}\,]_{t=0}^{t=\infty} + \int_0^\infty f(t)se^{-st}\,dt \\ &= sF(s) - f(0)\end{aligned}$$

[1] In practice, the line integral in (C.2) is seldom used to find the inverse Laplace transform. Instead, Laplace transform tables are used for most of the functions of interest.

[2] The second equality follows from the assumption that $f(t) = 0$ for $t < 0$.

proving the first property in (d) for $n = 1$.[3] The case $n > 1$ and the second property in (d) can be proved similarly.

Example C.1

The Laplace transform of the unit step function

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases}$$

is

$$U(s) = \int_0^t e^{-st}\, dt = [\, -se^{-st}\,]_{t=0}^{t=\infty} = \frac{1}{s}\ , \quad \text{Re}\, s > 0$$

By property (b),

$$u(t - t_0) \ \longleftrightarrow \ \frac{e^{-st_0}}{s}\ , \quad \text{Re}\, s > 0$$

and, by property (d)

$$tu(t) \ \longleftrightarrow \ -\ \frac{d}{ds}\,(\,\frac{1}{s}\,) = \frac{1}{s^2}$$

C.2 Some Laplace Transform Pairs

The Laplace transform of the unit step function obtained in Example C.1, together with the properties listed in the previous section, allows us to obtain the Laplace transform of many useful functions. For example, from the second property in (b), we obtain

$$e^{\sigma_0 t}u(t) \ \longleftrightarrow \ \frac{1}{s - \sigma_0}$$

and from property (d),

$$te^{\sigma_0 t}u(t) \ \longleftrightarrow \ \frac{1}{(s - \sigma_0)^2}$$

The Laplace transform of $e^{s_0 t}u(t)$, in turn, can be used to find Laplace transforms of sine and cosine functions. On noting that

$$e^{i\omega_0 t} = \cos\omega_0 t + i \sin\omega_0 t$$

we obtain

$$(\cos\omega_0 t + i \sin\omega_0 t)u(t) \ \longleftrightarrow \ \frac{1}{s - i\omega_0} = \frac{s + i\omega_0}{s^2 + \omega_0^2}$$

Thus

$$(\cos\omega_0 t)u(t) \ \longleftrightarrow \ \frac{s}{s^2 + \omega_0^2}$$

[3] Since f is exponential order α and Re $s > \alpha$

$$\lim_{t \to \infty} f(t)e^{-st} = 0$$

and

$$(\sin \omega_0 t)u(t) \longleftrightarrow \frac{\omega_0}{s^2 + \omega_0^2}$$

A list of some Laplace transform pairs, which can be derived similarly, is given in Table C.1.

C.3 Partial Fraction Expansions

A function F that is expressed as a ratio of two polynomials is called a ***rational function***. A rational function

$$F(s) = \frac{c(s)}{d(s)} = \frac{c_0 s^m + c_1 s^{m-1} + \cdots + c_m}{s^n + d_1 s^{n-1} + \cdots + d_n} \tag{C.3}$$

is said to be ***proper*** if $m \leq n$ and ***strictly proper*** if $m < n$.

The Laplace transforms in Table C.1 are simple strictly proper rational functions with denominators being first or second degree polynomials or powers of such polynomials. This observation suggests that if a rational function F can be expressed as a linear combination of such simple rational functions, then by linearity of the Laplace transform, the inverse Laplace transform of F can be obtained as the same linear combination of the inverse Laplace transforms of individual functions, which can be written down directly from the table. For example, the inverse Laplace transform of

$$\frac{s+2}{s^2+s} = \frac{2}{s} - \frac{1}{s+1}$$

can be written down using Table C.1 as

$$\mathcal{L}^{-1}\{\frac{s+2}{s^2+s}\} = (2 - e^{-t})u(t)$$

Consider a strictly proper rational function $F(s)$ expressed as in (C.3). Suppose that the denominator polynomial $d(s)$ is factored out as

$$d(s) = \prod_{i=1}^{k} (s - p_i)^{n_i}$$

where $p_i \in \mathbb{C}$ are distinct zeros of d with multiplicities n_i, $i = 1, \ldots, k$. p_i are called the ***poles*** of F. Then F can be expressed as

$$F(s) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{r_{ij}}{(s - p_i)^j} \tag{C.4}$$

where $r_{ij} \in \mathbb{C}$. This expression is known as the ***partial fraction expansion*** of F. The coefficients r_{ij} can be obtained by collecting the terms on the right-hand side of (C.4) over the common denominator d and equating the coefficients of the resulting numerator polynomial to those of c.

Table C.1: Some Laplace transform pairs

$f(t)$	$F(s)$
1	$\frac{1}{s}$
t^n	$\frac{n!}{s^{n+1}}$
$e^{\sigma_0 t}$	$\frac{1}{s-\sigma_0}$
$t^n e^{\sigma_0 t}$	$\frac{n!}{(s-\sigma_0)^{n+1}}$
$\cos\omega_0 t$	$\frac{s}{s^2+\omega_0^2}$
$\sin\omega_0 t$	$\frac{\omega_0}{s^2+\omega_0^2}$
$t\cos\omega_0 t$	$\frac{s^2-\omega_0^2}{(s^2+\omega_0^2)^2}$
$t\sin\omega_0 t$	$\frac{2\omega_0 s}{(s^2+\omega_0^2)^2}$
$e^{\sigma_0 t}\cos\omega_0 t$	$\frac{s-\sigma_0}{(s-\sigma_0)^2+\omega_0^2}$
$e^{\sigma_0 t}\sin\omega_0 t$	$\frac{\sigma_0}{s-\sigma_0)^2+\omega_0^2}$
$te^{\sigma_0 t}\cos\omega_0 t$	$\frac{(s-\sigma_0)^2-\omega_0^2}{((s-\sigma_0)^2+\omega_0^2)^2}$
$te^{\sigma_0 t}\sin\omega_0 t$	$\frac{2\omega_0(s-\sigma_0)}{((s-\sigma_0)^2+\omega_0^2)^2}$

MATLAB provides a built-in function, `residue`, to compute p_i and r_{ij}. The commands

```
>> c=[c0 c1 ...  cm]; d=[1 d1 ...  dn];
>> [r,p]=residue(c,d);
```

return the poles p_i in the array `p` (with each pole appearing as many times as its multiplicity) and the coefficients r_{ij} in the array `r`.

Example C.2

The strictly proper rational function

$$F(s) = \frac{2s^2+4s+1}{s^3+4s^2+5s+2} = \frac{2s^2+4s+1}{(s+2)(s+1)^2}$$

has the poles $p_1 = -2$ with $n_1 = 1$ and $p_2 = -1$ with $n_2 = 2$. Hence F has a partial fraction expansion of the form

$$F(s) = \frac{r_1}{s+2} + \frac{r_{21}}{s+1} + \frac{r_{22}}{(s+1)^2}$$

Reorganizing the above expression, we get

$$\begin{aligned} F(s) &= \frac{r_1(s+1)^2 + r_{21}(s+1)(s+2) + r_{22}(s+2)}{(s+2)(s+1)^2(s+2)} \\ &= \frac{(r_1+r_{21})s^2 + (2r_1+3r_{21}+r_{22})s + (r_1+2r_{21}+2r_{22})}{s^3+4s^2+5s+2} \\ &= \frac{2s^2+4s+1}{s^3+4s^2+5s+2} \end{aligned}$$

Equating the coefficients of the numerators of the last two expressions we obtain a system of three equatins in three unknows

$$\begin{bmatrix} 1 & 1 & 0 \\ 2 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} r_1 \\ r_{21} \\ r_{22} \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix}$$

the unique solution of which is easily computed as $r_1 = r_{12} = 1, r_{22} = -1$.

Alternatively, the MATLAB commands

```
>> c=[2 4 1]; d=[1 4 5 2];
>> [r,p]=residue(c,d);
```

produce

$$r = [1 \quad 1 \quad -1], \quad p = [-2 \quad -1 \quad -1]$$

Note that the coefficients r_{ij} associated with a multiple pole p_i appear in the order of increasing j in the array `r`.

Thus

$$F(s) = \frac{1}{s+2} + \frac{1}{s+1} - \frac{1}{(s+1)^2}$$

and its inverse Laplace transform is

$$f(t) = (e^{-2t} + e^{-t} - te^{-t})u(t)$$

Example C.3

To find the inverse Laplace transform of

$$G(s) = \frac{s^2 + 9s + 16}{s^3 + 5s^2 + 9s + 5}$$

we execute the MATLAB commands

```
>> c=[1 9 16]; d=[1 5 9 5];
>> [r,p]=residue(c,d)
```

which compute

$$r = [\,-1.5 - i \quad -1.5 + i \quad 4\,], \quad p = [\,-2 + i \quad -2 - i \quad -1\,]$$

Thus

$$G(s) = \frac{-1.5 - i}{s + 2 - i} + \frac{-1.5 + i}{s + 2 + i} + \frac{4}{s + 1}$$

and

$$\begin{aligned} g(t) &= (-1.5 - i)\, e^{(-2+i)t} + (-1.5 + i)\, e^{(-2-i)t} + 4\, e^{-t} \\ &= 2\,\mathrm{Re}\,\{\,(-1.5 - i)\, e^{(-2+i)t}\,\} + 4\, e^{-t} \\ &= e^{-2t}(2\sin t - 3\cos t) + 4\, e^{-t} \end{aligned}$$

C.4 Solution of Differential Equations by Laplace Transform

Consider an nth order linear differential equation with constant coefficients

$$y^{(n)} + a_1 y^{(n-1)} + \cdots + a_{n-1} y' + a_n y = u(t) \tag{C.5}$$

together with a set of n initial conditions

$$y(0) = y_0\,, \quad y'(0) = y_1\,, \quad \ldots\,, \quad y^{(n-1)}(0) = y_{n-1}$$

specified at $t_0 = 0$. Taking the Laplace transform of both sides of (C.5) and using the differentiation property, we obtain

$$\begin{aligned} s^n Y(s) - s^{n-1} y_0 - \cdots - s y_{n-2} - y_{n-1} &+ \\ s^{n-1} Y(s) - \cdots - s y_{n-3} - y_{n-2} &+ \\ \vdots & \\ s Y(s) - y_0 &+ \\ Y(s) &= U(s) \end{aligned}$$

Rearranging the terms, the above expression can be written as

$$Y(s) = \frac{b(s)}{a(s)} + \frac{1}{a(s)}\, U(s) \tag{C.6}$$

where

$$\begin{aligned} a(s) &= s^n + a_1 s^{n-1} + \cdots + a_n \\ b(s) &= y_0 s^{n-1} + (y_0 + y_1) s^{n-2} + \cdots + (y_0 + y_1 + \cdots + y_{n-1}) \end{aligned}$$

Taking the inverse Laplace transform of (C.6), the solution of the given initial-value problem is obtained as

$$y = y_o(t) + y_u(t)\,, \quad t \geq 0 \tag{C.7}$$

where

$$y_o(t) = \mathcal{L}^{-1}\{\, \frac{b(s)}{a(s)} \,\}$$

is part of the solution due to the initial conditions, and

$$y_u(t) = \mathcal{L}^{-1}\{\, \frac{1}{a(s)}\, U(s) \,\}$$

is the part due to the forcing term. Note the similarity between the expressions in (2.15) and (C.7).

Example C.4

Consider the differential equation

$$y'' + 2y' + 26y = 26u(t)\,, \quad y(0) = y'(0) = 0$$

where $u(t)$ is the unit step function.

Taking the Laplace transforms of both sides of the given differential equation and rearranging the terms, we get

$$Y(s) = \Phi(s) = \frac{26}{s(s^2 + 2s + 26)} \tag{C.8}$$

Expanding $Y(s)$ into partial fractions, we obtain

$$\begin{aligned} Y(s) &= \frac{26}{s(s+1-5i)(s+1+5i)} \\ &= \frac{1}{s} + \frac{-0.5 + 0.1\,i}{s+1-5i} + \frac{-0.5 - 0.1\,i}{s+1+5i} \end{aligned}$$

Thus

$$\begin{aligned} y = \phi(t) &= 1 + (-0.5 + 0.1\,i)\, e^{(-1+5i)t} + (-0.5 - 0.1\,i)\, e^{(-1-5i)t} \\ &= 1 + 2\,\mathrm{Re}\,\{\, (-0.5 + 0.1\,i)\, e^{(-1+5i)t} \,\} \\ &= 1 - e^{-t}(\cos 5t + 0.2\, \sin 5t)\,, \quad t \geq 0 \end{aligned} \tag{C.9}$$

the plot of which is shown in Figure C.1.

If the initial conditions were specified as $y(0) = 1, y'(0) = 0$, then the Laplace transform would yield

$$s^2 Y(s) - s + 2sY(s) - 2 + 26Y(s) = \frac{26}{s}$$

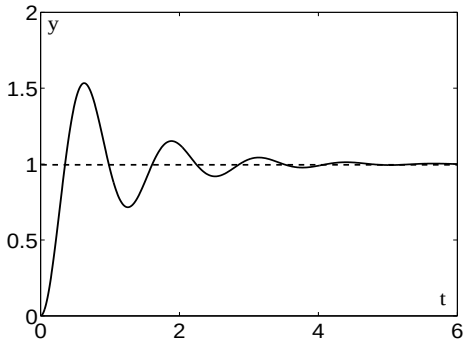

Figure C.1: Solution of the DE in Example C.4

or equivalently,

$$Y(s) = \frac{1}{s}$$

Then the solution would be

$$y = 1\,, \quad t \geq 0$$

Example C.5

Consider the same differential equation in the previous example with a different forcing function:

$$y'' + 2y' + 26y = 26f(t)\,, \quad y(0) = y'(0) = 0$$

where

$$f(t) = \begin{cases} 1, & 0 < t < 2 \\ 0, & t < 0 \text{ or } t > 2 \end{cases}$$

is a pulse of unit strength extending from $t = 0$ to $t = 2$.

Observing that

$$f(t) = u(t) - u(t-2)$$

we have

$$F(s) = U(s) - e^{-2s}U(s) = \frac{1 - e^{-2s}}{s}$$

Then the Laplace transform of the solution is obtained as

$$Y(s) = \frac{26(1 - e^{-2s})}{s(s^2 + 2s + 26)} = (1 - e^{-2s})\Phi(s)$$

where $\Phi(s)$ is given by (C.8). Taking the inverse Laplace transform, we compute the solution as

$$\begin{aligned} y &= \phi(t)u(t) - \phi(t-2)u(t-2) \\ &= \begin{cases} \phi(t), & 0 < t < 2 \\ \phi(t) - \phi(t-2), & t > 2 \end{cases} \\ &= \begin{cases} 1 - e^{-t}(\cos 5t + 0.2\,\sin 5t), & 0 < t < 2 \\ -e^{-t}(\cos 5t + 0.2\,\sin 5t) + \\ \qquad e^{-(t-2)}(\cos 5(t-2) + 0.2\,\sin 5(t-2)), & t > 2 \end{cases} \end{aligned}$$

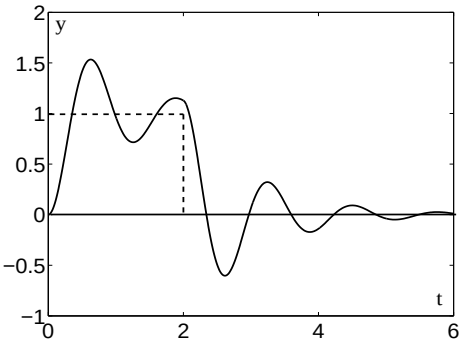

Figure C.2: Solution of the DE in Example C.5

The solution is plotted in Figure C.2.

The Laplace transform can also be used to solve systems of differential equations provided we properly define the Laplace transform of a vector-valued function. This is fairly straightforward: We define the Laplace transform of

$$\boldsymbol{f}(t) = \text{col}\,[\,f_1(t),\,\ldots\,,f_n(t)\,]$$

element-by-element as

$$\boldsymbol{F}(s) = \mathcal{L}\{\,\boldsymbol{f}(t)\,\} = \text{col}\,[\,F_1(s),\,\ldots\,,F_n(s)\,]$$

With this definition it is easy to prove that all the properties of the Laplace transform are also valid for the vector case. For example

$$A\boldsymbol{f}(t) + B\boldsymbol{g}(t) \;\longleftrightarrow\; A\boldsymbol{F}(s) + B\boldsymbol{G}(s)$$

and

$$\boldsymbol{f}'(t) \;\longleftrightarrow\; s\boldsymbol{F}(s) - \boldsymbol{f}(0)$$

Consider a SLDE with a constant coefficient matrix:

$$\boldsymbol{x}' = A\boldsymbol{x} + \boldsymbol{u}(t)\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_o \tag{C.10}$$

Taking the Laplace transform of both sides, we obtain

$$s\boldsymbol{X}(s) - \boldsymbol{x}_o = A\boldsymbol{X}(s) + \boldsymbol{U}(s)$$

which can be solved for $\boldsymbol{X}(s)$ as

$$\boldsymbol{X}(s) = (sI - A)^{-1}\boldsymbol{x}_o + (sI - A)^{-1}\boldsymbol{U}(s) \tag{C.11}$$

Then the solution is

$$\boldsymbol{x} = \mathcal{L}^{-1}\{\,(sI - A)^{-1}\,\}\boldsymbol{x}_o + \mathcal{L}^{-1}\{\,(sI - A)^{-1}\boldsymbol{U}(s)\,\} = \boldsymbol{\Phi}_o(t) + \boldsymbol{\Phi}_u(t)$$

When $u(t) = 0$, i.e., (C.10) is homogeneous, the solution expression above reduces to

$$\boldsymbol{x} = \mathcal{L}^{-1}\{\,(sI - A)^{-1}\,\}\boldsymbol{x}_o$$

Comparing the above expression with (6.19) we observe that

$$\mathcal{L}^{-1}\{\,(sI - A)^{-1}\,\} = e^{At}$$

We thus obtain an alternative formula to compute the matrix exponential function e^{At}.

Example C.6

Consider again Example 6.4, where

$$(sI - A)^{-1} = \begin{bmatrix} s+3 & 2 \\ 1 & s+2 \end{bmatrix}^{-1} = \frac{1}{(s+1)(s+4)} \begin{bmatrix} s+2 & -2 \\ -1 & s+3 \end{bmatrix}$$

We can compute e^{At} by taking the inverse Laplace transform of the elements of $(sI - A)^{-1}$ after expanding each of them into its partial fractions. However, a more elegant approach is to expand the matrix rational function $(sI - A)^{-1}$ into its partial fractions as

$$\begin{aligned} (sI - A)^{-1} &= \frac{1}{s+1} R_1 + \frac{1}{s+4} R_2 \\ &= \frac{1}{(s+1)(s+4)} [(s+4)R_1 + (s+1)R_2] \\ &= \frac{1}{(s+1)(s+4)} [(R_1 + R_2)s + (4R_1 + R_2)] \end{aligned}$$

Comparing the numerator polynomial matrices of the two expressions for $(sI - A)^{-1}$, we get

$$\begin{aligned} R_1 + R_2 &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ 4R_1 + R_2 &= \begin{bmatrix} 2 & -2 \\ -1 & 3 \end{bmatrix} \end{aligned}$$

from which we obtain

$$R_1 = \frac{1}{3} \begin{bmatrix} 1 & -2 \\ -1 & 2 \end{bmatrix}, \quad R_2 = \frac{1}{3} \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$$

Thus

$$e^{At} = e^{-t} R_1 + e^{-4t} R_2 = \frac{1}{3} \begin{bmatrix} e^{-t} + 2e^{-4t} & -2e^{-t} + 2e^{-4t} \\ -e^{-t} + e^{-4t} & 2e^{-t} + e^{-4t} \end{bmatrix}$$

and the solution corresponding to the given initial condition $\boldsymbol{x}_o = \text{col}\,[1, 2]$ is

$$\boldsymbol{x} = e^{At}\boldsymbol{x}_o = \begin{bmatrix} -e^{-t} + 2e^{-4t} \\ e^{-t} + e^{-4t} \end{bmatrix}$$

which is the same as the one obtained in Example 6.4.

Of course, we can obtain the solution corresponding to a given initial condition directly without computing e^{At}. By computing $\boldsymbol{X}(s)$ and expanding it into partial fractions as

$$\begin{aligned} \boldsymbol{X}(s) &= (sI - A)^{-1}\boldsymbol{x}_o = \frac{1}{(s+1)(s+4)} \begin{bmatrix} s+2 & -2 \\ -1 & s+3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= \frac{1}{(s+1)(s+4)} \begin{bmatrix} s-2 \\ 2s+5 \end{bmatrix} = \frac{1}{s+1} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \frac{1}{s+4} \begin{bmatrix} 2 \\ 1 \end{bmatrix} \end{aligned}$$

we get the same solution.

Appendix D
A Brief Tutorial on MATLAB

MATLAB is an interactive system and a programming language for general scientific and technical computation. When it is invoked (either by clicking on the Matlab icon or by typing the command `matlab` from the keyboard) the command prompt `>>` appears indicating that MATLAB is ready to accept command from the keyboard. Commands are terminated by "return" or "enter" keys. The `exit` or `quit` command ends MATLAB.

D.1 Defining Variables

The basic data element of MATLAB is a matrix that does not require dimensioning. Scalars and arrays (vectors) are treated as special matrices. A variable is a data element with a name, which can be any combination of upper and lowercase letters, digits and underscores, starting with a letter and length not exceeding 19. Variables are case sensitive, so `A` and `a` are different variables.

Variables are assigned numerical values by typing an expression or a formula or a function that utilizes arithmetic operations on numerical data or previously defined variables. For example, the commands

```
>> a=2+7; b=4*a;
>> c=sqrt(b);
```

assign the values 9, 36 and 6 to the variables `a`, `b` and `c`, respectively; and the command

```
>> c=c/3;
```

reassigns `c` the value 2. Note that more than one command, separated by commas or semicolons, may appear on a single line. When a command is not terminated by a semicolon the result of the operation is echoed on the screen:

```
>> d=sin(pi/c)
d =
    1
```

If the result of an operation is not assigned to a variable, it is assigned to a default variable `ans` (short for "answer"):

```
>> a+sqrt(-b)
ans =
 9.0000+6.0000i
```

The last example shows that MATLAB requires no special handling of complex numbers. In fact, the imaginary unit `i` is one of the special variables of MATLAB. Some others are `j` (same as `i`), `ans` (default variable name), `pi` (π), `eps` (smallest number such that when

added to 1 creates a floating-point number greater than 1), `inf` (∞) and `NaN` (not a number, e.g., $0/0$). It is recommended that these variables should not be used as variable names to avoid changing their values.

A matrix is defined by entering its elements row by row as

```
>> A=[1 2 3 4; 3 4 5 6; 5 6 7 8]
A =
     1     2     3     4
     3     4     5     6
     5     6     7     8
```

where elements in each row are separated by spaces (or commas), and the rows by semicolons. Thus the commands

```
>> x=[2 4 6 8]; y=[-3; 2; -1];
```

define a row vector `x`, and a column vector `y`. In particular, the command

```
start:increment:end
```

generates a row vector (an array) of equally spaced values with the values of `start` and `end` specifying the first and the last elements of the array. If the increment is omitted, it is assumed to be 1. Thus

```
>> arry=-3:2:9
arry =
   -3   -1    1    3    5    7    9
```

Note also:

```
>> B(1,2)=7, B(2,4)=2
B =
     0     7
B =
     0     7     0     0
     0     0     0     2
```

The command `size(A)` returns a 1×2 row vector consisting of the number of rows and the number of columns of `A`. For a row or column vector `x`, the command `length(x)` returns the number of elements of the vector. Thus

```
>> size(A), size(arry), length(arry)
ans =
     3     4
ans =
     1     7
ans =
     7
```

A particular element of a matrix (or a row or column vector) can be extracted as

```
>> a23=A(2,3), x3=x(3), x3new=x(1,3), y2=y(2)
a23 =
     5
x3 =
     6
```

```
x3new =
    6
y2 =
    2
```

To extract a submatrix of a matrix, the rows and columns of the submatrix are specified:

```
>> sub1=A(2,3:4), sub2=A([1 3],[2 3]), sub3=A(2,:)
sub1 =
    5    6
sub2 =
    2    3
    6    7
sub3 =
    3    4    5    6
```

Thus `sub1` consists of row 2 and columns 3 and 4 of `A`, `sub2` rows 1 and 3 and columns 2 and 3, and `sub3` row 2 and all columns. Similarly,

```
>> part=arry(2:5)
part =
   -1    1    3    5
```

Conversely, a matrix can be constructed from smaller blocks:

```
>> A1=[x;0:3], A2=[A y A(:,[3 1])]
A1 =
    2    4    6    8
    0    1    2    3
A2 =
    1    2    3    4   -3    3    1
    3    4    5    6    2    5    3
    5    6    7    8   -1    7    5
```

MATLAB has special commands for generating special matrices: `eye(n)` generates an identity matrix of order `n`, `zeros(m,n)` an $m \times n$ zero matrix, `ones(m,n)` an $m \times n$ matrix with all elements equal to 1. If `d` is a row or column matrix of length n, the command `diag(d)` generates an $n \times n$ diagonal matrix with the elements of `d` appearing on the diagonal; and if `A` is an $m \times n$ matrix, `diag(A)` gives a column vector of the diagonal elements of `A`.

All commands entered and variables defined in a session are stored in MATLAB's workspace, and can be recalled at any time. Typing the name of a variable returns its value:

```
>> A1
A1 =
    2    4    6    8
    0    1    2    3
```

The command `who` gives a list of all variables defined.

```
>> who
Your variables are:
A      a      c      sub2   x3new
A1     a23    d      sub3   y
```

```
A2      arry    part    x       y2
B       b       sub1    x3
```

The command `clear v_name_1 v_name_2` clears the variables `v_name_1` and `v_name_2` from the workspace, and `clear` clears all variables.

The command `save fn` saves the workspace in the binary file `fn.mat`, which can later be retrieved by the `load fn` command. Menu items *Save Workspace As...* and *Load Workspace* in the *File* menu serve the same purpose.

D.2 Arithmetic Operations

MATLAB utilizes the following arithmetic operators: + (addition), − (subtraction), ⋆ (multiplication), ^ (power operator), ′ (transpose), / and \ (right and left division).

These operators work on scalars or matrices:

```
>> i*sub1, (sub1-[8 5])*sub2
ans =
      0+5.0000i       0+6.0000i
ans =
     0    -2
```

where ⋆ denotes a scalar multiplication in the first command and matrix multiplication in the second. However,

```
>> sub2*sub1
???  Error using ==> *
Inner matrix dimensions must agree.
```

which indicates that the matrices are not compatible for the product.

Although addition of matrices requires that the matrices be of the same order, for convenience MATLAB also allows for addition of a scalar and a matrix by first enlarging the scalar to the size of the matrix. Thus

```
>> sub2+2
ans =
     4     5
     8     9
```

that is, `sub2+2` is equivalent to `sub2+2*ones(2,2)`.

The power operator requires a square matrix as operand:

```
>> C=[0 i; -i 0]^3
C =
       0               0+1.0000i
       0-1.0000i       0
```

The transpose operator takes the Hermitian adjoint of a matrix, which reduces to transpose when the matrix is real. Thus

```
>> [1-i;2+i]'
ans =
 1.0000+1.0000i 2.0000-1.0000i
```

Right division operator / works as usual when both operands or the divisor is a scalar:

```
>> C/5
```

```
ans =
      0               0+0.2000i
      0-0.2000i       0
```

However, care must be taken when "dividing" two matrices: The command `A/B` calculates a matrix `Y` such that `A=YB`. Obviously, this requires that `A` and `B` have the same number of columns. If the equation is inconsistent then `Y` is a least-squares solution. Thus

```
>> [0 2]/[1 2; 3 4]
ans =
    3   -1
```

calculates the exact solution of the consistent equation

$$[0\ \ 2] = Y \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

and

```
>> [2 4]/[6 2]
ans =
 0.5000
```

calculates a least-squares solution of the inconsistent equation

$$[2\ \ 4] = Y\,[6\ \ 2]$$

Similarly, the command `A\B` (left division) calculates a matrix `X` such that `AX=B`, provided that `A` and `B` have the same number of rows. Again, if the equation is inconsistent then `X` is a least-squares solution. Thus

```
>> [1 2; 3 4]\[0 2]
ans =
 2.0000
-1.0000
```

Note that `A\B = (B'/A')'`.

MATLAB also provides array versions of the above arithmetic operators that allow for element-by-element operations on arrays (row or column vectors). If `x` and `y` are arrays of the same length, then `x.*y` generates an array whose elements are obtained by multiplying corresponding elements of `x` and `y`. Array versions of right and left division and the power operator are defined similarly. For example,

```
>> x=[1 2 3]; y=[4 5 6];
>> x.*y
ans =
    4   10   18
>> x./y, x.\y
ans =
 0.2500 0.4000 0.5000
ans =
 4.0000 2.5000 2.0000
>> x./2, 2./x, x.\2
```

```
ans =
 0.5000 1.0000 1.5000
ans =
 2.0000 1.0000 0.6667
ans =
 2.0000 1.0000 0.6667
>> x.^y, y.^x
ans =
    1   32  729
ans =
    4   25  216
```

Array version of transpose operator takes the transpose (without conjugate) so that

```
>> [1-i; 2+3i].'
ans =
 1.0000-1.0000i 2.0000+3.0000i
```

D.3 Built-In Functions

MATLAB provides a number of elementary math functions that operate on individual elements of matrices. Among them are trigonometric, inverse trigonometric, hyperbolic, inverse hyperbolic, exponential, natural and common logarithmic functions, square root, absolute value, angle, conjugate, real and imaginary parts of complex numbers. For example,

```
>> u=(pi/4)*[1 -3];
>> v=sin(u)
v =
 0.7071-0.7071
>> w=sqrt(v)
w =
 0.8409              0+0.8409i
>> z=exp(w)
z =
 2.3184          0.6668+0.7452i
>> ang=(180/pi)*angle(w)
ang =
    0   90
```

MATLAB also provides many useful matrix functions, some of which are summarized below.

```
>> A=[1 2 3 4; 2 3 4 5; 3 5 7 9];
>> rank(A)
ans =
    2
>> rref(A)          % reduced row echelon form
ans =
```

```
     1    0   -1   -2
     0    1    2    3
     0    0    0    0
>> norm(A,1), norm(A,2), norm(A,inf)  % p norms
ans =
   18
ans =
15.7403
ans =
   24
>> % singular value decomposition:  A=USV'
>> [U,S,V]=svd(A)
U =
 0.3472 0.7390 0.5774
 0.4664-0.6702 0.5774
 0.8136 0.0688-0.5774
S =
15.7403      0      0      0
      0 0.4921      0      0
      0      0 0.0000      0
V =
 0.2364-0.8026 0.3025-0.4566
 0.3914-0.3831-0.0629 0.8343
 0.5465 0.0364-0.7815-0.2987
 0.7016 0.4558 0.5420-0.0790
```

The following matrix functions operate on square matrices:

```
>> A=[0 -3 1; 1 4 -2; 1 2 0];
>> det(A)
ans =
    4
>> inv(A)
ans =
 1.0000 0.5000 0.5000
-0.5000-0.2500 0.2500
-0.5000-0.7500 0.7500
>> % LU decomposition:  L*U=P*A
>> [L,U,P]=lu(A)
L =
 1.0000      0      0
      0 1.0000      0
 1.0000 0.6667 1.0000
U =
 1.0000 4.0000-2.0000
      0-3.0000 1.0000
```

```
      0      0 1.3333
P =
    0    1    0
    1    0    0
    0    0    1
>> % modal matrix and diagonal form:  A*P=P*D
>> [P,D]=eig(A)
P =
 0.5000+0.5000i 0.5000-0.5000i 0.7845
      0-0.5000i      0+0.5000i-0.5883
      0-0.5000i      0+0.5000i-0.1961
D =
 1.0000+1.0000i      0              0
      0         1.0000-1.0000i      0
      0              0         2.0000
```

Note that `eig` command computes the linearly independent eigenvectors of `A` but not the generalized eigenvectors. If A is not diagonalizable the `P` matrix will not be a modal matrix.

D.4 Programming in MATLAB

D.4.1 Flow Control

MATLAB commands that control flow of execution based on decision making are similar to those of most programming languages and are briefly summarized below.

For–End Structure

The general structure of a `for` loop is

```
for x=matrix
   commands
end
```

where the `commands` between the `for` and `end` statements are executed once for each column of the `matrix` with `x` assigned the value of the corresponding column. Usually `matrix` is an array, and `x` is a scalar. For example,

```
n=input('Enter n = ')
fact=1;
for k=1:n
   fact=fact*k;
end
```

calculates the factorial of `n`.

`For` loops can be nested as desired.

While–End Structure

The general structure of a `while` loop is

```
while expression
   commands
end
```

The `commands` between the `while` and `end` statements are executed as long as the `expression` is `True`. The `expression` may include the relational operators >, <, >=, <=, == (equal) and ~= (not equal), and/or logical operators `&` (AND), `|` (OR) and `~` (NOT). As an example,

```
>> n=1; x=1; series=1;
>> while x>0.000001
     series=series+x;
     n=n+1;
     x=x/n;
   end
>> format long
>> series
series =
   2.71828152557319
```

calculates e using the McLaurin series

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

for $x = 1$, truncated when $x^n/n! < 0.000001$.

A mistake in the expression controlling a `while` loop may result in a never-ending loop. For example, if the second command above is mistakenly typed as `while x>0` then the loop never ends. Such a run-away loop can be broken by [CTRL-C] keys.

If–End Structure and Variations

`If` structures allow for control of the flow of execution based on simple decision making. The basic structure of the `if` command is

```
if expression
   commands
end
```

where `commands` are executed if `expression` is `True` and skipped otherwise. The variation

```
if expression_1
   commands_1
elseif expression_2
   commands_2
...
elseif expression_k
   commands_k
else
   commands_last
```

```
end
```

allows for a choice among several sets of commands.

A `break` command within an if structure can be used to terminate a loop prematurely. As an example

```
>> x=1; series=1;
>> for n=1:1000
     if x<0.000001
       break
     end
     x=x/n; series=series+x;
   end
```

is equivalent to the sequence of commands in the `while` loop example.

D.4.2 M-Files

Rather than being typed on the keyboard, a sequence of MATLAB commands can be placed in a text file with an extension `.m`, which are then executed upon typing the name of the file at command prompt. Such a file is called a script file, or an m-file referring to its extension. A script file can be created by selecting the *M-file* option of the menu item *New* under the *File* menu, or by using any text editor. When a valid variable name is typed at command prompt, MATLAB first checks if it is the name of a current variable or a built-in command, and if not, looks for an m-file with that name. If such a file exists, the commands in it are executed as if they were typed in response to >> prompts.

The `input` command in an M-file allows the user to type a value from the keyboard to be assigned to a variable. As an example, suppose that the following set of commands are stored in the M-file `myfactorial.m`

```
n=input('Enter n = ')
fact=1;
for k=1:n
   fact=fact*k;
end
fact
```

When the command `myfactorial` is typed at the >> prompt, MATLAB starts executing the commands in the file starting with the first command, which types the prompt

```
Enter n =
```

and waits for the user to type an integer, which is assigned to the variable `n`. The program ends after the value of `fact`, computed by the `for` loop, is echoed on the screen. A typical session would be

```
>> myfactorial
Enter n = [5]
fact =
  120
```

where `[5]` denotes the number entered by the user (followed by a return). Of course, the program can be refined to provide suitable error messages when the keyboard entry is not an admissible input.

D.4.3 User Defined Functions

Each of MATLAB's built-in functions is a sequence of commands which operate on the variables passed to it, compute the required results, and pass those results back. For example, the function

```
[L,U,P]=lu(A)
```

accepts as input a square matrix `A`, computes its LU decomposition, and passes back the results in the matrices `L`, `U` and `P`. The commands executed by the function as well as any intermediate variables created by those commands are hidden.

MATLAB provides a structure for creating user-defined functions in the form of a text M-file. The general structure of a user-defined function is

```
[vo_1,...,vo_k]=fname(vi_1,...,vi_m)
commands
```

where `fname` is a user given name of the function and `commands` is a set of MATLAB commands evaluated to compute the output variables `vo_1,...,vo_k` using the input variables `vi_1,...,vi_m`. A single output variable need not be enclosed in brackets. The text of the function must be saved with the same name as the function itself and with an extension .m, i.e., as `fname.m`.

As an example, the following function finds the largest `k` elements of an array `v` and returns them in an array `u`.

```
function u=mymax(v,k)
for p=1:k
   [w,ind]=max(v);
   u(p)=w;
   v(ind)=-inf;
end
```

Its use is illustrated below:

```
>> u=[-7:3:5];
>> x=mymax(u,3)
x =
     5     2    -1
```

Note that the function `mymax` uses the built-in MATLAB function `max`, which finds the maximum element in an array and its position in the array. It should also be noted that unlike an M-file, a function does not interfere with MATLAB's workspace; it has its own separate workspace.

D.5 Simple Plots

The `plot` command of MATLAB plots an array against another of the same length:

```
>> t=0:0.01:2; x=cos(2*pi*t);
>> plot(t,x)
```

produces the graph in Figure D.1.

More than one graphs may be plotted on the same graph, with different line characteristics. Lines may be added, axes and tick marks may be redefined, axis labels and a title may be added as shown in D.2:

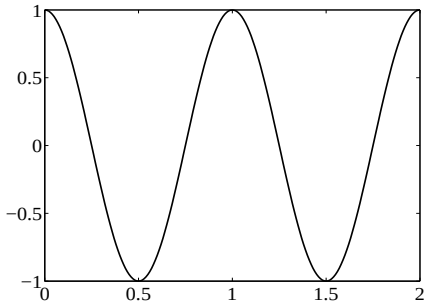

Figure D.1: A simple graph produced by MATLAB

```
>> newt=0:0.02:1; newx=sin(2*pi*newt);
>> plot(t,x,newt,newx,'o')
>> axis([-0.5 2.5 -1.25 1.25])
>> set(gca,'XTick',0:0.5:2,'YTick',-1:0.5:1)
>> line([-0.5 2.5],[0 0]), line([0 0],[-1.25 1.25])
>> xlabel('t'), ylabel('cos 2\pit (-) and sin 2\pit (o)')
>> title('A Simple MATLAB Plot')
```

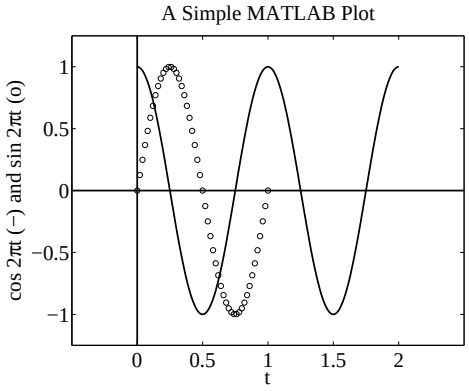

Figure D.2: A more complicated graph produced by MATLAB

D.6 Solving Ordinary Differential Equations

MATLAB provides two functions, `ode23` and `ode45`, for solving systems of first-order differential equation of the form

$$\boldsymbol{x}' = \boldsymbol{f}(t, \boldsymbol{x}), \quad \boldsymbol{x}(t_0) = \boldsymbol{x}_0$$

Although they use different numerical techniques, their formats are exactly the same:

```
[t,x]=ode23('myfnc', tspan, x0);
```

where `myfunc` is the name of a user-defined function that evaluates $\boldsymbol{f}(t, \boldsymbol{x})$ for a pair $(t, \boldsymbol{x})$ and returns it with name `xdot`, `tspan` is an array of strictly increasing or decreasing values of t_k at which the solution is to be found, and `x0` is a vector containing the initial value $\boldsymbol{x}_0$. The output array `t` contains a set of discrete points $t_k, k = 0, 1, \ldots, m$ in the range specified by `tspan`, and each column of the output matrix `x` contains the values of the corresponding component of the solution at t_k. If `tspan=[ti tf]`, then `ode23` use a variable step size to generate `t` with `t(1)=ti` and `t(m)<tf`.

As an example, the first order differential equation

$$y' = -2ty^2 \ , \quad y(0) = 1$$

has the exact solution (see Example 2.15)

$$y = \frac{1}{t^2 + 1}$$

The following set of commands evaluate the exact solution and plot it together with its difference from the solution obtained by the `ode23` function.

```
>> t=0:0.01:5;
>> y_e=1./(t.*t+1); [t,y_a]=ode23('myrhs',t,1);
>> subplot(211),plot(t,y_e)
>> Xlabel('t'),Ylabel('y_e')
>> subplot(212),plot(t,y_e-y_a')
>> Xlabel('t'),Ylabel('y_e-y_a')
```

where the MATLAB function

```
function xdot=myrhs(t,x)
xdot=-2*t.*x.*x;
```

which is saved as a text file with name `myrhs.m`, evaluates $f(t, y) = -2ty^2$. This example also illustrates the use of the `subplot(rcn)` command, which divides a figure area into an `r-by-c` array with `n` referring to the nth cell on which the current figure is to be plotted.

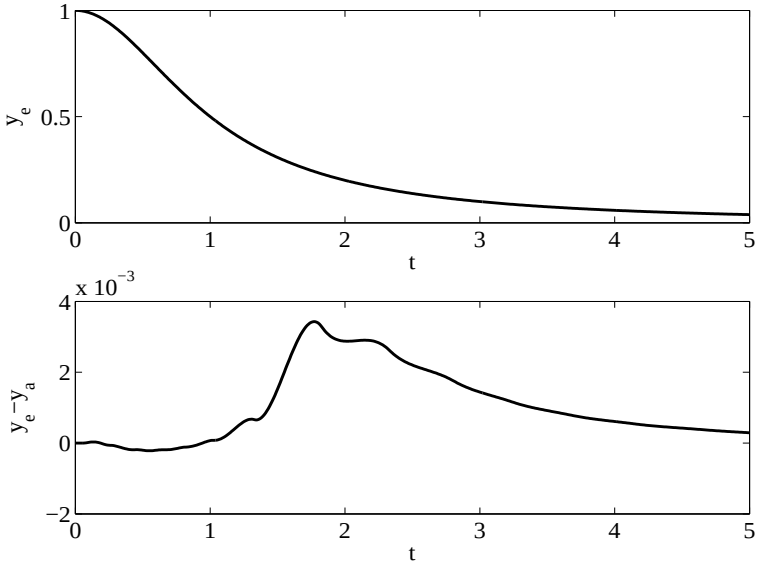

Figure D.3: Illustration of the `subplot` command

Index